



NTTS 2025

Book of Abstracts

Conference on New Techniques and Technologies for official
Statistics

Charlemagne Building, Brussels

11-13 March

Version 27.02.25

Foreword

Dear Readers,

As the editor of this Book of Abstracts, it is my great pleasure to present you this collection of contributions of the Eurostat's NTTS 2025 international conference (11 – 13 March in Brussels, Charlemagne Building) gathering innovative research, insightful discussions, and collaborative efforts from participants around the world.

This conference serves as a platform for the exchange of knowledge and ideas, fostering connections that drive progress in new techniques and technologies for statistics. I am confident that the topics presented here will inspire stimulating discussions and encourage new perspectives.

I would like to take this opportunity to express my heartfelt gratitude to my supervisors, Cristiano Tessitore and Dario Buono, for their invaluable guidance and support throughout the preparation of this book. I am also deeply thankful to the organizing committee and contributors who makes this event possible.

Finally, I extend my appreciation to all participants, including EMOS programme students, for their dedication and passion for advancing knowledge in the statistical field.

I hope you will find this Book of Abstracts both informative and inspiring, serving as a valuable tool for engaging discussions during the conference sessions.

Sincerely,

Klára Pavlíková

Editor, Book of Abstracts

About NTTS

New Techniques and Technologies for Statistics (NTTS) is an international biennial scientific conference series, organized by Eurostat, on new techniques and methods for official statistics, and the impact of new technologies on statistical collection, production and dissemination systems.

The purpose of the conference is both to allow the presentation of results from currently ongoing research and innovation projects in official statistics, and to stimulate and facilitate the preparation of new innovative projects with the aim of enhancing the quality and usefulness of official statistics. The NTTS conference series was initially triennial, and with the conference taking place in 1992 (Bonn), 1995 (Bonn), 1998 (Sorrento) and 2001 (Hersonissos, Crete). After a break of 8 years, the conference was re-launched on a biannual basis, and took place in Brussels in 2009, 2011, 2013, 2015, 2017 and 2019. NTTS 2021 was fully virtual, whereas NTTS 2023 took place in Brussels.

NTTS 2025 will be hosted again in Brussels in the Charlemagne Building.

Table of Contents

A blockchain solution enabling the reproducibility of findings based on official statistics	1
A Hybrid Machine Learning Approach for Carbon Price Forecasting to Support EU Climate Action	6
A long journey toward a shared statistical survey software suite	9
A multivariate approach to investigating socioeconomic challenges in residential areas	13
A population map of Europe by age	15
A risk-utility framework for anonymized georeferenced health microdata	19
A spatial Evaluation of Essential Citizenship Services at municipal level. The contribution of Network Analysis and Machine Learning Methodologies	23
A Statistical Analysis of Internet Standards Implementation on Dutch Enterprises' Websites	27
Accelerating Data-Driven Insights with EasyMorph's Intuitive Interface	32
Accessibility, JSON-stat and SDMX-JSON: Simplifying structures and harmonising metadata for dissemination	37
Advancing Statistical Dissemination: Insights from the EMOS Coding Lab on Cultural Goods Prices in Europe	40
AI Solutions in Istat Contact Centre for Knowledge Seekers, Data Users and Survey Respondents	50
AI-powered nomenclature codification	44
AM3DS: A new service to integrate administrative data in the French statistical information system, driven by metadata	53
An agricultural atlas for Europe – based on IFS 2020	62
An end-to-end statistical process to aggregate high-resolution temporal survey microdata: the monthly Labour Force Survey use case	66
An innovation score based on financial data and web-augmented firms' characteristics	71
An ML-based automatic coding machine as an integral element in an overall strategy for NACE implementation	75
An open innovation platform to share ideas and knowledge, and an open-source project for NSO's to work with their own sensitive data	80
Analysing EU unemployment and underemployment in a fuzzy perspective: the comparison between fuzzy and official rates	56
Analysis of voluntary survey data under missing not at random	61
Applicable AI in Official Statistics: A New Research Theme at Statistics Netherlands	83
Applying Machine Learning to educational level: An empirical evaluation	91
Applying a Non-Parametric "Benefit of the Doubt" Approach for Identifying Low-Quality Responses in Online Panels	87
Artificial Intelligence and Governance, driving innovation with responsibility	95
Automated Product Classification Using Machine Learning and Natural Language Processing: A Case Study of Scanner Data	98
Automatic analysis of metadata for secondary tabular data protection	101
Bayesian spatial modelling for prevalence of child mortality and morbidity: A bivariate Cauchit link model	104
Better enlighten local public decision-makers on the situation of the real estate market	111
Binding Surveys and Deep Learning Statistics on Arrival Ports for Maritime Mobility via AIS Data	114
Building a Comparable Measure of Consumption: Concepts and Measurement Challenges Faced by Emerging and Advanced Economies	118

Choice of Sampling Efforts in a Schnabel Census	122
Classification and Prediction of Rating Data with Flexible Uncertainty Trees	127
Combining Longitudinal Administrative Registries, Test Results and Stakeholders' Input to Protect Educational Trajectories in Uruguay	132
Combining Nonprobability Samples with Probability Samples to Correct for Selection Bias	137
Construction of a probabilistic scoring model for food allergy profiles in Tunisia	141
Cyclic Outlier Detection in AIS Data: A Direct Approach Based on Average Speed.	146
DATA CATALOGUE - a modern and harmonised approach to standardised data management in the NSI	150
Data Dialogue: Perspectives and guidelines on the use of Mobile Phone Data for Statistics, Development and Humanitarian Initiatives	152
Data Science in the Production of Official Statistics: Utilizing Reproducible Analytical Pipelines (RAPs) for Custom Declarations Processing at PCBS as a Case Study	155
Detecting and Correcting Erroneous OCR and Manual Diary Registrations in a Modernised Household Budget Survey	161
Detecting Change Points of Expectations and Judgments of Price Dynamics within the Italian Population	157
Detection and correction of quarterly effects in administrative versus survey turnover data	165
Development and Integration of Microservices for Smart End-to-End Surveys	170
Development of a Satellite Account for Culture at the Local Level - First results for Tuscany (Italy)	174
Development of crop pollination ecosystem service accounts in Ireland	179
Earth Observation and AI for Construction Statistics (EO4ConStat)	184
Earth Observation to accelerate ecosystem accounting	189
Efficiently Selecting Representative Audit Samples	193
Embracing Open Source: The European Commission's Journey	198
Enhance statistical literacy among non-expert users: Eurostat's experience	199
Enhancing data quality of a building and dwellings register through mapping building coordinates to building polygons using a Random Forest classifier	201
Enhancing European Business Statistics: A Harmonised Microdata Linking Approach	206
Enhancing housing surveys in Overseas Departments using Deep Learning and Satellite Data	210
Estimating CO2 emissions in Italy with local regression models	214
Evaluating an ESCO skill classifier	218
Evaluating the Completeness of Business Databases: A Comparison with Official Records Using Web Scraping Techniques	222
Experiences from experiments with smart surveys in the Smart Survey Implementation project	226
Farm register for agricultural sample surveys: toward a systematic integration of administrative sources	227
Finding the usual environment of tourists in Mobile Phone Data: a case study applied to the same-day visits	228
Flash estimates of income and poverty indicators: an empirical study predicting labour market transitions	233
Fostering Collaboration: Communication Strategies in the AIML4OS Project	237
Fostering the collaboration between Eurostat and the Academia	238
From NACE 08 to Nace 2025: Retraining an ML model in production using Large Language Models	242
From web to data: Quality and Methodological Considerations	245
From web to data: Selective scraping and WIN.The hackathon	248

From web to data: Winning Hackathon with Offline AI for E-Commerce Site Detection Strategies	251
Functional Clusterwise LDQ Regression for Assessing Poverty Risk Across European States	255
Future of Data Collection: Automated Data-Based Reporting	259
Gemma 2 Large Language Models for Aspect-Based Sentiment Analysis on Italian Texts: A Use Case for Tourism Statistics	263
Harmonised statistics on income and living conditions in Europe, innovations and development over time	267
Harnessing Machine Learning for Superior GDP Forecasting: A Comparative Study of Machine Learning versus Econometric Models	271
Households and their dwellings as communities in administrative data	275
How far would you walk to the supermarket? Estimating access to shops across the United Kingdom	279
How many diary days? Smart surveys can help to reduce burden of data collection for household statistics	283
"I don't know what I did last summer" – measurement error and memory effect in tourism statistics	287
Implementation of a new coding service at Insee	292
Improving AI Language Models via Specialised Training Corpora for Official Statistics	296
Improving estimates of short-stay accommodation occupancy in the EU using online platform data	300
Independent Address Identification Search Engine for National Statistical Institute Using ElasticSearch	303
ineapir: an R package to query statistical data and metadata from National Statistics Institute (INE Spain)	307
Integration of MNO data with survey data to produce commuting statistics	311
Integration of Mobile Network Operator (MNO) data with further data sources: How to treat errors and biases	315
Inveus. Remote access to data space for researchers	318
JDemetra+ 3.2.4: New (R) tools for (high-frequency) time series analysis	322
Joint Estimation of Common and Specific Dynamic Factors from Vectors of Time Series: Application to Electricity Prices in Several Markets in the EU	326
Jordan National Interactive Data Centre Data as a Service	332
Large language models for COICOP classification in Norway's household budget survey	336
Measuring data science skill levels in an NSI using text mining of production code	345
Methods for integrating survey data and big non-survey data	348
Micro-level bank card data for the use of official statistics	351
MILC vs. Tree-MILC: a simulation study comparing correction methods for multi-source statistics	357
Mixture models and graphical tools to analyse the perception of work-related stress across EU countries	362
Mobile Network Operator Data as a New Source for Transport Planning and Mobility Analyses	366
Mobile Network Operator data for Tourism Statistics: facing the challenge of cross-roaming	368
Mobility Research with a Smartphone App	372
Model explainability: a use case on FastText for automatic coding	375
Multi-Mode Protocol for the French Household Budget Survey: More Choices, More Respondents?	379
Municipal-Level Estimation of Tourism Perception: A Machine Learning-Based Approach to Small Area Estimation	383

New challenges demand new solutions	385
New Challenges in Official Statistics: Big Data Analytics and Multi-level Product Classification of Web-Scraped Data	390
New index for evaluation livestock farms digitalization	391
New results for time-dependent models: a perspective for seasonal adjustment	496
Nowcasting gas demand based on timely available data sources	399
One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics: Reflections on the First Year	402
Open Infrastructure and Open Source Datalab to Strengthen Collaboration and Innovation	405
Open source software sharing through the awesome list of official statistics software	407
Optimizing Text Classification with Deep Learning: The Impact of Preprocessing on ATECO Code Assignment in Italian Business Descriptions	411
Perceived health and life expectancy in EU countries: a machine learning approach for small samples	415
Predicting Travel Purpose in a Smartphone-Based Travel Survey	418
Prediction of crop yields at field scale from earth observation data	426
Privacy-Enhancing Technologies for Synthetic Data Creation with Deep Generative Models	430
Processing of receipts in Household budget survey	437
Protecting high-resolution grid data with additive noise while retaining fitness for use	440
R package for creating multi-resolution grids of census and sample data	443
Recent Guidelines for Statistical Disclosure Control	447
Redesigning the online French Census questionnaire: a user-centered methodology and its limitations	452
Regularizing Probability Sample Estimates Through an Angle-Based Similarity Approach	456
Research Network: Anonymization for Secure Data Use	462
RUMS – how to compare structures of enterprise groups?	466
SAS to R Training: A Collaborative Approach to Promote a Learning Culture in Statistical Programming	470
Small area estimation using incomplete auxiliary information	475
Smart Features in Surveys: Respondent Navigation and Its Impact on Outcomes - Insights from the French Time Use Survey Web App Test	478
Spatial Mobility and Housing Demand in the Dynamic Microsimulation Model MikroSim	482
Speeding up the release of the Total Market Production Index (TMPI)	485
Statistical Information Communication: past insights, future innovations	489
Statistical Insights about Elderly Hospitalization in India: Count Data Model	493
Statistical Interoperability Node	499
Streamlining the Supply and Analysis of Big Data at Research Data Centres	504
Sustainable open-source projects – key success factors	507
The Effect of Working Conditions on Health and Healthcare Use	511
The ESA "PEOPLE – Ecosystem Accounting" project: experiences from Greece	515
The eurostatRTool: a customisable innovative R tool for rapid dissemination of statistical information	519
The Journey Towards a Governance of Open Source Statistical Tools at Istat	524
The path of Earth Observation to official statistics	528
The role of official statistics in the fast-evolving data ecosystem (focus on European Data Strategy)	532
The use of NLP and Machine Learning in official statistics	536
The Uses of Administrative Data in the Danish Survey on Income and Living Conditions	540
The Visual Display of Skewed Data: Boxplot, Beeswarm and Beyond	543

Telling Stories with Data: Enhancing Museum Accessibility for Blind Visitors Through User Feedback	548
Timeliness criticality and its aggregated economic impact	551
Towards buildings demography	559
Towards dynamic land cover data for Austria using satellite imagery and time-series classification	563
Towards Statistical Disclosure Control for Complex Networks	568
Two methods for integrating smart surveys with traditional surveys	572
US Injuries Flood Prediction with Large Language Models Data Augmentation	576
Use of AI for built-up area land use mapping: A study on construction sites	580
Use of public data sources to increase the accuracy and completeness of the EuroGroups Register	586
Use of web scraped data for analysing labour market demand for ICT specialists	590
Using CI/CD methods for continuous benchmarking of statistical methods	596
Using Cloud and Open Source for Continuous Innovation in Statistical Production	599
Using machine learning to produce census population count from an administrative data system	601
Using R in statistical observation of entrepreneurial subjects	606
Using Shape-Based Time Series Clustering to construct Machine Learning Forecasting Models for Short Panel Data: A Case Study on the Sustainable Development Goals	614
Visualizing Official Statistics Using Reproducible Analytical Pipelines	618
WebQA Project	621
Weekly Nowcasting US Inflation with Enhanced Random Forests	626
When are you ready for a smart survey?	630
Where people live and work: a register-based spatial and longitudinal analysis for Italy	632
Which data to combine with Mobile Phone Data? Landscaping of the most promising sources and associated challenges	637
Who are the flood victims? Monitoring the effects of the disaster using Earth Observation, GIS and administrative data	641
Working with a mobile network operator (MNO) to create a privacy-conform method for a better access to MNO data	642

A blockchain solution enabling the reproducibility of findings based on official statistics

Martin Karlberg & Petre Turliu (Eurostat)

1. INTRODUCTION

1.1. Problem description

Suppose that an official statistics user (be it a researcher or policy analyst; we use ‘analyst’ below for brevity) downloads and subsequently processes official statistics datasets in order to underpin their research or policy analysis. When other users later try to carry out the same analysis, they may arrive at different results, as the published official statistics may have been revised in the meantime – the original dataset being ‘overwritten’ and no longer accessible. This will, in turn, lead to analysis outcomes different from those produced by the initial user, rendering impossible an independent replication of the analysis. While much of this paper is oriented towards Eurostat, the solution described is transferrable to other official statistics producers interested in reproducibility and to the traceability of ‘vintage’ data.

1.2. Versioning to ensure reproducibility

An obvious solution that springs to mind is for the official statistics producer to version all statistics that are published – thus, rather than ‘overwriting’ the previous version of a dataset, each version of dataset is, once published, saved (along with relevant metadata, such as the time of publication and possibly also version numbers). This creates a ‘statistical data warehouse’, with each dataset version being an invariable ‘snapshot’, never changing after its publication. However, this solution does require, inter alia, considerable storage resources, with near-identical instances of the same dataset being saved multiple times.

1.3. Selective versioning ensuring partial reproducibility

One could assume that it would be more important to prioritise revisions of a more systematic nature, such as ‘routine revisions’ and ‘major revisions’ (according to the terminology of the [Eurostat revision policy](#)). Along these lines, Eurostat has recently implemented a versioning approach in a very focused way, with so-called ‘vintages’ being published for a [subset of Principal European Economic Indicators](#).

1.4. User-managed reproducibility

Full versioning (see 1.2) is not currently being practiced for European statistics, and selective versioning (see 1.3) does not necessarily cover the datasets of interest of a specific user, notably in the area of social statistics. An analyst wishing to create reproducible findings does therefore have to download and store, in infrastructure under their control, the Eurostat datasets that they are using, thereby providing any interested stakeholder with the means necessary to independently replicate their analysis.

However, a main drawback with this approach is that there is nothing, apart from the analyst’s own assurance, that ensures that the dataset thus stored is a genuine Eurostat dataset. Barring the few instances of data having a revision policy (described under the ‘data revision’ concept of their metadata) whereby they are being ‘frozen’ after their dissemination, Eurostat data are likely to differ at least slightly from the data used by an analyst in the past, thus

rendering it impossible for a third party to verify the authenticity of the data used. In some cases, a major revision has been undertaken, leaving the currently disseminated dataset substantially different from the one being stored by the analyst, thus further aggravating the discrepancy. In other cases, the original dataset has gone offline – so there are no Eurostat data whatsoever with which the downloaded data can be compared.

1.5. Authenticable reproducibility based on blockchain technology

Based on ideas introduced by colleagues at Statistics Canada [1] and Tessitore [2], it is possible to augment the user-managed approach (see 1.4) with the one missing element: authentication. Simply put, Eurostat injects a ‘fingerprint’ of each published dataset version into ‘Eurostat blocks’ of a public blockchain (as detailed in Section 2 below). This gives analysts the necessary tools to demonstrate that the data they have used are genuine Eurostat data (by following the steps described in 3.1 below) – and allows users to check this (see 3.2 below).

This approach for ensuring authentication will be piloted during 2025 – and if successful, subsequently be rolled out for all Eurostat data.

2. METHODOLOGY

2.1. Choice of blockchain ledger

To implement blockchain-based authentication, a blockchain ledger needs to be used for the insertion and retrieval of hashes and associated metadata. Setting up a new blockchain infrastructure can be costly in terms of time and resources. Running costs may also be considerable. The European Blockchain Services Infrastructure (EBSI) is an initiative of the European Commission to leverage the power of blockchain for public good. The EBSI infrastructure is already up and running, and has been used in various successful initiatives, from micro-credentials to transcripts of records. Therefore, in line with the proposal of Tessitore [2], Eurostat will pilot its authenticable reproducibility initiative using EBSI.

2.2. Inner working of the blockchain-based authentication mechanism

To provide dataset update authentication information in the EBSI, the following steps are taken:

- (I) Whenever a Eurostat dataset is updated, the full dataset is hashed by the Eurostat dissemination IT infrastructure.
- (II) Following a cooldown period (to allow possible error corrections), the hash, including key metadata (as illustrated in Table 1) is added as an event to an existing ‘document’, the dataset, via the node.js command line interface (CLI).

Table 1. Foreseen information in the ‘Eurostat’ EBSI blocks (during piloting)

Eurostat dataset metadata attribute	Sample value (for a specific dataset version)
Hash of the dataset	3733cd977ff097f.....e83e4d5
Format of the dataset file (xlsx, csv, tsv)	tsv.gz
Associated DOI	10.2908/SDG_01_10
Precise timestamp for when the dataset was updated	30/08/2024 23:00:00
Dataset status	Unofficial pilot release

Note: the hash of the dataset has been truncated in order to simplify the graphical appearance of the above presentation

3. PRACTICAL APPLICATION

In this section we briefly present what a researcher needs to do to ensure that other stakeholders can check that the dataset they have used for their analysis is an authentic Eurostat dataset – and how users can perform a check of the authenticity of the dataset of a researcher. It should be noted that other obviously necessary elements of authenticable reproducibility (such as the sharing of source code) are not presented here, as we are focusing on the authenticability aspect.

3.1. What does a researcher need to do to ensure authenticability?

Simply put, a researcher or analyst wishing to ensure that other stakeholders can verify the authenticity of a Eurostat dataset that they have used for their research/analysis need to do the following:

- (1) Download the **full** dataset in **csv** or **tsv** format
As there is an extremely large number of possible subsets (somewhat simplistically put: 2 to the power of the number of selectable values for each dataset dimension), it is impossible to store them all. It should thus be noted that the ‘full dataset’ is the one including the maximum amount of embedded metadata (codes and labels).
- (2) Save that dataset – without modifying it in any way whatsoever
As EBSI only contains the ‘hashes’ the responsibility for storing the dataset remains with the researcher/analyst, and has to be done on infrastructure that they control (e.g. their GitHub site).
- (3) Make sure that the hash of the dataset corresponds to a ‘Eurostat block’ in the EBSI blockchain (following the procedure for users described in Section 3.2 below).
There is a risk of authentication failure (be it because of inadvertent manipulations following the download, because the dataset is very recent and thus not yet stable, or because a subset was downloaded rather than the full dataset). It is therefore advisable that the dataset used for the analysis is ‘locked in’ into the blockchain. User guidance is foreseen to be provided to address this risk.
- (4) Disseminate the dataset in such a way that it is accessible to the intended stakeholders
This would normally be in the public domain, but in case the analysis is sensitive, sharing with a restricted audience would also be a possibility.

3.2. What does a user need to do to check the authenticity of a dataset?

A user wishing to check that the data that a researcher/analyst has based their analysis on are indeed authentic Eurostat data, will have to do the following:

- (A) Download the dataset disseminated by the analyst (4).
- (B) Hash the file by computing a SHA-256 via a local tool such as PowerShell.
- (C) Send an API http call to the EBSI, using the hash as a parameter.

The response from the EBSI API will be either of the following:

- a message that the hash is **not** present in the EBSI blockchain (as illustrated in the example of Figure 1):

```
1 {  
2   "title": "Bad Request",  
3   "status": 400,  
4   "type": "about:blank",  
5   "detail": "[\"documentId must be 32 bytes encoded in hexadecimal and start with 0x\"]"  
6 }
```

Figure 1. Failed call to EBSI for a non-existing hash

- the EBSI block corresponding to the hash with its associated metadata, as illustrated in the sample response in Figure 2. Here we can note that:
 - The 'creator' field includes the digital ID (DID) of Eurostat (also possible to look up via [an EBSI service](#)), verifying that the hash corresponds to a Eurostat dataset.
 - The DOI allows users to retrieve rich metadata from elsewhere (to augment the 'bare bones' metadata provided in the 'Eurostat block' in EBSI)

Figure 2. Sample response for a dataset

4. DISCUSSION

As all Eurostat blocks contained in EBSI are in the public domain, it renders it possible for any interested user to see the revision history of any Eurostat dataset. However, as EBSI has not been designed for search, focusing mainly on verification, this will in practice (to avoid time-consuming searches) require that either Eurostat or a third party stores references to EBSI blocks relevant to a dataset. (In the future, this feature may be offered directly by EBSI.)

4.2. The comparative advantage of using EBSI

4.3. Conclusion

5. REFERENCES

- [1] Kathryn Fedchun, Lillian Klein and Didem Demirag 2022. *Investigating the Use of Blockchain to Authenticate Data from the Statistics Canada Website*. Statistics Canada Analytical Studies: Methods and References (19 September 2022);
<https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2022007-eng.htm>
- [2] Cristiano Tessitore 2023. The usage of blockchain technology for official statistics. *Book of abstracts of the 2023 conference on New Techniques and Technologies for Official Statistics (NTTS 2023)*; 449-451. Available from:
https://wayback.archive-it.org/12090/20231230192209/https://cros-legacy.ec.europa.eu/sites/default/files/book_of_abstracts.pdf.

A Hybrid Machine Learning Approach for Carbon Price Forecasting to Support EU Climate Action

Despoina Makariou (University of St Gallen), George Tzougas (Heriot Watt University), Rui Zhu, Andreas Tsanakas (University of London), Dimitris Christopoulos (Heriot Watt University) & Zezhun Chen (LSE)

1. INTRODUCTION

We extend the current statistical and machine learning literature concerning time series forecasting by developing a novel hybrid ARIMA-Long Short Term Memory Network (LSTM) model. The ARIMA component captures linear dependencies and short-term patterns in a time-series, while the LSTM component captures more complex nonlinear relationships and long-term patterns in the residuals. Our suggested methodology improves shortcomings in the current hybrid machine learning methods for time series forecasting where the data are decomposed into linear and non-linear models separately by ARIMA and LSTM. Such approaches are more computationally complex, not fully accounting for interactions between linear and non-linear components, and they are less interpretable.

To showcase the importance of our methodology for official statistics, we apply it in the context of carbon price forecasting for the EU Emissions Trading System (EU ETS) and the newly created UK Emissions Trading Scheme (UK ETS) post Brexit. Firstly, we use our proposed method to accurately forecast daily carbon prices for both the EU and UK Emissions Trading Systems (ETS) where we show that our modelling approach leads to better prediction accuracy results. Secondly, we provide insights into how these two carbon markets (i.e. EU ETS and UK ETS) differ post-Brexit. To the best of our knowledge this is the first time that this phenomenon is analysed. Our empirical investigation reveals distinct patterns in the best ARMA orders and coefficients of the EU and UK ETS markets post-Brexit, indicating that businesses operating in these markets could potentially encounter future variations in carbon pricing dynamics and risk exposures. As for policy implications, our findings suggest the need for coordinated policies to address differences in carbon pricing mechanisms, promote fair competition, and lead to efficient allocation of emission reduction goals across borders.

2. METHODOLOGY

Our proposed hybrid ARIMA-LSTM model combines the linear capabilities of ARIMA, which captures short-term patterns and linear dependencies in carbon price data, with the Long Short-Term Memory (LSTM) network to model non-linear dependencies and long-term patterns. We first applied an ARIMA model to forecast carbon prices, and the residuals were then modeled using the LSTM, resulting in a comprehensive model that captures both linear and non-linear relationships. This approach ensures a robust and computationally efficient forecast model, avoiding the complexity of separate linear and non-linear decompositions. The algorithm of our model is provided below.

Algorithm 1: The hybrid model of ARMA + LSTM.

Input: The training series $\{y_t\}_{t=1}^T$.

Output: The prediction $\hat{y}_t^H$ of y_t from the hybrid model.

1 Fit an ARMA model to $\{y_t\}_{t=1}^T$ and obtain the predictions $\{\hat{y}_t^A\}_{t=1}^T$.

2 Calculate the residuals of the ARMA model: $e_t^A = y_t - \hat{y}_t^A$.

3 Reformat the residuals to training an LSTM network.

 Initialisation $\mathbf{L} = \text{matrix}()$, $\mathbf{e}^A = \mathbf{c}()$

 for t in $\text{seq}(l+1, T)$ do

$\mathbf{l}_t = [e_{t-l}^A, e_{t-l+1}^A, \dots, e_{t-1}^A]$

$\mathbf{L} \leftarrow \text{cbind}(\mathbf{L}, \mathbf{l}_t)$

$\mathbf{e}^A \leftarrow [\mathbf{e}^A, e_t^A]$

4 Train an LSTM network f^L with the lag matrix $\mathbf{L}$ to predict $\mathbf{e}^A$.

5 Calculate the predicted residual from the LSTM network: $\hat{e}_t^L = f^L(\mathbf{l}_t)$.

6 The prediction of the hybrid model is $\hat{y}_t^H = \hat{y}_t^A + \hat{e}_t^L$.

Figure 1. Presentation of the algorithm associated with the suggested hybrid model of ARIMA+LSTM.

3. RESULTS AND PRACTICAL APPLICATION

The empirical investigation shows that while the EU ETS and UK ETS are highly correlated, they exhibit distinct ARMA orders and coefficients post-Brexit, indicating different pricing dynamics. Our hybrid model performs well in capturing long-term trends and structural shifts in the EU and UK carbon markets, revealing potential future disparities in carbon pricing and risk exposures. These findings have practical implications for policymakers and businesses involved in carbon trading, suggesting coordinated policies to harmonize carbon pricing mechanisms across these two markets.

3.1. The EU ETS and UK ETS exhibit different patterns post-Brexit, implying that businesses in these markets may face varying carbon pricing dynamics and risk exposures. The divergence between two Emissions Trading System (ETS) carbon price markets, such as the EU ETS and UK ETS post-Brexit, does not inherently indicate "carbon dumping," but it can create conditions that might encourage such behaviour thus this is an important result for policy purposes.

3.2. The hybrid ARIMA-LSTM model outperforms traditional forecasting methods by efficiently capturing both linear and non-linear dependencies in the carbon market data, making it a powerful tool for understanding long-term pricing trends.

3.3. Compared to other hybrid models in the literature, the proposed model reduces computational complexity while maintaining interpretability, making it suitable for practical policy analysis.

4. MAIN FINDINGS

The suggested hybrid ARIMA-LSTM model outperforms traditional forecasting methods by efficiently capturing both linear and non-linear dependencies in the carbon market data, making it a powerful tool for understanding long-term pricing trends. The EU ETS and UK ETS exhibit different patterns post-Brexit, implying that businesses in these markets may face varying carbon pricing dynamics and risk exposures. Compared to other hybrid models in the literature, the proposed model reduces computational complexity while maintaining interpretability, making it suitable for practical policy analysis.

5. RELEVANT REFERENCES

- [1] Byun, S. J. and Cho, H. (2013). Forecasting carbon futures volatility using GARCH models with energy volatilities. *Energy Economics*, 40:207–221.
- [2] Bartram, S. M., Hou, K., and Kim, S. (2022). Real effects of climate policy: Financial constraints and spillovers. *Journal of Financial Economics*, 143(2):668–696.
- [3] Berg, T., Saunders, A., Schafer, L., and Steffen, S. (2021). Brexit and the contraction of syndicated lending. *Journal of Financial Economics*, 141(1):66–82.
- [4] Zhu, B., Chevallier, J., Zhu, B., and Chevallier, J. (2017). Carbon price forecasting with a hybrid ARIMA and least squares support vector machines methodology. *Pricing and Forecasting Carbon Markets: Models and Empirical Analyses*, pages 87–107.
- [5] Ji, L., Zou, Y., He, K., and Zhu, B. (2019). Carbon futures price forecasting based with ARIMA-CNN-LSTM model. *Procedia Computer Science*, 162:33–38.
- [6] Yang, S., Chen, D., Li, S., and Wang, W. (2020). Carbon price forecasting based on modified ensemble empirical mode decomposition and long short- term memory optimized by improved whale optimization algorithm. *Science of the Total Environment*, 716:137117.

A long journey toward a shared statistical survey software suite

Lionel Hébrard & Juliette Fourcot (The Department of Statistics and Foresight Analysis of the French Ministry of Agriculture and Food)

Abstract: The Department of Statistics and Foresight Analysis of the French Ministry of Agriculture and Food runs several surveys a year to meet European statistical regulations and France specific needs. The Ministry IT teams have developed a survey software suite at the late 2000's when surveys became dematerialized. For many reasons developed hereby, this software suite is doomed and is meant to be replaced by the Insee suite, following a shared statistical services strategy. The paper describes the project genesis, its organisation and the main findings, in order to share feedback on a large-scale implementation of this type of projects, encouraged to multiply over time in the European Statistical System (SSE).

1. INTRODUCTION

The Department of Statistics and Foresight Analysis of the Ministry of Agriculture and Food (SSP) is in charge of producing and disseminating statistics on farming, fishing, forestry and food industry. It is in charge of data collection and dissemination, regarding European statistical regulations in this field. Whilst administrative data are extensively used, surveys remain by far the most common design for data collection. On average, the Department runs about twenty surveys a year, with a frequency ranging from monthly to once every 3 or 4 years, some collected directly through a web form (computer assisted web interview), other with the intermediation of an investigator through a face-to-face or a phone interview (computer assisted personal/telephone web interview). At the early 2000's, surveys were very little dematerialized. Interviewees filled questionnaires on paper, and workers had to register computer entries for each form. In order to reduce the workforce for the 2010 agricultural census, the Ministry has engaged a project to develop a specific software suite to replace paper-based interviews by computer-assisted interviews. They first worked on a specific survey, and later gained in generalization with a generic survey designer tool. Ten years later, and only 3 years after the general adoption of this suite, it revealed unable to maintain the tools in the long term. Several reasons were identified: a monolithic architecture that is difficult to update and to evolve, fewer overall resources (human and financial), and an increasing need for multi-mode survey which induced to revise the whole architecture of the tools.

At the same period, the French National Statistical Institute (Insee), was embracing a project for its household surveys mostly run through investigators, and an increasing trend toward multi-mode surveys. The goal was to extend the software suite developed for enterprise surveys (mostly CAWI), built on top of an active metadata architecture [1].

After an analysis of the functional coverage for SSP and a relatively brief analysis of gains for the Insee agreed to expand its project for the SSP and more generally for providing a survey service for other national authorities (ONA) of the French Statistical System.

This paper aims to present the project from the perspective of the beneficiary, with its specific constraints and needs.

2. PROJECT METHODOLOGY

2.1. Project genesis

Although Insee and the French ONAs share few statistical services, the labor cost survey run by the French Ministry of Labor has paved the way a couple of years ago, by benefiting from access to the enterprise survey software suite. The narrow relationship between Insee and the ONAs has produced many synergies on statistical and methodological sides, but fewer in the field of IT projects.

Since the 2010's, the French government strategy includes incentive to share digital commons and services. For example, in order to reduce the number of datacenters and improve them qualitatively, ministries were encouraged to either provide a hosting service or shutdown their computer centers. More recently, a few private clouds (run by public state) have been selected to provide a sovereign cloud offer for public services. On a service level, several use cases have been promoted and deployed, such as for human resources management or travel management.

In the ESS side, the I3S ESSNet objective was similar, but even within France statistical entities, the development of share services / tools remains scarce.

2.2. Project qualification

In 2018 the SSP stopped a technical maintenance project because of exceeding time and costs. The Insee software suite was identified in 2020 as a possible future, and the first step that took place in 2021 was to check the adequacy between the actual functional needs and the target features of the Insee survey tools. The SSP has benefit of a detailed knowledge of both systems thanks to an engineer acting as business project manager who was formerly developer working on the Insee survey tools.

Once the comparison was done, the SSP and Insee conclude for a partnership.

The project has then been split into several subprojects:

- Development of missing and additional features, by expressing needs toward Insee business team,
- Development of an interface to transmit input (period of collect, reminder dates, assistance information, list of statistical units and contact information...) and outputs (collected data, control state), but also identification and authorizations,
- Deployment: adoption of the new tools for each statistical survey,
- Identification and implementation of a trajectory for processes that are not surveys and for which the previous suite was used,
- Change management.

Taking various constraints into account, the actual planning of migrations leads the project to end in 2028/2029. Costs have been estimated, and deployment projects occupies the largest part.

3. MAIN FINDINGS

3.1. Expected benefits

Besides the ability to run surveys on the long term and make multi-mode protocols possible, the SSP is confident into Insee design of its survey tools because of the place given to

standards such as DDI and modern stats. The modular architecture would provide a better maintainability than the monolithic approach.

This alignment with statistical standards will ease the adoption of other tools such as the Insee reference frame for statistical metadata, "RMéS", which is meant to work with the survey tools by design. This frame is interesting for the SSP because it could help the better manage nomenclatures and facilitate the reuse of statistical metadata of the operation lifetime.

Other expected benefits were not identified at the earlier stage of the project:

- Introduction of a promotion portal to give attractive and comprehensive presentation of the survey for interviewees (what it is, the objectives, example of past dissemination products, frequently asked questions section, contact form)
- An accelerating factor for the implementation of the identity federation system in the Ministry of Agriculture and Food,
- A better management of data lifecycle: in the previous system, it was not easy to identified whether the microdata was a declared row data, or whether it was modified by an internal worker, or an automaton
- The possible access to an integrated voice server, that could reduce or facilitate human assistance.

3.2. Facilitating factors

Though the implementation of the project is just at its beginning, several factors are already identified as facilitating:

- Use of GSBPM vocabulary: the discussion on features between actors coming from very different domains is eased by the adoption since a couple of years ago of the GSBPM framework.
- Interpersonal networks: as mentioned above, the presence of a former Insee developer played a key role in the project to build relationships with actors and facilitate communication. More recently, some of the contacts within Insee is a former methodologist and frame manager of the SSP. We have considered this stake for several recruitments at SSP.

3.3. Remaining challenges

The impact on job and skills of this IT shift from specific development toward shared services is not yet fully identified by organizations. For example, development resources will decrease but interface and interoperability expertise become more important. The coordination tasks take more time and become more complex because the number of actors increases.

A second challenge is to design the future of some specific surveys. For example, the Farm Accountancy Data Network (FADN) actual survey includes a two steps validation workflow, whilst a 1step is sufficient for most surveys. Another example is the case of surveys that can not be considered as meeting French statistical regulation because they answer both administrative needs (regulation impact measurement, monitoring) and statistical needs.

Last but certainly not least is the challenge of building a realistic planning that takes deployment opportunities for each survey and the functionality roadmap. The further we look, the blurrier it is, which implies a careful posture especially for heavy surveys that require a long period of tool preparation.

4. REFERENCES

- [1] F. Cotton and T Dubois, « Pogues », a Questionnaire Design Tool, Courrier des statistiques #3 (2019).

A multivariate approach to investigating socioeconomic challenges in residential areas

Thomas Önskog (Statistics Sweden)

1. INTRODUCTION

The national statistical institute in [country] annually produces a classification of the 3,363 regional statistical areas (residential areas) in the country into five different area types based on a socio-economic index produced by the NSI. The socio-economic index weighs together the proportion in the residential area that has a low level of education, the proportion that has been long-term unemployed and the proportion that has a low income into an index between 0 and 100. Based on the mean and standard deviation of the socio-economic index, the residential areas are then divided into five different categories, where area type 1 corresponds to major socio-economic challenges and area type 5 to very small socio-economic challenges.

In a recent government assignment, the NSI has been asked to further develop the classification into area types based on socio-economic index in order to obtain a more developed picture of the problems in the country's various residential areas with the aim of identifying residential areas with specific or multiple problems and relate these to certain conditions in the area. The assignment specified that the list of problems to be included in the analysis should include the following seven areas: dependency on benefits, criminality, low voter turnout, weak connection to the labour market, weak school results, overcrowding and ill health.

2. DATA

For each of the seven areas listed above, we have, based on subject knowledge, found a register variable that provides a good description of the problem. For example, turnout in the most recent parliamentary election is used as a register variable for voter turnout and the number of reported crimes per 100 inhabitants as a register variable for crime. For school results, three different variables are weighted together to form the register variable for weak school results.

In addition to the registry variables for the seven problems, we also have found registry variables for a number of different conditions that are suspected to affect one or more of the problem variables. Such variables exist at municipality level (e.g. size, geographical location, and labour market in the municipality), at the residential area level (e.g. migration pattern and housing stock in the residential area) and for the population in the area (age, sex, country of birth, education, household type for the population in the area).

3. METHODOLOGY

The registry variables for the different problems have very different distributions (generally the distributions are very skewed) and in order to weigh the variables together without risking that the different distributions give any of the problems an unreasonably large weight, we standardize the variables by considering the percentiles for each variable. Moreover, for a

majority of the seven problem areas, there is no predetermined target value for the problem, except that the target for the problem in question is that it does not occur at all. This is another reason to study relative measures of the problem variables.

From the percentiles we can create the vector

$$J^{(k)} = (I_1^k, \dots, I_7^k),$$

where $I_i^k \in [1,100]$ denotes the percentile in residential area k for the registry variable connected with problem i . In the analysis, we focus on the 5% percent of the residential areas with the greatest challenges for each problem. In analogy with this, we define the vector

$$\hat{J}^{(k)} = (\chi_{[0,5]}(I_1^k), \dots, \chi_{[0,5]}(I_7^k)),$$

where $\chi_{[0,5]}(x)$ denotes a function taking value 1 if $x \in [0,5]$ and value 0 otherwise. A residential area with large challenges for each of the seven problems satisfies $\hat{J}^{(k)} = (1, \dots, 1)$, whereas a residential area with no challenges for any of the problems satisfies to $\hat{J}^{(k)} = (0, \dots, 0)$.

Based on the different values of the vector $\hat{J}^{(k)}$, we classify the residential areas into a number of classes, where the one corresponding to $\hat{J}^{(k)} = (0, \dots, 0)$ contain 83% of the residential areas in the country. A number of classes contain residential areas with only one specific problem. The areas in these classes are in general quite specific, for example including student areas (where the connection to the labour market is weak), residential areas in centres of large cities (where crime rates are high) or areas close to the country border (where voter turnout is low).

Moreover, we find a number of classes with large challenges for two or more areas. For a subset of these classes, containing the residential areas with the most general problems, we have investigated the effect of conditions (background variables) in the area. This investigation has been carried out in two different ways.

Firstly, we have investigated if the means of the background variables are significantly different in the classes with general problems compared to the class of residential areas where no problems are present. Secondly, we have carried out logistic regression analyses with the classes of residential areas with the most general problems as respondent variable and the background variables as covariates.

4. MAIN FINDINGS

Based on the analysis described above, we are able to identify residential areas with the largest general problems with respect to the seven problem areas identified. We note that the correlation between the areas identified using this methodology and the residential areas classified as area type 1 according to the socio-economic index is quite high, but that some aspects of the residential area are missed out when only considering the socio-economic index. Based on the investigation of means of background variables and the logistic regression analyses, we are able to identify the background variables that has the most power in explaining the risk for large challenges in a residential area.

A population map of Europe by age

Julien Gaffuri (Eurostat)

1. INTRODUCTION

Eurostat released the Census 2021 population grid on June 16th, 2024, in accordance with EU regulation 1799/2018 [1]. This new dataset provides a comprehensive overview of the European Union resident population on a 1km resolution grid. The dataset includes various indicators such as population by sex, age, current activity status, place of birth, and place of usual residence. An interactive map¹ showcasing these indicators is available on the Eurostat website, offering an efficient way to explore the dataset for different places and zoom levels.

2. METHODOLOGY

To complement the interactive map, we propose a printable A0 map poster highlighting key indicators from the dataset, focusing on total population by age group. The indicators of interest are:

- Total population
- Population by age groups: under 15 years, 15-64, 65 years and over.

Visualizing four indicators on a single map efficiently is a challenge. Guided by the principles of graphic semiotics [2], we employed **size** as a visual variable to represent total population and **colour hue** for age groups.

- **Size:** Each grid cell is represented by a circle, with its size proportional to its total population (Figure 1, left).
- **Colour hue:** Given the three age groups composing the total population, we opted for a ternary map legend [3,4] for colour hue (Figure 1, right). Seven classes are defined: three for predominant age groups (green, blue, and red), three for under-represented age groups, and a central class for balanced cases (grey colour). The central point corresponds to a balanced composition (15%, 60%, and 25%) for the respective age groups (under 15 years, 15-64, 65 years and over). While a more detailed classification with additional classes was possible, a simplified version was preferred to reduce complexity and enhance legibility.

¹ https://ec.europa.eu/assets/estat/E/E4/gisco/website/census_2021_grid_map/index.html

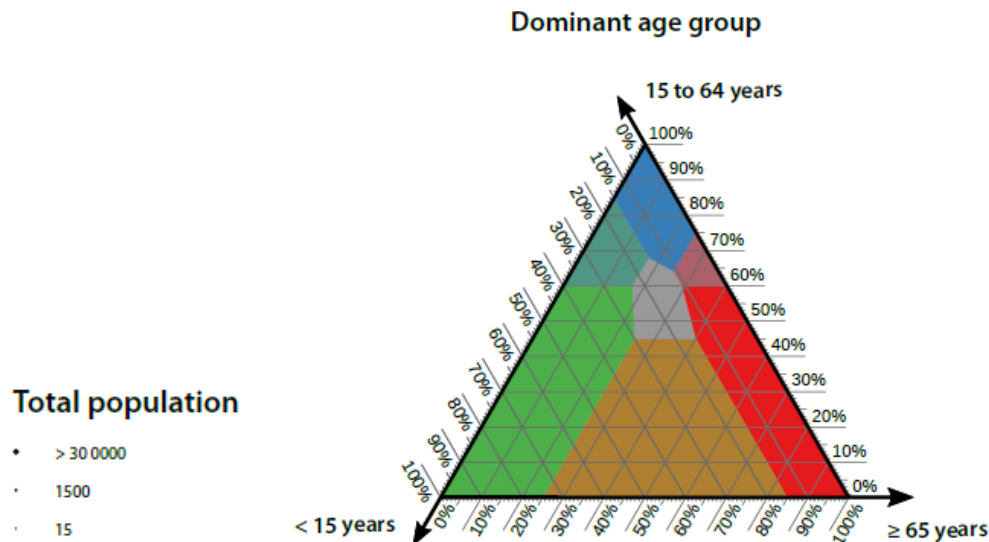


Figure 1. Map legend

The map size was limited to A0 format, with the requirement to show the entire European union territory, in ETRS89-LAEA map projection. The most detailed scale for this map was thus set to 1:4'500'000 (1cm for 45km). At this scale, a 1km square measures 0.22mm on the map, which is unfortunately far below legibility threshold. Consequently, the grid resolution was increased to 5km so that each grid cell can be depicted within a 1,1 mm square. This size is sufficient to show the circles and allow size variations to be represented.

Insets were added for the overseas territories covered in the dataset. Map labels showing main cities help readers to identify main population clusters. Only few labels were selected in order to keep the map uncluttered.

3. RESULT

The resulting map shows a truly striking image of the European union population by age. The overall situation within large regions can be easily assessed. Spatial patterns can easily be visualised.

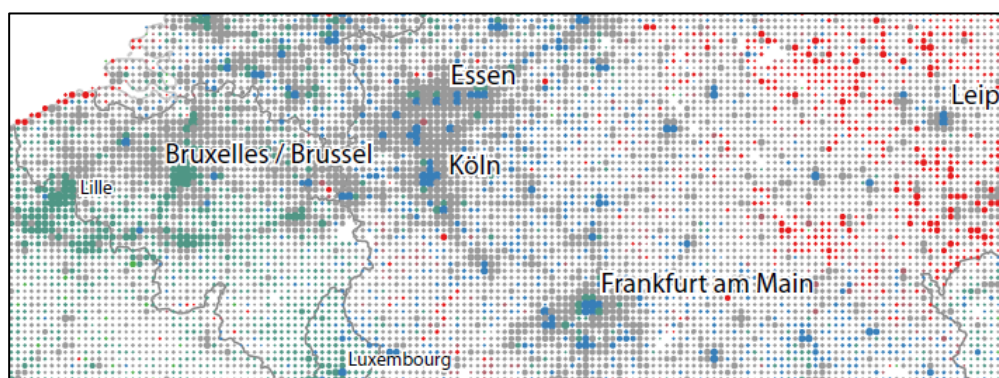


Figure 2. Map extract on Belgium and Germany

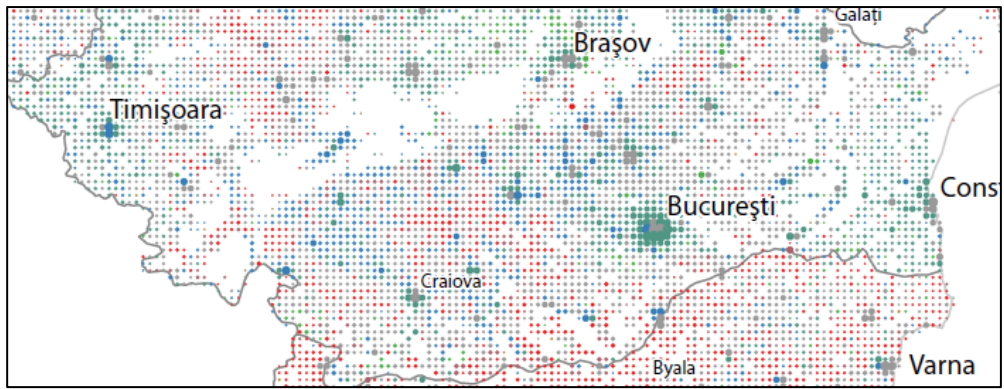


Figure 3. Map extract on Romania

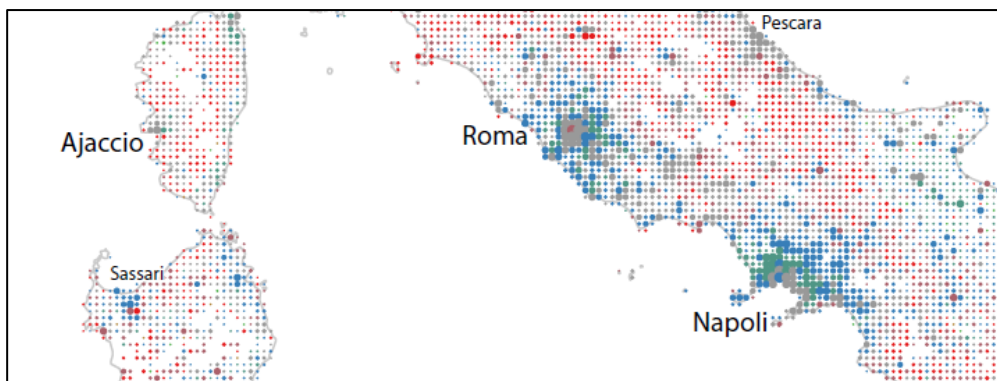


Figure 4. Map extract on Central Italy

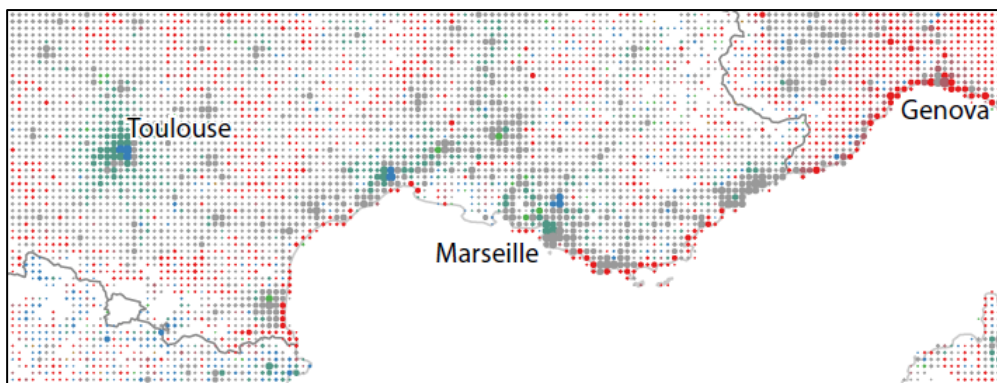


Figure 5. Map extract on the south of France

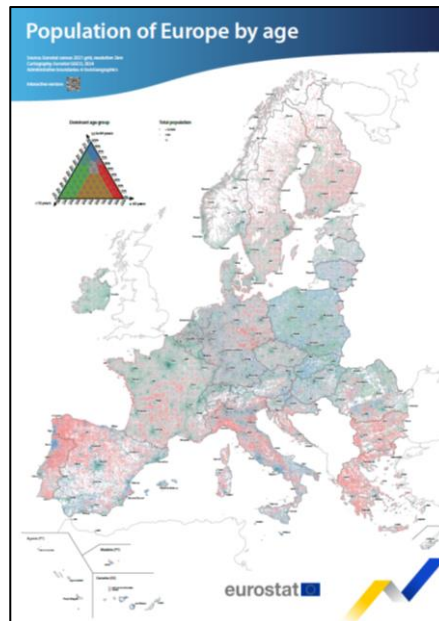


Figure 6. Map overview (original format is A0)

4. REFERENCES

- [1] Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km² grid. URL: https://eur-lex.europa.eu/eli/reg_impl/2018/1799/oj
- [2] J. Bertin, *Sémiologie graphique*, Paris, Mouton/Gauthier-Villars, 1967.
- [3] W.A. Deer, R.A. Howie, J. Zussman, *An Introduction to the Rock-Forming Minerals*. 1966. London: Longman
- [4] W. S. Cleveland, *Visualizing Data*, 1993, Hobart Press. ISBN 0-9634884-0-6.

A risk-utility framework for anonymized georeferenced health microdata

Lydia Jehmlich, Simon Cremer & Rainer Lenz (TH Köln)

1. INTRODUCTION

Whether for the use of targeted advertising measures or tracing the spatial spread of viruses such as the recent corona virus: geodata can - depending on the attributes it is provided with - hold enormous added value for society, science and research. However, the desired information can often not be extracted despite the inherent analytical content. The reason for this is that access to personal georeferenced datasets is severely restricted, as these are subject to statutory data protection. One way out of this dilemma is to apply a suitable anonymization method that guarantees data protection without significantly reducing the analytical validity of data. Georeferenced microdata has extraordinary potential for research and teaching, public administration and business. Key questions about the future and sustainability of our society can only be answered with high-quality and accessible geodata. Anonymization of health data is a major challenge due to the high degree of individuality of personal and household data. According to the GDPR (Art. 9), these data are considered particularly sensitive and worthy of protection.

The what is called k-anonymity provides information on the degree to which data sets can be re-identified by combining the quasi-identifiers. A data set is k-anonymous if each sequence of values appears with at least k occurrences in the data. The concept of k-anonymity can also be applied to spatial datasets. Spatial k-anonymity is the most widely used metric for measuring anonymity in the masking of sensitive geodata. It basically describes the number of masked points that are closer or equally close to the original location as the masked original point itself. Thus, at least k-1 masked points are closer to the original address than the associated masked point. As a result, the probability of a data attacker locating the original location is at most 1/k. Data anonymization is always accompanied by a reduction of information. If a specific research objective is already being pursued with the data to be anonymized, the method can be specifically adapted to this objective. Typical questions here deal with the population's access to education, medical care (specialists, hospitals, pharmacies, etc.) or cultural facilities.

2. METHODOLOGY

This section presents selected anonymization methods which have proven to be particularly suitable in the preselection for further studies on the anonymization of health data, as they offer scope for combinations or more specific adaptations. They can be roughly divided into two categories: aggregation and perturbation. In the following subsections those methods are presented, which – in the opinion of the authors - are particularly promising for anonymizing health microdata.

2.1. Data Aggregation

In principle, a large number of prominent clustering methods, especially hierarchical ones, can be used to aggregate geodata. This also includes any administrative area codes, which naturally represent a partitioning of the area. Aggregation is the most frequently used approach for masking geodata (e.g. see [2]). Applying methods of aggregation, the number of different data points decreases after masking. In other words, they are aggregated, which gives the method its name. Clustering can be carried out according to spatial, but also non-spatial criteria such as age or gender. Spatial clusters can be formed, for example, according to administrative units or using grid lines, where different strategies can be distinguished here. Most promising are grid masking and areal aggregation. With grid masking, all original data points are aligned to uniform grid cells [3]. The decision maker can select an appropriate size of the grid cells. A regularly selected grid cell has a side length of 100 m, 125 m or 250 m. Grid masking can further be divided into two different types of application. In grid line masking, the original points are moved to the nearest edge of the grid cell surrounding them. An actual aggregation occurs when several points have the same nearest grid line and are combined into one single point. The other type of application is called grid centroid masking. Here, all points in the interior of a cell are moved to the corresponding cell centroid. Instead of a large number of data points being reflected by one point, the surrounding area can also represent a cloud of data points. In area aggregation “an appropriate areal unit is defined, and then the grouped data of all or some of the cases located within that area are provided” [1]. One prominent example of this is the mapping of the corona incidence at district or federal state level as a choropleth map.

2.2. Data Perturbation

The most widely used approach for coding an original position is to apply a random perturbation. In contrast to aggregative methods, the number of data points generally remains the same for both deterministic and stochastic perturbation. The following data perturbing methods are considered to be particularly suitable for the anonymization of health data. Naturally, it is possible and even recommended to use the methods adaptively depending on the population density. Namely, we present the four approaches of Voronoi Masking, Donut Masking, Bimodal Gaussian Noise and Data Swapping.

Voronoi diagrams have been used in mathematics since the beginning of the 20th century. Seidl et al. (2015) were the first to use Voronoi diagrams in the context of masking geodata [3]. According to them, Voronoi polygons are placed around the points to be masked. Each point defines the center of gravity of its surrounding polygon. In addition, the polygons have the property that their boundary lines always run exactly in the middle of two points. After the polygons have been created, each point is placed on the nearest edge of its surrounding polygon. The donut method is a continuation of the random distribution within a circular region [2]. In addition to the circular region with the radius r_{max} , into which the point is randomly displaced, an inner ring with the radius r_{min} is defined around the original location, beyond which the displacement must extend. Thus, two radii, which resemble the shape of a donut, determine the minimum and maximum distance between which the new coordinates of the point have to lie. If, in addition to the torus shape, there are further restrictions with regard to the new placement, so-called "eaten donuts" may result. The application of a suitable bimodal Gaussian distribution is in some ways related to donut masking. Instead of the uniformly distributed probability, where the masked point is placed between some inner and the outer radius, a bimodal Gaussian distribution is used for the random distance function [4]. A distance d_i to the original point is chosen at random from a normal distribution with mean value p and standard deviation σ , resulting in masked coordinates (x_i^{mask}, y_i^{mask}) . Generally, different values of p are selected and several calculation runs are carried out so that an optimum result can be achieved. The displacement angle θ_i is also selected randomly, namely from the

uniform distribution over $[0, 2\pi]$. Hence, it follows: $(x_i^{mask}, y_i^{mask}) = (x_i^{orig} + d_i * \cos(\theta_i), y_i^{orig} + d_i * \sin(\theta_i))$ Data swapping encompasses various approaches. The conventional approach can also be applied to geodata. This involves swapping "records from one place to another, so that information from an individual with a certain set of key attributes is exchanged with the information from another individual, located in a different geographical area, but who matches the same attributes" [5].

3. RESULTS AND PRACTICAL APPLICATION

In order to be able to evaluate a method, the conflict of objectives between data anonymity and data quality must be taken into account. Although aggregation methods generally promise a higher level of data protection, the original data is heavily altered, resulting in a loss of spatial information. Another major problem with aggregation is the "modifiable areal unit problem (MAUP)". This describes the fact that aggregated data values are always dependent on the boundaries of the aggregated units.



Figure 1. MAUP using the example of hospital density

As can be seen in the example in Figure 1, boundaries are decisive in the evaluation of aggregation values. In the example on the left, there is a low (light gray) and high density (dark gray) of hospitals with horizontal boundaries. If the boundary is drawn vertically (right example), there is a medium density (medium gray) in addition to a high density. In terms of hospital construction, less need for action could be interpreted in the case on the right if the original data is not available. An administrative boundary does not usually represent a barrier for individuals in terms of access to health-related services in everyday life [6]. The advantage of Voronoi masking is that the displacement distance in densely populated areas is smaller than in sparsely populated areas. Thus, the pattern of anonymized points closely resembles the original distribution. This can lead to smaller displacement distances than with concealment techniques that do not take such patterns into account [3]. However, it must be critically questioned whether Voronoi masking can be meaningful at all in a complete data set, as re-identification would often be possible here or the spatial k -anonymity could in some areas reduce to $k = 2$. A larger k can then only be achieved if the procedure is run several times. The strength of data swapping obviously lies in maintaining the geographical distribution of points, while multivariate correlations can be affected depending on specific users' interests.

4. MAIN FINDINGS

As a whole, no method can be named independently as better or worse. However, they do have different strengths and weaknesses. In order to be able to evaluate methods, valid instruments must first be described so that anonymity and quality can be measured. This paper is intended to serve as a basis for a possible elaboration of these dimensions. Initial approaches are made as to how, for example, k -anonymity can be expanded with the concept of l -diversity. The overarching goal is to derive recommendations for action and tools for data

holders to generate research data with the highest possible analysis potential. A scientifically sound portfolio of methods can therefore only be developed in a discursive process involving all stakeholders in order to meet the strict data protection and diverse user requirements.

5. REFERENCES

- [1] M.P. Kwan, I. Casas, B.C. Schmitz (2004), Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica*, 39 (2), 15-28, DOI: <https://doi.org/10.3138/X204-4223-57MK-8273>
- [2] K.H. Hampton, M.K. Fitch, W.B. Allshouse, I.A. Doherty, D.C. Gesink, P-A. Leone, W.C. Miller (2010), Mapping Health Data: Improved Privacy Protection With Donut Method Geomasking, *American Journal of Epidemiology*, 172 (9), 1062–1069, DOI: <https://doi.org/10.1093/aje/kwq248>
- [3] D.E. Seidl, G. Paulus, P. Jankowski, M. Regenfelder (2015), Spatial obfuscation methods for privacy protection of household-level data, *Applied Geography*, 63, 253-263, DOI: <https://doi.org/10.1016/j.apgeog.2015.07.001>
- [4] C.A. Cassa, S.J. Grannis, J.M. Overhage, K.D. Mandl (2006), A Context-sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection, *Journal of the American Medical Informatics Association*, 13 (2), 160-165. DOI: <https://doi.org/10.1197/jamia.M1920>
- [5] M.P. Gutmann, K. Witkowski, C. Colyer, J. McFarland O'Rourke, J. McNally (2008), Providing Spatial Data for Secondary Analysis: Issues and Current Practices Relating to Confidentiality, *Population Research and Policy Review* 27 (6), 639-665. DOI: 10.1007/s11113-008-9095-4
- [6] D. Koller, D. Wohlrab, G. Sedlmeir, J. Augustin (2020), Geografische Ansätze in der Gesundheitsberichterstattung, *Bundesgesundheitsblatt*, 63, 1108–1117, DOI: <https://doi.org/10.1007/s00103-020-03208-6>

A spatial Evaluation of Essential Citizenship Services at municipal level. The contribution of Network Analysis and Machine Learning Methodologies

Angela Stefania Bergantino, Mario Intini (Department of Economics, Management and Business Law, University of Bari Aldo Moro) & Gianluca Monturano (Department of Economics, Management and Business Law, University of Bari Aldo Moro & Department of Economics, University of Modena and Reggio Emilia)

1. INTRODUCTION

This research focuses on evaluating the accessibility of essential services in Italian municipalities. Using advanced machine learning and network analysis techniques, we assess the current state of accessibility and predict future developments, aiming to provide valuable insights for improving the distribution of services across the country.

1.1. Literature Review and Research Context

Inequalities in access to services reflect broader socioeconomic gaps, affecting factors such as education, health outcomes, and economic opportunities. Key studies [1] highlight the crucial role of geographic accessibility in economic productivity and social cohesion. Rodriguez-Pose [2] and Barca [3] emphasize the importance of targeted regional strategies, while Florida [4] and Acemoglu and Autor [5] explore the dynamics between regional inequalities and technological changes. So far, most assessments of accessibility to services have relied predominantly on causality analyses using counterfactual econometric estimators. However, predictive approaches, less used so far, are becoming increasingly essential to anticipate future trends and dynamics. There is a growing need for data-driven methodologies that integrate complex systems to better predict changes in accessibility and guide policy making.

This project aims to contribute to the existing literature through a polycentric study of the Italian territory, overcoming traditional territorial dichotomies such as North-South and coast-hinterland, providing a more granular analysis of service accessibility [6]. Using unsupervised machine learning [7] and network analysis, we aim to identify underserved areas and develop innovative strategies to improve accessibility and territorial resilience.

Our approach promotes sustainable development and equitable access to services, thus contributing to more effective public policies and urban planning.

2. DATA, METHODS, AND EMPIRICAL STRATEGY

2.1. Data Collection and Spatial Mapping

The methodological process begins with a detailed mapping of Italian municipalities, aimed at identifying and locating essential public services such as schools, hospitals, and railway stations. We use geospatial data to determine the exact coordinates of these services across Italian municipalities, thus establishing the foundation for subsequent analyses.

2.2. Network Analysis with Spatial Calculation of Road Distances

We calculate the road distances between services using and iterating functions from the Open Source Routing Machine (OSRM) package developed within the R programming environment (Giraud & Lambert, 2018; Ferster et al., 2022).

The generic formula for calculating the distance in a graph, using and implementing OSRM, can be expressed with the following basic equation:

$$d_{street}(u, v) = \min(\sum_{i=1}^n l(e_i))$$

- where $d_{street}(u, v)$ represents a path in the graph between two services (nodes) u and v ; $l(e_i)$ the length of the arc; e_i the shortest path u, v .

Practically, through distance-network analysis, we calculate the road distances between the selected services for each Italian municipality (see Table 1 for an example based on simulated random data). Consequently:

- For municipalities with all services, the graph connects the distances between services within the same municipality (see the graph in Table 1 of Town1).
- For municipalities with at least one missing service, the remote network algorithm connects the incomplete municipality to the nearest municipality with the missing service, within 20 minutes, optimizing the shortest road route (see graphs in Table 1).

Table 1. Network Database

From	To	Distance	Duration)	Health	Education	Transport	Connection
Municipality1_hospital1	Municipality1_school1	0.5	2	1	1	0	Direct
Municipality1_hospital1	Municipality1_station1	0.6	3	1	0	1	Direct
Municipality1_hospital1	Municipality1_school2	0.7	2	1	1	0	Direct
Municipality1_school1	Municipality1_station1	0.3	1	0	1	1	Direct
Municipality1_school1	Municipality1_school2	0.4	1	0	2	0	Direct
Municipality1_station1	Municipality1_school2	0.8	2	0	1	1	Direct
Municipality2_school	Municipality2_station	0.6	1.5	0	1	1	Direct
Municipality4_service3	Municipality1_hospital1	2.5	5	1	0	0	Indirect
Municipality5	Municipality1_hospital1	3	6	1	0	1	Indirect
Municipality5	Municipality1_school1	3.2	6.5	0	1	1	Indirect
Municipality5	Municipality1_station1	3.4	7	0	0	1	Indirect
Municipality5	Municipality4_service3	4	8	1	0	0	Indirect

2.3. Machine Learning and Clustering for Service Accessibility

From the distance-network analysis, we derive municipal-level average engineering features, which we combine with granular socio-economic and institutional variables (e.g., per capita income, resident population, number of local businesses, mayor's educational background). We use these datasets to cluster Italian municipalities based on service accessibility. To achieve this, we employ unsupervised machine learning algorithms, which by design do not require labeled data. Specifically, we implement algorithms such as:

- **K-Means:** minimizes the variance within clusters. $S = \sum_{i=1}^k \sum_{x \in C_i} (x - \mu_i)^2$. Where μ_i is the centroid of the cluster C_i .
- **DBSCAN:** identifies clusters based on point density. $DBSCAN(D, \epsilon, MinPts)$. Where D is a dataset, ϵ it is the proximity radius, $MinPts$ is the minimum number of points in a neighborhood to form a dense region.
- **Hierarchical Clustering:** creates a hierarchy of clusters that can be visualized as a dendrogram. There is no single formula, as the method proceeds through levels of aggregation based on distance or similarity.
- **Spectral Clustering:** uses the spectral properties (eigenvalues) of the dataset's similarity matrix to reduce the dimensionality before clustering. $L = D^{\frac{1}{2}} - (D - W)D^{\frac{1}{2}}$. Where L è the normalized Laplacian matrix, D is the degree matrix and W is the adjacency matrix.
- **Mean Shift:** proceeds towards the data density peaks, based on the kernel density estimate $m(x) = \frac{\sum_{xi \in N(x)} K(xi - x)xi}{\sum_{xi \in N(x)} K(xi - x)}$. Where $m(x)$ is the mean shift vector, N it's the neighborhood of x , and K is a kernel.

3. PRELIMINARY RESULTS

The map in Figure 1, which applies the K-Means clustering algorithm with $k=3$ selected through the Elbow Method, divides Italian municipalities into three distinct clusters based on accessibility to citizenship services, such as schools, hospitals and transport infrastructure. From the analysis of the map, the clusters corresponding to the largest urban areas, such as Rome and Milan, are highlighted, which show a high density of services, indicating better accessibility. These urban clusters, densely populated and with a wide range of services, contrast markedly with those of more rural or less densely populated areas, where services are more scattered and less accessible. This geographical distinction highlights the importance of targeted policies that address the specific needs of different areas. For example, it may be necessary to improve the transport infrastructure in rural areas or increase the number of services in municipalities with less accessibility. Furthermore, it may also be appropriate to strengthen existing services in large cities to maintain and improve the effectiveness of access to services in the face of growing demand. In summary, the classification of municipalities into clusters based on the accessibility of services provides a useful guide for the development of public policies and urban planning, aiming to ensure fair and effective access to essential services for all citizens, regardless of their geographical location.

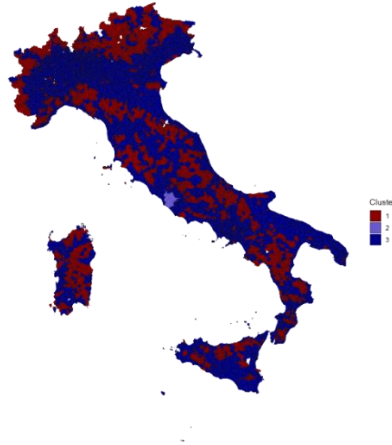


Figure 1. Cluster Analysis of Citizenship Service Accessibility

4. REFERENCES

- [1] Krugman, P. (1991). *Geography and Trade*. MIT Press.
- [2] Rodriguez-Pose, A. (2018). *The revenge of the places that don't matter (and what to do about it)*. *Cambridge Journal of Regions, Economy and Society*, 11(1), 189–209.
- [3] Barca, F. (2009). *An agenda for a reformed Cohesion Policy: A place-based approach to meeting European Union challenges and expectations*. Bruxelles.
- [4] Florida, R. (2002). *The Rise of the Creative Class*. Basic Books.
- [5] Acemoglu, D., & Autor, D. (2011). *Skills, tasks and technologies: Implications for employment and earnings*. In *Handbook of Labor Economics*, vol. 4, pp. 1043–1171.
- [6] Bergantino, A. S., Buonarota, M., Buongiorno, A., & Intini, M. (2023). *Regional multimodal accessibility: Policies and strategies for sustainable tourism destinations in coastal areas*. *Research in Transportation Business & Management*, 48, 100872.
- [7] Cerulli, G. (2023). *Optimal Policy Learning with Observational Data in Multi-Action Scenarios: Estimation, Risk Preference, and Potential Failures*. arXiv preprint arXiv:2403.20250.

A Statistical Analysis of Internet Standards Implementation on Dutch Enterprises' Websites

Eelco van Vliet & V.J.E.Z. Mastenbroek (Statistics Netherlands)

The use of modern and up-to-date Internet standards for websites is crucial for enhancing the safety, connectivity, and reliability of the internet. This research presents a methodology to measure the adoption of these standards by enterprise websites in the Netherlands. To achieve this, websites collected through the '*ICT Usage in Enterprises*' survey [3] are scanned and rated with a score between 0% and 100% using the Internet.nl testing tool [2]. This approach enables the calculation of a weighted average score for enterprises in the Netherlands, broken down by business size and economic activity. Additionally, insights are provided into the underlying test results. The findings reveal that the average test score has improved over the past three years, rising from 60.3% in October 2020 to 65.1% in April 2023.

1. INTRODUCTION

The adoption of modern and up-to-date Internet standards for websites is crucial for enhancing the safety, connectivity, and reliability of the internet. This paper presents a methodology to assess the implementation of these standards on enterprise websites in the Netherlands. A detailed explanation of the methodology is provided in Section 2, followed by key results in Section 3. Lastly, the main conclusions and recommendations are summarized in Section 4.

2. METHODOLOGY

This research is based on the Eurostat-coordinated survey '*ICT usage in enterprises*.' In this survey, companies are asked whether they have a website. In the Netherlands, around 75% of companies with two or more employees report having a website. Companies that confirm having a website are then asked to provide the URL. If a company operates multiple websites, the one with the highest traffic is requested. Through this process, a representative sample of enterprise websites in the Netherlands was obtained.

In the next step, all the collected URLs are scanned using the test tool [Internet.nl](https://internet.nl) [2]. Internet.nl evaluates the correct implementation of modern Internet standards that enhance the reliability of online services. A score of 100% indicates that a website, email service, or internet connection fully complies with the testing criteria. These criteria are based on the Internet Standards from the 'comply-or-explain' list of the Dutch Standardisation Forum, the security recommendations of the Dutch National Cyber Security Centre (NCSC), and the relevant RFCs (Request for Comments) issued by the Internet Engineering Task Force (IETF) [4]. In addition to the overall score, the results of all individual tests —such as the use of DNSSEC or HTTPS— are also provided.

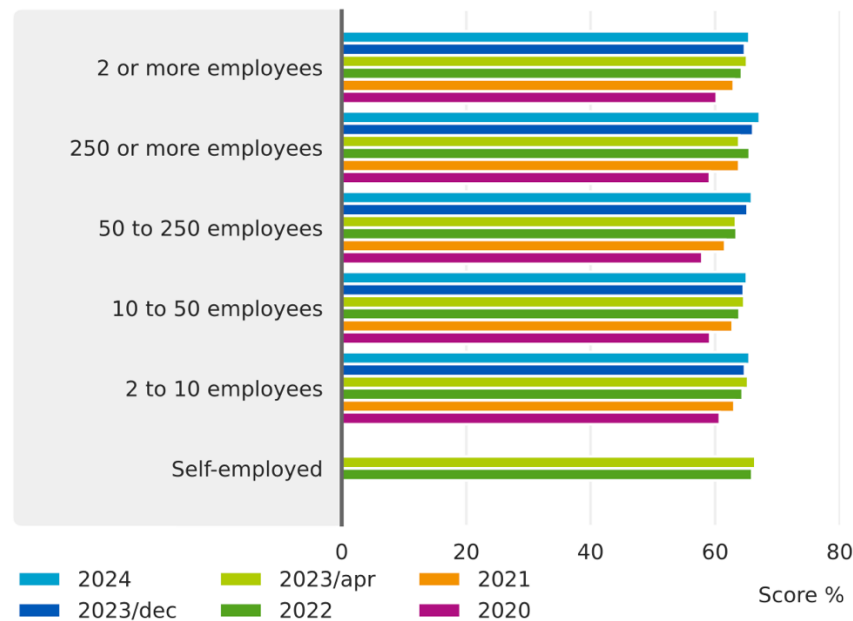


Figure 1. Final mean score per business size class

Once the sample URLs are scanned, a weighted average score is calculated for each business size category and industry (see Appendix A see more details on this calculation). The sample weights are based on the four-digit NACE classification and ten business size categories¹. This provides a comprehensive overview of the compliance levels of company websites in the Netherlands, categorized by business size and industry.

3. RESULTS

3.1. Final mean test score

Figure 1 shows the final scores of the websites of Dutch enterprises per size class over a period of three years. For each year, the scan is based on the same sample of URLs obtained from that year's the survey *ICT-usage in enterprises*. Firstly, it can be seen that a positive trend over the years is measured for all size classes. It can be seen that the average test score for companies with two or more employees has increased over the past three years from 60.3% in October 2023 to 65.1% in April 2023. Another observation is that the differences between the size classes are not very significant: both small and large companies have a similar final test score of around 60%.

3.2. Subtest scores: HSTS and DNSSEC

Although the final test scores do not differ significantly between small and large companies, Figure 2 reveals distinct variations in the underlying tests. Figure 2a shows the percentage of enterprises implementing an HSTS (HTTP Strict Transport Security) policy. HSTS is a web security mechanism that protects websites from cyberattacks, such as man-in-the-middle and protocol downgrade attacks, by ensuring that browsers only connect through secure HTTPS rather than insecure HTTP. The data indicates that large enterprises have a considerably higher adoption rate of HSTS compared to smaller enterprises.

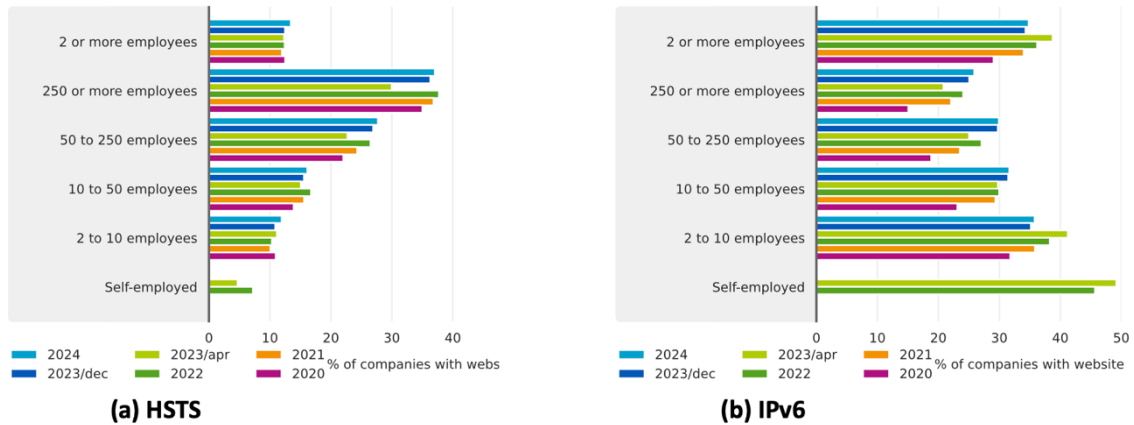


Figure 2. Percentage of enterprises per business size class succeeding for test if Web server offers HSTS policy (a) and if IPv6 is supported (b).

In contrast, Figure 2b shows that smaller enterprises perform better in supporting IPv6. IPv6 (Internet Protocol version 6) is the latest version of the Internet Protocol, responsible for routing and addressing data on the internet. It was introduced to replace IPv4, the previous version, due to the increasing demand for IP addresses and the limitations of IPv4's address space. IPv6 allows the internet to continue expanding and accommodate the growing number of devices and users globally. The reason smaller companies tend to have better IPv6 support is that they typically host their websites on external servers, rather than managing their own. Upgrading an internal server to support IPv6 can involve hardware replacements, which are costly and may not provide immediate security benefits. See [1] for an overview of all the test results.

3.3. Distribution of the final test scores

The test results from the previous section show that small and large companies perform differently on various subtests. However, these differences are averaged out in the overall final score, resulting in a relatively similar score for both groups.

Figure 3 presents the distribution of scores across all companies in the Netherlands. The same weighting procedure was used to calculate this distribution, giving more weight to smaller companies due to their higher weighting factors. The distribution reveals a wide range of scores, from near zero to a perfect 100%. This indicates that many low-scoring websites have significant room for improvement in complying with Internet standards. Continuous monitoring of these results is therefore essential.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1. Conclusions

In this study, a weighted average score for the adoption of Internet standards was assessed over a three-year period using the Internet.nl testing tool [2]. The results indicate an increase in the average score, from 60.3% in October 2020 to 65.1% in April 2023. While the overall score remains consistent across different company sizes, significant differences in the underlying subtests highlight areas for improvement within specific categories

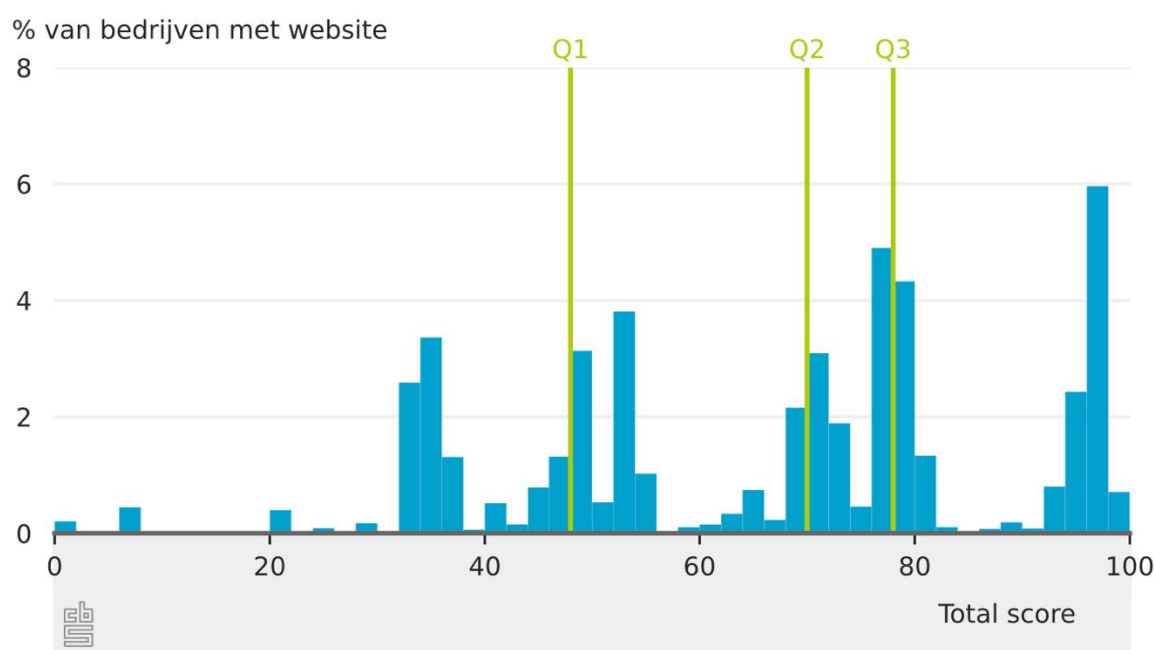


Figure 3. Final mean score per enterprise size class

4.2. Recommendations

This paper highlights the importance of measuring compliance with Internet standards. Effective monitoring can be established through the ICT survey, as URLs can be directly linked to enterprises, enabling statistical analysis. While this study focuses on Dutch enterprises, the same approach can easily be applied to other European countries participating in the survey. This would allow for cross-country comparisons of compliance levels within Europe and could inform improved policies and support efforts to enhance website adherence to Internet standards.

5. REFERENCES

- [1] CBS. [Toepassing van internetstandaarden voor websites van bedrijven](#), 2024.
- [2] ECP. [Platform Internetstandaarden](#), 2023.
- [3] Eurostat. [ICT usage in enterprises](#), 2024.
- [4] IEFT. [Internet Engineering Task Force](#), 2024.

A. WEIGHTED SAMPLE MEAN

The weighted sample mean is calculated using weighting factors W_j based on the sample size n_j and population size N_j for a stratum j , as follows:

$$W_j = \frac{N_j}{n_j}, \quad (1)$$

Assuming there are M defined strata, the total population of enterprises N is derived by summing the population of each stratum:

$$N = \sum_{j=1}^M N_j \quad (2)$$

Using this total population N , the normalized weighting factors can be expressed by:

$$w'_j = \frac{w_j}{N}, \quad (3)$$

By definition, this satisfies

$$\sum_{j=1}^M w'_j n_j = 1. \quad (4)$$

With the normalized weighting factors w'_j defined for each stratum j , the mean score $\bar{S}$ across all enterprises and strata can be calculated as:

$$\bar{S} = \sum_{j=1}^M w'_j \left(\sum_{i=1}^{n_j} S_{i,j} \right) \quad (5)$$

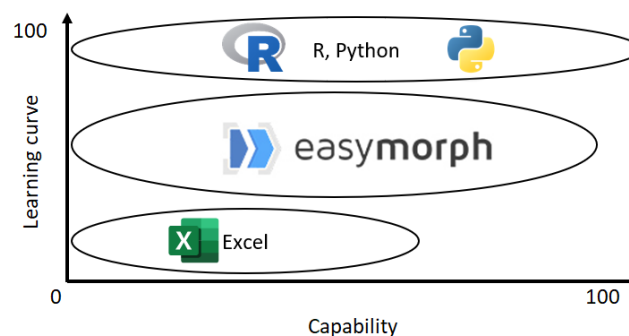
where $S_{i,j}$ represents the score of enterprise i within stratum j .

Accelerating Data-Driven Insights with EasyMorph's Intuitive Interface

Kam Chat (Novo Nordisk A/S)

1. INTRODUCTION

While specialized statistical software like R and Python are powerful tools for statisticians and data scientists, their steep learning curve often limits their accessibility to a wider range of users. Entry-level tools like Excel, on the other hand, may fall short when it comes to complex data manipulations. EasyMorph offers a compelling solution by providing a full-stack data platform that bridging this gap. Its visual interface and low-code approach empower users of all skill levels to perform sophisticated data analysis without the need for extensive programming knowledge. By lowering the barrier to entry, EasyMorph enables organizations to democratize data and foster a data-driven culture.



1.1. Redefining Olympic Dominance: A Population-Weighted Analysis of 2024 Medals

The 2024 Paris Olympics were the biggest sports event of the year. We now know the final medal count! Here are some of the top countries:

Order	NOCs	G	S	B	Total	
1	United States of America	40	44	42	126	+
2	People's Republic of China	40	27	24	91	+
3	Japan	20	12	13	45	+
4	Australia	18	19	16	53	+
5	France	16	26	22	64	+
6	Netherlands	15	7	12	34	+
7	Great Britain	14	22	29	65	+
8	Republic of Korea	13	9	10	32	+
9	Italy	12	13	15	40	+
10	Germany	12	13	8	33	+



Source: Reference [1] and [2]

But, how do these countries do if we look at how many medals they won compared to their population size? Which countries are the real medal powerhouses? And how does the whole European Union do as a team?

2. METHODOLOGY

Objective: To make a ranking of countries by total Olympic 2024 medal per capita. Treat the European Union as a country, add into this ranking.

Data Source:

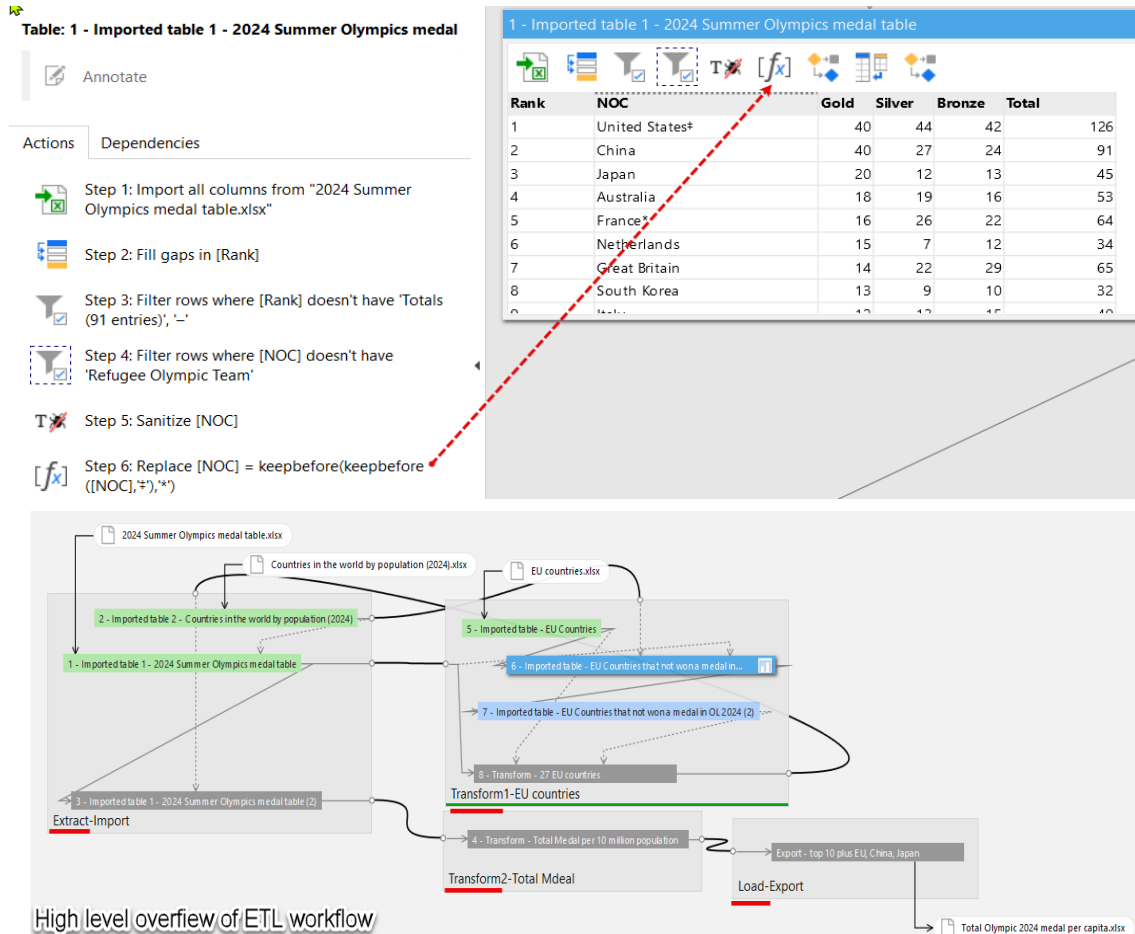
- 1) 2024 Summer Olympics medal table, reference [3].
- 2) Countries in the world by population, reference [4].
- 3) The European Union countries list, reference [5].

Extract these tables in Excel, save into local disk. Import data into EasyMorph.

Name	Date modified	Type	Size
 2024 Summer Olympics medal table.xlsx	08/12/2024 20:25	Microsoft Excel Worksheet	22 KB
 Countries in the world by population (2024).xlsx	08/12/2024 20:25	Microsoft Excel Worksheet	40 KB
 EU countries.xlsx	08/12/2024 20:25	Microsoft Excel Worksheet	11 KB

2.1. Workflow in EasyMorph

Use EasyMorph, a modern and visualized interface full-stack data tool to design an extract, transform and load (ETL) workflow.



3. RESULTS AND PRACTICAL APPLICATION

Rank of Total Medal per capita (per 10 million people) for the top 10 countries, the European Union, the United States, China, and Japan.

NOC	Total	Population in million	Total Medal per 10 million population	Rank	Rank by Total Medal
Grenada	2	0.12	170.64	1	80
Dominica	1	0.07	151.05	2	62
Saint Lucia	2	0.18	111.27	3	55
New Zealand	20	5.21	38.36	4	11
Bahrain	4	1.61	24.89	5	33
Jamaica	6	2.84	21.13	6	44
Australia	53	26.71	19.84	7	4
Hungary	19	9.68	19.64	8	14
Cape Verde	1	0.52	19.05	9	84
Netherlands	34	18.23	18.65	10	6
EU Countries	309	450.26	6.86	36	
United States	126	345.43	3.65	49	1
Japan	45	123.75	3.64	50	3
China	91	1419.32	0.64	77	2

Here is the key measure, to calculate medal per capita. Population set to 10 million for better human reading.

Rank	NOC	Gold	Silver	Bronze	Total	NOC2	Population (2024)	Rank by Total M...	Population in mill...	Total Medal per 10 million population
1	United States	40	44	43	126	United States	345426571	1	345.43	3.65
2	China	40	27	24	91	China	1419321278	2	1419.32	0.64
3	Japan	20	12	13	45	Japan	123753041	3	123.75	3.64
4	Australia	18	19	16	53	Australia	267142205	4	26.71	19.84
5	France	16	26	22	64	France	66548530	5	66.55	9.62
6	Netherlands	15	7	12	34	Netherlands	18228742	6	18.23	18.65
7	Great Britain	14	22	29	65	United Kingdom	69138192	7	69.14	9.40
8	South Korea	13	9	10	32	South Korea	51717590	8	51.72	6.19
9	Italy	12	13	15	40	Italy	59342867	9	59.34	6.74
10	Germany	12	13	8	33	Germany	84552242	10	84.55	3.90

Calculate new column(s)

Add one or more columns calculated as expressions. [More](#)

New column name

Population in million = [Population (2024)]/1000000

Total Medal per 10 million population = [Total]/[Population (2024)]*10000000

[Add more columns](#)

There are some data cleansing efforts needed for “2024 Summer Olympics medal table”. For example, “United States‡” and “France*” need to remove the last symbol, so that it only showing as “United States” and “France”.

There may be many values with such issue, so let’s ask Google Gemini AI how to handle this in Python. As per AI, you need to know regular expression, which is an advanced skill.



I'd be glad to help you remove those strange marks from your NOC column in Python. Here's a combined approach that leverages the strengths of both `pandas` and regular expressions:

Python

```
import pandas as pd
import re

# Load your Excel file into a pandas DataFrame
df = pd.read_excel("your_excel_file.xlsx")

# Define a regular expression pattern to match the strange marks
pattern = r'[\+\*]$\s' # Matches any of the marks at the end of the string

# Remove the marks using the 'str.replace' method with the regular expression
df['NOC'] = df['NOC'].str.replace(pattern, '', regex=True)

# Save the modified DataFrame back to Excel
df.to_excel("your_excel_file_cleaned.xlsx", index=False)
```

While you just need one simple action in EasyMorph. Modify column(s), and set it like this: NOC = keepbefore(keepbefore([NOC],'+'),'*').

Back to list

Modify column(s)

Replace values in one or more columns using expressions. [More](#)

Column to replace

NOC = keepbefore(keepbefore([NOC],'+'),'*')

Replace more columns

Remove Revert Apply

2024 Summer Olympics medal table No Label • Last Modified: Fri at 15:12

File Home Insert Draw Page Layout Formulas Data Review View

AutoSave On

https://sivantos-my.sharepoint.com

B1

NOC

	A	B	C	D	E	F
	Rank	NOC	Gold	Silver	Bronze	Total
1	1	United States+	40	44	42	126
2	2	China	40	27	24	91
3	3	Japan	20	12	13	45
4	4	Australia	18	19	16	53
5	5	France*	16	26	22	64
6	6	Netherlands	15	7	12	34
7	7	Great Britain	14	22	29	65

4. MAIN FINDINGS

- 1) Grenada, Dominica and Saint Lucia are top 3 countries measured by total medal per capita.
- 2) The top 3 countries by total medal count (United States, China, and Japan) rank only 49th, 77th, and 50th, respectively, when measured by total medal per capita.
- 3) Australia and Netherlands are the only two countries remain in top 10 in both ranking methods.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my exceptional manager, talented colleagues, and the entire management team for their invaluable support and encouragement in introducing EasyMorph to our company.

5. REFERENCES

- [1] <https://olympics.com/en/paris-2024/medals> Olympics Official 2024 Paris Medal
- [2] <https://www.facebook.com/EuropeanCommission/posts/hats-off-to-all-european-athletes-for-their-stellar-performance-at-the-paris2024/899177312247399/>, European Commission's post on Facebook
- [3] https://en.wikipedia.org/wiki/2024_Summer_Olympics_medal_table
- [4] <https://www.worldometers.info/world-population/population-by-country/>, Countries in the world by population
- [5] https://european-union.europa.eu/easy-read_en, the EU countries list

Accessibility, JSON-stat and SDMX-JSON: Simplifying structures and harmonising metadata for dissemination

Rolf Mach (Statistisches Bundesamt - Destatis)

1. INTRODUCTION

Our competence centre for development of our widely used dissemination database made a complete redesign of its web frontend. During the project we faced some interesting challenges in using RESTful Services with JSON-stat to optimize the accessibility of the dissemination products.

1.1. Problem definition

Historically, dissemination tables are quite complex for several good reasons: providing regional and hierarchical context, aligning values for coherence and explaining the meaning to our customers. Unfortunately, tables with all these features are too complex for handicapped customers and often also for the “standard JSON-stat”. Additionally, our departments for subject statistic were asked to replace printed products with online publications. This added more and more complex tables to the dissemination system, in which we emphasize the harmonization of metadata and delivery performance – a classical mismatch.

1.2. Overview

We provide an overview of the lacks of JSON-stat and SDMX-JSON for complex tables and our extensions to both, list the pain points for accessibility – regarding our legal requirements and some hidden problems coming from complex tables. Specifically, we discuss challenges of “squaring the circle” – delivering contextual information, coherent values and meaningful dissemination products with the requirements of speed, harmonized metadata and accessibility.

2. METHODOLOGY

We focus on the defacto standard JSON-stat and its fulfilment for the dissemination of complex tables. Also, in consideration will be SDMX-JSON 3.0.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Extension of JSON-stat

In the software development project, JSON-stat was extended by the following means:

- Elements for classification structures
- Dynamic sums
- Subordinate and secondary orders of classifications
- Alternative labelling for classification attributes
- Explanatory texts and definitions
- Bilingual texts

Plus, the division into 'data' and 'structure'.

These extensions serve exclusively to support the required technical structures.

3.2. Reduction of payload and improvement of performance

By usage of JSON-stat, payload could be reduced be 60 percent. Additionally, a cache for table structure and data was implemented which allows the delivery of dissemination products by far faster than the mean average times measured in 2023/24 for internet applications:

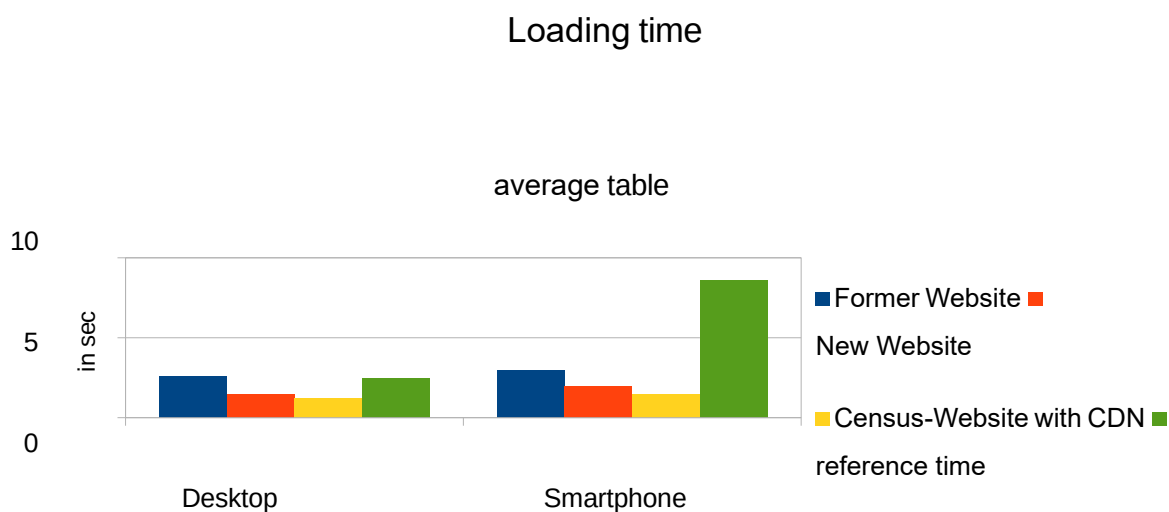


Figure 1. Loading for an average table compared to reference times 2023/24

(Reference 1)

4. MAIN FINDINGS

Defacto standard JSON-stat does not provide sufficient means for complex tables, but its provided extensions were able to fulfil the following goals:

- Significant reduction of payload by 60 %
- Definition of complex structures for required technical structures.

JSON-stat tools do not provide barrier-free publication functions in combination with hierarchical classifications, but our new implemented Javascript (Typescript) libraries match these requirements and additional display quite complex structures.

5. REFERENCES

- [1] Cassie Bottorff and Daniel Pöhler: „Website-Statistiken und -Trends 2023/2023“
<https://www.forbes.com/advisor/de/business/website-statistiken/>

Advancing Statistical Dissemination: Insights from the EMOS Coding Lab on Cultural Goods Prices in Europe

Sona Yavrumyan (University of Bologna)

1. INTRODUCTION

The increasing emphasis on dissemination and transparency of official statistics aligns with the broader objectives of the European data strategy. I took part in the EMOS Coding Lab 2024 organized by Eurostat, aimed to replicate Statistics Explained articles from Eurostat using reproducible methods in R. By leveraging the Eurostat API, I retrieved data, conducted calculations, reproduced statistical tables and visualizations on the article "[Consumer prices of cultural goods and services](#)" [1]. Emphasis was put on streamlining the dissemination process through literate programming, integrating code, data, and narrative text in a clear manner. I explored pricing mechanisms, demand dynamics, and efficiency considerations within the context of public economics, examining how these factors influence consumer behaviour and market outcomes in addition to the project purpose. I maximized the automation of the code which was noted as particularly useful during the final presentation by those who worked on manually updating the articles each year, so the project's goal of streamlining the dissemination of statistics was successfully achieved. My full work can be found on the [Eurostat GitHub Domain](#) [2].

2. METHODOLOGY

2.1. Data Collection

The data for this study was sourced from the Eurostat database using the *restatapi* [3] package in R [4], specifically focused on the Harmonised Index of Consumer Prices (HICP) dataset, identified by the code "prc_hicp_aind". This dataset provides annual HICP data from 2013 to 2023 for the European Union (EU), with an emphasis on the cultural goods and services sector.

The analysis concentrated on six key cultural categories, as defined by the Classification of Individual Consumption by Purpose (COICOP). These categories include: equipment for the reception, recording, and reproduction of sound and pictures, information processing equipment, recording media, cultural services, books, and newspapers and periodicals. The dataset included the annual average rate of change in prices for these goods and services, with the geographic scope restricted to the European Union, and Türkiye, Albania, North Macedonia, Montenegro, Kosovo, Iceland, Norway, Switzerland, Serbia [1] [2].

2.2. Data Transformation and Processing

The collected data was initially reshaped into a wide format to facilitate data cleaning and ensure consistency across COICOP categories. This process involved verifying the completeness of the data, aligning COICOP codes with their corresponding labels. Once cleaned, the dataset was transformed into a long format to enable detailed analysis. This format allowed for easier manipulation of the data, particularly in calculating average rates of

change across different time periods. Descriptive labels were assigned to the COICOP categories to improve the clarity and interpretability of the results.

To capture temporal dynamics and maintain the original structure of the article, two specific periods were defined for analysis: a medium-term period covering the five years from 2018 to 2023, and a short-term period focusing on the most recent year, 2022 to 2023. The average rate of change for each period was computed as the arithmetic mean of the annual percentage changes in prices across the relevant years. These averages allowed for the comparison between medium-term and short-term inflationary pressures in the cultural goods and services sector [1].

2.3. Data Analysis and Visualization

A comparative analysis was conducted on the average rates of change for the selected cultural goods and services through three visualization processes. A summary table was drawn to display the average percentage change for each category over the two defined periods: 2018–2023 and 2022–2023 including all mentioned countries (Table 1). A line chart (Figure 1) allowed for category-wise comparisons, and a horizontal bar chart (Figure 2) for a view of inflationary trend shifts.

3. RESULTS AND PRACTICAL APPLICATION

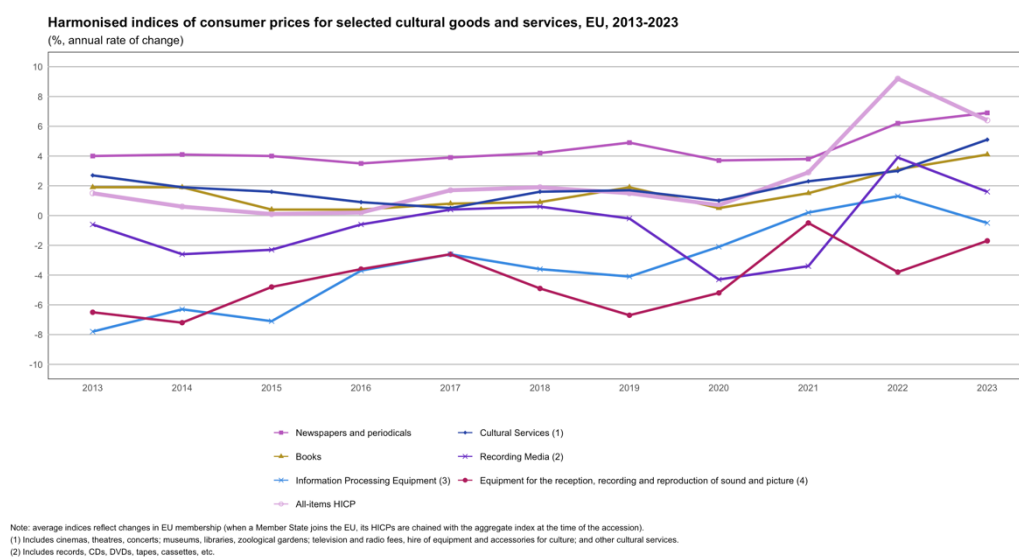


Figure 1. The all-items HICP and HICPs for cultural goods and services, EU, 2013-2023

eurostat

Prices of newspapers and periodicals have increased at a faster rate than other cultural goods tracked by the HICP. Over the past decade, prices for books and cultural services have risen modestly, while prices for recording media have generally decreased, with increases occurring only in 2017, 2022, and 2023. Similarly, the prices of information processing equipment and equipment for the reception, recording, and reproduction of sound and picture exhibited a comparable trend, with only slight increases noted in 2021 and 2022. These trends are detailed in Figure 1 [1].



Figure 2. Harmonised indices of consumer prices for selected cultural goods and services, EU, annual average



In 2023, the 'newspapers and periodicals' category recorded the highest price increase among cultural products at the EU level, with an average HICP increase of 5.1% per year, outpacing the all-items HICP by 1 percentage point. This indicates that the prices of newspapers and periodicals rose faster than headline inflation [1]. While the rates of change for 'cultural services' and 'books' were also positive, they remained lower than the all-items HICP. Among the other three groups of cultural products that experienced price declines, the most significant drop was seen in the 'equipment for the reception, recording, and reproduction of sound and picture' sub-index, which fell at an average annual rate of 3.6% between 2018 and 2023. From 2022 to 2023, this sub-index decreased by 1.7%, while prices for 'information processing equipment' declined by 0.5% [2]. Notably, of all the culture-related HICPs that increased, only the prices of newspapers and periodicals exceeded the growth rate of the all-items HICP (see Figure 2).

Table 1. The all-items HICP and HICPs for cultural goods and services, annual average rates of change 2018-2023 and 2022-2023 (HEADER ONLY)

	All-items HICP		Newspapers and periodicals		Cultural Services (1)		Books		Recording Media (2)		Information processing equipment (3)		Equipment for the reception, recording and reproduction of sound and picture (4)	
	2018-2023	2022-2023	2018-2023	2022-2023	2018-2023	2022-2023	2018-2023	2022-2023	2018-2023	2022-2023	2018-2023	2022-2023	2018-2023	2022-2023
EU	4.1	6.4	5.1	6.9	2.6	5.1	2.2	4.1	-0.5	1.6	-1.0	-0.5	-3.6	-1.7
Belgium	3.5	2.3	6.3	7.0	3.3	6.1	3.0	4.2	-1.5	-5.5	0.7	-1.3	-7.8	-12.5
Bulgaria	5.6	8.6	11.2	12.5	3.7	10.3	5.0	3.8	-3.2	-2.6	-2.4	-3.8	-9.0	-12.3
Czech Republic	7.2	12.0	6.5	12.2	4.2	7.1	4.3	6.0	1.5	5.3	1.4	3.8	-0.6	5.0
Denmark	3.0	3.4	4.9	7.0	3.5	4.8	7.7	3.3	0.0	3.8	-5.9	-1.2	-1.4	-0.8
Germany	3.9	6.0	5.2	6.6	2.0	4.2	2.8	5.5	1.6	4.1	-0.4	-2.5	-1.9	0.2
Estonia	6.9	9.1	6.3	6.4	8.0	12.8	2.4	6.6	-1.1	13.0	-8.3	-10.6	0.2	-4.4
Ireland	3.2	5.2	5.3	4.1	1.3	1.9	0.4	-4.5	2.6	4.5	-7.0	-2.6	-5.3	-1.4
Greece	2.7	4.2	0.6	2.5	1.0	3.8	-0.4	1.4	-2.0	0.3	-3.6	0.0	-6.6	-3.9
Spain	3.0	3.4	3.7	4.0	1.8	4.4	1.9	3.6	-1.8	3.9	-5.6	-6.7	-2.7	-2.8

Between 2018 and 2023, all EU countries experienced an increase in the prices of newspapers and periodicals, with the annual average rate of change surpassing headline inflation in 20 Member States. From 2022 to 2023 [2], this price growth outpaced inflation in 16 countries, notably in Hungary (21.0%) and Romania (19.6%). In contrast, the prices of cultural services rose more slowly than overall inflation in all EU Member States except Estonia and Denmark, while the average price of books fell in four countries, most significantly in Portugal. Meanwhile, the price index for recording media declined in 17 Member States but saw increases in 2023, particularly in Estonia and Sweden. Lastly, information processing equipment prices decreased in 16 countries over the same five-year period, with significant declines observed in Estonia and Spain [1].

4. MAIN FINDINGS

This project showcased the significant role of transparency and dissemination in the context of official statistics, aligning with the objectives outlined in the European data strategy. By using the Eurostat API and implementing reproducible methods through R, I streamlined the processes of data retrieval, analysis, and visualization, thereby enhancing the accessibility of public statistical information. The analysis reiterated the findings - significant trends in consumer prices for cultural goods and services: notably, the prices of newspapers and periodicals exhibited a sharp increase, surpassing the overall inflation rates across various EU Member States. Conversely, the growth in prices for cultural services was slower, while recording media prices experienced modest declines. These results highlight the varying market dynamics within the cultural sector and contribute to a more nuanced understanding of consumer behaviour and pricing mechanisms also in light of public economic policy [5]. Ultimately, this project facilitated improved clarity and dissemination of statistical data, benefiting users and stakeholders across Europe.

5. REFERENCES

- [1] Eurostat (2024). Consumer prices of cultural goods and services. Retrieved September 30, 2024, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Consumer_prices_of_cultural_goods_and_services
- [2] EMOS Coding Lab Group 3 (2024). HICP for cultural goods and services. In Eurostat, emos-codinglab-2024. GitHub. <https://github.com/eurostat/emos-codinglab-2024/tree/main/code/group3/HICP>
- [3] Eurostat. (n.d.). API - Getting started. European Commission Wiki. Retrieved September 30, 2024, from <https://wikis.ec.europa.eu/display/EUROSTATHELP/API+-+Getting+started>
- [4] Wickham, H. (2016). ggplot2: Create elegant data visualisations using the grammar of graphics. <https://ggplot2.tidyverse.org>
- [5] Eurostat. (2007). Handbook on data quality assessment methods and tools. Office for Official Publications of the European Communities. <https://ec.europa.eu/eurostat/documents/64157/4373903/05-Handbook-on-data-quality-assessment-methods-and-tools.pdf/c8bbb146-4d59-4a69-b7c4-218c43952214>

AI-powered nomenclature codification

Jeremy Lombaerts, Christophe Van Dijck (Sopra Steria), Youri Baeyens & Ann Depoorter (Statistics Belgium)

1. INTRODUCTION

The Belgian national statistical institute (NSI) has recently started to explore the use of artificial intelligence (AI) to address various challenges in the processing of data. As a first objective, we aimed to simplify the codification of free-text input in surveys against international nomenclatures such as ISCO, ISCED, ISCED_F and NACE. In an initial proof-of-concept (PoC) stage, we evaluated the use of generative AI (GenAI) to automatically determine the correct codes for survey input on the professional occupation of civilians against ISCO, the International Standard Classification of Occupations [1]. The current manual process, search-find lookup in nomenclature documentation, is time-consuming and error-prone due to the large amount of occupation codes (400+) that often only differentiate on nuanced differences between occupations.

The tool aims to streamline the process of identifying corresponding ISCO code(s) given free-text job descriptions acquired through surveys. Because Belgium has multiple national languages, the tool should be able to process multilingual descriptions, even in the presence of spelling and grammar errors. Because of the wide variability in input descriptions, the tool should not rely on predefined keywords to establish a direct link with the nomenclature. Within the context of this proof-of-concept, we aim to develop a tool that empowers codification agents with mastery over the entire nomenclature, including rarely used codes. In the future, an enhanced version could be employed directly in the field: a dialogue among the respondent, the investigator, and the AI could refine the respondent's description until it uniquely matches a code from the nomenclature.

GenAI, specifically large language models (LLMs), have been used to classify free-text input. One approach is to use so-called zero or few-shot learning methods where the classification task is included in the prompt instruction [2], [3], [4]. However, these techniques are limited to small to medium sized class-sets and struggle with nuanced differentiation as the number of classes increases. This challenge is compounded when the distinctions between classes are subtle, often leading to misclassifications without sufficient examples. Alternatively, pre-trained language models can be finetuned to behave as classifiers [5], [6], but this approach quickly becomes costly and is difficult to scale across multiple nomenclatures, requiring substantial computational resources and extensive labeled data for each class. Therefore, we set out to develop a method that optimally utilizes the information stored in the nomenclature (description and examples of codes) without the need to train dedicated language models.

2. METHODOLOGY

Tooling

The NSI prioritized the use of open-source tools to ensure transparency in the research phase and to provide flexibility and accessibility in subsequent deployment phases, allowing easy adaptation to different deployment environments or changes in technology. The implementation of algorithms and UI was done using Python 3.11, in combination with open-

source third-party packages such as Numpy, Pandas, and Langchain. For nomenclature storage and retrieval, we used a ChromaDB vector database and to create the single-page web application demonstrator Streamlit was used. An Azure OpenAI service was used to deploy and run different AI models, including the LLMs as well as an embedding model, to maximize implementation speed and flexibility during development.

Nomenclature

The ISCO nomenclature was obtained as an Excel file [1]. The nomenclature describes the different occupation codes, each with a description and several associated keywords. The Excel file was parsed into a JSON format for easier querying and manipulation, paying close attention to the structure of the nomenclature to aid in the retrieval process by retaining the hierarchical structure of the nomenclature, adding exclusion criteria and adding multilingual keywords.

The nomenclature data was then converted into a ChromaDB vector database by embedding the description of each code using the Azure OpenAI *text-embedding-3-large* embedding model.

Methods and algorithms

Our method for codification consists of a semantic retrieval step, followed by a semantic evaluation step. In the first retrieval step, a semantic search is performed to retrieve a fixed number of codes that best match the input job description. The semantic search retrieves entries in the nomenclature vector database that are the closest to the embedded job description. Different parameters were evaluated, and we found that a top 5 search under cosine-similarity provided the best balance between success rate and computational cost and efficiency.

A semantic search will always return the top results for a specific input job description. To remove incorrect codes and select the best matching one(s), a final semantic evaluation step was applied. In this step, the match between the job description and the retrieved code is evaluated by scoring each code based on its description, associated keywords, and optional exclusion criteria. Evaluation and scoring are done by instructing an LLM to score the match between a code and the input job description against a scoring table. This table rewards a code the highest score if the job description fully matches the code's description and one or multiple keywords and assigns a low score if the job description matches one of the code's exclusion criteria. Techniques such as chain-of-thought and structured output parsing have been applied to maximize the accuracy and consistency of the LLM responses.

Validation and evaluation

An initial codification test dataset was provided by the NSI, containing roughly 160,000 manual codifications, each consisting of a description and an assigned ISCO code. These texts, mostly written in Dutch or French and sometimes English or German, varied in quality with some being very short and non-descriptive, some containing spelling and/or grammatical errors, and some matching incorrect outputs.

From this dataset, we created a curated subset of 844 manually reviewed codifications to serve as a validation set. This subset included exactly one Dutch and one French instance for each of the 436 ISCO codes, except for codes absent from the initial dataset. While this sample ensured a uniform representation of ISCO codes, it should be noted that this sample is not representative of the overall population due to the varying frequency of different occupations (e.g. office cleaners are much more common than deep aircraft maintainers).

The performance of the proposed method was quantified as the success rate, defined as instances where one of the codes returned by the tool matched with the target code. This approach accommodates ambiguous inputs or closely related outputs (e.g., '*Accountant 2411*'

vs. 'Accounting Assistant 3313'), preventing unfair penalization of the pipeline against ambiguous inputs. In addition, computational time and cost were calculated to evaluate the real-world applicability of GenAI for nomenclature codification. Different LLM models were considered during the validation: *GPT-3.5 Turbo*, *GPT-4*, *GPT-4o*, and *GPT-4o-mini*. However, *GPT-4* was not fully validated due to cost considerations.

An upcoming end-user validation phase aims to gather feedback and assess the system's usability in real-world scenarios, refining the system to better meet the needs of day-to-day codification tasks. To this end, a demonstrator was deployed on Azure Container Instances, allowing users to input free-form text and receive ISCO code suggestions with argumentation.

3. RESULTS AND PRACTICAL APPLICATION

Table 1. Validation results

Model	Success rate	Runtime	Cost / 1000 codes	Notes
GPT-3.5 Turbo	74%	4.0 s	€ 10,25	High accuracy, but also high recall, and returns too many codes.
GPT-4o-mini	72%	1.6 s	€ 1,13	Good trade-off between speed, cost, and performance.
GPT-4o	67%	4.3 s	€ 36,64	Slightly lower accuracy, slightly lower recall, less codes returned.
GPT-4	-	8.4 s	€ 198,17	Not properly validated due to unreasonably high cost, but 'feels' like the best model.

These results are encouraging but warrant a nuanced interpretation. They underscore the necessity for further improvements, particularly given the persistence of low-quality inputs, codification errors, and the occurrence of rarely used codes within the validation set. These issues not only skew the results but also further emphasize the need for refinement. Key areas of improvement include enhancing the accuracy and precision of the document retrieval step, which could significantly improve the overall performance of the model and developing a more robust method to identify poorly described inputs. Currently, the model often still attempts to return a code even when the input quality is poor, introducing potential inaccuracies or misinterpretations. By bolstering the model's ability to detect and appropriately filter out such inputs prior to codification, we could improve the quality and reliability of the results (no output is much better than wrong output!).

3.1. Main Findings

Discussion

The initial results of this proof-of-concept are both promising and encouraging, demonstrating the potential of AI to significantly enhance the codification process. The model showcases its flexibility by accommodating various languages, handling poorly formulated descriptions, and understanding different word forms and relationships. Additionally, its ability to compensate for encoding weaknesses contributes to the improvement of codification quality, particularly for rarely used codes. This empowers codification agents with mastery over the entire nomenclature, instead of relying solely on knowledge of common codes. Furthermore, the AI model's potential to be used for direct field codification presents an opportunity to refine low

quality descriptions immediately, further improving the accuracy and efficiency of codification work.

Future Work

The solution developed in this study is a proof-of-concept, aimed at demonstrating the potential utility of AI in improving codification work. It should be clarified that the developed tool is a demonstrator. As such, it serves to illustrate the concept and does not encompass the full range of features that a fully developed product should have.

In the next phase, day-to-day codification agents will get the opportunity to test this tool. Their hands-on experiences and feedback will be invaluable in identifying practical issues, potential enhancements, and additional features that might be needed.

The feedback and insights gained from the testing phase will be translated into requirements for potential future development towards a production-ready software tool. As the tool evolves and improves, it will continue to accommodate more precise descriptions and better handle ambiguities, thereby improving the quality of the results and the quality and consistency of codification in general.

Considerations

Despite these promising developments, there are critical points of attention that need addressing:

- 1) **Model Certainty:** Future developments will aim to enhance the model's confidence in its results, potentially by assigning a probability to each selected code.
- 2) **Public Perception:** As AI is not always perceived positively by the public, future work will need to address these concerns, emphasizing the role of AI to improve codification quality rather than being a threat to jobs.
- 3) **Carbon Footprint:** Given the significant CO₂ emissions associated with digital technology, future developments will need to consider the environmental impact, potentially opting for less carbon-intensive models, where a trade-off needs to be made between accuracy and ecology.

Conclusion

The Initial results of this PoC are encouraging. The tool has demonstrated its capability to enhance the quality of codification tasks, particularly for rarely encountered codes. It empowers codification agents with a comprehensive understanding of the entire nomenclature. The proposed approach allows for a judicious and conscious use of AI. It highlights the importance of maintaining a balance between achieving optimal results and the responsible use of AI, considering factors such as computational cost and ecological footprint.

Technological advancements are progressing rapidly, requiring continuous monitoring of developments and flexible architectures able to accommodate new models and advancements. Furthermore, these advancements are paving the way to make fine-tuning LLM's with the nomenclature economically viable, potentially removing the need to provide the nomenclature directly through prompt engineering.

In conclusion, the PoC demonstrated promising potential for the application of AI in improving the codification processes. Future work will focus on refining the model, fine-tuning it for better performance, and integrating user feedback for more practical and user-friendly implementations.

4. REFERENCES

- [1] 'International Standard Classification of Occupations (ISCO)', *International Labour Organization*. [Online]. Available: <https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/>
- [2] S. Parikh, Q. Vohra, P. Tumbade, and M. Tiwari, 'Exploring Zero and Few-shot Techniques for Intent Classification', May 11, 2023, *arXiv*: arXiv:2305.07157. Accessed: Sep. 25, 2024. [Online]. Available: <http://arxiv.org/abs/2305.07157>
- [3] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, 'Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification', Apr. 18, 2023, *arXiv*: arXiv:2303.07142. Accessed: Sep. 25, 2024. [Online]. Available: <http://arxiv.org/abs/2303.07142>
- [4] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, 'TabLLM: Few-shot Classification of Tabular Data with Large Language Models', Mar. 17, 2023, *arXiv*: arXiv:2210.10723. Accessed: Sep. 25, 2024. [Online]. Available: <http://arxiv.org/abs/2210.10723>
- [5] H. Yu, Z. Yang, K. Pelrine, J. F. Godbout, and R. Rabbany, 'Open, Closed, or Small Language Models for Text Classification?', Aug. 19, 2023, *arXiv*: arXiv:2308.10092. Accessed: Sep. 25, 2024. [Online]. Available: <http://arxiv.org/abs/2308.10092>
- [6] Y. Zhang *et al.*, 'Pushing The Limit of LLM Capacity for Text Classification', Feb. 16, 2024, *arXiv*: arXiv:2402.07470. Accessed: Sep. 25, 2024. [Online]. Available: <http://arxiv.org/abs/2402.07470>

Appendix A: Example of a Dialogue with AI

This brief example illustrates what a dialogue between the respondent, the investigator, and the AI might look like. Let's suppose that, when talking to the investigator, the respondent gives this description of their professional occupation:

“Je travaille pour un institut de statistique. J'exploite des banques de données spatiales pour réaliser toutes sortes d'études.”

Note that the description is given in French, while the official nomenclature is available only in English. Regardless, the language model can seamlessly make the link and proposes 3 codes with a brief justification for the choices:

3314 (8/10): The user's job involves working for a statistics institute and exploiting spatial databases to conduct various studies, which closely aligns with the tasks and responsibilities outlined in the job description for code 3314. This includes planning and performing statistical calculations and data analysis.

2120 (8/10): The user's job description mentions working for a statistics institute and exploiting spatial databases to conduct various studies, which aligns well with the tasks of evaluating, processing, analyzing, and interpreting statistical data as described in the code. Although the description is general, it corresponds well to the responsibilities of a statistician.

2529 (6/10): The user's job involves working with spatial databases, which loosely relates to the tasks of database professionals. Although not explicitly mentioned in the keywords or inclusions, the general responsibilities of data handling and analysis align with the tasks outlined for code 2529. Therefore, it scores a 6 for being related but not directly classified.

At this stage, the AI hesitates between three codes. The level of responsibility appears to be important for the conclusion, as well as the distinction between the roles of a statistician and a computer scientist. The investigator, therefore, asks the respondent to specify whether they are a statistician or a computer scientist and to clarify their level of responsibility. The response might be:

“Je travaille pour un institut de statistique. J'exploite des banques de données spatiales pour réaliser toutes sortes d'études. Je suis statisticien et je coordonne le travail de l'équipe.”

The AI can then conclude with certainty that the correct code is 2120:

2120 (9/10): The user's job description mentions that they are a 'statisticien' (statistician), which is directly included in the job titles under the inclusions section. They also coordinate a team and conduct studies using statistical data, aligning well with the responsibilities and tasks outlined in the code description.

Appendix B: Knowledge of Obscure Codes

This second example shows the potential of AI assisted codification. Suppose a respondent gives as description of their professional occupation:

“I work on an off-shore oil rig as a welder where I weld both on the platform and underwater.”

A codification agent might classify the input under code “7212 Welders and Flame Cutters” because it is a code they are probably familiar with and also the only code that appears when naively searching for ‘welder’. But equally accurate could be “7541 Underwater Divers” because the keywords of this code mention “Diver (construction and maintenance of structures)”. It is then the codification agent's job to ask follow-up question if possible and determine whether to put emphasis on the welding aspect of the job or the underwater aspect.

AI Solutions in Istat Contact Centre for Knowledge Seekers, Data Users and Survey Respondents

Paola Bosso, Gabriella Fazzi & Roberta Roncati (Istat)

1. INTRODUCTION

Istat serves two types of users: “Knowledge Seekers,” individuals who seek information, and “Respondents,” those who are involved in surveys. Both user groups may occasionally need to contact Istat, either because they cannot find the information they are looking for or because they need reassurance regarding the survey process. Since July 2024, Istat has introduced a unified Contact Platform, allowing all users to connect through a single point of access.

Understanding the need for timely and efficient access to information, Istat prioritizes helping users find answers without requiring direct assistance. Users generally prefer resolving issues independently, and when they can do so seamlessly, their satisfaction increases. This often leads to greater engagement with statistical data and higher response rates for surveys. However, when users are unable to find the information they need, they are more likely to seek help. To address this, Istat has introduced tools that enable users to resolve common issues without having to contact support staff, employing a strategy called “case deflection.” This method reduces the number of support requests and prevents potential user frustration, ultimately enhancing the overall user experience.

To improve case deflection, Istat has integrated a comprehensive knowledge base into the Contact platform. This knowledge base offers quick answers, detailed articles, and procedural guides to help users find solutions to frequently asked questions.

For Knowledge Seekers, the content in the knowledge base is curated by frontline operators who are well-versed in the recurring questions and issues users face. These operators, having firsthand experience in responding to users, are best positioned to identify solutions to common inquiries. As a result, the knowledge base empowers users to explore available resources, enhancing their self-sufficiency.

For Respondents, the knowledge base is supported by a robust database of questions and answers collected since 2018. This data is analyzed using supervised machine learning techniques. The process consists of three main phases:

- Training: A classification model is built and refined using labeled tickets, which contain data about previous user interactions.
- Testing: The model’s performance is evaluated by testing it against a portion of the labeled dataset to ensure accuracy.
- Classification: The trained model is applied to new, unlabeled tickets, allowing Istat to categorize recurring issues and link them with relevant FAQ entries or knowledge base articles.

This process ensures that both Knowledge Seekers and Respondents receive accurate, relevant answers to their questions, without needing to contact an operator. By enabling users to solve problems on their own, Istat significantly reduces the workload on its operators, freeing up resources for other critical projects and initiatives. The overarching goal is to foster a culture of knowledge sharing, leveraging tools like collaborative support systems and artificial intelligence to enhance both user satisfaction and operator productivity.

2. METHODOLOGY

To further streamline navigation within the Contact Centre and improve the user experience, Istat added a “Search Among Services” feature to the homepage. This search bar allows users to quickly locate the services they need by typing common terms. However, the search bar alone is insufficient, as it does not provide real-time interaction with the platform's users.

To address this gap, Istat introduced a generative AI solution designed to offer personalized guidance based on user inquiries. This solution leverages machine learning techniques to analyze unstructured data, allowing the platform to provide tailored assistance. Istat developed and launched a chatbot—a virtual assistant that interacts with users, providing standardized responses, documents, and links to relevant services. The chatbot is powered by Salesforce's AI platform, "Einstein," and is programmed to analyze conversations, identify user needs, and either offer specific answers or direct users to the appropriate human operator when necessary.

The development of the chatbot focused on two primary areas:

- **Training:** The chatbot's responses were enhanced through the careful input of “dialogue intents,” which are designed to improve the accuracy and reliability of its responses.
- **Knowledge Cards:** Articles or “knowledge cards” were implemented to enable the chatbot to offer users detailed informational content without requiring them to submit a support ticket.

The chatbot primarily handles first-level support tasks but can escalate more complex cases to human operators when required. It is also integrated with web sessions, enabling it to collect relevant information from users and provide a seamless multichannel support experience.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Enhancing Operator Efficiency

In addition to providing users with accurate and timely responses, the system adapts operator support based on user needs, ensuring that responses are always current and consistent. The introduction of predictive AI systems further aids operators by streamlining the process of managing support tickets. Operators no longer need to search across multiple systems, tabs, windows, and documents to resolve user queries. Instead, they are equipped with tools that enable them to offer accurate, timely, and coordinated responses.

This ambitious project involves multiple aspects of statistical research and data science, all aimed at improving Istat's understanding of user queries and integrating information from various sources. The development of statistical information shared through the Contact Centre is essentially a capacity-building exercise—one that focuses on providing operators with the tools and resources necessary to meet user demands effectively and efficiently. While the investment in this system is significant, the improvements in time savings and service quality are expected to deliver considerable long-term benefits.

3.2. Efficiency and Impact

The integration of AI into Istat's Contact Centre brings several important benefits:

- **Personalized Assistance:** AI allows Istat to offer tailored responses based on user needs, enhancing the overall support experience.
- **Improved Timeliness:** Responses are provided more quickly, reducing the time users spend waiting for assistance.

- **Higher User Satisfaction:** When users can resolve issues independently or with minimal delay, their satisfaction with Istat's services increases.
- **Increased Efficiency:** AI systems streamline the management of support requests, allowing operators to handle inquiries more effectively.
- **Reduced Operational Costs:** By reducing the need for human intervention in routine cases, Istat can lower the costs associated with operating its Contact Centre.

Two primary metrics are used to evaluate the impact of AI on Istat's operations: case deflection and time savings. Case deflection measures the reduction in support requests requiring human intervention. Time savings reflect the reduction in effort required by operators to handle tickets. These improvements translate into significant cost savings and efficiency gains over time.

By automating key interaction processes and applying AI across multiple communication channels, Istat has not only enhanced the user experience but also optimized its internal operations. Machine learning algorithms play a key role in managing frequent requests, reducing wait times, and further improving the efficiency of the support system.

In conclusion, Istat's integration of AI and the enhancement of its Contact Centre is a forward-thinking approach to modernizing user support. This initiative underscores the importance of innovation in public service, demonstrating how technology can improve both operational efficiency and the quality of services provided to the public. Through continued investment in AI and data-driven solutions, Istat is well-positioned to meet the evolving needs of its users while ensuring that its Contact Centre remains a valuable resource for years to come.

AM3DS: A new service to integrate administrative data in the French statistical information system, driven by metadata

Olivier Haag (Insee)

1. INTRODUCTION

1.1. Context: The implementation of statistical registers based on administrative data

Insee (French NSI) launched a project called Resil (from the French: Statistical Register of individual and dwellings) in 2020. The aim of this project is to build register from administrative data in order to produce frame and to facilitate the production of statistics from external sources.

The first step of this project was the development of a shared integration system of administrative data called AM3DS (from the French: reception and provision of administrative data).

1.2. Why was this new tool developed?

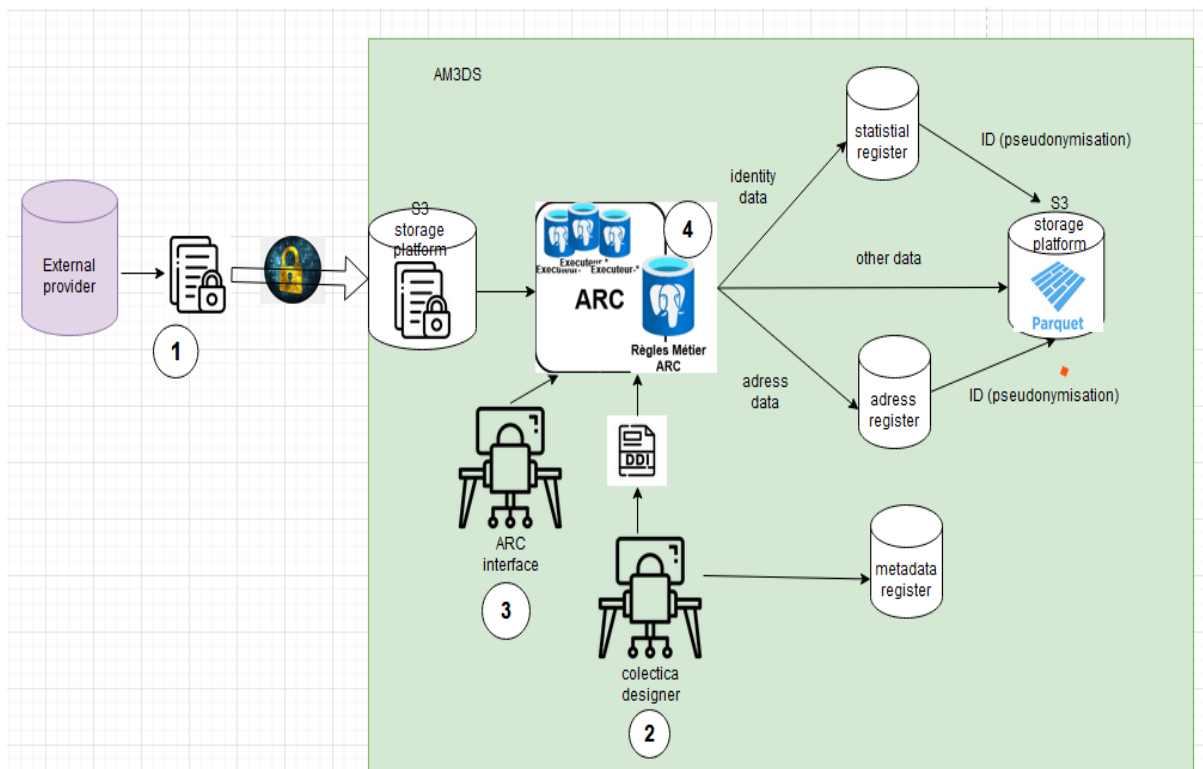
The aim is to implement a general data integration framework based on an automated approach to structured data delivered by an external producer. This involves setting up a pipeline with checkpoints to ensure that the succession of tasks (renaming, data restructuring, recoding, pseudonymisation, etc.) is carried out correctly and to stop the process as soon as any problems are encountered. In addition, the use of standards and active metadata upstream of this pipeline allows the designer to be as autonomous as possible, making it easier to adapt to changes in external sources.

This tool is based on the software ARC (from the French: Acquisition - Réception – Contrôle) already developed by Insee for a specific administrative source and extended to other data frames. Besides, the application became more efficient and robust, and a number of new functions have been added, the main one being to connect ARC to the INSEE metadata register: ARC is connected to Colectica designer by DDI files in order to be driven by metadata.

This tool provides users with raw statistical data. It makes it possible to desynchronise the collection and processing of administrative data.

2. METHODOLOGY

Figure 1. The workflow of the service



- 1) The external producer sends to Insee the encrypted data and associated metadata. Data format depends on each producers (csv, xml, parquet)
- 2) The responsible of the service AM3DS enters the metadata (concept, definition, format, classification, data model etc.) into Colectica designer. This information will feed the French statistical metadata register and ARC (via a DDI File). Automatic format controls are directly generated in VTL from these descriptions.
- 3) The responsible enters in the ARC interface the mapping between the external file and the data model described in Colectica.
- 4) The input data is then transformed and uploaded into parquet files according to the data model in our S3 storage platform. The data is divided into 3 types:
 - the ID data (name, surname date of birth, etc.) are sent to our individual register in order to be replaced by a non significant ID after an identification process. The data stored is therefore pseudonymised;
 - the address data (street name, municipality, etc) are sent to our address register in order to be replaced by a non significant ID after an identification process.
 - The other data is directly uploaded into parquet files.

After this process the administrative data is accessible, structured by statistical unit and easily exploitable to produce statistics. Statistician users can pick them up if they have the appropriate right to do so.

3. RESULTS AND PRACTICAL APPLICATION

This service is in production and allows to provide statistical user by monthly or annual data in order to produce short terms or structural statistics.

- Every month, more than 50 millions records containing information on wages, other income and social security benefits are uploaded.
- 4 times a years tax data (about a thousand of variables) about 60 millions people and 30 millions of dwellings are uploaded too.

4. MAIN FINDINGS

By the past the same external source used by different INSEE units could be formatted more than once, this one-stop shop for data will make it possible to be more efficient.

This centralized system enable as well to professionalise the reception of sources and therefore facilitate contacts with data producers. It adapts to different input file formats and allows data structuring to be adapted to users' needs in order to provide raw statistical data.

Metadata is entered as soon as it enters the INSEE information system, enabling it to be reused throughout the statistics production chain.

The link between the data warehouse and the statistical register means that pseudonymised data is available, making it less sensitive. It also makes it possible to comply with the principle of minimisation, as identity data is present in only one place in the INSEE information system.

This link between the warehouse and the register also facilitates data matching, which will make it possible to produce more data without increasing the statistical burden.

Finally, ARC is an open source software which has been consolidated and containerized with the support of Eurostat as part of ESSNET I3S.

5. REFERENCES

- [1] Durr, J-M, Dupont F., Haag O et Lefebvre OI., 2022. "Setting up statistical registers of individuals and dwellings in France: Approach and first steps". In : Statistical Journal of the IAOS (SJIAOS). Volume 38, n° 1, pp. 215-223.
<https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210916>
- [2] Cotton F., Haag O. 2023 "L'intégration des données administratives dans un processus statistique". In courrier des statistiques.
<https://www.insee.fr/fr/information/7635829?sommaire=7635842>. English version will be available soon.
- [3] Lefebvre O, Soulier M., Tortosa T. 2024 "L'accueil des données administratives : un processus structurant". In courrier des statistiques.
<https://www.insee.fr/fr/information/8203046?sommaire=8203072> English version will be available soon.

Analysing EU unemployment and underemployment in a fuzzy perspective: the comparison between fuzzy and official rates

Bruno Cheli, Alessandra Coli (University of Pisa) & Andrea Regoli (University of Naples Parthenope)

1. INTRODUCTION

The 13th International Conference of Labour Statisticians (ICLS) [1] established the definitions of employment and unemployment that remain in use today. According to this labour force framework, individuals of working age who are engaged in paid work during a specified short reference period are classified as employed, while those who are unable to work despite seeking employment are classified as unemployed. This framework accounts for contributions from all workers, including those in irregular employment, but it has some notable limitations. Namely, It does not consider aspects such as personal fulfilment or income adequacy, both of which significantly impact well-being. Critics argue that this broad definition of employment may obscure vulnerable segments of the labour market and provide a narrow and inadequate perspective on unemployment.

To address these limitations, Hauser [2, 3] and Clogg [4] developed the so-called Labour Utilization Framework, incorporating elements such as low-paid employment and underemployment. Based on these advancements, the 18th ICLS [5] introduced a more comprehensive conceptualisation of employment, that considers underlying constraints associated with employment and recognizes the existence of a continuum in the space embracing employment, unemployment, and inactivity.

This paper proposes a fuzzy approach [6] to redefine the concepts of employment and unemployment by incorporating both the actual hours worked and the satisfaction derived from those hours. This methodology enables a more nuanced classification of employment status, aligning with the concept of a continuum between employment and unemployment or inactivity. The approach is applied to Labour Force Survey data from 29 European countries, and the resulting fuzzy unemployment rates are compared with the corresponding official rates computed by Eurostat.

2. METHODOLOGY

The basic assumption of our proposal is that workers are employed to a certain degree, represented by a membership function μ_E in the fuzzy subset E of the employed, measured on a scale from 0 to 1, whereby 1 means full membership to the set of employed persons and 0 full non-membership.

Contextually, the membership function μ_U in the fuzzy subset U of the unemployed is defined. Assuming that the fuzzy set U of the unemployed is the complement of the fuzzy set E of the employed in the space of the labour force, the membership function in the fuzzy set U is given by $\mu_U = 1 - \mu_E$.

According to the ILO employment status, inactive individuals are not part of the labour force; therefore, their membership functions μ_E and μ_U are set equal to 0.

Fig. 1 illustrates the calculation of the μ_E membership function, whose values depend on the number of hours worked by the individuals and their satisfaction with those hours. An upper threshold for worked hours is established, serving as a limit based on statutory or collectively agreed standards. This threshold roughly corresponds to the average hours worked by full-time employees and may vary across countries.

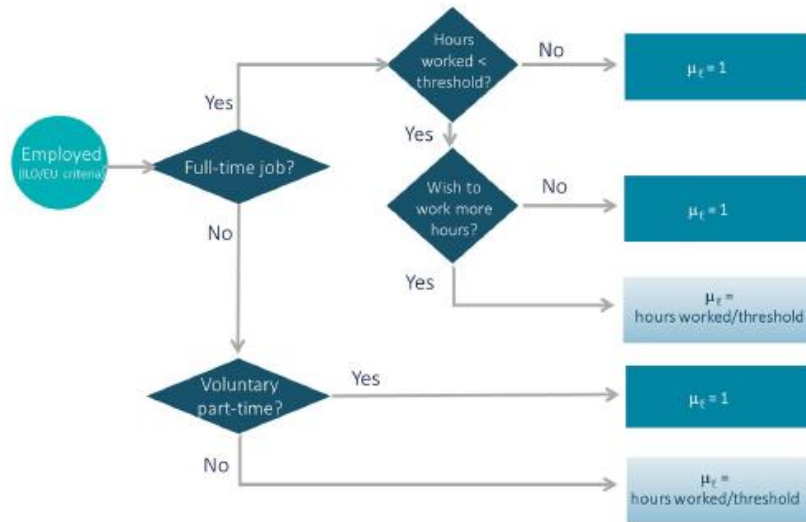


Figure 1. Specification of the membership function μ_E for individuals who are employed according to the ILO/EU criteria, Source: [7]

Denoting by x the number of hours worked and by t the threshold, the membership functions μ_E and μ_U are defined as follows.

$\mu_E = 1$ (and therefore $\mu_U = 0$) for:

- Full-time workers who work a number of hours equal to or greater than the threshold ($x \geq t$).
- Full-time workers who do not wish to work additional hours, even if they work fewer hours than the threshold ($x < t$).
- Voluntary part-time workers ($x < t$).

$0 < \mu_E < 1$ (and therefore $0 < \mu_U < 1$) for:

- Involuntary part-time workers
- Full-time workers who work below the threshold and wish to work additional hours.

This segment of the labour force represents the most vulnerable group of workers. These individuals are likely facing economic difficulties, as they have expressed a willingness to work additional hours. For this category of workers, the value of the membership function μ_E is calculated as the ratio of the number of hours worked to the threshold value, as follows:

$$\mu_E(x) = \frac{x}{t} \text{ for } 0 < x < t$$

$\mu_E = 0$ (and therefore $\mu_U = 1$) for the individuals classified as unemployed based on ILO criteria, that is: $\mu_E(x) = 0$ for $x = 0$.

The fuzzy employment rate (FER) is calculated as the weighted arithmetic mean of individual membership values $\mu_{E,i}$ with weights w_i representing the sample weights from the survey providing the employment data:

$$FER = \frac{\sum_{i=1}^n \mu_{E,i} \cdot w_i}{\sum_{i=1}^n w_i}$$

This mean is calculated for all sampled individuals aged 15 to 64 and can be directly compared to the official employment rate for the same age group.

The weighted mean of the μ_U membership function across the sampled individuals yields the fuzzy unemployment rate (FUR), which can be directly compared to the official unemployment rate for the same age group.

$$FUR = \frac{\sum_{i=1}^n \mu_{U,i} \cdot w_i}{\sum_{i=1}^n \mu_{E,i} w_i + \sum_{i=1}^n \mu_{U,i} w_i}$$

3. RESULTS AND PRACTICAL APPLICATION

We applied the fuzzy approach to data from the European Union Labour Force Survey (EU-LFS), for the reference year 2019 (Eurostat 2019)¹. The analysis includes 29 European countries: 25 EU countries, the United Kingdom, Switzerland, Norway and Iceland². Fig. 2 compares official and fuzzy unemployment rates across countries, with countries ordered by decreasing values of the official rate, from Greece (17.5%) to Czechia (2.1%). The increase of the fuzzy rate over the official one ranges from 4.8% in Czechia to 85.0% in Romania. Countries with the largest difference (more than 40 per cent) include Cyprus, Germany, Ireland, Italy, Norway, the United Kingdom, and Romania. On the other hand, there are eight countries with an increase of less than 15 percent compared to the official measure, namely Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Malta, and Slovakia.

¹ The responsibility for all conclusions drawn from the data lies entirely with the authors.

² Slovenia and Lithuania were ultimately excluded: Slovenia due to insufficient data on the reasons for part-time work, and Lithuania because over 50% of self-employed and family workers reported zero hours usually worked.

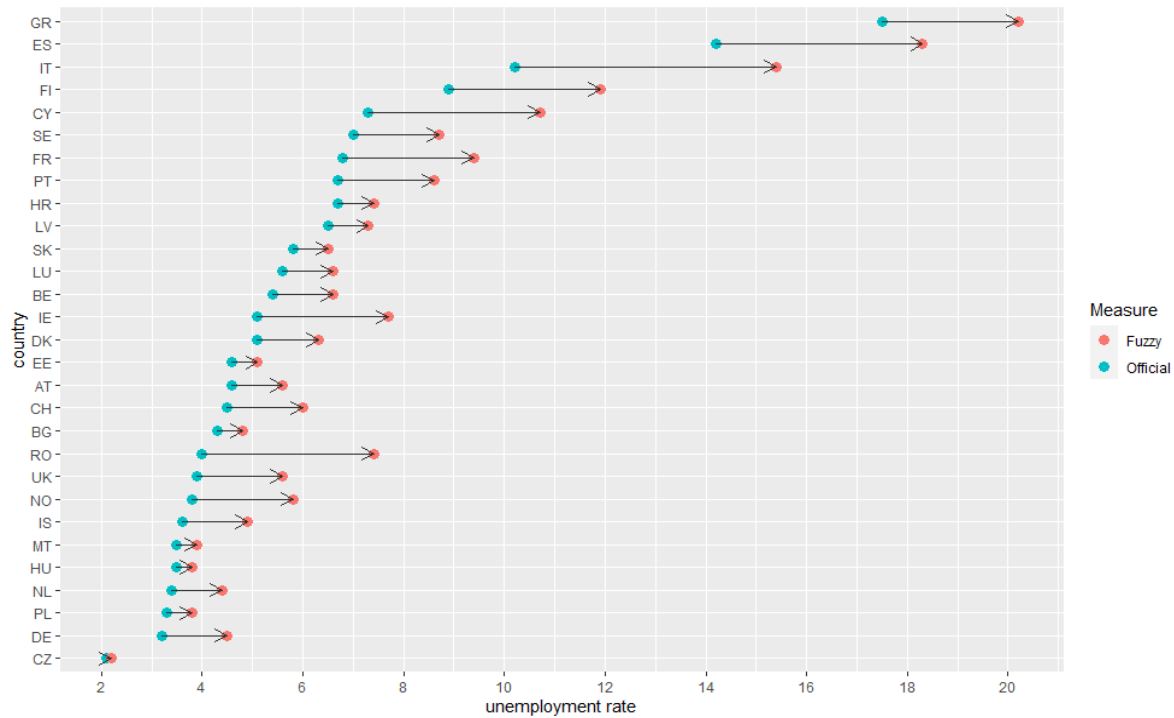


Figure 2. Comparison between official and fuzzy unemployment rates (15–64 years) by country - year 2019.

Note: our elaborations on EU-LFS 2019 data.

Source: [7]

4. MAIN FINDINGS

Among the employed there are workers who need additional hours of work to improve their economic well-being. Our proposal to redefine employment status as employed to a certain degree (and symmetrically as unemployed to a certain degree) leads to the definition of fuzzy rates that incorporate this dimension of weakness and vulnerability. The application shows how the transition to fuzzy measures differs across countries and how there are significant differences behind similar official unemployment rates. In future research, we intend to extend our analysis to other vulnerability factors, such as low wages and insecurity.

5. REFERENCES

- [1] ILO Statistics of labour force, employment, unemployment and underemployment, Report II of the 13th International Conference of Labour Statisticians, Geneva (1982).
- [2] P.M. Hauser, The measurement of labour utilisation, *Malaysian Economic Review*, 19(1), 1–15 (1974).
- [3] P. M. Hauser, The measurement of labour utilization—More empirical results. *Malaysian Economic Review*, 22(1), 10–25 (1977).
- [4] C. C. Clogg, *Measuring underemployment: Demographic indicators for the United States*, Academic Press, (1979).
- [5] ILO, Beyond unemployment: Measurement of other forms of labour underutilization, working group on labour underutilization, 18th International Conference of Labour Statisticians, Geneva, 24 November–5 (2008).
- [6] L. A. Zadeh, Fuzzy sets, *Information and Control*, 8, 338–353, (1965).

- [7] B. Cheli, A. Coli, A. Regoli, Measuring employment and unemployment from a fuzzy perspective. A comparative analysis across European countries, Soc Indic Res 175, 467–486 (2024).

Analysis of voluntary survey data under missing not at random

Jae Kwang Kim (Iowa State University) & Yonghyun Kwon (Korea Military Academy)

Incorporating the auxiliary information into the survey estimation is a fundamental problem in survey sampling. Calibration weighting is a popular tool for incorporating the auxiliary information. The calibration weighting method of Deville and Sarndal (1992) uses a distance measure between the design weights and the final weights to solve the optimization problem with calibration constraints.

In this paper, we propose a new framework using generalized entropy as the objective function for optimization. Design weights are used in the constraints, rather than in the objective function, to achieve design consistency. The new calibration framework is attractive as it is general and can produce more efficient calibration weights than the classical calibration weights. Furthermore, we identify the optimal choice of the generalized entropy function that achieves the minimum variance among the different choices of the generalized entropy function under the same constraints. Asymptotic properties, such as design consistency and asymptotic normality, are presented rigorously. An extension of the proposed method to doubly robust propensity score estimation will also be presented.

An agricultural atlas for Europe – based on IFS 2020

Marijn van der Velde (European Commission - Joint Research Centre), Nicolas Lampach (Eurostat), Jon Skoien (ARHS Developments), Helena Ramos, Julien Gaffuri (Eurostat), Renate Koeble (ARHS Developments)
& Linda See (IIASA)

1. INTRODUCTION

The Integrated Farm Statistics census is a European wide census, which last took place in 2020. More than 300 variables on agriculture and farm structure were collected from 9.03 million farmers in the EU (and EFTA countries Iceland, Switzerland, and Norway). Despite the large number of data, IFS records are traditionally aggregated only to NUTS 2, NUTS 1 or national levels, which are then published on the Eurostat website. Higher resolution products would be useful for a range of applications, but confidentiality regulations have so far restricted this. Here we will present a set of maps from IFS with the highest possible resolution, based on a method by [1].

A few attempts have been made to create interoperable agricultural atlas with information and data of single agricultural indicators at a finer level of resolution. A notable example is the Agrarian Atlas disseminated by the German statistical office, which provides a range of agricultural variables on a 5 km grid (<https://agraratlas.statistikportal.de/>). The layers are accessible through a Web Coverage Service (WCS) but only as aggregated classes, not with real numbers. Gridded agricultural data are also available for the UK at resolutions of 2, 5 and 10 km hosted by the Edinburgh Data (EDINA) portal (<https://agcensus.edina.ac.uk/>) but free access is only available for academic institutions; otherwise the data must be purchased.

There are some other examples of agricultural data at 10 km grid based on data originally disseminated at the level of administrative zone. This includes FAO's Gridded Livestock of the World, which provides gridded livestock numbers for 2010 and 2015 [2],[3] and crop types from the SPAM (Spatial Production Allocation Model) data set for 2010 [4]. Gridded data on crop types are also available from the EarthStat website for the year 2000 [5]. It has been updated for the year 2020 as the CROPGRIDS product [6], and includes other gridded layers such as nutrient application for major crops. The World Bank provides a catalogue of gridded data sets including global gridded agricultural gross domestic product [7]. The disadvantage of these downscaled data sets is the uncertainty related to estimated values from models instead of employing aggregations from actual census data.

During the previous CAP period (2014-2022), the Common Monitoring and Evaluation Framework (CMEF) developed a set of 'Context Indicators' [8], some of which have used data from the IFS [9] and can be viewed on the EU's Agridata Portal (Further information can be found in the dashboard, see

https://agridata.ec.europa.eu/extensions/DataPortal/context_indicators.html. Most of the information related to context indicators is presented at national level, although some of the maps display information at NUTS2 administrative zones. However, NUTS2 regions are also too coarse, and they mask the large structural disparities within NUTS2 regions that are not identifiable using such data. Hence, to more effectively monitor and evaluate the implementation of the CAP and to guide the design of future funding schemes, more highly resolved spatial information from agricultural censuses and surveys is needed. This would

also benefit many other applications and models that need high resolution inputs related to the agricultural sector.

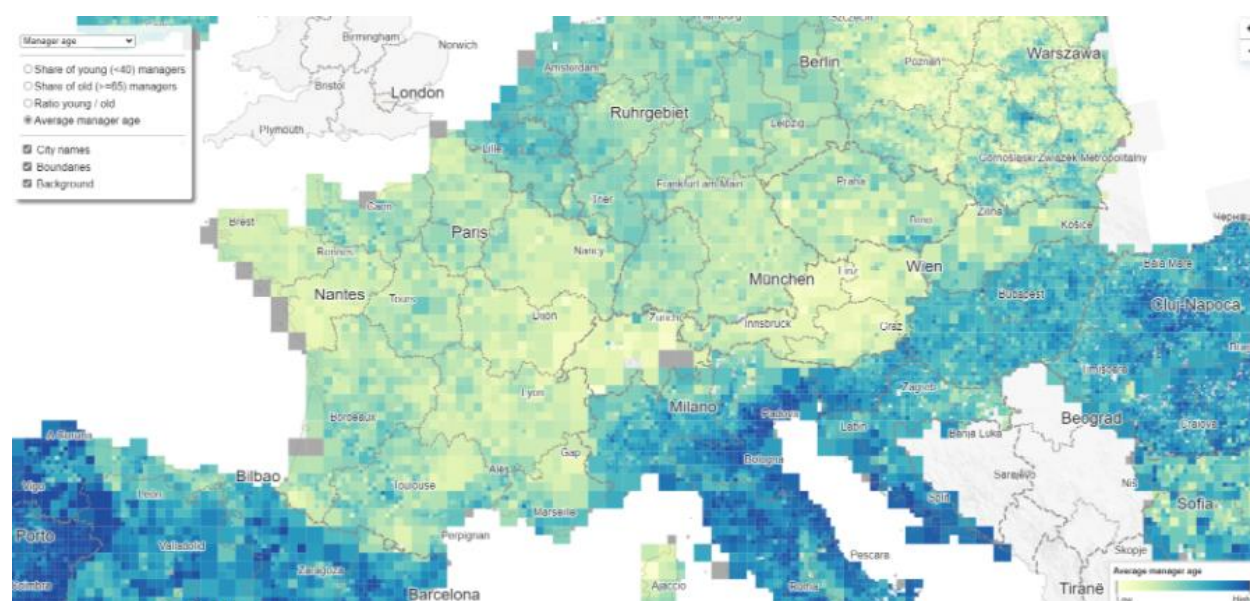
2. METHODOLOGY

Recent methodological advances have resulted in a method for gridding agricultural census data that employs a flexible, multi-sized grid cell approach. Such a procedure can present the data at a much higher resolution than NUTS 2 while respecting the confidentiality regulations that safeguard the identities of individual farms (Skøien et al., 2024). The resolution of the grid varies from a minimum of 1 km (based on the 1 km INSPIRE Statistical Units Grid for pan-European data) to 80 km, which is the maximum size that nullifies the disclosure risk while maximizing the utility of the information content presented.

By applying this new approach to variables in the 2020 IFS, we now have an unprecedented and harmonized view of European agriculture compared to previous analyses undertaken at the coarser NUTS2 level. To illustrate this innovation, we will present selected key variables from the 2020 IFS that are relevant to the CAP, which includes information on structural components (i.e., agricultural holdings, land use, livestock patterns and labour input), the demographics of farmers (i.e., age, gender and skills) and agricultural production methods (i.e., irrigation and organic farming). These data sets are publicly accessible through an interactive maps viewer called Gridviz. (<https://eurostat.github.io/gridviz/>). It can be retrieved via an API from Geographic Information System of European Commission (GISCO). The agricultural atlas provides new opportunities for policy evaluation, assessment, analysis, and the creation of land use related maps at a high spatial resolution.

3. RESULTS AND PRACTICAL APPLICATION

Below are two examples as shown in Gridviz. The first one shows the average age of farm managers, whereas the second one shows the number of farms. In both of them it is possible to see how the resolution is higher in agriculturally dense areas, and lower in mountain regions, around big cities etc, reflecting the spatial distribution of farms.



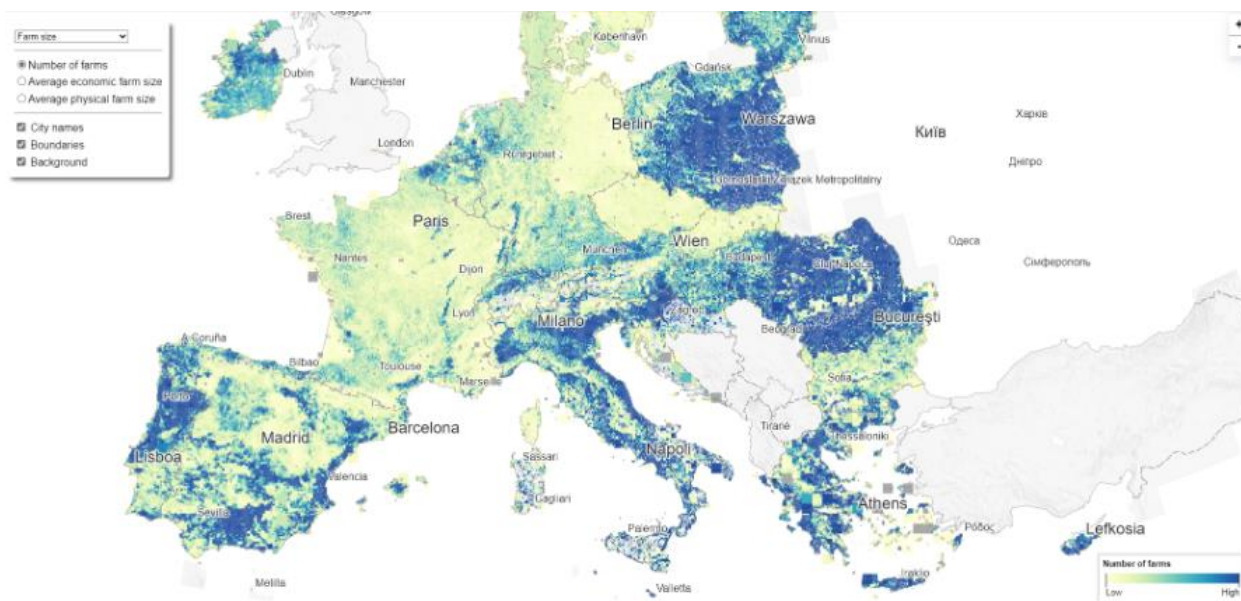


Figure 1. Examples of the multi-resolution grid available from Gridviz. Upper panel: Age of manager. Lower panel: Number of farms

4. MAIN FINDINGS

We have shown some examples of multi-resolution maps of agricultural data that will be available through the Gridviz portal. This will make it easier to more effectively monitor and evaluate the implementation of the CAP and to guide the design of future funding schemes. The maps will also benefit many other applications and models that need high resolution inputs related to the agricultural sector.

5. REFERENCES

- [1] Skøien, J. O., Lampart, N., Ramos, H., Seljak, R., Koeble, R., See, L., and van der Velde, M.: R package for creating multi-resolution grids of census and sample data, 2024. Presented at New Techniques and Technologies for Statistics conference, NTTS, 2025.
- [2] Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T. P., Vanwambeke, S. O., Wint, G. R. W., and Robinson, T. P.: Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010, *Scientific Data*, 5, 180–227, <https://doi.org/10.1038/sdata.2018.227>, 2018.
- [3] Gilbert, M., Cinardi, G., Da Re, D., Wint, W. G. R., Wisser, D., and Robinson, T. P.: Global cattle distribution in 2015 (5 minutes of arc), <https://doi.org/10.7910/DVN/LHBICE>, 2022.
- [4] Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K. B., Fritz, S., Xiong, W., Lu, M., Wu, W., and Yang, P.: A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps, *Earth System Science Data*, 12, 3545–3572, <https://doi.org/10.5194/essd-12-3545-2020>, number: 4 Publisher: Copernicus GmbH, 2020.

- [5] Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A.: Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000, *Global Biogeochemical Cycles*, 22, n/a–n/a, <https://doi.org/10.1029/2007GB002952>, number: 1, 2008.
- [6] Tang, F. H. M., Nguyen, T. H., Conchedda, G., Casse, L., Tubiello, F. N., and Maggi, F.: CROPGRIDS: a global geo-referenced dataset of crops, *Scientific Data*, 11, 413, <https://doi.org/10.1038/s41597-024-03247-7>, 2024.
- [7] World Bank: Global Gridded Agricultural Gross Domestic Product (AgGDP). <https://datacatalog.worldbank.org/search/dataset/0061507/Global-Gridded-Agricultural-Gross-Domestic-Product-AgGDP->, 2023.
- [8] The European Commission: CAP Indicators. Agri-food Data Portal. https://agridata.ec.europa.eu/extensions/DataPortal/cap_indicators.html, n.d.
- [9] The European Commission: Principal quality characteristics of EU countries' Agricultural Censuses for 2020 – 2024 edition, Publications, Office of the European Union, 2024.

An end-to-end statistical process to aggregate high-resolution temporal survey microdata: the monthly Labour Force Survey use case

María de Blas, Carlos Saez, Luis Sanguiao, Sandra Barragán (SG of Methodology and Sampling Desing, Statistics Spain) & Guillermo Gorgas (SG of Labour Market Statistics, Statistics Spain)

1. INTRODUCTION

The traditional statistical process using survey data is not enough to produce timely and granular official statistics. To address this, statistical offices are innovating by using new methods and data sources, including administrative and privately-held digital data. In this work, it is shown that even with traditional survey data, novel statistical methods can significantly improve these quality dimensions. They demonstrate that the trade-off between accuracy and temporal granularity can be mitigated by combining finite-population methodology with statistical learning and time series treatment. The use case to illustrate these ideas makes use of survey microdata from the Spanish Labour Force Survey. We show how calibration techniques of survey sampling weights, statistical learning modelling, and time series filtering and benchmarking techniques can produce more granular information.

The combination of survey microdata with process paradata and these techniques can enhance the performance of traditional statistical outputs. For future work, the incorporation of new data sources can further expand the potential of these statistical products.

In previous works we conducted preliminary analysis to know the feasibility of the methodology developed. Here, we present a second stage of the project that has been conducted to prepare a Minimum Viable Product (MVP) ready for deployment in production. The implementation of this MVP is the key of this contribution. An end-to-end statistical process has been designed with the aim to compute monthly aggregates for the Labour Force Survey by using the advantage of having survey microdata at a weekly time resolution.

2. METHODOLOGY

This section consists of two parts: a) methodology to compute the monthly aggregates by using survey microdata with weekly periodicity, b) the design and implementation of the statistical production process.

2.1. Monthly Aggregates

Computation of time-disaggregated design weights

We take the quarterly sampling design for the *LFS* as our starting point, producing quarterly design weights, $d_k^{[Q]} = 1/\pi_k^{[Q]}$, and calibrated sampling weights, $\omega_k^{[Q]}$, for the usual linear estimator $\hat{Y}_d^{[Q]} = \sum_{k \in S_d} \omega_k^{[Q]} y_k$ for the domain total, $Y_d = \sum_{k \in U_d} y_k$. Interviews are weekly, so that every respondent in the rotating panel is assigned a given week w within each calendar quarter Q . This procedure is executed semiautomatically by sampling experts inheriting interview week assignments from preceding waves and assigning new units to fit a balanced

data collection field work across the national territory. These assignments involve both, the calendar-quarter weeks $v = 1, \dots, 13$ and the so-called rotation turn τ for field work balancing.

As a first step, we train a random forest model using (v, τ) as the categorical target variable and some identification and design variables as regressors x . The model produces probabilities $P(v, \tau|x)$ for each value (v, τ) in terms of the regressors, so that we can compute the marginal probabilities $P_a(v|x)$. Since the regressors can be considered as partial identification variables for each unit, we can write just $P_k(v)$. Then, by computing weekly design weights for week v we reason as follows:

$$d_k^{[W]} = \frac{1}{\pi_k^{[W]}} = \frac{1}{\pi_k^{[Q]} \times P(k \rightarrow v | s^{[Q]} \ni k)} = \frac{d_k^{[Q]}}{P_k(v)}$$

As a matter of fact, this computation can be generalized to any group of weeks within a given calendar-quarter. Usual calibration techniques are applied to these weekly design weeks.

Monthly Aggregates and Variance Computation

Once weekly calibrated sampling weights, $\omega_k^{[W]}$, are computed we can produce estimates with weekly and monthly time scopes as following:

- Weekly aggregate for week v : $\hat{Y}_v^{[W]} = \sum_{k \in s_w} \omega_k^{[W]} y_k$.
- Monthly aggregate for month m based on weekly aggregates:
 $\hat{Y}_m^M = \frac{1}{n_m^W} \sum_{v \in W_m} \hat{Y}_v^{[W]}$, where W_m is the subset of weeks in month m , with cardinal n_m^W .
- Estimate variance of monthly aggregate:

$$\hat{V}(\hat{Y}_m^M) = \widehat{V}_{JK}(\hat{Y}_m^M) + \frac{1}{n_m^W} \sum_{s \in W_m} \hat{V}(\hat{Y}_v^{[W]}).$$

Filtering and benchmarking of time series of estimates

These monthly raw estimates $\hat{Y}_m^{[M]}$ still contain a lot of variability (noise) since the original sampling design was thought to be quarterly. We propose to use them as building blocks within filtering models. A first trivial elementary model can be proposed by setting ARIMA-model-based decomposition: $Y_m^{[M]} = T + S_{13} + I$.

ARIMA models for T , S_{13} and I are given by the canonical decomposition for the Y model. S_{13} contains all AR roots associated to the repeating sample, T contains the remaining AR roots, and I is considered noise. We keep the Wiener-Kolmogorov filtered estimation for T , both S_{13} and I are discarded.

Filtered values are provisional monthly estimates that are corrected by using benchmarking methods once the quarterly aggregates are obtained. Then, the definitive monthly values can be disseminated:

$$\hat{Y}_q^{[Q]} = \frac{1}{3} \sum_{m \in Q} \hat{Y}_m^{[M]}$$

2.2. Implementation of the statistical production process

The design of the statistical production process has been based on the following principles:

- **Modularity**: Production steps are modular. They can be isolated, analysed, transformed and recombined in useful ways reducing complexity and facilitating modifications and scalability.

- **Standardization:** Production steps, functions and processes are standardized ensuring homogeneity, interpretability and components reutilization.
- **Parametrization:** A number of parameters control the particularities of the production process which allows making important modifications of the process by simply adjusting some parameters.
- **Version Control:** Code, functions and processes changes are tracked and controlled over time. Ensuring version control enables changes traceability which greatly facilitates rollback and recovery and enhances branching and experimentation without damaging the integrity of the process.
- **Scalability:** The process is designed to be scalable in terms of data and new productions steps. In fact, modularity, standardization and parametrization, altogether, give rise to a highly scalable process.

To help achieve these goals, language *R* is employed as the main programming language in the production process. This decision goes along with all previous criteria, specifically considering that it is the predominant coding language in international official statistics and allows to take advantage of already existing developments in the form of *R* packages. Indeed, several internal *R* packages were used; namely *fastReadfwf*, for reading the data, and *Reptime*, for working with dates. Moreover, all the ad-hoc functions created for the monthly *LFS* production process were systematically compiled in a new *R* package called *EPAML*. Isolating and gathering the main components of the process in a single *R* package greatly improves code modularization and version control, making it easier to implement changes and revert to previous versions.

In addition, *C++* language has also been incorporated in the process. Calibration computations are very time and computing power intensive. As a consequence, calibration has been partially implemented in *C++* to obtain a faster and lighter process. This is especially relevant for variance calculations since jackknife loops call the same functions hundreds of times.

3. RESULTS

3.1. Statistical production process

The resulting *LFS* monthly process consists of four different modular interconnected processes which, in turn, are composed of more subprocesses. These four main processes compose together the complete production process shown in the Figure 1.

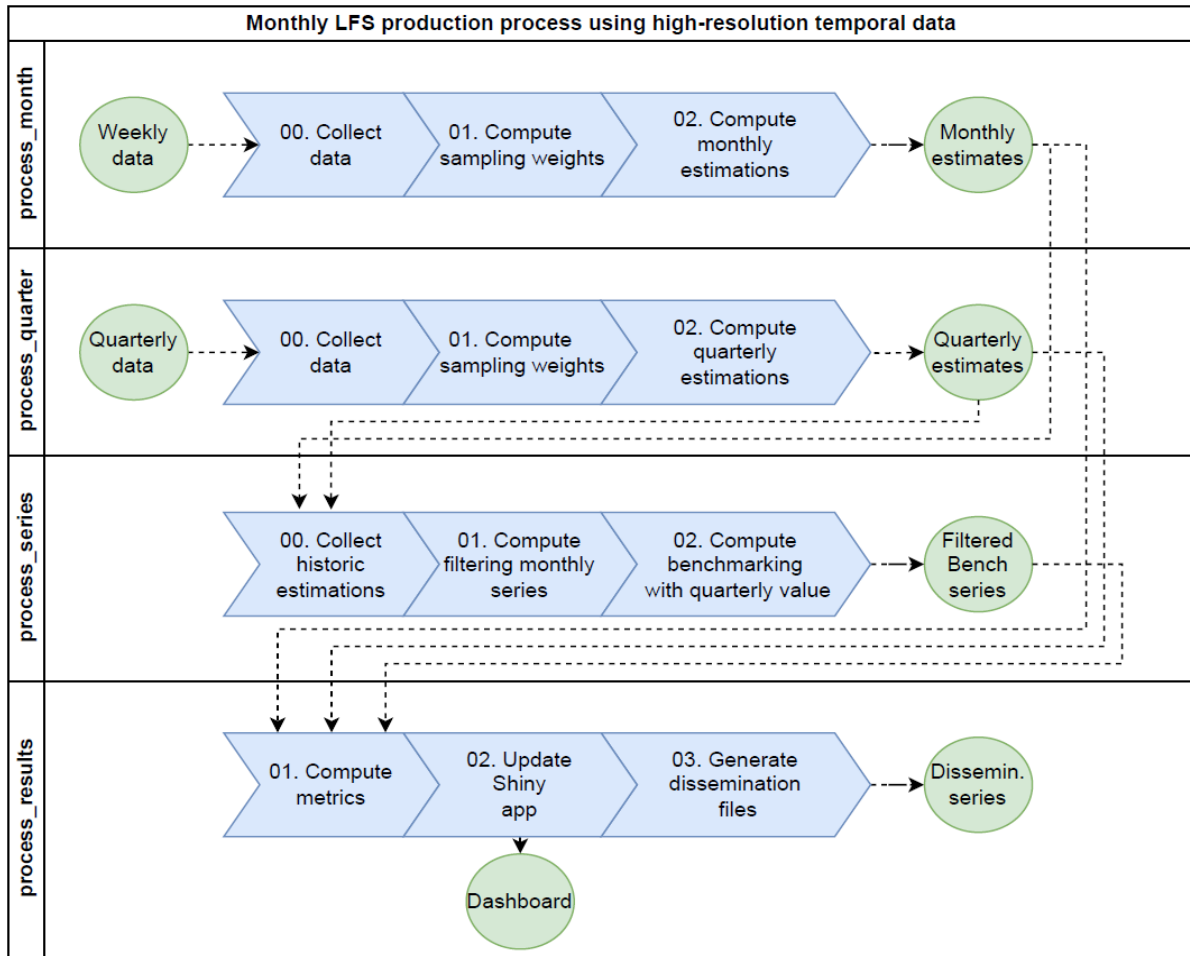


Figure 1. Diagram of the statistical production process by taking into consideration all the subprocesses to obtain the monthly and the quarterly aggregates from the collection to the dissemination.

3.2. Monthly aggregates

The main output from the preceding proposal can be depicted as in Figure 2, in this case using the time series of employed males aged 16 – 24 in the whole national territory. Similar results are obtained for both sexes in age groups 16 – 24, 25 – 74, 75₊ for employed, unemployed and inactive resident people. This figure represents the quarterly values (*quarter – ORIG_{edited}*), the monthly values directly computed (*monthJK_{edited}*), the filtered provisional values (*month – JK – ARIMAseason – prov_{edited}*) and the definitive benchmarked values (*month – JK – ARIMAseason – bench_{edited}*).

4. MAIN FINDINGS

Survey microdata in combination with process paradata and novel statistical methods can notably improve the performance of some classical statistical products by statistical offices. A bottom-up approach focused on the microdata stands up as a serious alternative to the top-down approach based on the application of econometric models upon quarterly times series. Once the novel statistical methodology is decided for production, the statistical process will need some adjustments. In our opinion, these changes will not be disruptive at all.

Finally, the trade-o between temporal granularity and accuracy can be stretched out so that traditional data sources are not condemned to sacrifice one of them. This paves the way for an even greater quality improvement with the integration of new data sources. We can sum up the main conclusions in the following points:

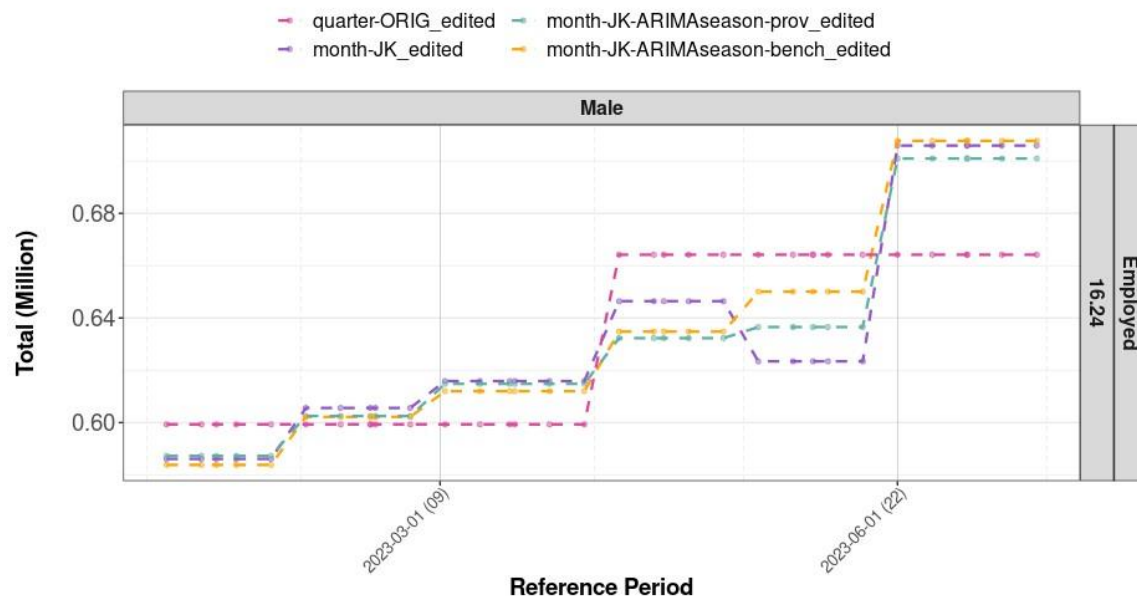


Figure 2. Monthly and quarterly series for 1st and 2nd quarters of year 2023.

Methodological challenges achieved:

- Time disaggregation of sampling weights is solved with statistical learning methods.
- Calibration techniques applied to time-disaggregated design weights has the problem of small sample size per calibration group: geographical aggregation of marginal totals solves the issue.
- Large variability in high-frequency time series is dealt with filtering models.
- Coherence between monthly and quarterly aggregates is achieved through benchmarking methods.

Computational challenges achieved:

- Computation time is reduced by efficient code implementation of the jackknife algorithm to estimate variance.
- A modular design in a minimum viable product has been implemented with a deployment ready for production.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Dr. David Salgado (Statistics Spain) for his valuable insights, guidance and technical contributions during the development of this project.

An innovation score based on financial data and web-augmented firms' characteristics

Carlo Bottai (Università di Milano-Bicocca), Lisa Crosato (Università Ca' Foscari - Venezia), Marco Guerzoni & Caterina Liberati (Università di Milano-Bicocca)

1. INTRODUCTION

Over the past half-century, the innovation economics literature has made significant progress in measuring this phenomenon. To do so, balance sheets, surveys, patents, and product catalogues have been used as the main data sources. However, each of these sources has limitations in capturing such latent features, particularly when applied to small and medium-sized enterprises (SMEs), as underlined by the OECD [1]. In line with some recent literature [2-6], this study suggests that SMEs' corporate websites can serve as a new data source to improve indicators of the innovative activity of a firm. Specifically, a growing number of papers have attempted to use corporate websites to analyze firms' innovative activity [7-9]. However, this study proposes an extensive use of the HTML code structure of these web-pages, instead of focusing on the semantic analysis of their text.

The paper first introduces the topic and the main motivations and advantages of focusing on the HTML structure of corporate websites to detect whether a firm is actively innovative. Second, we highlight the key importance of data preprocessing when dealing with such unstructured sources, like information scraped from corporate websites. Specifically, the paper proposes an original method to assess the quality of the matching between traditional firm-level data sources and unconventional web-based ones. Lastly, we demonstrate how the HTML code structure of corporate websites can be used to distinguish innovative SMEs from non-innovative ones. Thereafter, we use a machine learning algorithm exploiting both HTML data and firms' structural characteristics, to predict the innovation intensity of a firm, using data about innovative Start-ups and SMEs as ground truth. This method allows us to make point predictions at the firm level.

1.1. Data Collection

To collect information about the Italian manufacturing SMEs in 2018, we exploit the Aida database, by Bureau van Dijk (BvD). The database provides us, for each company, (a) its tax identification number (*codice fiscale*); (b) business name; (c) 'status' (active, bankrupt, in liquidation, etc.); (d) industrial sector (NACE); (e) geographical location (NUTS 2); (f) business address (street name, number, and postcode); (g) and telephone number. Using information from this data source, we also classify each company as micro, small, or medium-sized, as defined by the Eurostat.

Then, we enrich the conventional firm-level data just described with web-based information scraped from the corporate website of each firm included in the sample, following the procedure described in [10-11].

First, we search for the 2018 archived version of each firm's website URL, as reported in Aida, on the Wayback Machine of the Internet Archive (<https://web.archive.org/>). Secondly, we evaluate whether the URL provided by Aida correctly matches the corresponding firm. Starting with the 131,332 manufacturing SMEs, we have excluded certain observations due to missing web addresses in Aida, the absence of a 2018 snapshot archived by the Wayback Machine, or because our algorithm failed to positively confirm the attribution of the website reported in

Aida to the respective firm. Therefore, our final sample was composed of 28,778 Italian manufacturing companies.

2. METHODOLOGY

The starting point of our methodology consists of downloading the list of these innovative SMEs, as defined by the Italian Startup Act, from Aida.

We train and test different machine learning algorithms to predict whether a firm enrolled in the business register as an innovative Start-up or SME using metrics based on the HTML code of their respective websites. We evaluate different models in terms of standard accuracy metrics and analyze their output in terms of Explainable A.I. We are not interested in the categorical prediction, but in the predicted probability, which is a continuous number ranging from 0 to 1. The distribution of the predicted probability is then employed to build an innovation score, not directly, but following standard statistical manipulation to ensure desirable properties.

3. RESULTS AND PRACTICAL APPLICATION

In our study, we evaluated the predictive performance of different datasets to assess firms' innovation intensity. Specifically, we tested three distinct datasets and a reference model:

- 1) **FIN+FIXED Dataset:** This dataset comprised financial variables combined with fixed variables such as sector, company size, and age. This model achieved the highest predictive performance; however, it relies on financial data, which are not always updated or available.
- 2) **FIXED Variables Dataset:** This dataset included only the fixed variables (sector, size, and age), ensuring these features are always available.
- 3) **WEB Variables Dataset:** This dataset relied solely on variables derived from web scraping, which are also always accessible.
- 4) **WEB+FIXED:** This model combined the FIXED variables with the WEB variables.

For each dataset, we experimented with several machine learning models, including Recursive Partitioning Trees, Random Forest, AdaBoost, and eXtreme Gradient Boosting (XGBoost) with a custom gradient. We selected the best-performing model from each set based on their predictive accuracy.

Our results, illustrated in the accompanying ROC curves, indicate that the FIN+FIXED Dataset achieved the highest predictive performance (Figure 1 panel a). The WEB Variables Dataset alone provided limited predictive power. However, the Reference Model (FIXED + WEB) approached the performance of the FIN+FIXED Dataset, despite relying only on always-available data. This demonstrates that combining fixed variables with web scraping variables offers a strong predictive capability, making it a valuable alternative when financial data are not accessible. Overall, while the FIN+FIXED model remains the most effective, the FIXED + WEB combination provides a reliable and consistently available approach for timely assessments of firms' innovative activities.

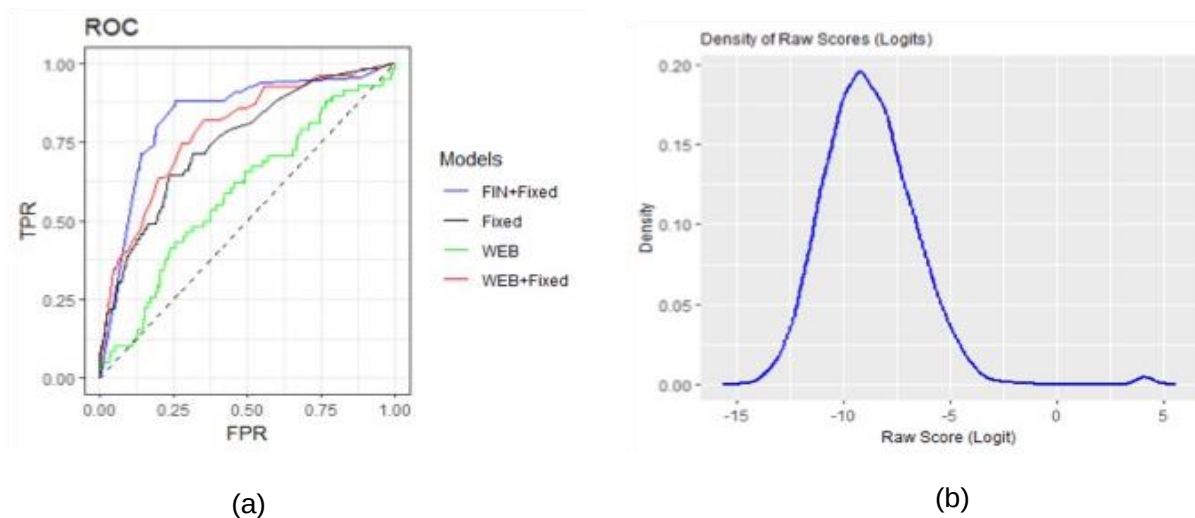


Figure 1. Panel a) ROC curves based on the four sets of predictors. Panel b) Innovation score distribution

Building on our proposed approach, we proceeded to apply the model trained on the FIXED + WEB dataset to generate predictions for all firms within our comprehensive dataset. This application is visually represented in the accompanying graph, which illustrates the distribution and intensity of predicted innovation across the entire population of firms analyzed (Figure 1 panel b).

This scalability highlights the practicality of our approach, enabling wide-ranging assessments of innovation activities across diverse enterprises. Furthermore, the ability to utilize always-available data sources enhances the model's applicability for real-time monitoring and strategic decision-making.

Our results complement conventional data and official statistics in at least two aspects—they are timely and able to cover micro-enterprises—both key for innovation policy [12-13]. On the other hand, they use the same units of observation, the firm, of official statistics on which policy tools, like the European Innovation Scoreboard, are built and so can easily go hand-in-hand with these.

4. MAIN FINDINGS

Building upon previous studies in this research area, our results confirm that variables derived from web scraping corporate websites can significantly enhance predictive accuracy when combined with traditional variables such as company sector, age, and size. Together, these variables achieve a predictive performance similar to that of financial indicators but have the distinct advantage of being available much earlier. This early accessibility allows for more timely and proactive assessments of a firm's innovative activities. By leveraging the HTML code structure of corporate websites, we demonstrate that it is possible to effectively predict the innovation intensity of firms, offering valuable insights, especially for SMEs where financial data may be limited or delayed.

5. REFERENCES

- [1] OECD. Oslo Manual: OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. OECD Publishing (1992).
- [2] S. K. Arora, Y. Li, J. Youtie, and P. Shapira. Using the Wayback Machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology*, 67(2016), 1904–1915
- [3] D. Blázquez, J. Domènech, and A. Debón. Do corporate websites' changes reflect firms' survival?. *Online Information Review*, 42(2018), 956–970.
- [4] P. J. H. Daas, and S. van der Doef. Detecting innovative companies via their website. *Statistical Journal of the IAOS*, 36(2020), 1239–1251.
- [5] L. Crosato, J. Domènech, and C. Liberati. Predicting SME's default: Are their websites informative?. *Economics Letters*, 204 (2021), 109888.
- [6] L. Crosato, L., J. Domènech, and C. Liberati. Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research*, 342 (2024), 1671-1686.
- [7] A. Gök, A. Waterworth, and P. Shapira. Use of web mining in studying innovation. *Scientometrics*, 102 (2015), 653–671.
- [8] M. Héroux-Vaillancourt, C. Beaudry, and C. Rietsch. Using web content analysis to create innovation indicators—What do we really measure?. *Quantitative Science Studies*, 1(2020), 1601–1637
- [9] J. Kinne, and J. Axenbeck. Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 125 (2020), 2011–2041
- [10] C. Bottai, L. Crosato, J. Domenech, M. Guerzoni, and C. Liberati. Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (2022), 338–334.
- [11] C. Bottai, L. Crosato, J. Domenech, M. Guerzoni, and C. Liberati. Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs. *Technological Forecasting and Social Change*, 207 (2024), 123597.
- [12] S. Roper and N. Hewitt-Dundas. Investigating a neglected part of Schumpeter's creative army: What drives new-to-the-market innovation in micro-enterprises?. *Small Business Economics*, 49 (2017), 559–577.
- [13] H. Hollanders. A critical assessment of the European Innovation Scoreboard. In *Handbook of Innovation Indicators and Measurement*, (2023) Ch. 21. Cheltenham, UK: Edward Elgar Publishing.

An ML-based automatic coding machine as an integral element in an overall strategy for NACE implementation

Adrián Pérez Bote (Statistics Spain, INE), Carlos Sáez Calvo (Instituto Nacional de Estadística) & David Salgado (Statistics Spain, INE)

1. INTRODUCTION

Statistics Spain (INE) has designed an overarching strategy to implement the new version of CNAE, the so-called CNAE-2025, the national version of the economic activity classification emerging from the new NACE revision, NACE Rev 2.1. CNAE-2025 is identical to NACE Rev.2.1, except for 10 additional classes to account for specificities in the Spanish economy.

The main goal of the strategy is to set up the coding and recoding activities at the data generation processes, including those of administrative registers feeding our statistical information systems. In this line, this strategy has the following pillars:

- 1) To collect data in the form of text-code pairs, both to feed the statistical business register and to be used as train-test data.
- 2) To build and deploy a machine-learning-based automatic coding machine for statistical classifications, currently especially focused on CNAE-2025.
- 3) To make CNAE-2025 legally compulsory in all administrative registers, including variables about economic activities.

The combination of these pillars is thought to extend and standardize the use of CNAE-2025 in those statistical and administrative data registers dealing with the concept of economic activity, thus allowing the continuous improvement of data quality, either statistical or administrative. The development and deployment of an automatic coding machine will facilitate the incorporation of CNAE codes and description already in the data generation phase, so that even administrative data shared for statistical purposes will enter our office with a high-quality standard.

This contribution will focus on points 1 and 2, but we should bear in mind that 3 is also key to the success of the described strategy.

1.1. Functionalities

- The automatic coding machine is conceived to be used for both statistical and administrative purposes.
- We use the described framework to build ML classifiers both for CNAE-2009 (corresponding to NACE Rev 2) and CNAE-2025 (corresponding to NACE Rev 2.1).
- The automatic coding machine will provide a conversion from CNAE-2009 to CNAE-2025.

2. METHODOLOGY

We decided to implement a “classic” NLP classifier, in particular the Fasttext model [1], over other possible strategies¹:

- Rule-based approaches provide worse performance and are more human-labour intensive.
- LLMs show more difficulties regarding data protection, control of model versioning, monetary cost and latency in production.

We build a training set, made of text-code pairs, by collecting data from the following sources:

- Historical databases of CNAE-2009 coming from different business statistics, both short-term and structural.
- Synthetic data of CNAE-2009 and CNAE-2025: we build synthetic descriptions based on titles and explanatory notes of the classifications. We apply data augmentation to enlarge this subset.
- Ad-hoc surveys covering CNAE-2009 and CNAE-2025: we add new modules to existing surveys (Structural Business Statistics and Innovation Survey) and launch a new survey to collect economic activities in free text, along with their codes for both the old and new CNAE versions. Since we will not use this sample to compute aggregates, we do not use a probabilistic sampling design. Instead, we optimize the sampling for the most difficult classes to code.

To get the CNAE-2025 codes for the historical databases, we use correspondence tables from CNAE-2009 and CNAE-2025, as well as semantic rules when the correspondence is 1:N. The most ambiguous or difficult instances are left unmapped and are removed from the CNAE-2025 dataset.

Ensuring data quality is key to the proper performance of the NLP model:

- Synthetic data adds volume to all classes, which is very relevant for scarce ones. At the same time, it ensures perfect class labeling.
- Specific surveys produce high-quality data, as the companies in the sample are requested only this information, which is revised by coding clerical staff at Statistics Spain (INE).
- We remove suspicious training instances in the preprocessing step.
- We measure the quality of the historical databases by applying manual labeling to a sample of each of these databases.
- Both the descriptions and the labels of the test set are reviewed manually to ensure the highest quality.

With all this, we get a training set of approximately 3 million instances. We get a random 10% subset for validation and manually label around 500 samples for the test set. CNAE-2009 and CNAE-2025 both have between 600 and 700 classes, and the dataset is imbalanced, even after adding synthetic data: the most common class accounts for 10.1% of the instances, while the less common only for 0.00013%. There is a big tail of minority classes.

We use the validation set to get some performance metrics per class.

Finally, we train the two models over the concatenation of the training set and the validation set and obtain metrics over the test set (see section 3).

¹ We acknowledge the strategic advice at this point from INSEE, already pioneering the use of Fasttext with the same purpose for the French *Nomenclature d'activités française*.

All the methodology is implemented in Python functions, classes, and libraries. A wrapper class is built for the Fasttext model class. We have used internal git repositories (Gitlab) to manage code versioning.

2.1. Ongoing tasks

The following tasks to improve the performance of the automatic coding machine are in an early phase, so results are still in their way:

- Metaoptimization: a metaoptimization procedure is to be run to choose the metaparameter combination that provides best results over the validation set.
- Conformal prediction: Fasttext provides scores for its predictions, but they cannot be read as rigorous probabilities. We are testing conformal prediction [2] and some variations over the basic method to be able to output sets of classes that guarantee a given confidence level.
- Confident learning: even after the quality efforts described above, the training set has some misclassified instances or inconclusive texts. Confident learning [3] can help with this issue, at least quantifying it and, hopefully, cleaning some problematic samples.

3. RESULTS AND PRACTICAL APPLICATION

We are developing a web interface and a web server that will allow both internal and external users to access the automatic coding machine, both manually and programmatically. Besides this, we offer “mode” as a parameter with two possible values:

- assistance: all results with a score higher than 1% will be shown, with a maximum of 15.
- codification: only the class with the highest score is returned.

The results of the two coding machines (CNAE-2009 and CNAE-2025) with their double modes are shown in figures 1 and 2. These represent results not including the 90000 instances coming from the ad-hoc survey currently in its collection phase for the next two months. This is expected to impact both models, especially for the CNAE-2025 (corresponding to NACE Rev 2.1).

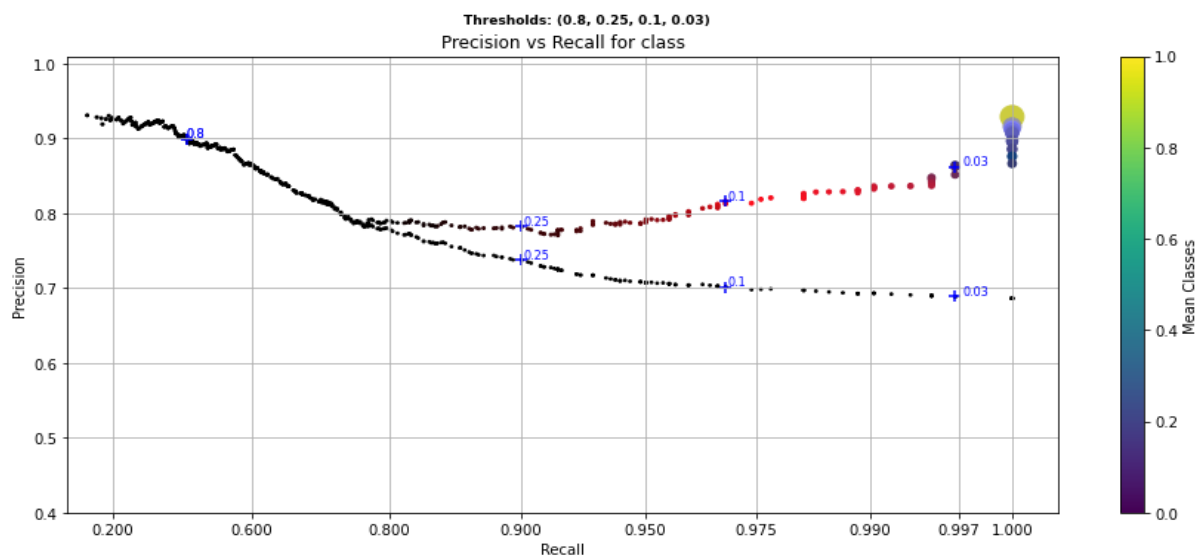


Figure 1. Here we overlap two precision vs recall curves for CNAE-2009. The lower one is computed in the “classical” way, getting just one prediction per instance. The upper one, however, considers all the predictions with score higher than the threshold, and they are

considered “hits” if any of them matches with the ground truth. The colour and size of the dots indicates the mean number of predictions.

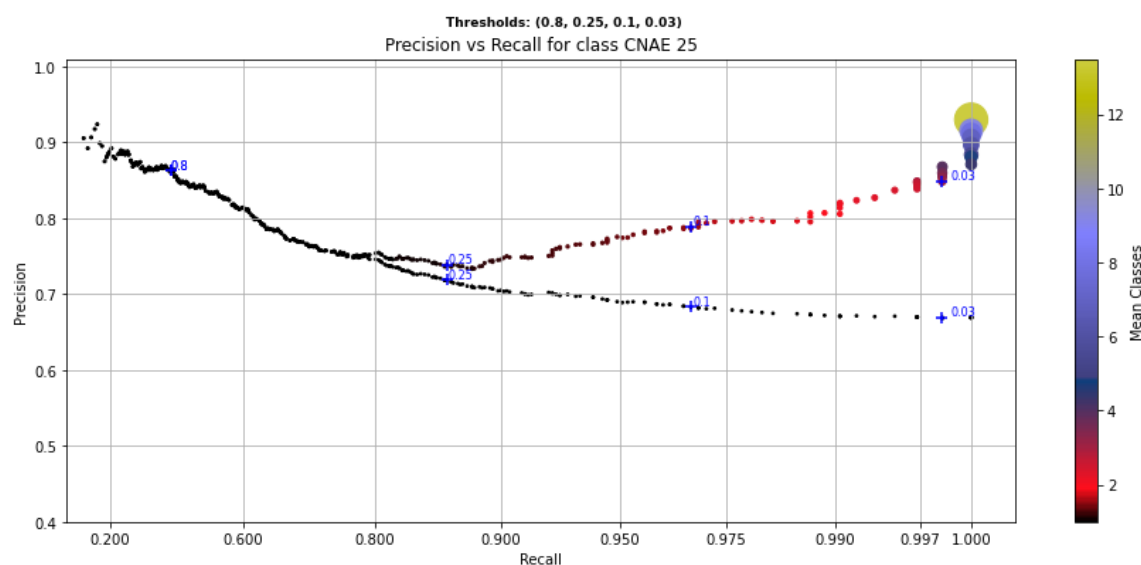


Figure 2. Analogous to Figure 1 but showing here the results for CNAE-2025 (corresponding to NACE Rev 2.1).

The web server and interface will be openly available for external users. Administrative-data holding institutions will be encouraged to use this statistical service.

3.1. Comparison with the current rule-based coding machine (AUTOCOD)

Table 1 shows that the performance of the Fasttext approach is more accurate than our current rule-based version (so-called AUTOCOD). The recall of the rule-based coding algorithm is fixed, while that of Fasttext can be adjusted. The same happens with the possible number of outputs.

Table 1. CNAE-2009 Fasttext coding vs rule-based coding

	Precision	Recall
Fasttext	0.69	0.94
Rule-based	0.45	0.94

4. MAIN FINDINGS

- “Classic” NLP models, and particularly Fasttext, show a huge performance improvement over the rule-based previous solution.
- This kind of solution allows the implementation of automatic coders as well as a coding assistance tool.
- Ensuring data quality both for training and for testing is key and a still open challenge.
- Providing a legal support for the incorporation of CNAE-2025 in all administrative registers, combined with the external use of the automatic coding machine will increase the quality of administrative data for statistical purposes.

5. REFERENCES

- [1] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: vol. 2:427 - 431, 2016
- [2] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. Foundations and Trends® in Machine Learning, vol. 16, issue 4: 494 - 591, 2022
- [3] Curtis G. Northcutt, Lu Jiang, Isaac L. Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence Research, vol. 70:1373 - 1411, 2022

An open innovation platform to share ideas and knowledge, and an open-source project for NSO's to work with their own sensitive data

Frédéric Comte (Insee)

1. INTRODUCTION

As the demand for advanced computational power continues to grow, especially with the increasing integration of machine learning (ML) and artificial intelligence (AI) into data analysis, organizations face new challenges. National Statistical Organizations (NSOs), in particular, must provide modern development environments at scale to support these advanced technologies. **SSPCloud** offers a common platform where NSOs can share knowledge, collaborate, and experiment with new ideas. This open innovation platform enables NSOs to develop, test, and refine methodologies and tools, addressing the rising computational demands. Additionally, **Onyxia**, an open-source software, allows each NSO to create its own datalabs under their own governance and security rules. This ensures that organizations can manage sensitive data and maintain autonomy while benefiting from the shared innovations within SSPCloud.

2. ADDRESSING COMPUTATIONAL CHALLENGES THROUGH A COMMON SANDBOX FOR NSOs

As the demand for advanced data processing capabilities continues to grow, particularly with the increased use of machine learning (ML) and artificial intelligence (AI), National Statistical Organizations (NSOs) must find innovative ways to collaborate and share resources. SSPCloud provides a common sandbox environment where NSOs across Europe can come together to share knowledge, ideas, and best practices in a non-sensitive, open space that drives innovation.

Sharing a platform for sensitive data is a highly complex and resource-intensive project, requiring intricate governance, security measures, and compliance with strict regulations. However, by focusing on a **non-sensitive sandbox** like SSPCloud, NSOs can bypass many of these challenges while still reaping significant benefits. This approach enables organizations to share computational power, tools, and methods without the added risk of exposing sensitive information, making it a more pragmatic and scalable solution.

One of the key advantages of SSPCloud is its ability to address the challenges associated with scarce computational resources, such as GPUs. These resources are vital for processing large datasets and running advanced ML/AI models, yet they are often limited and expensive. By leveraging SSPCloud's shared infrastructure, NSOs can efficiently utilize these resources, making them accessible to multiple organizations without the need for duplicative investments.

Moreover, SSPCloud offers a unique opportunity for **shared training**. By standardizing training materials and methodologies, NSOs can improve the skill sets of their teams, ensuring that everyone has access to the same level of expertise. This shared approach helps accelerate learning and enhances collaboration across the statistical system.

A further benefit is **reproducibility**. With SSPCloud, NSOs can work in an environment where research and analysis can be reproduced and validated across different organizations, ensuring transparency and reliability in statistical findings. This is crucial for maintaining the integrity and credibility of data-driven decisions.

Additionally, SSPCloud serves as a **showroom for innovations**. It provides a platform where NSOs can experiment with and showcase new ideas, tools, and methodologies. This not only fosters innovation but also enables organizations to learn from each other's breakthroughs, creating a dynamic ecosystem of continuous improvement.

In summary, SSPCloud offers a reasonable and effective solution for collaboration and innovation, allowing NSOs to share resources and ideas without the complexities of managing sensitive data, while still addressing the key challenges of modern data processing.

3. ONYXIA – AN OPEN-SOURCE SOLUTION FOR SECURE, AUTONOMOUS DATALABS

While SSPCloud provides National Statistical Organizations (NSOs) with a platform for sharing open data and facilitating training, it doesn't fully address their need to deploy machine learning models in production on sensitive data. This is where **Onyxia** comes into play: as an open-source software project, Onyxia empowers organizations to create and manage their own secure datalabs, offering the autonomy and control necessary to innovate with advanced analytics on sensitive data within their own governance and security frameworks.

One of the core principles of Onyxia is **to avoid lock-in**, not only from proprietary software vendors **but even from Onyxia itself**. Onyxia is designed to be agnostic and highly flexible, capable of deploying nearly any data analysis tool you require. Whether you need Jupyter notebooks, RStudio, SAS, Spark, TensorFlow, or any other specialized software, Onyxia can seamlessly integrate and manage these applications within your datalab environment. Additionally, Onyxia enhances transparency by emulating a terminal that shows the equivalent commands a user would have executed manually. This feature allows you to see exactly what operations are being performed behind the scenes, fostering a deeper understanding and trust in the system. By providing both flexibility in tool deployment and transparency in operations, Onyxia ensures that you're not constrained by proprietary formats or tied to specific technologies, allowing you to move your applications and data freely as your needs evolve.

If your organization is already utilizing cloud technologies like **Kubernetes** and **S3**, Onyxia can enhance your control and customization capabilities. Onyxia integrates smoothly with your existing cloud infrastructure, allowing you to leverage your current investments. It abstracts the complexities of managing cloud environments, making it easier for your teams to deploy and maintain datalabs, and enables you to tailor your computational environments to meet specific project needs without the limitations imposed by proprietary software.

For organizations planning to transition to cloud technologies, Onyxia serves as an ideal starting point. Its user-friendly interface lowers the barrier to entry, enabling teams without deep technical expertise to utilize advanced cloud features. It allows for a flexible transition, adopting cloud technologies at your own pace, with the ability to scale and adapt as your organization's needs evolve. As an open-source solution, Onyxia reduces initial costs associated with proprietary software licenses.

4. CONCLUSION

In conclusion, the One-Stop Shop for AI and Machine Learning for Official Statistics (AIML4OS), specifically Work Package 3 (WP3), plays a pivotal role in advancing the capabilities of National Statistical Organizations (NSOs) by enhancing the SSPCloud experience, particularly in the utilization of generative and by assisting NSOs in deploying Onyxia and expanding its functionalities. This focused effort enables NSOs to leverage cutting-edge technologies within collaborative platforms and secure environments, empowering them to innovate, handle sensitive data effectively, and deploy machine learning models in production with greater ease and efficiency.

Applicable AI in Official Statistics: A New Research Theme at Statistics Netherlands

Marco Puts (Centraal Bureau voor de Statistiek)

1. INTRODUCTION

In order to fulfil the growing demand for data-driven decision-making, Artificial Intelligence (AI) has been integrated across a wide range of fields, including official statistics. The rise of artificial intelligence, however, raises concerns about its appropriateness and reliability, especially when deployed in production. Our methodology department at Statistics Netherlands introduces Applicable AI as a new research theme to address these issues. As a central research question, we ask: how do we ensure that AI models not only perform well on test sets but also remain relevant to their intended applications? While it may seem straightforward, this question is complex. Since AI models can perform well on test data, they may struggle to generalize in real-world applications, where the data differs significantly from the training set. Additionally, ensuring that AI systems adhere to ethical standards, such as transparency, fairness, and bias prevention, requires careful consideration of their underlying algorithms and data sources. Finally, AI models should be evaluated for both their internal accuracy and external validity, i.e. how relevant and applicable they are to a broader population or domain they are intended for.

As we proceed, we will examine how rigorous methodological frameworks can be integrated into AI development to address these challenges. The focus will be on enhancing external validity, increasing ethical transparency, and ensuring that AI models remain adaptable to real-world statistical applications.

2. QUALITY OF OFFICIAL STATISTICS

Before exploring AI applicability, it is crucial to define what constitutes a "good product", or in our case, a "good model". We should begin our applied research by examining the standards that are used to evaluate quality: What are the ethical and quality standards for AI and machine learning models? And how do these affect our choice of methodology and technique?

According to the European Statistics Code of Practice [1], official statistics adhere to strict standards of quality and ethics. According to [3] and [6], these normative frameworks raise concerns about transparency, accuracy, and comparability when applied to machine learning. For AI to be effectively integrated into official statistics, these challenges must be addressed. Similarly to the European Statistics Code of Practice, relevant frameworks, such as Responsible AI and Trustworthy AI [7], offer detailed guidelines. While maintaining data confidentiality and privacy, both systems emphasize high-quality output, transparency, and independence. However, the emphasis of these frameworks on the ethical implications with respect to, a.o., biases could be beneficial for official statistics.

3. METHODOLOGY: AI AND MACHINE LEARNING IN OFFICIAL STATISTICS

There is no doubt that AI has demonstrated remarkable performance across a wide range of tasks, but its statistical underpinnings are often overlooked. Using the Total Survey Error (TSE) Model [2] survey data quality can be evaluated, accounting for both measurement error and non-response bias as part of the overall evaluation. The TSE model can also be utilized when creating training datasets [4]. The Total Machine Learning Error Model [5] (see Figure 1) has recently been introduced to identify potential errors that may occur in machine learning models. The total machine learning error model encompasses both representational and measurement dimensions, which is like the TSE model. It is important to remember that training a machine learning model is estimating a set of parameters of the model, and hence, the training process is actually an estimation process.

As the training set represents a sampling of a population, it is possible to introduce coverage and sampling errors because of its being a sample. As a result of these errors, a model will not be as representative as it should be. Training data is comprised of cases that are described by measurements. When supervised learning is performed, the variables involved are features (independent variables, X) and targets (dependent variables, y). All these measurements are subject to errors. As a final step, the model is chosen, along with the assumptions that go along with it.

Moreover, the model's assumptions and any errors in the training set must be explicitly acknowledged to ensure accurate results within a specific domain. The training data must be representative of the problem domain. This model concludes that while independent test sets assess internal validity, external validity is often neglected, which is critical for official statistics.

Models are trained on populations samples, which introduces sampling errors. These errors are influenced by the size of the training set and the distribution of classes in classification tasks, and they are affected by the absolute size of the training set. Training in an unbalanced manner exacerbates sampling error in rare event classes, where the rare event class comprises less than 1% of the data and the other class comprises more than 99%. In contrast, balancing the training set may result in insufficient data for the dominant class, resulting in high sampling error and poor generalizability of the model.

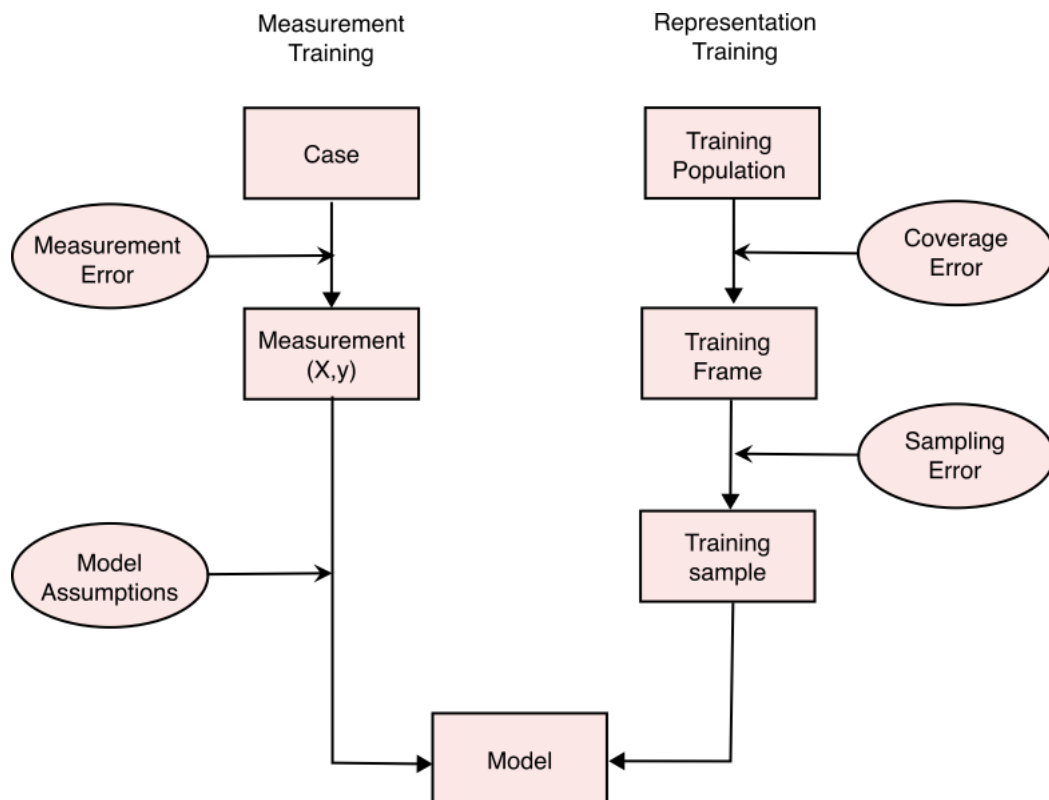


Figure 1. The Total Learning Error Model. The model consists of representation and measurement of the training data. Training set (training sample) is seen as a sample of the training population. The training set contains measurements and a model is trained with respect to model assumptions

4. AI APPLICATION VS. APPLICABLE AI

By applying AI to real-world statistical problems, we aim to ensure that AI models are suitable for their intended purposes. Often, AI is deployed purely technically, without considering the methodological challenges associated with data modeling. A new methodology is needed to ensure that AI will be applicable in the future, guiding statisticians as well as data scientists to make sure that it will be fit for purpose in official statistics.

By developing a comprehensive understanding of Applicable AI within a statistical context, we will contribute to the development of reliable and responsible AI solutions that are aligned with the needs of statisticians and policymakers.

5. DISCUSSION

It is important for Statistics Netherlands to focus on the need for AI models that are both high-performing and purpose-driven as part of the methodology research program. While AI models excel in isolated testing environments, they may not meet the stringent standards required for official statistical applications, where accuracy, transparency, and ethical concerns are paramount. Using principles from the European Statistics Code of Practice, responsible AI frameworks, and advanced methodological approaches, we aim to bridge the gap between AI's performance and its applications in real-world settings.

According to our research, AI models should align with rigorous quality standards found in official statistics, so that policymakers, researchers, and the public can trust these systems. As part of this effort, model accuracy needs to be improved, transparent decision-making processes need to be improved, biases need to be addressed, and data privacy regulations must be adhered to. Moreover, we strive to make AI models generalizable across a variety of data sets and capable of solving a variety of statistical challenges.

Through the development of AI systems that meet these criteria, we hope to improve the quality and utility of official statistics by creating reliable, trustworthy, and applicable AI solutions. In order to ensure AI technologies are fit for purpose in addressing complex societal issues through evidence-based decision-making, statisticians will need to leverage its potential while adhering to ethical standards.

6. REFERENCES

- [1] European Statistical System Committee. (2018). *European statistics code of practice: For the national statistical authorities and Eurostat (EU statistical authority)*. Publications Office of the European Union. <https://doi.org/10.2785/798269>
- [2] Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Wiley.
- [3] Puts, M. and P. Daas (2021, July). Machine learning from the perspective of
- [4] official statistic. *The Survey Statistician* 84, 12–17.
- [5] Puts, M. J.H., A. da Silva, L. D. Consiglio, I. Choi, D. Salgado, C. Clarke, S. Jones, and A. Baily (2022). ONS-UNECE Machine Learning Group 2022. Quality of Training Data. Theme Group Report. Technical report, UNECE.
- [6] Puts, M.J.H., Salgado, D., Daas, P. (2024) Leveraging Machine Learning for Official Statistics: A Statistical Manifesto. To appear in the proceedings of the conference on Foundations and Advances of Machine Learning in Official Statistics, Wiesbaden, 2024. <https://arxiv.org/abs/2409.04365>
- [7] Yung, W. & Tam, Siu-Ming & Buelens, Bart & Chipman, Hugh & Dumpert, Florian & Ascari, Gabrielle & Rocci, Fabiana & Burger, Joep & Choi, InKyung. (2021). A quality framework for statistical algorithms. *Statistical Journal of the IAOS*. 38. 1-18. 10.3233/SJI-210875.
- [8] See website <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Applying a Non-Parametric “Benefit of the Doubt” Approach for Identifying Low-Quality Responses in Online Panels

Marco De la Cruz, Femke Dekeulenaer (IPSOS), Katarzyna Bankowska & Athanasios Tsiortas (European Central Bank)

1. INTRODUCTION

In recent years, there has been a significant upsurge in the use of online panel surveys for research purposes, including among statistical offices. This trend has been driven by the advantages of online surveys, such as lower costs, faster data collection and the ability to reach a wider audience. However, the increasing reliance on high-frequency online surveys, has also brought new potential data-quality issues (e.g., multitasking while completing the survey or completing the survey in a distracting environment) (Zwarun & Hall, 2024). As there are no universally specified thresholds for assessing quality, alternative approaches can be explored. Using data from the European Central Bank’s Consumer Expectations Survey (CES), this paper presents a non-parametric Benefit-of-the-doubt (BoD) model for identifying low-quality responses by integrating existing indicators, such as speeding, excessive use of non-response options, satisficing, response inconsistency and straightlining.

The BoD composite indicator of low-quality responses combines different indicators by assigning weights to the existing unidimensional indicators in an endogenous way so that each respondent is evaluated under the best possible light (Verbunt et al., 2022, Cherchye et al. 2007). This allows for a more benevolent, yet objective and comprehensive assessment of response quality and allows the identification of patterns and behaviours that may not be captured by single-based indicators. We compare the results from the BoD composite indicator and the individual indicators to evaluate their alignment in flagging low-quality responses.

Despite the widespread use of BoD to construct composite indicators in various fields, its application to assess data quality in online panels remains relatively underexplored. As such, this study contributes to the development of alternative methods for assessing survey data quality and offers insights into the benefits and limitations of the BoD approach in the context of high-frequency surveys.

2. METHODOLOGY

The study utilizes data from one specific wave of the European Central Bank’s Consumer Expectation Survey. The CES is a high-frequency survey that collects information on consumers’ perceptions and expectations about the economy, divided into two key survey modules: core and quarterly. Our selected wave covers both these modules. A set of three quality indicators, capturing different aspects of response behaviour, is used to construct the BoD model. The quality indicators considered in this study are: 1. Speeding - measured by the overall module response time; 2. Response patterns - including “unchanged” category answer across main questions, straightlining in grid questions, and consistency of open-ended and

probabilistic bin forecasts¹; and 3. Percentage of non-response – use of “don’t know” or “prefer not to answer”, whenever these are available in a question. We summed up the indicators in point 2. and reversed the score when necessary to create a single ‘global key score’. Note that the ‘percentage of non-response’ was also reversed in this process to ensure that higher values in each indicator correspond to higher quality.

2.1. Implementation of the Benefit-of-the-Doubt approach

To evaluate overall response quality, we employ the BoD method introduced by Cherchye et al. (2007). It allows aggregation of different dimensions - in our case, quality indicators - by assigning weights that present each respondent in the most favourable light. This method offers three key advantages: 1. it is objective, as the weights are endogenously assigned; 2. it is fair, as it evaluates the responses under the best possible light; and 3. it is easy to interpret, as it delivers scores between 0 and 1, with higher scores corresponding to better quality. Technically, the BoD approach is equivalent to the input-oriented Data Envelopment Analysis (DEA) model proposed by Charnes et al. (1978), assuming constant returns to scale and a unitary input level for all observations. Like the DEA, the BoD framework can impose weight restrictions to accommodate experts’ or researchers’ preferences for giving more or less weight to certain indicators. The BoD model is a non-parametric optimization problem, typically expressed as follows:

$$\max(u_r) \text{ s.t. } \theta = \sum_{r=1}^s u_r y_{j,r} \text{ and } \sum_{r=1}^s u_r y_{j,r} \leq 1 \quad j = 1, \dots, n \text{ and } u_r \geq 0 \quad r = 1, \dots, s$$

Where y_{it} is the value of the unit j (in our case, a respondent) along the dimension r (for example speeding), u_r is the weight assigned to the r th dimension. Note that θ represents a quality indicator since all the dimensions considered are positively associated with quality (higher values for interview duration, percentage of responses, and the global key corresponding to better quality). To evaluate the BoD model’s performance, the BoD quality scores are compared with the single-indicator measures of response quality.

3. RESULTS AND PRACTICAL APPLICATION

The BoD scores were computed separately for two groups of respondents (the NPS and PS groups)² to assess the quality of survey responses. As shown in the table, both groups achieved high quality scores in terms of the quality measures included in the BoD scores, with the NPS group slightly outperforming the PS group. The mean quality scores for the PS and NPS groups were 0.9826 and 0.9893, respectively.

The majority of respondents in both groups were of very high quality, with 74.1% of respondents in the NPS group and 71.5% in the PS group achieving the highest possible score of 1. Setting a threshold of less than 0.9 for detecting ‘low’ quality responses in the BoD scores for PS and NPS groups would mean that 6.5 and 2.3 percent respectively of respondents would be labelled as ‘low’ quality.

¹ The first group of variables indicates whether the sum of responses in bins is not equal to 100, and the other group indicates whether the point estimate provided by the respondent is not within the bin range of the respective questions.

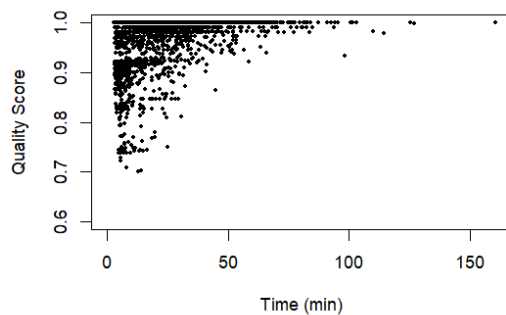
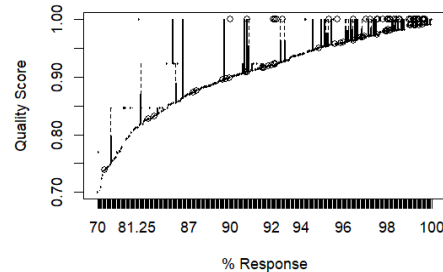
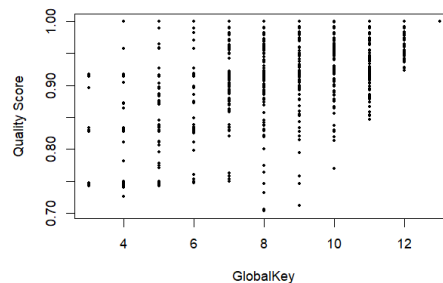
² The PS group is recruited into the CES panel via probability sampling methods and the NPS groups via non-probability sampling methods.

Table 1. Summary of the BoD quality indicator for PS and NPS

	Min	1 st Qu.	Median	Mean	3 rd Qu.	Max.	% low Qua.
PS	0.703	0.99	1.00	0.98	1.00	1.00	6.5
NPS	0.627	0.99	1.00	0.98	1.00	1.00	2.3

Note: Low quality is defined here with 0.9 cut-off.

To have insight on what is the contribution of each single indicator (duration, global key, percentage of response) on the BoD, we compare the results of this analysis with more classical quality measurements. In Figures 1, 2, and 3, we report the relationships between the BoD and each of the single indicators for the PS sample. The figures show a clear positive relation between the BoD and the single indicators. In the case of survey length, the relation seems to be logarithmic (very steep for low values of duration and almost steady after a certain threshold). The relation between the BoD and the percentage of answered questions and between the BoD and the global key appears instead to be linear. The concentration of points around higher quality scores as the indicator values increase can be attributed to the optimization process of the BoD model. By assigning weights that maximize the quality scores, the model effectively captures the positive relationship between the indicators and response quality. The dispersion of points for intermediate values of the single indicators can be explained by the presence of other indicators in the BoD model which allows for compensations.

Figure 1. Relationships BoD quality scores and time for the PS**Figure 2.** Relationships BoD quality scores and percentage response for the PS**Figure 3.** Relationships BoD quality scores and global key for the PS

To further test the correspondence of the BoD quality index and the single indicators, we compared the respondents flagged as low quality by the BoD indicator to those identified by the single-based indicators (see Table 2). This enabled us to assess not only the accuracy of the BoD indicator but also to identify any low-quality cases that were not detected by the other measures. The BoD quality score appears to be a useful measure. While it flags a significant proportion of the same respondents as the single-based indicators (61% of those flagged as

low quality by Global Key are also flagged by BoD, 97% overlap with % Response, and 37% overlap with Duration), it also identifies additional unique cases not flagged by these other measures. This suggests that the BoD provides added value by capturing aspects of survey data quality that may be missed by the single-based indicators.

Table 2. Low-Quality Respondent Cross-Identification by Indicators

PS	BoD (363)	Global Key (198)	% response (368)	Duration (538)
'Low' quality BoD	1	0.61	0.97	0.37
Global key	0.33	1	0.32	0.18
% response	0.98	0.61	1	0.36
Duration	0.55	0.48	0.53	1

Note: The numbers in parentheses represent the raw values of low-quality cases identified for each indicator. For the Global Key, cases with a score below 7 are flagged as low quality. For the % Response measure, cases with less than 90% response are flagged. For the Duration measure, low quality is defined based on specified thresholds rules for NPS and PS.

4. MAIN FINDINGS

In this study, the BoD approach is considered for the construction of a comprehensive quality indicator. Results show that the BoD approach is in general more benevolent than considering single indicators, as it allows for compensations between indicators. Interestingly, the BoD has also a strong positive relation with each of the indicator, which can be fine-tuned by imposing specific weight restrictions. This possibility makes the model particularly appetible in institutional contexts. By contrasting the respondents identified as low quality from the BoD and the single quality indicators (global key, percentage of responses and duration) it is also possible to verify the extent to which the different quality measures overlap.

5. REFERENCES

- [1] Cherchye, L., Moesen, W., Rogge, N., & Puyenbroeck, T. V. (2007). An introduction to 'benefit of the doubt' composite indicators. *Social indicators research*, 82, 111-145.
- [2] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, 2(6), 429-444.
- [3] Verbunt, P., Rogge, N., & Puyenbrock, T.V. (2022). Non-parametric estimation and evaluation of capability sets. *Oxford Economic Papers*, 74(1), 265-296.
- [4] Zwarun, L. & Hall, A. (2014). What's going on? Age, distraction, and multitasking during online survey taking. *Computers in Human Behavior*, 41, 236-244.

Applying Machine Learning to educational level: An empirical evaluation

De Fausti Fabrizio, Di Zio Marco, Filippini Romina & Simona Toti (Italian National Institute of Statistics, Istat)

Abstract: This paper presents the results of an empirical evaluation of Random Forest, Recurrent Neural Network and Long Short-Term Memory applied to administrative longitudinal data. These models are used to predict the attained level of education of the Italian population. For the assessment of the methods, the estimates are compared with those obtained with the Istat official census procedure and the ones based on administrative data.

1. INTRODUCTION

Since 2018, the population census has yearly carried out, and it strongly relies on the integration of the information stored into registers with data specifically collected through a sample survey. The Base Register of Individuals (BRI) is the population register at the basis of the new population census. BRI, primarily derived from administrative sources, is constructed at individual level and includes demographic information such as gender, age, place of birth and citizenship. However, it does not contain all the data traditionally collected by the census, such as the attained educational level (ALE). The Ministry of Education (MIUR) provides administrative data on ALE. They refer to people enrolled in a school course from 2011 onward, capturing only students enrolled in educational programs held in Italy. It excludes courses like Fine Arts, Drama, Dance and Music academic diplomas, as well as other training and vocational programs managed by Italian Regions which are not required to report data to MIUR. Timeliness is another critical issue, since MIUR data are typically available in Istat with a delay of 1 or 2 years compared to the BRI reference time.

The official procedure for estimating ALE relies on different prediction models, combined to address sub-populations characterized by varying amounts of information [1]. For the subset of individuals with available administrative information, ALE at time t is predicted using data at time $t-2$ and $t-1$, from which the probability of achieving a new educational level is estimated. ALE is predicted by using the estimated conditional probabilities of attaining a certain level of education given some characteristics of the people. In the procedure currently in use, log-linear models are used to estimate those probabilities. In [2], ML methods are analysed as an alternative to the official procedure, the results of the experiments showed similar performances. Here, we study ML models such as Random Forest (RF), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) that leverage the longitudinal information available from administrative sources. It is worthwhile noting that, once the data time-lag has elapsed, updated administrative data become available. This is particularly useful to assess the performance of the methods. In fact, in 2023, administrative data pertaining to 2021 are available and can be used as the gold standard for evaluating results generated by the ML models and the official procedure.

2. METHODS FOR THE ESTIMATION OF ALE

The general idea is to estimate a model for the prediction of ALE given the values of known covariates X . Considering $I^{(t)}$ as the target variable (i.e., ALE at time t), the conditional probabilities $h(I^{(t)} | X)$ are (in the current procedure) estimated through LL models, and I_t is predicted by randomly selecting a value from this distribution. As an alternative to LL, we aim to study the performance of machine learning methods that consider longitudinal information for the estimation of $h(I^{(t)} | X)$. RNN, where a recurrent connection is introduced in the network, is a possible approach to dynamic ML ([3], [4]). Each neuron in an RNN maintains a hidden state, which captures information about previous inputs in the sequence and influences future predictions. This capability makes RNNs particularly effective for tasks involving sequential data processing. If the series are more than 10 time points, RNN can suffer from difficulty in capturing long-term dependencies and is vulnerable to vanishing or exploding gradient problems. LSTM is a specialized type of RNN architecture designed to address the limitations of RNNs in capturing long-range dependencies and mitigating the vanishing gradient problem [5]. Finally, we consider the RF [6] which is an ensemble learning method used for classification, regression, and other tasks. RF do not explicitly accounts for time, it operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees.

3. THE STUDY

The dataset for the application consists of approximately 475,000 individuals, aged 9 or older, residing in the Emilia Romagna region (NUTS 2) in 2021 with complete longitudinal administrative information on ALE from 2015 to 2021, classified into 7 modalities (11.5% of the total reference population). The covariates considered are gender, age, province of residence and the school enrolment. In the LL model, only the course attended in the previous year is considered in X . In the ML methods the training set is composed by all the individuals with complete longitudinal information from 2015 to 2020 ($t-1$), meaning that entire series of available longitudinal information is taken into account. The estimated conditional probabilities are then applied to one-year-forward shifted data (test set), and ALE in 2021 (t) for each unit in the dataset is obtained by randomly selecting a value from the estimated ALE probability distribution.

The ALE ($\hat{I}$) estimated through the LL and the ML methods are compared with the data from MIUR (I), considered as the target ALE value (Table 1). Since the predictions are obtained through a random draw from the estimated distributions, the procedures are repeated 100 times to take into account the variability, and the results are computed averaging over those repetitions.

Table 1. Estimated absolute values (a.v.) and percentage (%) distribution of ALE with RNN, LSTM, RF, LL and administrative data (AD) in 2021*.

	RNN		LSTM		RF		LL		AD	
ALE	a.v. (.000)	%	a.v. (.000)	%	a.v. (.000)	%	a.v. (.000)	%	a.v. (.000)	%
Primary ed.	120.1	25.3	120.1	25.2	120.3	25.3	120.2	25.3	120.6	25.4
Lower sec. ed.	200.0	42.1	200.1	42.0	200.0	42.1	200.0	42.1	200.4	42.1
Upper sec. ed.	109.1	22.9	109.9	23.1	109.2	23.0	109.3	23.1	109.1	23.0
Bachelor's deg.	31.4	6.6	31.3	6.6	31.1	6.6	31.0	6.5	30.7	6.5
Master deg.	14.6	3.1	14.0	3.0	14.5	3.1	14.5	3.0	14.4	3.0
PhD	0.2	0.0	0.2	0.0	0.2	0.0	0.1	0.0	0.2	0.0
Total	475.4	100.0	475.4	100.0	475.4	100.0	475.4	100.0	475.4	100.0

*The class “no educational attainment” represents a residual class and is not included in the table. This accounts for the discrepancy in the total amounts.

All the methods provide an estimated distribution that is close to the AD ALE distribution, considered as target variable. Relative errors RR_i for each modality i are computed: $RR_i = \frac{(\hat{I}_i - I_i)}{I_i}$, where I_i is the AD absolute frequency of modality i and $\hat{I}_i$ is the corresponding estimated value. Table 2 shows the mean of the relative errors $m(RR_i)$ computed over the 100 repetitions. Examining the mean of the relative errors, the RF method produces the best results, with $m(RR_i)$ being the lowest among all methods. We notice high values of RR_i for the PhD category, likely due to the very low frequency of this modality. MLs show a sensible improvement, especially in this class, with respect to LL.

Table 2. Mean relative error $m(RR_i)$ and standard deviation (std) computed over 100 runs for RF, RNN, LSTM

	RF		RNN		LSTM		LL	
ALE	$m(RR_i)$	(std)	$m(RR_i)$	(std)	$m(RR_i)$	(std)	$m(RR_i)$	(std)
Primary ed.	0.235	(0.132)	0.416	(0.042)	0.435	(0.032)	0.214	(0.017)
Lower sec. ed.	0.176	(0.110)	0.164	(0.079)	0.151	(0.071)	0.329	(0.014)
Upper sec. ed.	0.097	(0.059)	0.694	(0.462)	0.719	(0.561)	0.458	(0.078)
Bachelor's deg.	1.343	(0.417)	1.781	(1.222)	1.996	(1.689)	1.026	(0.317)
Master deg.	0.874	(0.429)	2.534	(1.720)	3.278	(1.380)	0.534	(0.361)
PhD	6.844	(4.529)	7.898	(5.744)	7.244	(4.936)	35.924	(2.072)
Mean	1.595		2.247		2.304		6.414	

Table 3 shows the f1 score for each ALE class and the global f1 score. The f1 score is given by $f1 = TP / (TP + 0.5(TP + TN))$ where TP is the number of correctly predicted instances belonging to the positive class, FP is the number of instances erroneously predicted as positive and FN is the number of instances erroneously predicted as negative. The global f1 score (also known as micro f1) requires TP, FP, FN, and the partial TP, FP, FN obtained by defining each time one of the 7 classes as positive. Those measures are computed for a micro-level evaluation. The f1 score ranges from 0 and 1, with 1 indicating a perfect match. Table 3 shows similar scores, with a preference for RNN and LSTM. This preference increases with higher ALE classes. Specifically, for modalities up to “upper secondary education” the f1 scores are close each other. However, for the “bachelor and master degree” classes there is a more sensible difference, and for ‘PhD’ class the f1 score for RNN and LSTM is significantly higher.

Table 3. F1 score computed over 100 runs

ALE	RF	RNN	LSTM	LL
Primary ed.	0.9950	0.9965	0.9966	0.9959
Lower sec. ed.	0.9889	0.9919	0.9920	0.9913
Upper sec. ed.	0.9138	0.9289	0.9286	0.9209
Bachelor's deg.	0.6910	0.7344	0.7340	0.7103
Master deg.	0.7063	0.7618	0.7613	0.7578
PhD	0.6305	0.8160	0.8484	0.6059
Global f1	0.9450	0.9548	0.9548	0.9507

The code is developed using Python, while LL models are developed in Sas.

4. MAIN FINDINGS

The paper shows the results of an empirical evaluation about the use of RF, RNN and LSTM for longitudinal data used for the prediction of ALE. To highlight the advantage offered by ML methods, the results are compared with those obtained from the official procedure, which considers only the most recent year as a covariate. The results show that RF is preferable when distributional accuracy is our main interest. On the other hand, when the interest is in the micro-prediction, LSTM has a slight better performance. In general, all the ML models show a better behaviour with respect to the current procedure. It should be noticed that the data used are not characterised by a long series, which is an important factor to consider when interpreting and generalizing the results. The good performance of RF, that are not introduced in literature mainly for dealing with the time dimension of the data, could be explained by the fact they estimate the conditional distribution without any particular constraint used for taking into account the time. This is certainly a problem when a long time series is the input of our case and in this situation models as RNN and LSTM, specifically introduced for these problems, can be more efficient. Further studies will be devoted to the cases presenting those characteristics.

5. REFERENCES

- [1] Di Zio, M., Filippini R., Rocchetti G. (2019). An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. *Rivista di Statistica Ufficiale*, N. 2-3/2019.
- [2] De Fausti, F., Di Zio, M., Filippini, R., Toti, S., & Zardetto, D. (2022). Multilayer perceptron models for the estimation of the attained level of education in the Italian Permanent Census. *Statistical Journal of the IAOS*, 38(2), 637-646.
- [3] De Mulder, Wim, Steven Bethard, and Marie-Francine Moens. "A survey on the application of recurrent neural networks to statistical language modeling." *Computer Speech & Language* 30.1 (2015): 61-98.
- [4] Sherstinsky, Alex (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404: 132306.
- [5] Hochreiter, Sepp, and Jürgen Schmidhuber (1997). Long short-term memory. *Neural computation* 9.8: 1735-1780.
- [6] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Artificial Intelligence and Governance, driving innovation with responsibility

Massimo Fedeli, Francesca Orecchini & Sara Letardi (Istat)

1. INTRODUCTION

Thanks to its many practical applications, such as the ability to analyse large amounts of data in a short period of time, Artificial Intelligence (AI) is becoming increasingly important in all sectors and at various organisational levels. In particular, for those who are in charge of directing and steering the strategic choices of an organisation, it is crucial to have at their disposal tools that enable them to speed up the decision-making process and make it more accurate, without however compromising its reliability and adherence to ethical values. Indeed, in an ever-changing context characterised by increasingly complex dynamics, decision-making capabilities must evolve accordingly, reflecting awareness of the present situation rather than knowledge of the past.

At the strategic level, new opportunities for improving the services offered by public bodies can be identified through the use of AI solutions. Through the analysis of users' behaviour and preferences, it is indeed possible to optimise and personalise services, making them more responsive to users' needs.

Furthermore, AI is able to identify inefficiencies in processes and in the management of human and material resources, providing solutions at the top for a better allocation of resources. This not only reduces operational costs, but also improves the overall efficiency of operations, enabling a more strategic use of available resources.

Despite the opportunities described, not all organisations are ready to introduce it into their processes. A Gartner survey reveals that many companies still show mistrust and concern in the face of this emerging technology. Although AI offers numerous opportunities, its challenges cannot be ignored.

Indeed, the adoption of artificial intelligence entails a number of risks and critical issues related to its use, which decision-makers must take into account when evaluating and managing projects. Among the most relevant risks for an official statistical agency are those related to 'reputation', i.e. the possibility of disseminating inaccurate data, and those related to the 'black box', which make the decision-making process less transparent due to the complexity of the models used by AI. Other risks include the potentially uncontrolled costs of AI platforms and the possible lack of appropriate professional skills.

Among the main critical issues are the reliability of the data used to train AI models, potentially subject to bias, and those related to cyber security, such as the unwitting dissemination of confidential material (sensitive data, code, documents). Vulnerability to cyber attacks could also compromise the quality of decisions, especially in areas such as healthcare, finance and public administration, where sensitive data are processed. Therefore, the reliability of AI must be ensured through the use of robust protection systems and constant updates, in order to avoid privacy breaches and protect critical information.

Ultimately, it is essential that organisations address the challenges related to artificial intelligence with a balanced view, implementing effective governance systems that allow them to better manage both the opportunities and risks associated with this revolutionary technology.

2. METHODOLOGY

Several Artificial Intelligence projects have been launched at Istat, with the aim of promoting innovation in methods and technologies for producing and disseminating statistical information and increasing administrative efficiency, such as: chat bots to query databases, generative collaborative tools, multilingual translators for editorial production. In order to preserve the ethical values of the Institute, such as transparency, accountability and sustainability, a structured process to elaborate guidelines, best practices and monitoring tools was added to the innovation path, with the aim of introducing AI in a responsible manner and ensuring regulatory compliance [1].

The process consisted of several gradual steps:

- 1) Analysis of international national guidelines and Best Practices;
- 2) Reconnaissance and monitoring of the experimental projects already underway, in order to identify critical issues and points of attention;
- 3) Analysis of current legislation (AI ACT);
- 4) Identification of guidelines and best practices, listening to all stakeholders.

This process led to the definition of a governance framework peculiar to the Institute, characterised by three distinct and interacting areas: the technological area, the regulatory area and the human resources area. Indeed, it is essential to provide the Institute's structures that develop, apply or use artificial intelligence technologies with practical guidelines that can translate the principles and ethical values referred to into well-defined actions. Moreover, in light of these elements, it is clear that investing in training and skills development is crucial to ensuring transparency and accountability in the use of AI.

3. RESULTS AND PRACTICAL APPLICATION

The three areas just described constitute the backbone for AI governance at Istat and are characterised by peculiar tools and lines of action. For example, the technological area includes the analysis and evaluation of existing and developing AI applications at Istat and the tools to ensure that these solutions are developed, put into operation and monitored in accordance with the regulations and with the values and strategic objectives of the institution. The regulatory and normative area includes processes, procedures and guidelines required to implement the ethical and legislative principles of AI governance. Finally, the human resources area encompasses, in turn, aspects related to the definition of training plans needed to enhance the professional skills of human capital or to communication, such as those needed to accompany change. Once the framework has been defined, the next step concerns its activation, through an initial pilot phase that envisages the implementation of the activities described in the various areas of the framework and the use of monitoring tools to carry out periodic reviews of the solutions introduced.

4. MAIN FINDINGS

At the Institute, several agreements and experimental projects have been initiated to explore and develop artificial intelligence to support the statistical production process. In addition to those already mentioned, we mention the participation in research programmes for the development of Italian Large Language Models (LLM), carried out through agreements with public and private entities. The projects undertaken are fully in line with the Italian Artificial Intelligence Strategy, which promotes the development of advanced skills and technologies, the modernisation of the Public Administration and the collaboration with public and private entities for the adoption of AI-based solutions. With these initiatives, Istat actively contributes to the objectives of the strategy, promoting technological innovation and the responsible and advanced use of artificial intelligence in the Italian context [2].

Parallel to these first projects, at Istat, the path to activating the framework has begun, with the preparatory activities of collecting and analysing structured forms to collect information on individual projects and the assessment, in the light of the AI ACT principles, of regulatory compliance. In addition, training courses for staff on basic literacy, specialised training on international regulatory issues and specific courses for decision-makers were launched for the various competence profiles.

Driving the innovation brought by AI requires flexibility and speed of change, which is why the framework developed will be constantly updated to take into account both technological developments and regulatory adaptations.

5. REFERENCES

- [1] MÄNTYMÄKI, MATTI & MINKKINEN, MATTI & BIRKSTEDT, TEEMU & VILJANEN, MIKA 2022. Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance. [arXiv:2206.00335](https://arxiv.org/abs/2206.00335)
- [2] AGID, Strategia italiana per l'Intelligenza Artificiale: https://www.agid.gov.it/sites/agid/files/2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-2026.pdf
- [3] Gartner: <https://www.gartner.com/smarterwithgartner/would-you-let-artificial-intelligence-make-your-pay-decisions>

Automated Product Classification Using Machine Learning and Natural Language Processing: A Case Study of Scanner Data

Egle Kondrataite & Tomas Rudys (Methodology and Data Science group, State Data Agency, Statistics Lithuania)

1. INTRODUCTION

In recent years the availability of new or alternative data sources has shown a wide range of possible applications in different areas of official statistics. In price statistics, scanner data collected from private owners such as supermarkets are important for future improvements in calculating the consumer price index (CPI). Traditional methods for CPI calculation heavily rely on manual product classification, a process that is time consuming and resource intensive. By using machine learning (ML) and natural language processing (NLP) techniques, we have developed a fully automated pipeline that integrates, processes and classifies product-level data into the categories of two classifications of individual consumption according to purpose – ECOICOP and COICOP2018. The scope of this automation includes food, non-alcoholic beverages, alcohol and tobacco products, contributing to the efficiency and accuracy of CPI calculations while significantly reducing human effort and integrating new big data sources.

2. METHODOLOGY

To automate the classification of products into ECOICOP and COICOP2018 categories using scanner data, we used a supervised machine learning approach combined with NLP techniques to process product descriptions and train classification models. The problem was divided into two main blocks:

1) Food and non-alcohol beverages:

- 61 classes for ECOICOP,
- 65 classes for COICOP2018.

2) Alcohol and tobacco:

- 13 classes for ECOICOP,
- 8 classes for COICOP2018.

2.1. Data

A comprehensive dataset was manually classified by a team of 13 experts, covering more than 43,000 food and non-alcoholic beverage products and nearly 7,000 alcohol and tobacco products. The product description consisted of the product name, the internal category name and additional text information about the product provided by the supermarket. This text was used as the input for classification. The datasets were split into 80% training data and 20% test data.

2.2. Text pre-processing

To process the product descriptions for classification, we experimented with a number of NLP techniques to extract meaningful features from the data. Text pre-processing techniques:

- Transforming to lowercase,
- Removing whitespaces, special symbols, punctuation and/or numbers,
- Lemmatization.

To transform text into numerical features that ML models can understand, we used Count Vectorizer and TF-IDF text vectorization techniques.

2.3. Classification

For classification, we experimented with different combinations of NLP techniques and ML algorithms, which included Logistic Regression (LR), Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN). To evaluate the effectiveness of different combinations of NLP and ML models, we tested different approaches and compared their performance in terms of accuracy, precision, recall and f1-score.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Experimentation results

The results provided important insights into which methods produce the highest classification accuracy and efficiency. Our experiments with different NLP techniques revealed several important findings. Due to the lexical characteristics of the language, minimal text pre-processing gave better results. In particular, applying lemmatization had a negative impact on the performance. In addition, the TF-IDF vectorization method performed better when combined with k-NN and SVM, while the Count Vectorizer performed better when combined with the LR algorithm.

The best performing combinations for food and non-alcohol beverages were the Count Vectorizer combined with LR and TF-IDF combined with SVM, providing high accuracy of 97.53 percent, precision of 97.54, recall of 97.53 and f1-score of 97.51.

Table 1. Best performing combinations for Food and Non-alcohol Beverages classification into ECOICOP

Classification algorithm	Vectorization algorithm	Accuracy	Precision	Recall	F1-score
k-NN	TF-IDF	94.28	94.35	94.28	94.26
SVM	TF-IDF	97.53	97.54	97.53	97.51
LR	Count Vectorizer	97.53	97.54	97.53	97.51

The best performing combination for alcohol and tobacco was the Count Vectorizer combined with LR, where all metrics reached 99.91 percent.

Table 2. Best performing combinations for Alcohol and Tobacco classification into ECOICOP

Classification algorithm	Vectorization algorithm	Accuracy	Precision	Recall	F1-score
k-NN	TF-IDF	94.68	93.62	94.68	94.04
SVM	TF-IDF	99.58	99.59	99.58	99.58
LR	Count Vectorizer	99.91	99.91	99.91	99.91

Similar results were also achieved for the classification into COICOP2018 categories.

3.2. Practical application

For the practical implementation of the pipeline and to ensure a smooth workflow, we used the Palantir Foundry platform. The models and processing steps were developed in Python, using several key libraries such as scikit-learn for machine learning, pandas for data manipulation, and spaCy for natural language processing. This combination of tools allowed for efficient integration, deployment and automation of the classification models.

4. MAIN FINDINGS

The automated classification pipeline showed high accuracy and efficiency in categorizing products into both ECOICOP and COICOP2018 classes. Our experiments showed that minimal text pre-processing, such as keeping the original word, gave better results due to the specific lexical features of the data. For vectorization techniques, TF-IDF performed better with k-NN and SVM, while the Count Vectorizer performed best with Logistic Regression. These results guided the final model selection, resulting in a reliable solution capable of classifying new products with minimal human intervention.

Our work serves as a practical case study of the application of ML and NLP in official statistics, improving methods of statistical production, while laying the basis for more advanced solutions.

In the future, we aim to expand the automated classification pipeline into additional product categories, such as clothing and household items. To achieve this, we plan to explore advanced methods, including large language models (LLMs) and semantic search techniques, which have shown promising results in improving classification accuracy and reducing human resource dependency.

Automatic analysis of metadata for secondary tabular data protection

Clara Baudry (Insee)

1. INTRODUCTION

The R package `rtauargus` allows users to call Tau-Argus from R. It leverages the renowned SDC software Tau-Argus to automatically handle linked tables, extending its functionality beyond separate table processing [1]. It detects common cells between tables and flags it accordingly. Nevertheless, for the package to detect common cells, a list of linked tables must be provided. The constitution of this list of tables demands a certain level of expertise in the understanding of how tables are linked from a confidentiality viewpoint [2]. Many statisticians in national institutes end up protecting tables only once a year, thus they hardly get to reach the level of expertise needed. In order to deal with this issue, a new tool has been developed. It takes the description of the tables to be published – their metadata – and analyses it providing lists of tables that should be treated conjointly (i.e. clusters). This joint treatment can easily be done with `rtauargus`. The goal of this tool is to help all staff working on tabular data protection, from the data producer to the statistical disclosure control expert.

2. METHODOLOGY

A step-by-step process may be used to deal with the protection of a set of tabular data [3]. The first step is to analyse the set of tabular data that will be published. In fact, the list of published tables is not the same as the list of tables to protect. It has been the role of confidentiality experts to conduct this analysis but it would be more efficient if the producers of the tables could do it by themselves. In order to lessen the burden of the analysis, a tool that analyses the metadata has been developed.

The tool uses the metadata file and creates an output file that describes the tables that should be treated and how they should be treated. In fact, the tables can either be treated independently or together in clusters. When tables are treated together it means that the program takes into account the cells that are common across tables and applies the same flag to those common cells.

2.1. Identify hierarchies

The first step of the tool is to identify the hierarchies across tables, i.e. the dimensions of the tables. The idea is that all the spanning variables that are part of the same hierarchical variable are renamed after the latter. This way the tool understands that it is actually only one spanning variable. For example, `nuts2` and `nuts3` should be viewed as one hierarchical spanning variable `HRC_NUTS` with two levels. This step also treats hierarchies on indicators (e.g. balance sheet) or field by creating a new spanning variable that takes the indicators as modalities.

2.2. Split in clusters

This step gathers tables in groups that should be treated together, clusters. The tables that should be treated together are the tables that have common cells. Usually, those tables share the same indicator and the same field with at least one common spanning variable. Each cluster is independent from the others.

2.3. Detect tables included in other tables

Since the description of tables is not done by confidentiality experts, data producers often end up describing tables separately, even if they are included in each other in terms of confidentiality. For example, if a table that presents the turnover by nace and another one that presents the same turnover by nace x nuts are published, then protecting the table turnover by nace x nuts is sufficient. In fact, when nuts = total then the table ends up being turnover by nace. In this step, the tool analyses the spanning variables to detect tables included in other tables.

The result of this analysis can be seen and understood using the option of inclusion graphs created by the tool.

2.4. Regroup tables that are included in each other

Once the inclusions are detected, the tables that are included in each other are regrouped in one table to treat.

2.5. Create additional tables for the unnested hierarchies

There could be cases of unnested hierarchies between different tables. In that case, an extra table must be created for which the unnested hierarchy is applied.

Finally, the result of all those steps is summarized in a single table with a new column “cluster” that indicates which tables must be treated together.

3. RESULTS AND PRACTICAL APPLICATION

3.1. A theoretical example to understand the tool

The article presents a theoretical example of tabular data needing protection. Each step of the automatic analysis is presented through this example.

3.2. Examples of Eurostat publications

This tool was initially developed to help producers protect tables by themselves. But there are some publications that are too complex for them, even tricky for the experts. The tool comes in handy for the experts when they face large publications composed of many tables linked to each other. In fact, when there are a lot of tables it is easy to forget some links, having an automated approach prevents those mistakes.

The tool has been tested for the analysis of some Eurostat publications:

- Business demography
- Structural business statistics
- ICT usage in enterprises

For example, Figure 1 is an extract of the output file of the analysis of the business demography data for Eurostat. Each row describes a table and the cluster column identifies which table should be treated jointly. The tables should be treated together because they share common margins.

ent_act.N.eff_empl	tot_ent3_taille_jur	ent_act.N	EMPL	HRC_NAF	HRC_NUMBER_EMPL	cat_jur	hrc_naf	hrc_number_empl
ent_act.N.eff_empl	tot_ent3_taille_jur_alt1	ent_act.N	EMPL	HRC_NAF	HRC_NUMBER_EMPL	cat_jur	hrc_naf_alt1	hrc_number_empl
naiss.N.eff_empl_brth	tot_ent6_jur	naiss.N	EMPL_BRTH	HRC_NAF	cat_jur		hrc_naf	
naiss.N.eff_empl_brth	tot_ent6_jur_alt1	naiss.N	EMPL_BRTH	HRC_NAF	cat_jur		hrc_naf_alt1	
naiss.N.eff_empl_brth	tot_ent6_taille	naiss.N	EMPL_BRTH	HRC_NAF	taille		hrc_naf	
naiss.N.eff_empl_brth	tot_ent6_taille_alt1	naiss.N	EMPL_BRTH	HRC_NAF	taille		hrc_naf_alt1	
deces.N.eff_empl_dth	tot_ent12_jur	deces.N	EMPL_DTH	HRC_NAF	cat_jur		hrc_naf	
deces.N.eff_empl_dth	tot_ent12_jur_alt1	deces.N	EMPL_DTH	HRC_NAF	cat_jur		hrc_naf_alt1	
deces.N.eff_empl_dth	tot_ent12_taille	deces.N	EMPL_DTH	HRC_NAF	taille		hrc_naf	
deces.N.eff_empl_dth	tot_ent12_taille_alt1	deces.N	EMPL_DTH	HRC_NAF	taille		hrc_naf_alt1	
deces.N-1.eff_empl_dth	tot_ent9_jur	deces.N-1	EMPL_DTH	HRC_NAF	cat_jur		hrc_naf	
deces.N-1.eff_empl_dth	tot_ent9_jur_alt1	deces.N-1	EMPL_DTH	HRC_NAF	cat_jur		hrc_naf_alt1	

Figure 1. Extract of the result file after processing the metadata

The ten columns in Figure 1 are as followed:

1. Name of the cluster
2. Name of the table
3. Field of the table (population, year for example)
4. Response variable (also called indicator)
5. First spanning variable of the table
6. Second spanning variable
7. Third spanning variable
8. Name of the hierarchy for the first spanning variable
9. Name of the hierarchy for the second spanning variable
10. Name of the hierarchy for the third spanning variable

4. MAIN FINDINGS

The automated analysis developed allows the confidentiality expert to double check their analysis, especially for publications with many tables. But, the end goal of this function is for it to be used by data producers so that they internalize the protection of the data, taking this task off the methodology expert plate.

This automatic analysis was developped as a tool to be used upstream of rtuargus and thus Tau-Argus. However, this upstream work is also relevant for other secondary tabular data protection methods. In fact, the identification of common cells and hierarchies is necessary for any tabular data protection method.

5. REFERENCES

- [1] J. Jamme and N. Rastout (2023, September 21). *Protect several linked tables at once with rtuargus*. rtuargus. https://inseefr.github.io/rtuargus/articles/protect_multi_tables.html
- [2] A. Hundepool et al., Éd., Handbook on *Statistical disclosure control*, 2nd Edition, 2024, chapter 5. <https://sdctools.github.io/HandbookSDC/05-frequency-tables.html>
- [3] J. Jamme and C. Baudry, A step-by-step process to deal with the protection of a set of tabular data, *European Conference on Quality in Official Statistics, Estoril, 2024*. <https://airdrive.eventsair.com/eventsairwesteuprod/production-leading-public/a32757074124462ca4c54d43b9a22e1d>

Bayesian spatial modelling for prevalence of child mortality and morbidity: A bivariate Cauchit link model

Oluwafunmilayo Dawodu, Osafu Egbon, Francisco Louzada &
Ricardo Ehlers (University of Sao Paulo)

1. INTRODUCTION

The onus of child death and illness in an emerging world, particularly in sub-Saharan African countries has persisted as a major issue with an increase of 27 per cent from 1990 - 2021. As the population changes, less than 10% of children survive before their 60-month birthday, roughly 15 times higher than developed nations [7]. This discrepancy solely depends on the rate of child deaths recorded among the countries. In 2021, Nigeria, Africa's most populous nation is prominent with this substantial problem by recording the highest mortality rate of 852,298 compared to Andorra where 2 deaths were recorded [7]. However, modelling child mortality and morbidity has gained interest recently but integrating spatial modelling techniques remains relatively limited despite its prominence in terms of socio-demographics. Hence, this paper seeks to bridge this void by exploring the nexus between socio-demographic and the joint prevalence of child mortality and morbidity in Nigeria, taking into account spatial factors, and thereby contributing to a more comprehensive understanding of this issue.

To achieve this, we considered four different baselines of link functions aside from power and reversal power link functions defined by [1] from a Bayesian perspective to investigate which model best fits the dataset on child mortality and morbidity binary response. We modelled the spatial pattern of the joint effects of child mortality and morbidity across all 36 states plus the Federal Capital Territory, Abuja, Nigeria. We selected the model with the least value of the Widely Applicable Information Criterion (WAIC) and proposed its Bayesian bi-variate model, namely the Bayesian bivariate Cauchit link function with the notion that child mortality and morbidity can be related. As reported in the study of [4] a life-threatening illness can lead to mortality, especially in environments where healthcare is limited or inadequate.

Therefore, we modelled the geographical variation of the disease by suggesting a bivariate intrinsic conditional autoregressive (biCAR) prior through a Gaussian Markov random field for localized smoothing of estimates within the structural additive regression model as well as developing an efficient Rstan algorithm model. Prior to this, a model selection criterion called model challenger was conducted from the Bayesian point of view by comparing the Baseline link function including probit, logit and complementary log-log (Cloglog) and Cauchit model for the response variables (child mortality and morbidity).

2. MATERIAL AND METHOD

2.1. Material

Three phases of the Nigeria Demographic and Health Survey (NDHS) conducted in 2008, 2013, and 2018 covered three consecutive 5-year periods. These surveys encompass Nigeria's six geopolitical zones or regions, which include the 36 states and the Federal Capital Territory, Abuja. The survey was mainly carried out among women of reproductive age (15-49 years) as defined by the World Health Organisation (WHO) having children < 60 months old.

For proper guidelines, the enumerators were monitored to strictly adhere to the standard procedure, allowing them to collect data from various locations. In addition, a two-stage sampling technique was employed to draw a sample from the population and housing census.

All variables including the response and the spatial effect covariates were selected based on past studies carried out by [2, 5]. Previous studies have primarily focused on univariate analyses using link functions such as the logit and probit models. However, none of these studies have addressed the joint effects of mortality and morbidity in children under 5 years using a bivariate Cauchit link function. In our study, we considered four different link functions, as presented in Table 1, and selected the one with the lowest WAIC based on a Bayesian approach. After identifying the optimal link function, the selected variables included sex, place of residence, mother's education, father's education, wealth index, religion, and birth order. These variables were found to be significant predictors of both child mortality and morbidity.

Prior to variable selection, some variables were recoded to ensure compatibility with our model. For example, a child who had a fever, cough, or diarrhoea in the last 2 weeks was coded as 1 for child morbidity and 0 otherwise. Similarly, a child who was not alive was coded as 1 for child mortality, and 0 if alive. In total, 41,168 observations were included in the analysis.

Table 1. Link function for Bayesian optimal model and Variable selection with their model diagnostic

Distribution	pdf	Link Function	Mortality WAIC	Morbidity WAIC	Type
Gaussian	$\phi(\mu) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-\frac{t^2}{2}} dt$	probit	2935.177	3290.779	Symmetric
Logistic	$\Lambda(\mu) = \frac{1}{1 + e^{-\mu}}$	logit	2923.786	3287.471	Symmetric
Cauchy	$CT(\mu) = \frac{1}{\pi} \arctan(\mu) + \frac{1}{2}$	Cauchit	73.740	73.848	Asymmetric
Reversal-Gumbel	$\nu(\mu) = 1 - e^{-e^{\mu}}$	Cloglog	2924.609	3287.829	Asymmetric

$\mu = \alpha + X_i^T \beta$, we set the prior distribution for the four distributions at mean = 0 and variance = 1.

2.2. Method

In this section, the joint prevalence of the Bayesian Bivariate Cauchit link function (BBCLF) will be considered using some covariates selected from the Bayesian optimal Variable selection discussed in subsection 2.1. We consider that $Y_{ijk}^* \sim \phi_c(\mu, \Sigma_v)$ is the latent continuous variable of the binary outcomes Y_i and Y_j which follows a standard Cauchit distribution defined

by the Cumulative distribution function (CDF) of a bivariate Cauchy distribution and Σ_v is a 2×2 symmetric correlation matrix which models the dependence between the latent variables Y_{ijk}^* defined as:

$$\Sigma_v = \begin{pmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{pmatrix} \quad \forall |\rho| = \rho_{12} = \rho_{21} < 1.$$

Each univariate standard Cauchy link function introduces a spatial effect to the linear predictor $H(\eta_{ijk}) = H(\alpha_{ij} + X_{ij}^T \beta + \phi_{ik})$ written as: $H(\eta) = \frac{1}{\pi} \tan^{-1}(\eta_{ijk}) + \frac{1}{2}$ $j = 1, 2; i$ is the number of child living in state k ,

$X_{ij}^T = (X_{i1}, X_{i2}, \dots, X_{ip})$, a $p \times 1$ vector covariates, $\beta = (\beta_1, \dots, \beta_p)^T$ is a vector of coefficient of covariates associated with vector of response variable, α_{ij} is a constant; and $\phi_{(ijk)}$ represents the spatial covariates in the model representing the 36 states plus FCT Abuja hypothesizing that one or more states share a common boundary. Then, we defined a Bayesian Bivariate Cauchy Link Function (BCLF) as an extension of the Univariate Cauchy Link Function (UCLF), using the Cauchy copula to model the dependencies and capture the association between two binary response variables, particularly accounting for heavy tails and strong correlations.

The outcome variables $Y_j = (Y_1, Y_2)$ have all possible responses in probability $[0, 1]$, denoted as: $y \in \{y_{11}, y_{10}, y_{01}, y_{00}\}^T$ such that for $\gamma_{00} = P(Y_1 = 0, Y_2 = 0) = 1 - \gamma_{01} - \gamma_{10} - \gamma_{11}$, we have the latent variables Y_j with structural additive linear predictor defined using the probability of a bivariate Cauchy cumulative distribution as:

$$\gamma_{11} = P(y_1 = 1, y_2 = 1 | X_1, X_2) = \frac{1}{2\pi} \left[\tan^{-1}(\eta_{1ik}) + \tan^{-1}(\eta_{2ik}) + \tan^{-1} \left\{ \frac{(\eta_{1ik})(\eta_{2ik})}{\sqrt{1 + \eta_{1ik}^2 + \eta_{2ik}^2}} \right\} \right] + \frac{1}{4} \dots \dots \dots (1)$$

Equation 1 holds if $\rho = 0$. If otherwise, $\rho \neq 0$, then we adopt the Lemma stated by [3] to introduce a correlation parameter ρ to Equation 2 below.

Lemma 2.1: if $(Y_1, Y_2) \sim N_2(0, 0, 1, 1, \rho)$ and $X \sim N(0, 1)$ are independent, then

$$P\{Y_1 < y_1, Y_2 < y_2, X > 0\} = \frac{1}{4\pi} \left\{ \cos^{-1} \left(-\frac{\rho + y_1 y_2}{\sqrt{1 + y_1^2} \sqrt{1 + y_2^2}} \right) + \tan^{-1}(y_1) + \tan^{-1}(y_2) \right\}$$

Using Lemma 2.1 for $y_1 = 1$ and $y_2 = 1$, the bivariate Cauchy cumulative distribution function and its correlation parameter are written as:

$$\gamma_{11} = P(y_1 = 1, y_2 = 1, \rho | X_1, X_2, \rho) = \frac{1}{2\pi} \left[\tan^{-1}(\eta_{1ik}) + \tan^{-1}(\eta_{2ik}) + \tan^{-1} \left\{ \frac{\eta_{1ik} \eta_{2ik} + \rho \sqrt{(1 + \eta_{1ik}^2)(1 + \eta_{2ik}^2)}}{\sqrt{1 + \eta_{1ik}^2 + \eta_{2ik}^2}} \right\} \right] + \frac{1}{4} \dots \dots \dots (2)$$

Now, we show the binary response of BBCLF using the Lemma adopted from [6] such that we suppose that $Y_j = (Y_1, Y_2) = (\gamma_{11}, \gamma_{01}, \gamma_{10}, \gamma_{00})$ follows independently and identically bivariate Cauchy distribution, then the joint probability is given as: $p(y_j | X) = P(\gamma_{11} = y_{11}, \gamma_{01} = y_{01}, \gamma_{10} = y_{10}, \gamma_{00} = y_{00})$. We also introduce Cauchy copula as a measure of dependence between the two outcome variables. The log-likelihood function is expressed as:

$$\begin{aligned}
l(\theta) = & \sum_{i=1}^n (1 - y_1)(1 - y_2) \cdot \log[C[\Phi_c(-\eta_{1ik}), \Phi_c(-\eta_{2ik}), p]] + \sum_{i=1}^n (y_1)(1 - y_2) \cdot \log[C[1, \Phi_c(-\eta_{1ik}), -p] - C[\Phi_c(-\eta_{1ik}), \Phi_c(-\eta_{2ik}), -p]] \\
& + \sum_{i=1}^n (1 - y_1)(1 - y_2) \cdot \log[C[\Phi_c(-\eta_{1ik}), 1, -p] - C[\Phi_c(-\eta_{1ik}), \Phi_c(-\eta_{2ik}), -p]] + \\
& \sum_{i=1}^n (y_1)(y_2) \cdot \log[1 - C[\Phi_c(-\eta_{1ik}), 1, p] - C[1, \Phi_c(-\eta_{2ik}), p] + C[\Phi_c(-\eta_{1ik}), \Phi_c(-\eta_{2ik}), p]]
\end{aligned}$$

Where Φ_c is the standard Cauchy distribution, $\eta_{ijk} = \alpha_{ij} + X_{ij}^T \beta + \phi_{ik}$, θ is the parameters used in the study.

2.3. Posterior Distribution

A full Bayesian model is developed using a prior and posterior distribution. The choice of the prior distribution is influenced by the dataset used. In this study, we considered the weakly informative prior for β 's such that we adopt $\beta \sim N(\mu_{\beta_0} = 0, \sigma_{\beta_0} = 1)$, $\phi \sim \text{unif}(-1, 1)$ and bivariate conditional auto-regressive (CAR) models assigned to the spatial effects.

The posterior distribution is formulated as:

$$\pi(\theta) \propto \text{log}l(\theta) \times \pi(\beta_0) \times \pi(\beta_{10}) \times \pi(\beta_{20}) \times \exp\left(\frac{-1}{2} \phi_k [(D - W) \otimes Q_k^{-1}] \phi_k\right)$$

Where $1(\cdot)$ is an indicator function, $\pi(\theta)$ is the marginal posterior distribution to be estimated, $l(\theta)$ is the likelihood function, $\pi(\cdot)$ is the prior specification assigned to each of the parameters, and $\exp\left(\frac{-1}{2} \phi_k [(D - W) \otimes Q_k^{-1}] \phi_k\right)$ represents the spatial prior assigned to the 37 states in Nigeria such that D is a 37×37 diagonal matrix, W is a 37×37 adjacency matrix such that if, $W(d, d') = 1$, then the state has a common boundary and if otherwise $W(d, d') = 0$. The spatial element d is a neighbour to state d' , and Q_k^{-1} is a sparse matrix.

3. RESULTS, FINDINGS AND CONCLUSION

The results of the proposed Bayesian bivariate Cauchit link function, analyzed using RStan, are presented in Table 2 and Figure 1. The table presents the results of the fixed-effects model explaining the joint prevalence of child mortality and morbidity in Nigeria. The estimates in Figure 1 illustrate the spatial pattern of the joint prevalence of these diseases. The yellow portion of the legend indicates a lower risk, the pink portion represents a moderate risk, and the blue/purple areas depict a high risk. As the colour shifts from yellow to blue, the risk of joint occurrence increases across all six regions, indicating that higher probability values correspond to a greater joint prevalence of the disease. The spatial pattern reveals probability values ranging from 0.48 to 0.52, indicating the potential for both prevalences to occur across all states in Nigeria, though these probabilities vary by state. As shown in Figure 1, children living in Taraba, Sokoto, Yobe, Kwara, and FCT Abuja in the Northern part of Nigeria, as well as Osun State in the South West and Anambra in the South East, are particularly susceptible to joint prevalence. Other states exhibit lower prevalence, though the differences between them are minimal.

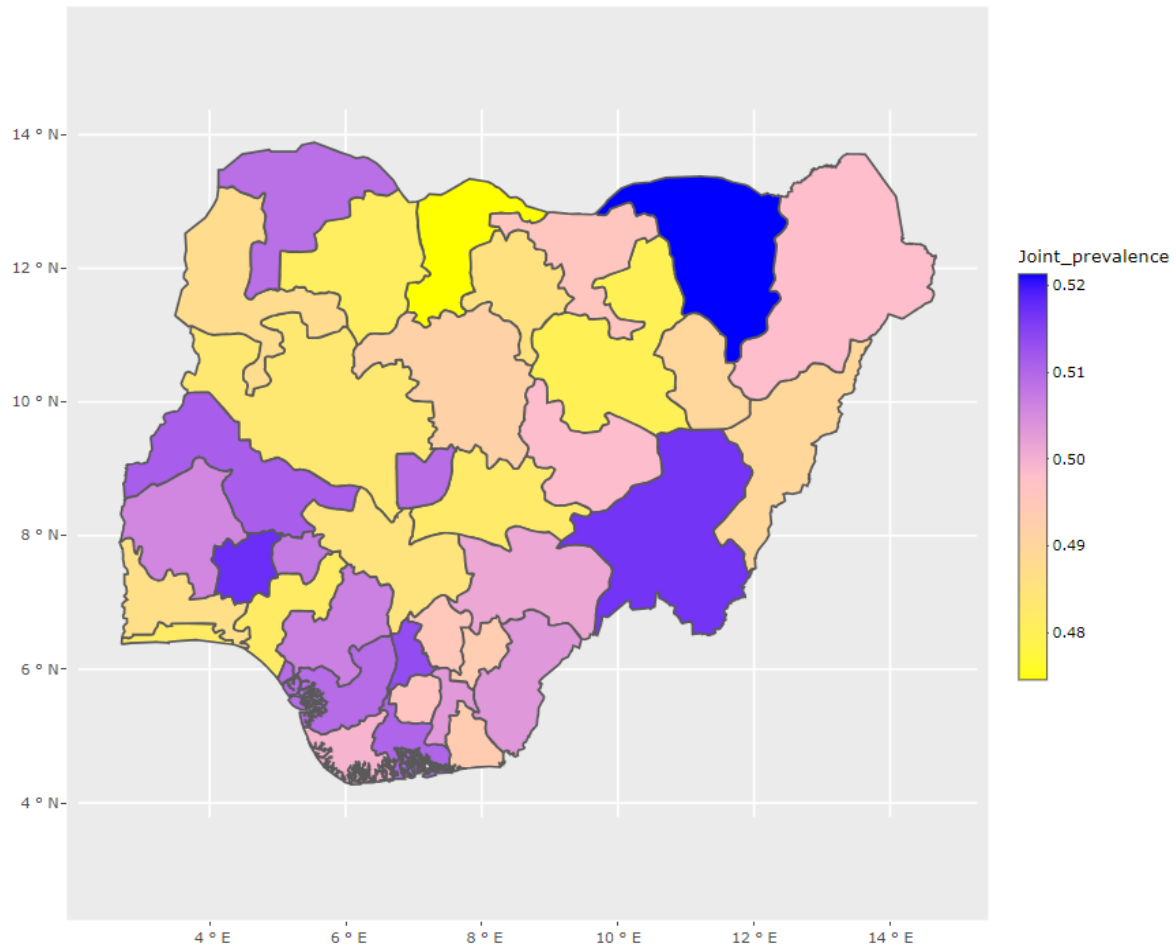


Figure 1. The spatial pattern of the joint prevalence of child mortality and morbidity

We observed higher estimates for certain variables when modelling the probability distribution, suggesting that a higher estimated mean corresponds to an increased risk or joint risk of mortality and morbidity. The estimates suggest that male children have a higher prevalence of mortality and morbidity. Furthermore, children whose mothers and fathers have higher education levels exhibit a lower prevalence of morbidity and mortality compared to those with parents at other education levels. Regarding the wealth index, the joint prevalence results show that children of mothers in the richest category have a lower prevalence of the disease compared to others. However, children from Islam and Christian religions faced a lower risk of joint mortality and it is uncertain the children who experience both mortality and morbidity. Therefore, findings discovered that Yobe, Taraba, Sokoto and some states in the Northern part of Nigeria need an intervention on the joint risk to expedite the amelioration process in achieving the reduction of child mortality by 2030.

Table 2. Joint prevalence of child morbidity and mortality

Variable	P.Meann	CI(2.5%, 97.5%)	Variable	P.Mean	CI(2.5%, 97.5%)
Constant	0.499	-0.134, 1.130			
Gender (Male)	1.000		Residence (Rural)	1.000	
Female	0.497	-0.113, 1.120	Urban	0.4973	0.437, 0.733
Mother's Education (No Educ.)	1.000		Father's Education (No Educ.)	1.000	
Primary	0.529	0.421, 0.814	Primary	0.501	0.489, 0.819
Secondary	0.505	0.492, 0.828	Secondary	0.495	0.464, 0.833
High	0.493	-0.139, 1.117	High	0.502	-0.124, 1.129
Religion (Traditionalist)	1.000		Birth order (1st birth)	1.000	
Islam	0.498	-0.121, 1.127	2nd & 3rd	0.500	-0.122, 1.123
Christian	0.499	-1.121, 1.127	4th above	0.515	-0.101, 1.141
Wealth Index (Poorest)	1.000		Wealth Index (Poorest)	1.000	
Poorer	0.502	0.495, 0.929	Rich/Richer	0.494	-0.154, 1.137
Middle	0.503	-0.102, 1.126	Richest	0.491	-0.154, 1.137

4. REFERENCES

- [1] Bazan, J. L., Torres-Aviles, F., Suzuki, A. K., & Louzada, F. (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, 33(1), 22 - 34. <https://doi.org/10.1002/asmb.2215>
- [2] Dadras, O., Hazratzai, M., & Dadras, F. (n.d.). The association of child marriage with morbidities and mortality among children under 5 years in Afghanistan: findings from a national survey. *BMC Public Health*, 23(1), 23 - 32. doi: 10.1186/s12889-023-14977-5
- [3] Jamalizadeh, A., & Balakrishnan, N. (2008). On a Generalization of Bivariate Cauchy Distribution. *Communications in Statistics - Theory and Methods*, 37(4), 469-474. 10.1080/03610920701469160
- [4] Kruk, M. E., Gage, A. D., Joseph, N. T., Danaei, G., García-Saisó, S., & Salomon, J. A. (2018). Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries. *The Lancet*, 392(10160), 2203-2212.
- [5] Pintu, P. (2020). Child marriage and its association with morbidity and mortality of children under 5 years old: Evidence from India. *Journal of Public Health*, 28, 331-338. <https://doi.org/10.1007/s10389-019-01038-8>
- [6] Purhadi, P., & Fathurahman, M. (2021). A logit model for bivariate binary responses. *Symmetry*, 13(2), 326. <https://doi.org/10.3390/sym13020326>
- [7] UNICEF. (2023, January). *Under-five mortality*. Retrieved August 17, 2024, from <https://data.unicef.org/topic/child-survival/under-five-mortality/>

Better enlighten local public decision-makers on the situation of the real estate market

Matthieu Wyckaert, H  l  ne Soual & Philippe Bertrand (Insee)

1. INTRODUCTION

Applying national official statistics to the local level is not always easy. Often it is the granularity of the input data which will or will not make it possible to produce statistics that are fine enough to provide relevant statistics. In France, the order of notaries has been producing statistics on housing prices for a long time with their *Perval* database. With their help, INSEE produces a reference index for the price of old housing at the national level, the house price index [1]. Faced with the surge in real estate prices, public authorities have taken an interest in local real estate markets, potential factors of local imbalances. This resulted in the provision of new open data in 2019, with the aim of ensuring the transparency of local real estate markets [2].

Drawing on its experience in handling this open data and its local network, INSEE has invested in the subject. The expertise center of the regional directorate of INSEE in Lyon (“le Pôle de service de l’action régionale Synthèse locales” in French) worked in partnership with public stakeholders and INSEE research services to produce a shared tool for knowledge of local real estate markets. This work is an asset to shed light on the question of household location decisions in the territories and provide a better understanding of peri-urbanization, and the trade-offs between transport and housing costs, in connection with household income. This paper deals with this project which landed in 2023.

2. METHODOLOGY

This work matches several administrative data sources and several surveys to explore old and new real estate markets.

2.1. Tidying, cleaning, matching and geolocation of data

The cornerstone of the process implemented is the « Real Estate Transaction » datafile, « *Compte de Valeurs Foncières* » in French (DVF) [3]. It contains disaggregated information on every French real estate transaction (amount paid, housing attributes, legal regime of the sale, buyers). It is produced by the Central State tax services (DGFiP). This dataset contains information for 3.4 million transactions, from 2010 to 2023. With the help of Cerema, a research center on the domain, additional information on the properties and the land is added from the “MAJIC” cadastral data produced by the DGFiP. Data cleaning operations are then carried out to identify and process both transactions not included in the field (garage sales for example) and outliers (such as sales of “exceptional” goods). The transactions retained following this processing are then precisely geolocated and then classified in a geographical area to respect the confidentiality of the data disseminated.

To produce indicators on new housing, two surveys are used: the survey on the price of land and buildings and the survey on the marketing of new housing carried out by the Ministerial Statistical Office (MSO) in charge of housing policy, according to the same process.

2.2. Price modeling taking into account spatial autocorrelation

Here we seek to estimate the price per square meter of housing by taking into account their location and the characteristics of the property. We will explain the choice of the model among several candidates according to several criteria. Among the possible models, and after the Hausmann specification test, the Spatial Error Model was chosen [4]. It can be written in the following standard form, by using the following two-stage formulation:

$$Y = X.\beta + u$$

$$u = \lambda.Wu + \varepsilon$$

Where Y the vector of variables to explain, ε the vector of error with $\varepsilon_i \rightarrow N(0, \sigma^2)$

3. RESULTS AND PRACTICAL APPLICATION

Many results can be produced from this data. We will detail two of them: the price per square meter across the entire French territory (data from Moselle and Alsace are being retrieved) and the link between public transport average and price.

In mainland France excluding Alsace-Moselle, over the period 2020-2022, the median price per m² of an old house amounts to €2,040, that of an old apartment reaches €3,170.

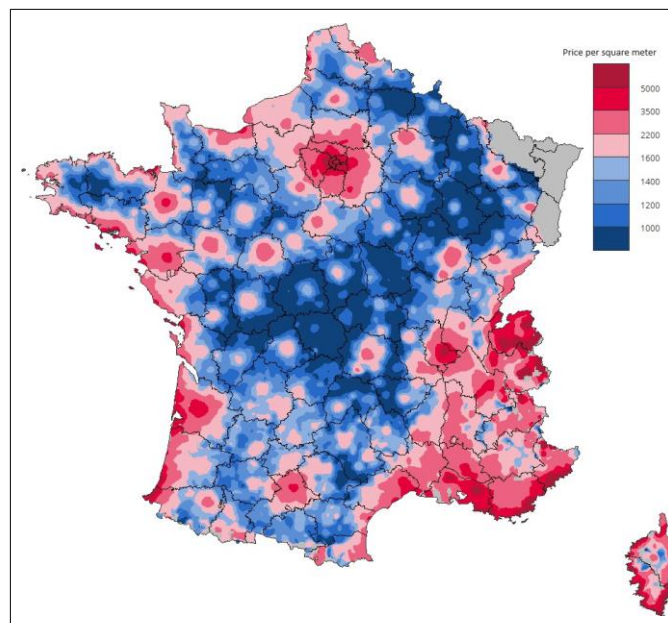


Figure 1. Price per m² of old houses sold over the period 2020-2022

Figure 1 shows the disparity in housing prices by location. Concerning houses, the price per m² is particularly low in the municipalities of “the diagonal of low densities”, also called the “diagonal of emptiness”, going from the south of the Massif Central to the Ardennes to the East of “Hauts de France”.

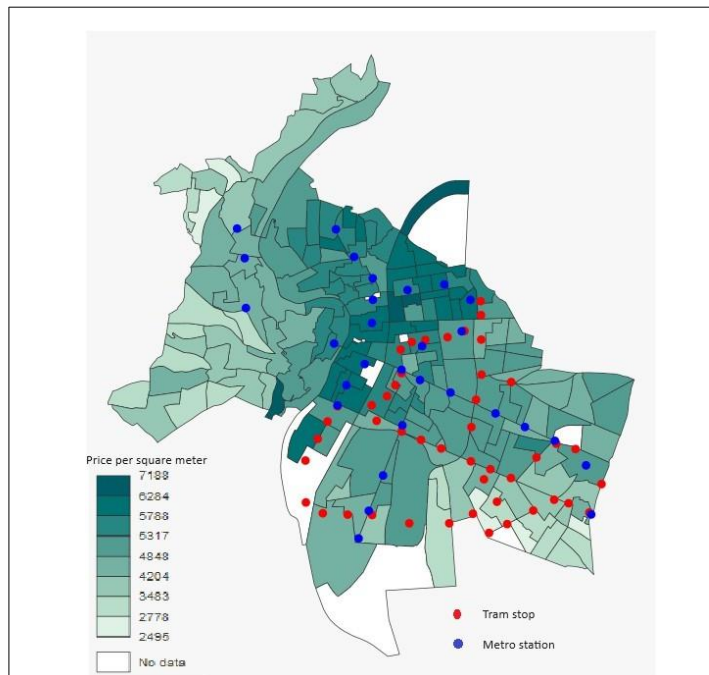


Figure 2. Price per square meter of apartments in the city of Lyon and transport infrastructure

Figure 2 illustrates with the municipality of Lyon the positive relationship between access to metro and tram stations.

4. MAIN FINDINGS

This work illustrates the importance of data quality as input to an analysis. Data cleaning makes it possible to consolidate the statistics produced and to better understand the mechanisms that contributed to the production of the initial data. In this vein, and in order to further improve the possibilities offered with real estate transactions, it is planned to enrich the analysis with the use of the “Fidelimmo” datafiles currently being produced at INSEE.

In view of this success, it is planned to complete the analysis with work on rents with a new tax source (tax application “GMBI”) at the same level of detail as the work on real estate markets.

5. REFERENCES

- [1] INSEE, Les indices Notaires-Insee des prix des logements anciens, Insee méthodes n°312 (2019).
- [2] JORF N°0302 du 30 décembre 2018, Décret n° 2018-1350 du 28 décembre 2018 relatif à la publication sous forme électronique des informations portant sur les valeurs foncières déclarées à l'occasion des mutations immobilières (2018), <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000037884472>.
- [3] Casanova Enault L., Boulay G. & Coulon M., The Big day for geographers? Opportunities and precautions of open DVF datafiles to work on property market, Cybergeog: European Journal of Geography (2019).
- [4] Loonis V & allii, Handbook of Spatial Analysis, Theory and Application with R, Insee Méthodes n°131, chapter 7 (2018).

Binding Surveys and Deep Learning Statistics on Arrival Ports for Maritime Mobility via AIS Data

Francesco Pugliese, Mauro Bruno, Giulio Massacci & Francesco Ortame Italian
National Institute of Statistics (Istat)

1. INTRODUCTION

Over the past decade, maritime data monitoring has expanded rapidly, with systems like AIS (Automatic Identification System)¹ playing a key role in vessel position prediction and trajectory forecasting. These tasks are essential for operations such as search and rescue, traffic control, and pollution tracking. Although there is a vast amount of AIS data, forecasting vessel routes remains difficult due to the complexity of movement patterns.

The process of data imputation represents a significant ongoing challenge in the field of machine learning (ML) [1]. AIS signals contain numerous lacks in both covariates and target labels (arrival ports), and Machine Learning methods. Particularly, deep learning represents a potential solution to these datasets problems [2]. These methods are capable of handling large volumes of high-dimensional data, often surpassing traditional methods [3], especially when dealing with complex patterns.

In this study, our goal is to predict arrival ports of ships based on incomplete² time series of AIS signals representing their routes. This prediction allows us to integrate incomplete routes into Official Statistics on maritime traffic and mobility [4]. To achieve this, we trained several neural networks, with the most promising results obtained by harnessing Long Short-Term Memory (LSTM) networks and one-dimensional Convolutional Networks (CONV-1D). The problematic routes account for approximately 11% of all ship routes, and correctly labelling them would enhance the reliability of maritime mobility statistics related to arrivals, and potentially, departures. Our research introduces a novel approach to Port Classification, showing its potential applications in Official Statistics.

2. METHODOLOGY

To train our models, we extracted 65,300 ship routes from the first quarter of 2022. Within these ones, 58,000 have a valid label, while 7,300 are unlabelled. The labelled routes form our training set, and the unlabelled routes constitute our blind test set which will be involved into the final imputation step. Each route is represented as a time series with four signals:

- a) *Latitude*: coordinate indicating the North-South position of the vessel.
- b) *Longitude*: coordinate indicating the East-West position of the vessel.
- c) *SOG* (Speed Over Ground): measure of the vessel's speed relative to the ground
- d) *COG* (Course Over Ground): directional angle relative to the true North.

¹ AIS Data, transmitted via VHF (Very High Frequency) signals, provides details including vessel location, speed, cruise direction, and identity.

² A route is incomplete when it lacks signals near a port that would indicate an arrival. These routes cannot be classified through deterministic algorithms that rely on strict conditions regarding position, speed, and direction.

These signals are the input features for our models, while the arrival port serves as the target variable. We consider a total of 93 arrival ports.

We defined the experimental design as follows:

Data Preprocessing. 1) We split the labelled data into training, validation, and test sets. 2) To simulate the blind test set, we artificially introduced missing signals at the end of the test routes by randomly removing between 10% and 30% of the signals, with a minimum of 1 signal removed per route³. 3) We applied a two-step position filter to retain only routes within the Mediterranean Sea: first, we dropped routes whose last signal falls outside the Mediterranean, then we removed any signals outside the Mediterranean from the remaining routes. 4) We calculated the Haversine distance between consecutive signals and removed any with a distance of zero (i.e., duplicate positions). 5) Finally, a length filter was applied to keep only routes containing between 4 and 30 signals⁴. At the end of this procedure, we ended up with approximately 4,800 labelled test routes, and 3,300 unlabelled blind test routes, accounting for 5% of all ship routes. We only applied steps 3 to 5 to the training and validation splits, resulting in 49,000 routes.

Models and training. We trained two different models for the task: 1) An LSTM network, a type of recurrent neural network (RNN) designed to handle and learn temporal dependencies in sequential data and 2) A CONV-1D network, a neural network designed to process sequential data by applying convolutions along the time dimension of the sequence. The LSTM has a hidden size of 64, two layers, and a dropout rate of 0.5. It was trained for 500 epochs with a learning rate of $1e-4$ and a weight decay of $1e-5$. The CONV-1D models consists of two 1D convolutional layers, each followed by max-pooling. The first layer has 64 filters with a kernel size of 3, and the second layer has 128 filters with the same kernel size. The dropout rate is 0.5.

Prediction. After training, we applied the models to predict arrival ports for the unlabelled routes in the blind test set. We evaluated the quality of these predictions using two approaches: a) *Empirical assessment*, we filtered the routes based on their predicted ports and manually verify the plausibility of the classifications, and b) *Correlation analysis*, we compared the correlation between the number of arrivals associated to the labelled routes alone and Eurostat data, with the correlation between Eurostat data on Maritime Transport and the combined arrivals from both labelled routes and machine learning predictions. A higher correlation in the latter case would indicate more accurate estimates and improved predictive performance.

3. RESULTS AND PRACTICAL APPLICATION

Here, we present the performance achieved by the models and we show examples of their predictions on previously unlabelled routes.

In Table 1 we show accuracy measures for both models, highlighting similar performances. Given the high dimensionality of the classification space, we consider these results as promising, since they represent an improvement over previous experimentation prior to extensive data cleaning and pre-processing, where predictions on routes with missing data were quasi-random.

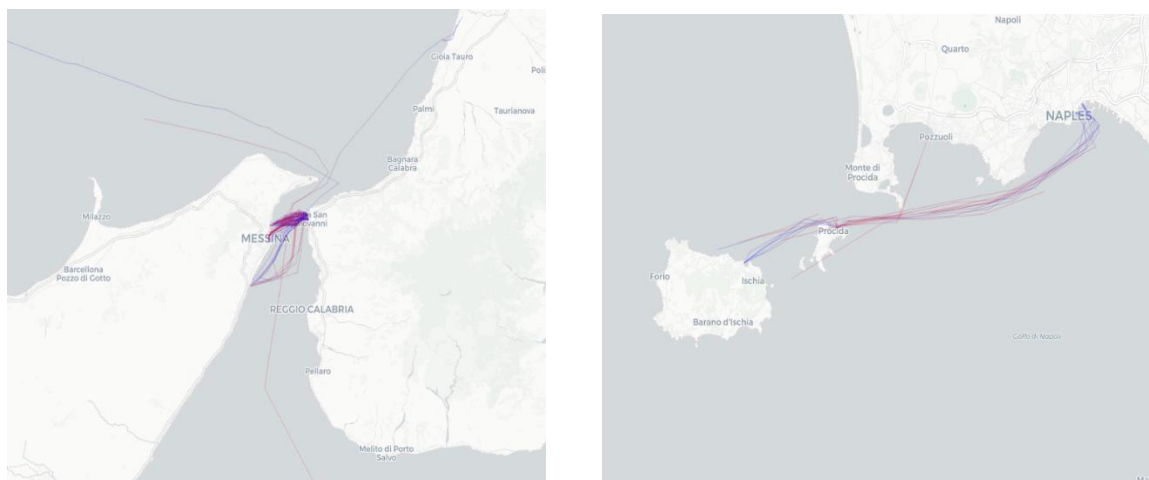
³ This is in line with what we observe in the blind test set, where most routes seem to only lack the very last signals.

⁴ If a route has more than 30 signals, only the last 30 signals are retained.

Table 1. Model Performance on Test Set

Model	F1-score (macro)	Accuracy
LSTM	0.5581	0.8126
CONV-1D	0.5567	0.8269

Below, we illustrate our approach to empirically validating the results. To evaluate the performance of the models on unlabelled routes (the blind test set), we filtered them by selecting specific arrival ports of interest and extracted only those routes whose predicted labels match the chosen ports. Figure 1 provides an example of this process, focusing on the ports of Messina and Procida, both of which are characterized by numerous routes with missing signals.

**Figure 1.** LSTM predictions for the ports of Messina (left) and Procida (right).

The first signals of a sequence are shown in blue, while the last signals are shown in red. Aside from a few mistakes, the LSTM network seems effective in assigning destination ports given incomplete trajectories.

We then validated our predictions by aggregating the ground truth (labelled routes) with the deep learning predictions on the blind test set. Next, we computed the total number of arrivals for the first quarter of 2022 and correlated these figures with Eurostat Maritime statistics. If the models accurately classifies the unlabelled routes, we should expect to see an increased correlation when compared to the scenario where we only correlate the ground truth with Eurostat data. Figure 2 depicts the positive impact of predicting unlabelled routes with the CONV-1D, increasing the correlation from exclusively labelled statistics (left side) to labelled statistics plus Deep Learning predictions (right side) against Eurostat data.

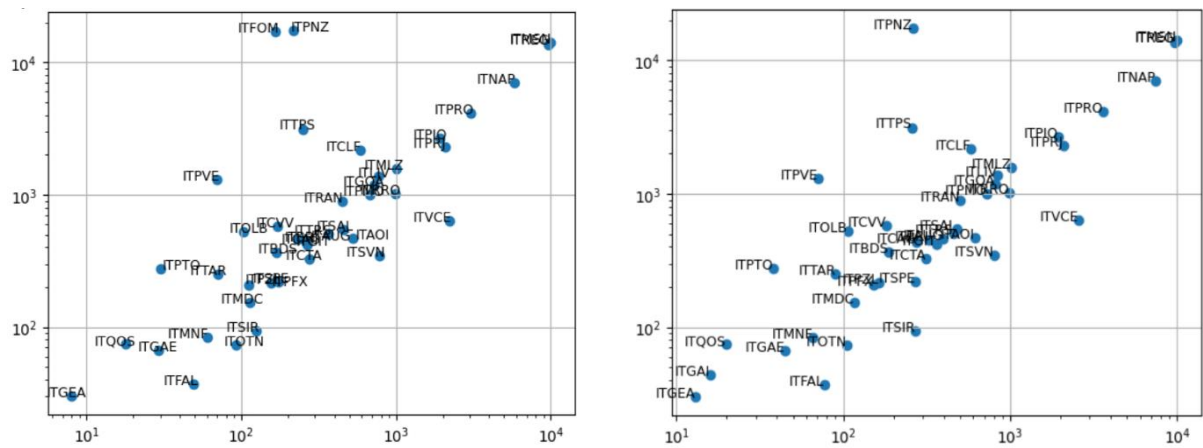


Figure 2. Correlation between the Labelled Statistics (x-axis) versus Eurostat data (y-axis) (left) and Labelled plus Deep Learning Prediction on Unlabelled routes versus Eurostat data (right).

The Pearson correlation coefficient between Eurostat data and AIS is 0.72 when including Deep Learning predictions (right) compared to 0.59, when only labelled routes are included (left).

4. CONCLUSIONS

These promising results, corroborated by empirical evidence, highlight a path to bind Official Statistics surveys with Big Data gathered from AIS sources. This encourages research towards the direction of integrating costly surveys conducted by human surveyors with inexpensive, accurate and fast Artificial Intelligence techniques.

Future developments will focus on exploring different training techniques including data augmentation (namely applying the same filtering techniques described in section 2 to the training set), attention mechanisms and the implementation of Transformer-based models.

5. REFERENCES

- [1] Richman, M. B., Trafalis, T. B., & Adrianto, I. (2009). Missing data imputation through machine learning algorithms. In *Artificial intelligence methods in the environmental sciences* (pp. 153-169). Dordrecht: Springer Netherlands.
- [2] Ipsen, N. B., Mattei, P. A., & Frellsen, J. (2022). How to deal with missing data in supervised deep learning?. In *10th International Conference on Learning Representations*.
- [3] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, 42, 146-157.
- [4] Tu, Enmei, et al. "Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology." *IEEE Transactions on Intelligent Transportation Systems* 19.5 (2017): 1559-1582.

Building a Comparable Measure of Consumption: Concepts and Measurement Challenges Faced by Emerging and Advanced Economies

Teresa Munzi (Luxembourg Income Study, LIS)

1. INTRODUCTION

In 2022, the Luxembourg Income Study (LIS) started to explore the feasibility of creating a Luxembourg Consumption Study (LCS) database. The aim of such exploration was to assess the feasibility of extending comparative distributional analysis by using a consumption-based well-being measure throughout high to middle and low-income countries. Such a measure would extend the potential joint distributional study of income, consumption and wealth immensely. This paper aims to take stock of the different conceptual elements of consumption as defined and applied in emerging (Deaton and Zaidi 2002, Mancini and Vecchi 2023) and advanced countries (OECD 2024; Zwijsenburg et al. 2021), and data collection efforts based on household surveys. We contribute to the discussion of how to guide statistical authorities in building a consumption-based economic well-being measure at the household level. The purpose of this paper is twofold: (1) to further clarify the conceptual framework for defining a comparable consumption-based well-being concept; and (2) to provide an empirical, descriptive, distributional analysis by consumption components and demographic groups across low, middle and high-income countries.

2. METHODOLOGY

This comparative work is based on nine country case studies: Mali, Laos, Palestine, Peru, Georgia, Italy, France, the United Kingdom, and the United States (U.S.A.). On the basis of these data, we provide the conceptual framework (aggregation plan and variable detail) for a potential Luxembourg Consumption Study database.

The conceptual framework to construct an LCS consumption aggregate (CA) can be summarized as follows:

$$CA = x^F + x^{NFND} + \hat{x}^{DUR} + \hat{x}^{HOUS}$$

where:

x^F denotes the food consumption aggregate:

$$x^{FOOD} = x^P + x^O + x^{IK} + x^{FAFH}$$

x^P denotes purchased food.

x^O denotes own-produced food.

x^{IK} denotes food received in kind.

x^{FAFH} denotes food prepared away from home.

x^{NFND} denotes non-food non-durable expenditures, where ‘lumpy’ expenditures are excluded:

$$x^{NFND} = \sum_{j \in J} x_j - x^{LUMPY}$$

$\hat{x}^{DUR}$ is the estimated consumption flow from durable goods:

$$\hat{x}^{DUR} = \sum_{k \in K} x_k^{CF}$$

$\hat{x}^{HOUS}$ denotes housing expenditures:

$$\hat{x}^{HOUS} = \begin{cases} x_{\text{rent}} & \text{if renter} \\ x_{\text{imputed rent}} & \text{if non-renter} \end{cases}$$

The conceptual difficulties in creating a fully comparable consumption concept, which stem largely from limitations in applying fully comparable definitions in cross-national data harmonisation, due to insufficient information in the survey data are further reviewed. We focus on three major groups: 1) the valuation of non-monetary consumption, 2) maintenance and repair, and 3) the treatment of insurance.

Finally, an empirical section first presents the core differences in the analyzed surveys and then provides a distributional analysis. We first present Gini coefficients of inequality for three different measures of well-being, using both square root and per capita equivalence scales. By doing this, we aim to shed light on how the choice of well-being measure and equivalence scales used to adjust for household size shape country rankings of inequality. Second, we look into the joint analysis of two out of the three measures of well-being: the LCS measure of household final consumption expenditure and disposable household income. Finally, we apply a decomposition analysis of the Gini coefficient of the of household final consumption expenditure in order to unpack the contributions of main groups of goods and services to overall consumption inequality.

3. RESULTS AND PRACTICAL APPLICATION

This study highlights the potential of using harmonized consumption expenditure data for cross-country inequality analysis. Based on the first findings from the comparisons of the data across eight countries, a reduced form of the full consumption aggregate presented above was used, which excludes the estimated consumption flow from durable goods, in order to present some preliminary comparisons of consumption-based economic well-being across low-, middle- and high-income countries:

$$CA1 = x^F + x^{NFND} + \hat{x}^{HOUS}$$

Table 1. Gini Index with different consumption/income aggregates

	Mali	Laos	Pales tine	Peru	Geor gia	Italy	Franc e	UK	US
CA1	31.7	37.8	32	33.1	27.1	27.9	25.7	28.4	28.8
CA1 + durables	31.6	45.3	35.5	34.3	27.7	29	27.7	31.3	32.6
Disposable income	36.5	-	42.6	43.9	34.6	-	29.2	28.2	40.4

The paper presents Gini coefficients for three measures of well-being (household consumption excluding durables, household consumption including the purchase value of durables, and disposable household income) across different countries using square root and per capita equivalence scales (the latter is not shown here). The coefficients for consumption (both excluding and including durables) are consistently smaller than those for disposable income, possibly due the role of consumption smoothing through mechanisms like access to credit and savings. The effect of including purchase value of durables (to bring the concept in line with the aggregate consumption concept used in National Accounts) on inequality is varied, but in most of the cases rather modest. On the other hand, the same distributional analysis carried out based on the LIS concept of Disposable Income (DHI), provides rather different results.

The joint distributional analysis of consumption and income shows how correlation of the two measures at the bottom of the income and consumption distribution, may indicate lack of access to savings and credit among the poorest individuals, while concentration of individuals at the highest income and consumption quintile demonstrates that high-income households have a disproportionate impact on overall consumption. Richest individuals not only have higher consumption levels, but also a more varied consumption mix: they spend significantly more on luxury goods and services such as recreation, sport and culture and restaurants and accommodation.

4. MAIN FINDINGS

The main conclusions of the study, which highlights the potential of using harmonized consumption expenditure data for cross-country inequality analysis, can be summarised as follows:

One of the key lessons is that the purpose of the survey is an integral part in defining a complete measure of well-being. While in consumption surveys, not all expenses are fully measured, in expenditure surveys, not all forms of consumption may be collected or valued. This became a clear obstacle in the harmonization work, which poses the need for a clear procedure to select surveys for inclusion in the LCS database. In particular, the information available for estimating the consumption flow of durables varies considerably across

countries, and is typically very limited. Additional efforts need to be done to construct consumption flows with the collected information.

The conceptual discussions in this paper, as well as the exploration of data availability have proven that the border between inclusion and exclusion from consumption is sometimes challenging to draw. This is particularly true for semi-durables and major durables, minor versus major shelter repairs, and accounting for insurance. The question of whether health and education expenditures should be included in an economic well-being measure is extremely delicate, and some very careful considerations need to take place.

Finally, the decomposition of the Gini coefficient highlights how the structure of consumption and its impact on inequality shifts with economic development, with basic needs driving inequality in poorer countries and more diverse consumption patterns driving it in wealthier nations. To the best of our knowledge, this is the first joint analysis of consumption patterns across low, middle and high-income countries, and these first result call for a much more detailed analysis with many more countries.

5. REFERENCES

- [1] A. Deaton and S. Zaidi, Guidelines for Constructing Consumption Aggregates for Welfare Analysis, LSMS Working Paper no. 135 (2002), World Bank.
- [2] G. Mancini and G. Vecchi, On the Construction of a Consumption Aggregate for Inequality and Poverty Analysis, 2023. World Bank.
- [3] OECD, OECD Handbook on the Compilation of Household Distributional Results on Income, Consumption and Saving in Line with National Accounts Totals, OECD Publishing (2024).
- [4] J. Zwijnenburg, S. Bournot, D. Grahn, and E. Guidetti, "Expert Group on Disparities in a National Accounts Framework - Results from the third Exercise", N. 2021/1, OECD Publishing (2021).

Choice of Sampling Efforts in a Schnabel Census

Su Na Chin (University of Southampton)

1. INTRODUCTION

Population size estimation plays a critical role in various scientific and practical fields such as ecology, conservation, and official statistics [1,2]. The Schnabel census [3] is a widely used multi-sample capture-recapture method for estimating population size. In this method, individuals from a population are captured, marked, and then released. Over multiple occasions, the process is repeated, and the number of recaptures is tracked to infer the total population size. Despite the method's importance, the topic of sampling effort has not been sufficiently explored. Sampling effort is critical for ensuring the feasibility of population studies, especially when resources such as labour, funding, and time are limited.

This paper investigates the necessary sampling effort in Schnabel census studies, accounting for varying capture success rates and population heterogeneity. By understanding how the number of capture occasions affects population estimates, researchers can design more effective and efficient studies.

2. THE IDEA OF SAMPLING EFFORT

In a Schnabel census, the observed count of individuals captured, n follows a binomial distribution, $n \sim \text{Bin}(N, 1 - p_0)$, where p_0 represents the probability of missing observations. The variance of n is given by (1).

$$\text{Var}(n) = Np_0(1 - p_0) \quad (1)$$

Using Horvitz-Thompson [4] estimator, $\hat{N} = n/(1 - p_0)$ leads to

$$\text{Var}(\hat{N}) = \frac{Np_0}{(1-p_0)} \quad (2)$$

Relationship in (2) highlights that the probability of missing observations directly influences the accuracy of the population estimate. Assuming a confidence level of $(1-\alpha)\%$, the margin of error for the estimate accommodates a variation within $\kappa \times 100\%$ of the true N , expressed as (3)

$$z_{\alpha/2} \sqrt{Np_0/(1 - p_0)} = \kappa N. \quad (3)$$

Figure 1 depicts the relationship between p_0 , κ , and N , when $1 - \alpha = 0.95$. The graph shows that maintaining p_0 below 0.5 keeps the uncertainty of the population size estimate within 20% of the true N for $N \geq 100$. This makes p_0 an effective tool for controlling uncertainty.

3. SAMPLING EFFORT FOR DIFFERENT CAPTURE SUCCESS RATES

3.1. Homogeneous Population

Given a homogeneous population, the individual capture counts X over T capture occasions follow a binomial distribution, with p.d.f

$$p_x = \binom{T}{x} \theta^x (1 - \theta)^{T-x}$$

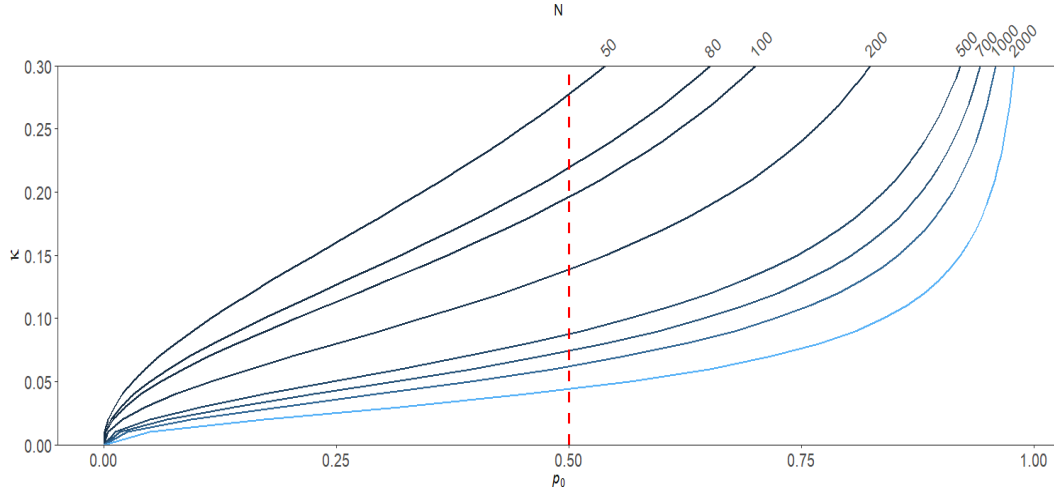


Figure 1. The relationship between p_0 , κ , and N , when $1 - \alpha = 0.95$. The red breaking line marks the position when $p_0 = 0.5$.

Replacing $p_0 = (1 - \theta)^{TT}$ in the variance formula in (2) yields

$$\text{Var}(\hat{N}) = N \frac{(1-\theta)^T}{1-(1-\theta)^T} \quad (4)$$

Reducing N to lower variance in (4) is not feasible due to N being unknown. However, increasing T can make $(1 - \theta)^T$ approach zero, thus minimizing prediction variance. Solving T from $pp_0 = (1 - \theta)^T$ results in

$$T = \frac{\ln(p_0)}{\ln(1 - \theta)}.$$

Hence, number of sampling occasions T required depends on the desired capture success rate and the detectability of the population. Researchers can adjust T according to the desired capture success rate, $1 - p_0$, to effectively control uncertainty. Contour plot in Figure 2 illustrates the relationship between T , θ and

pp_0 . Lower capture probabilities require more sampling occasions to achieve the same level of precision. This highlights the exponential relationship between sampling effort and capture success rate [5].

3.2. Allowing Heterogeneity: Mixture Models

In real-world scenarios, the assumption that all individuals have equal capture probabilities may not hold. To address population heterogeneity, this study incorporates the beta-binomial and binomial mixture models, which allow for varying capture probabilities among individuals.

Beta-binomial model: When capture counts X follow a Beta-Binomial(T, α, β) distribution, the optimal sampling effort T can be determined through the equation

$$p_0 = \frac{B(\alpha, T+\beta)}{B(\alpha, \beta)}$$

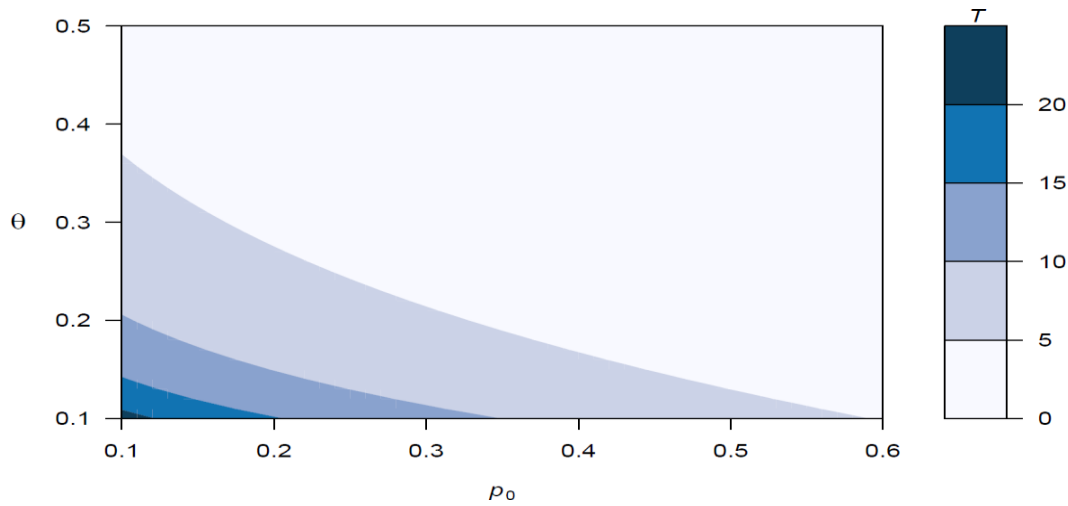


Figure 2. Contour plot on T in dependence of p_0 and θ .

Binomial mixture model: For capture counts following a binomial mixture distribution, the optimal sampling effort T can be found using the equation

$$p_0 = \sum_{j=1}^k w_j (1 - \theta_j)^T$$

where non-negative weights w_j correspond to θ_j , and $\sum_{j=1}^k w_j = 1$.

4. REAL DATA EXAMPLE

A classic study by [6] on cottontail rabbits provides a real-world example of the Schnabel census. In this study, 135 rabbits were captured over 18 consecutive nights, with 76 individuals captured at least once. The recorded capture frequencies were: $f_1=43$, $f_2=16$, $f_3=8$, $f_4=6$, $f_5=0$, $f_6=2$, and $f_7=1$. The zero- truncated binomial, beta-binomial, and binomial mixture models [7] were fitted to the data, with the binomial mixture model providing the best fit. The Akaike

Information Criterion (AIC) and Bayesian Information Criterion (BIC) confirmed that the two-binomial mixture model offered the most accurate estimate of the rabbit population.

Using the Newton-Raphson method, the study estimates the optimal number of capture occasions T for different desired capture success rates as shown in Table

1. For example, to achieve a 50% success rate, 17 capture occasions are needed. As the success rate increases to 90%, the required number of occasions rises significantly to 70. Bootstrap methods were employed to assess the variability in required sampling efforts.

5. CONCLUSION

This study provides valuable insights into the optimal sampling efforts required for accurate population estimation in Schnabel census studies. The exponential relationship between capture success rates and the number of sampling occasions highlights the trade-off between precision and resource use. The beta-binomial and binomial mixture models offer flexible alternatives for addressing population heterogeneity, improving the accuracy of population estimates in real-world scenarios.

By applying these methods, researchers can design population monitoring programs that balance accuracy, feasibility, and resource constraints, making them more sustainable and effective for long-term ecological studies, conservation efforts, and official statistics.

Table 1. Required capture occasions for various levels of desired capture success rate.

Desired capture success rate, $1-p_0$	Optimal sampling effort, T	25 th - 75 th percentile
0.4	12	8-17
0.5	17	11-24
0.6	24	15-34
0.7	34	21-47
0.8	47	29-69
0.9	70	42-98

6. REFERENCES

- [1] D. Böhning, P.G.M. Heijden, and J. Bunge, *Capture-recapture Methods for the Social and Medical Sciences*, Chapman and Hall/CRC, Boca Raton, FL (2018).
- [2] R.S. McCrea and B.J.T. Morgan, *Analysis of Capture-recapture Data*, CRC Press, Boca Raton, FL (2015).
- [3] Z.E. Schnabel, "The Estimation of the Total Fish Population of a Lake," *The American Mathematical Monthly*, 45 (1938), 348–352.
- [4] D.G. Horvitz and D.J. Thompson, "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47(260) (1952) 663–685. <https://doi.org/10.2307/2280784>
- [5] Z. Kordjazi, S. Frusher, C. Buxton, C. Gardner, and T. Bird, "The influence of mark-recapture sampling effort on estimates of rock lobster survival," *PLoS ONE*, 11 (2016), e0151683.

- [6] W.R. Edwards and L. Eberhardt, "Estimating Cottontail Abundance from Livetrapping Data," *The Journal of Wildlife Management*, 31 (1967), 87–96.
- [7] D. Böhning, E. Dietz, R. Kuhnert, and D. Schön, "Mixture models for capture- recapture count data," *Statistical Methods and Applications*, 14 (2005), 29–43.

Classification and Prediction of Rating Data with Flexible Uncertainty Trees

Rosaria Simone (Department of Political Science – University of Naples Federico II, Italy)

1. INTRODUCTION

In the setting of ordinal data analysis, the contribution aims at highlighting the extent by which uncertainty specification improves both explicative and predictive performance of a given preference model assumed for the underlying latent trait, following the rationale of mixture models with uncertainty. The focus will be on model-based trees implemented to perform classification of responses and to derive profiles in terms of feeling and possibly different uncertainty components. In this setting, residuals' diagnostics for ordinal data models can be exploited to implement flexible uncertainty trees with better explicative and predictive performance than classical model-based trees relying on a single maintained model, yielding a more effective learning process.

2. METHODOLOGY

This paragraph is devoted to a brief methodological overview of mixture models with uncertainty, residuals' diagnostics for ordinal data models and the resulting proposal of flexible uncertainty tree for ratings.

2.1. Mixture models with uncertainty

The original idea of specifying a discrete model to explain feeling and uncertainty of ordinal evaluations goes back to [1]. Over the last 20 years, several methodological extensions and applications have enriched the literature on categorical data in this framework: see [2] for an updated overview, with discussions and rejoinder. For a response R collected over m ordered categories, the class of mixture models with uncertainty considers a discrete mixture of a shifted Binomial distribution with feeling parameter ξ in $[0,1]$ and a discrete distribution for uncertainty u_r :

$$Pr(R = r) = \pi b_r(\xi) + (1 - \pi) u_r, \quad r = 1, \dots, m.$$

A relevant feature of the model is that the distribution u_r can be specified in different ways, in particular in such a way to account for inflated frequencies at some shelter category (by choosing a degenerate distribution) and for heterogeneity of the distribution (in which case, u_r is the discrete uniform model). Given the possible specifications of uncertainty, it is of interest to focus on the mixing weight π of the feeling component, indicating the relative importance of the latent sentiment with respect to the discrete noise modeled with the uncertainty component. Parameters can be promptly estimated with the R package 'CUB' available on CRAN.

For the baseline model specification, covariates can be included by letting feeling and uncertainty parameters be subject-specific on the basis of a logit link. Then, model-based trees can be designed with a recursive search of binary splits that either maximize the deviance

between father and children levels or that maximize the dissimilarity between children nodes. In particular, the (normalized) dissimilarity index:

$$\text{Diss}(\mathbf{p}, \mathbf{f}) = \frac{1}{2} \sum_{r=1}^m |f_r - p_r| \text{ in } [0,1]$$

allows to assess the extent by which a discrete model $\mathbf{p}$ matches category-wise an observed (relative) frequency distribution $\mathbf{f}$, thus can be used as a fitting measure. We resort also to the Rank Probability Score (RPS) as a scoring rule specifically designed for ordinal distributions to assess predictive performance of selected models, since it is based on the cumulative distribution function. The model that has the lowest total RPS score, summed over all observations in the test set, is the one providing the best prediction performance.

2.2. CUBREMOT

Assuming the CUB paradigm, model-based trees have been implemented in [3,4]. The resulting CUBREMOT (CUB Regression Model Trees) approach has proven to be effective for the identification of the main drivers of judgments, perceptions, evaluations and self-assessments collected on ordinal scales. Parsimony and effective graphical representation of results allow the characterization of response profiles at the terminal nodes, possibly distinguished in terms of both feeling (stress perception, health assessment, and so on) and uncertainty.

With respect to fluctuation tests usually advocated in the setting of MOB to pursue the partitioning algorithm [5], CUBREMOT assumes devoted splitting rules to grow the tree, which can be based either on log-likelihood deviance, on dissimilarity measure between response profiles corresponding to tree descendants, or on uncertainty reduction [6]. In the latter case, the procedure to grow the so-called uncertainty tree selects the covariate that implies the highest reduction in the uncertainty component for the response at each step of the partitioning algorithm, thus yielding – in the end – response profiles as homogeneous as possible with respect to feeling measurement.

The main assumption of the model-based approach to tree methods is that the maintained model is valid and provides good fitting at each partitioning level, which may not be the case. Motivated by this circumstance, in [6] a proposal for a flexible uncertainty tree has been advanced to yield a procedure that - starting from a baseline Binomial classification tree - selects at each step the best uncertainty specification to provide a more analytical description of the response profiles learnt from the tree. To this purpose, suitable diagnostics check have to be implemented to address adequately the local selection of the best fitting model and to assess the extent by which the maintained model can be assumed for the current split.

2.3. RESIDUALS' DIAGNOSTICS FOR ORDINAL DATA MODELS AND FLEXIBLE UNCERTAINTY TREE

In [7], a general procedure to perform residuals diagnostics for ordinal data models has been advanced. The idea is grounded on the definition of surrogate residuals, built by jittering the cumulative distribution between adjacent scores on the probability scale. On this basis, a necessary condition for a model to be correctly specified is provided, based on the uniform distribution of surrogate residuals built for an estimated model if it is correctly specified. This procedure has been then exploited in [6] to perform local diagnostics of a model-based trees for rating data: starting from a benchmark binomial tree, residuals' check is performed locally to first assess the extent by which the maintained model can be assumed as correctly specified, and then to run a local model selection for the best fitting model (within the class of mixture models with uncertainty) restricted to those extensions that in turn fulfill the necessary condition for being correctly specified.

3. RESULTS AND PRACTICAL APPLICATION

For the sake of illustration, results on two case studies discussed in [6] are briefly reported and commented. First, with reference to the ALLBUS German General Social Survey of 2012 (GESIS - Leibniz-Institut für Sozialwissenschaften, 2016), we consider the perceived trust expressed by $n = 2692$ respondents towards Press and Television, as institutions, collected on a rating scale with $m = 7$ ordered categories (1 = “no trust at all”, 7 = “a great deal of trust”), after list-wise omission of missing values of the considered set of variables. Figure 1 displays a table summarizing the results for the local model selection performed on the dissimilarity binomial tree for ratings on Trust for Television.

Table 11 Summarizing results for the local model selection on the (dissimilarity) binomial tree: Trust for Television

Node	Best model	$\hat{\pi}$	$1 - \hat{\xi}$	$\hat{\delta}$	$\hat{\phi}$	Shelter	Split with	Diss(Bin, f)	Diss(Best, f)	n
1	CUB + she	0.85	0.373	0.09	—	1	L: internet=0; R: internet=1	0.113	0.028	2692
2	BetaBin + she	0.94	0.450	0.06	0.05	1	L: west=0; R: west=1	0.107	0.033	702
3	CUB + she	0.87	0.349	0.094	—	1	L: Age _c ≤ 3; R: Age _c ≥ 4	0.111	0.029	1990
4	Bin + she	0.91	0.444	0.09	—	1	—	0.125	0.051	295
5	BetaBin + she	0.95	0.456	0.05	0.079	1	—	0.112	0.029	407
6	CUB + she	0.85	0.337	0.104	—	1	L: univ=0; R: univ=1	0.120	0.030	1635
7	Bin + she	0.97	0.392	0.033	—	1	—	0.064	0.040	355
12	CUB + she	0.84	0.366	0.103	—	1	L: leftright ≤ 4; R: leftright ≥ 5	0.128	0.032	996
13	Bin + she	0.90	0.297	0.103	—	1	—	0.103	0.034	639
24	BetaBin + she	0.87	0.283	0.13	0.116	4	—	0.167	0.022	312
25	CUB+she	0.83	0.373	0.092	—	1	—	0.115	0.014	684

Figure 1. Flexible uncertainty tree grown to classify ratings on Trust for Television

Overall, this procedure allows to obtain better explicative and prediction performance, as summarized in the table reported in Figure 2.

Table 13 Indicators of prediction performance: average dissimilarity between observed distribution of the test sets classified into terminal nodes and corresponding binomial and best mixture with uncertainty; total RPS and weighted average of RPS (averages are weighted with sample sizes)

	Dissimilarity		Average RPS		Total RPS	
	Binomial	Best mixture	Binomial	Best mixture	Binomial	Best mixture
Press	0.114	0.081	65.38	64.93	326.85	324.77
Tv	0.085	0.079	64.34	63.75	320.46	317.83

Figure 2. Summarizing results of explicative and predictive performance of the flexible uncertainty tree grown for ratings on perceived Trust for Press and Television - ALLBUS GESIS SOCIAL SURVEY

Finally, we report the results of the implementation of the (deviance) flexible uncertainty tree grown starting from a baseline CUBREMOT to classify and predict ratings on satisfaction for the PhD course of Italian PhDs, collected within a survey issued by the Italian Statistical Institute to investigate their satisfaction for professional placement (available at <https://www.istat.it/it/archivio/87536>): see Figure 3.

4. MAIN FINDINGS

Results indicate that different uncertainty components and different shelter effects are detected at varying partitioning levels, supporting the proposal of flexible uncertainty trees as a valid tool to run local model search for the best fitting and predictive performance, and thus enhancing a more analytical learning of response profiles for a given rating variable.

Table 14 Summarizing results for the local model selection on the (deviance) CUBREMOT for Ph.D. overall satisfaction

Node	Best model	$\hat{\pi}$	$1 - \hat{\xi}$	$\hat{\delta}$	$\hat{\phi}$	Shelter	Split with	Diss(CUB, f)	Diss(Best, f)
1	CUB+she	0.52	0.664	0.037	—	6	L: research= 0, R: research= 1	0.038	0.027
2	CUB+she	0.39	0.604	0.040	—	6	L: stem= 0; R: stem= 1	0.039	0.025
3	CUB+she	0.67	0.683	0.016	—	2	L: stem= 0; R: stem= 1	0.045	0.036
4	CUB+she	0.39	0.564	0.030	—	2	L: north= 0; R: north= 1	0.037	0.022
5	CUB+she	0.45	0.631	0.082	—	6	—	0.059	0.014
6	CUB+she	0.52	0.662	0.020	—	8	L: abroad= 0; R: abroad= 1	0.053	0.044
7	CUB+she	0.79	0.695	0.016	—	2	L: gender= 0; R: gender= 1	0.046	0.036
8	CUB	0.34	0.610	0.000	—	—	—	0.027	0.027
9	CUB+she	0.44	0.516	0.034	—	2	—	0.071	0.055
12	CUB+she	0.63	0.669	0.033	—	8	L: gender= 0; R: gender= 1	0.043	0.028
13	CUB+she	0.41	0.643	0.048	—	2	—	0.076	0.052
14	CUB+she	0.77	0.718	0.016	—	2	L: north= 0; R: north= 1	0.042	0.034
15	CUB+she	0.73	0.649	0.079	—	6	—	0.065	0.028
24	CUB+she	0.62	0.708	0.072	—	8	—	0.088	0.051
25	CUB+she	0.60	0.638	0.025	—	5	—	0.031	0.025
28	CUB+she	0.63	0.739	0.026	—	4	—	0.033	0.021
29	CUB+she	0.75	0.706	0.058	—	6	—	0.064	0.044

Figure 3. Flexible uncertainty tree to explain satisfaction of Italian PhDs

5. REFERENCES

- [1] Piccolo D. On the Moments of a Mixture of Uniform and Shifted Binomial random variables. QUADERNI DI STATISTICA (2003), 5: 85-104
- [2] Piccolo D., Simone R. The class of CUB models: statistical foundations, inferential issues and empirical evidence. STATISTICAL METHODS AND APPLICATIONS (2019), 28: 389-435.
- [3] Cappelli C., Simone R., Di Iorio F. CUBREMOT: a model-based tree for ordinal responses. EXPERT SYSTEMS WITH APPLICATIONS (2019), 124:39--49.
- [4] Simone R., Cappelli C., Di Iorio F. Modelling marginal ranking distributions: the uncertainty tree. PATTERN RECOGNITION LETTERS (2019). 125: 278--288.
- [5] Zeileis A., Hothorn T., Hornik H. Model-Based Recursive Partitioning. JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS (2008), 17(2), 492-514.
- [6] Simone R. Uncertainty Diagnostics of Binomial Trees for Ordered Rating Data. JOURNAL OF CLASSIFICATION (2023), 40:79-105.

- [7] Liu D., Zhang, H. Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION (2018), 113: 845-854.

Combining Longitudinal Administrative Registries, Test Results and Stakeholders' Input to Protect Educational Trajectories in Uruguay

Maria Castelao, María Seijas & Jennifer Viñas Forcade (INEED)

1. INTRODUCTION

Uruguay is classified as a high-income economy [1]. Its Human Development Index ranks fourth in Latin America [2]. Nevertheless, low educational attainment and its unequal socioeconomic distribution hinder opportunities for further development [3]. Only 40% of Uruguayan students reach the last year of mandatory secondary education timely and 20% are no longer enrolled in the educational system by then [4]. Recent national efforts to reduce grade repetition, absenteeism and dropout rates include an early-alert system, a curriculum reform and experiments with parental nudges [5]. Secondary school graduation has hence increased 13% but remains low (51.6% by age 23) [4].

To further inform national policies, we characterized the trajectories of a representative sample of Uruguayan public schools' 9th grade students who took part in Aristas (large-scale math and reading test) in 2018. The correlation between students' pre-test trajectories and their test performance is also analysed.

2. METHODOLOGY

The assembled longitudinal database (n=6809) results from identifying all students from the Aristas 2018 public school sample (n=8504), excluding those without scores in both tests (n=1464), with special needs (n=95), or no sampling weights (n=2). These were matched by student ID to administrative registries of their 2013-2018 school enrolment data. Students without 2013 enrolment data (n=134) were also excluded from the analyses.

We ran a weighted sequence analysis with optimal matching [6] using the *TraMineR* R package. These sequences were used for cluster analysis, combining Ward and Partitioning Around Medoids methods in the *WeightedCluster* R package. The Average Silhouette Width [7] and literature on educational trajectories in Uruguay [8] were used to define the number of clusters (k =4) in our typology of educational trajectories.

To analyse the correlation between students' pre-test trajectories and their test results, we estimated linear regression models with clustered standard errors¹ in the *estimatr* R package. Six models were estimated with each math and reading test results as dependent variables, plus an interaction model for reading.² We controlled for students' sex, 2013 old-for-grade status, 2017 academic achievement, 2018 family SES, 2018 school SES, and 2018 school track (Table 1). Multicollinearity was checked with bivariate correlations and the variance

¹ We needed to account for students being nested in schools, but weighted multilevel regression analysis wasn't possible, as the Aristas 2018 database doesn't include school weights.

² This interaction was not significant for math and is therefore not included in the math models.

inflation factor. To allow for non-linearity, we used the random forest machine learning technique and partial dependence plots by trajectory type.

Table 1. Description of the clusters

Variable	Whole sample	C1: optimal trajectory	C2: one-time repetition 8-9 grade	C3: multiple repetition 8-9 grade	C4: transition issues
Trajectory type	6809 (100%)	5092 (76.2%)	722 (10.3%)	442 (6.1%)	553 (7.4%)
Aristas math score	292 (44.9)	297 (45.4)	275 (38.4)	278 (38)	270 (36.6)
Aristas read score	294 (47.7)	299 (47.8)	281 (43.7)	279 (46.5)	271 (40.5)
Sex: female (%)	48.8	50.5	45.1	41	39.7
Sex: male (%)	51.2	49.2	54.9	59	60.3
Old-for-grade 2013 (%)	17.4	13.5	22.9	23.1	45.8
Math achievement 2017	6.85 (2.56)	7.36 (2.36)	4.5 (2.17)	4.72 (2.87)	6.51 (1.98)
Reading achievement 2017	7.37 (2.3)	7.9 (2.01)	5.15 (2.23)	5.17 (2.73)	6.9 (1.77)
Student SES 2018	-0.387 (0.849)	-0.414 (0.823)	-0.454 (0.907)	-0.218 (0.891)	-0.16 (0.947)
School SES 2018	0.387 (0.885)	0.516 (0.814)	0.203 (0.89)	-0.114 (0.962)	-0.269 (1.03)
<i>Student track 2018 (%):</i>					
regular/ academic	80.5	87	74.8	55.8	40.2
standard vocational	12.6	12.1	18	14.6	9.3
2nd chance vocational	6.9	0.8	7.2	29.5	50.5

3. RESULTS

119 different 2013-2018 students' trajectories were grouped in 4 clusters (Figure 1 & Table 1). C1: optimal trajectory (one grade per year), C2: repeated once the 8th/ 9th grade, C3: multiple repetitions of the 8th/ 9th grade, C4: grade repetition or temporary dropout (returnees) in the primary to secondary school transition.

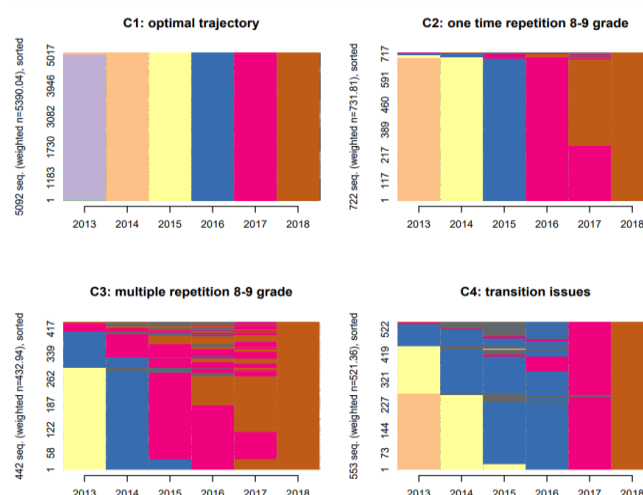


Figure 1. Cluster visualization of students' trajectories.

Regression analyses (Tables 2-3) show students who experience grade repetition or temporary dropout (C2-C3-C4) obtain lower test results than those in the optimal trajectory (C1). Grade repetition in primary school (old-for-grade in 2013) is associated with lower scores. Boys outperform girls in math. Girls outperform boys in reading, with no gender differences among boys and girls with the same previous-year grades in Spanish. When previous-year grades are controlled for, only those with transition difficulties (C4) perform significantly lower than optimal trajectory students (C1).

Student and school SES are positively associated with test results. When controlling for SES, multiple-repetition students (C3) outperform optimal trajectory peers (C1) in math, even after controlling for track (this is further analysed with machine learning below). Consistent with previous literature [9], vocational school students underperform those in regular academic schools in math. In reading, students in vocational second-chance programs underperform those attending standard programs in regular and vocational schools. Nevertheless, the model with a trajectory*track interaction shows students attending vocational second-chance programs tend to perform worse in all trajectory types but C4. This is: returnees' reading performance benefits from attending this remedial program.

Machine learning analyses focused on the behaviour of the multiple-repetition group (C3). The non-linear relation between students' previous-year grades and their test results varies greatly by cluster, but C3 underperforms C1 at all times (Figure 2). We conclude the surprising C3 results in linear regressions derive from the distribution of control variables among clusters. This is: we are comparing C1 & C3 students similar in all covariates, for which these groups' composition greatly differs. We are hence comparing the lower-performance tail of C1 with the higher-performance head of C3. It's also worth noting that working with a 9th grade sample means observing pre-test trajectories of students who actually made it to 9th grade, excluding those who left school before this grade-level. In a follow-up study, we will further analyse the post-test (2019-2024) trajectories of the four groups, putting these results in a whole-trajectory perspective.

Table 2. Math regression models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
C2: one-time repetition 8-9 grade	-21,650*** (2,301)	-21,942*** (2,276)	-20,159*** (2,235)	-1,946 (2,397)	1,348 (2,285)	2,019 (2,297)
C3: multiple repetition 8-9 grade	-18,773*** (2,683)	-19,202*** (2,649)	-17,765*** (2,712)	-0,083 (3,268)	6,309* (3,076)	8,410** (3,043)
C4: transition issues	-27,199*** (2,504)	-27,621*** (2,489)	-21,805*** (2,485)	-17,464*** (2,468)	-9,638*** (2,419)	-6,568*** (2,386)
sex: male		3,992** (1,163)	4,900*** (1,141)	7,459*** (1,042)	6,893*** (1,079)	7,223*** (1,086)
old-for-grade 2013			-18,766*** (2,020)	-15,410*** (1,937)	-10,929*** (1,836)	-10,023*** (1,868)
achievement 2017				6,579*** (0,496)	6,492*** (0,488)	6,588*** (0,487)
student SES 2018					3,939*** (0,885)	4,036*** (0,883)
school SES 2018					12,914*** (1,269)	11,749*** (1,381)
track: standard vocational						-6,060* (2,430)
track: 2nd.chance vocational						-8,754* (3,819)
Constant	297,119*** (1,535)	295,155*** (1,635)	297,164*** (1,651)	246,602*** (3,881)	241,938*** (4,156)	242,388*** (4,142)
Observations	5,946	5,946	5,946	5,946	5,946	5,946
R ²	0,046	0,048	0,071	0,189	0,232	0,235
Adjusted R ²	0,045	0,047	0,070	0,188	0,231	0,234
Residual Std. Error	44,654 (df = 5942)	44,612 (df = 5941)	44,070 (df = 5940)	41,175 (df = 5939)	40,068 (df = 5937)	40,009 (df = 5935)
F Statistic	95,235*** (df = 3; 5942)	74,640*** (df = 4; 5941)	90,778*** (df = 5; 5940)	230,923*** (df = 6; 5939)	224,753*** (df = 8; 5937)	182,270*** (df = 10; 5935)

Note:

+ p<0,1; * p<0,05; ** p<0,01; *** p<0,001

Table 3. Reading regression models

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
C2: one-time repetition 8-9 grade	-17,856*** (3,244)	-17,510*** (3,225)	-15,446*** (3,051)	2,616 (3,043)	5,475* (3,005)	5,797* (3,042)	5,683 (3,941)
C3: multiple repetition 8-9 grade	-19,740*** (3,391)	-19,249*** (3,399)	-16,983*** (3,449)	0,317 (3,406)	6,243* (3,666)	7,187* (3,624)	5,512 (5,420)
C4: transition issues	-27,466*** (2,867)	-26,903*** (2,837)	-19,598*** (2,780)	-14,620*** (2,731)	-7,541* (3,044)	-6,162* (2,750)	-9,031* (4,002)
sex: male		-5,233** (1,575)	-4,232* (1,572)	1,415 (1,564)	0,579 (1,526)	0,945 (1,563)	0,968 (1,552)
old-for-grade 2013			-22,514*** (2,121)	-18,340*** (2,165)	-14,294*** (2,115)	-13,677*** (2,119)	-13,658*** (2,140)
achievement 2017				6,892*** (0,480)	6,672*** (0,481)	6,630*** (0,481)	6,623*** (0,479)
student SES 2018					4,705*** (0,893)	4,770*** (0,888)	4,685*** (0,890)
school SES 2018					12,176*** (1,754)	10,907*** (1,982)	10,853*** (1,978)
track: standard vocational						-8,333* (3,316)	-8,981* (3,297)
track: 2nd.chance vocational						-5,854 (4,129)	-16,478* (7,624)
C2* standard vocational							-0,171 (6,137)
C3* standard vocational							13,950 (8,736)
C4* standard vocational							-2,990 (6,479)
C2* 2nd.chance vocational							11,325 (8,931)
C3* 2nd.chance vocational							8,734 (9,414)
C4* 2nd.chance vocational							16,680* (7,525)
Constant	298,995*** (1,681)	301,588*** (1,905)	304,055*** (1,887)	245,979*** (4,056)	243,253*** (4,173)	245,069*** (4,356)	245,272*** (4,328)
Observations	5,851	5,851	5,851	5,851	5,851	5,851	5,851
R ²	0,037	0,040	0,069	0,153	0,184	0,187	0,189
Adjusted R ²	0,036	0,039	0,068	0,152	0,183	0,186	0,187
Residual Std. Error	47,910 (df = 5847)	47,840 (df = 5846)	47,103 (df = 5845)	44,945 (df = 5844)	44,099 (df = 5842)	44,024 (df = 5840)	44,010 (df = 5834)
F Statistic	73,906*** (df = 3; 5847)	60,126*** (df = 4; 5846)	86,692*** (df = 5; 5845)	175,324*** (df = 6; 5844)	165,114*** (df = 8; 5842)	134,745*** (df = 10; 5840)	84,880*** (df = 16; 5834)

Note:

+ p<0,1; * p<0,05; ** p<0,01; *** p<0,001

4. MAIN FINDINGS

Uruguayan students' educational trajectories from primary to lower secondary education can be classified into four distinctive groups, for which individual and school characteristics are differentially associated with academic performance. Knowing which group each student belongs to can help inform actions undertaken inside schools to support students' academic performance, as well as their remaining in school and advancing towards secondary school completion. At the national level, this may also help inform allocation of resources according to the characteristics of each school's population in terms of students' trajectories.

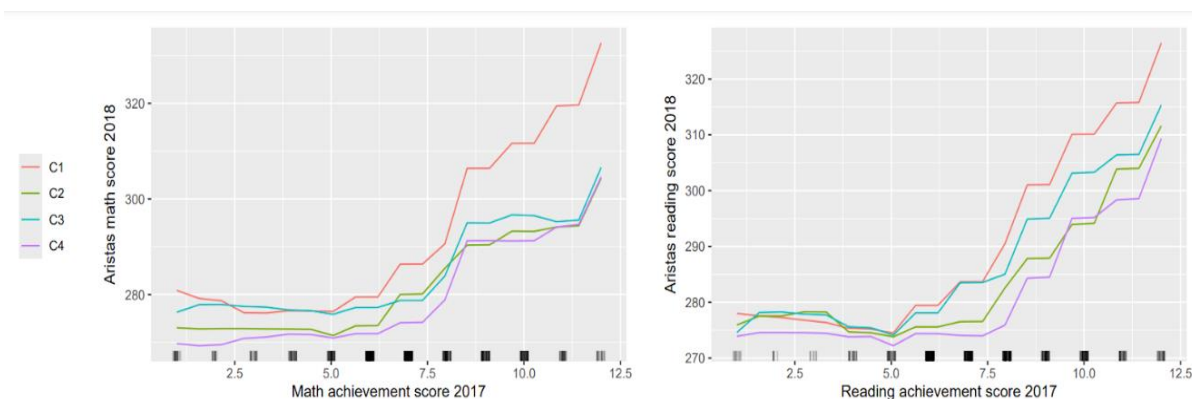


Figure 2. Partial dependence plots. Variation of Aristas score by 2017 achievement scores for each cluster.

5. REFERENCES

- [1] World Bank, Country and Lending Groups (2020), available in: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>.
- [2] UNDP, Human Development Report (2019).

- [3] OCDE and ECLAC, Multi-dimensional Review of Uruguay: Volume 1, 2014, available in: <https://doi.org/10.1787/9789264209459-en>.
- [4] ANEP, Trayectorias educativas, Observatorio de la Educación, 2024, in: <https://observatorio.anep.edu.uy/proceso/trayectorias-educativas>
- [5] D. de León, G. Suarez, L. Rosich, C. Muniz, I. González and I. Sánchez, ¿Cuánto faltó? Favoreciendo la asistencia escolar a partir de intervenciones comportamentales, *Revista Argentina de Ciencias del Comportamiento* (2024).
- [6] H. MacIndoe and A. Abbott. 2004. Sequence analysis and optimal matching techniques for social science data. In *Handbook of Data Analysis*, M. Hardy and A. Bryman (Eds.). Sage, 387–405.
- [7] M. Studer. *WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R* (2013). Available in: <https://doi.org/10.12682/lives.2296-1658.2013.24>.
- [8] S. Cardozo, Trayectorias educativas en la educación media PISA-L 2009-2014. INEE (2016).
- [9] D. Jacobs and A. Rea, *Gaspillage de talents.: Les écarts de performances dans l'enseignement secondaire entre élèves issus de l'immigration et les autres d'après l'étude PISA 2009*. Fondation Roi Baudouin (2011).

Combining Nonprobability Samples with Probability Samples to Correct for Selection Bias

Ton de Waal, Sander Scholtus (Statistics Netherlands), An-Chiao Liu (Utrecht University) & Sofia Villalobos-Aliste (Vrije Universiteit Amsterdam)

1. INTRODUCTION

National statistical institutes traditionally use probability samples to produce estimates for population parameters of interest. In probability sampling one draws units from the target population according to a sampling design for which the inclusion probability of each unit is known. This enables one to obtain unbiased estimates for population parameters. However, major drawbacks of probability samples are that it is time-consuming and expensive to collect the data, which also implies that such samples are often rather small and estimators based on them have a high sampling variance. Nowadays, a large variety of datasets that are not based on sampling designs and for which the inclusion probabilities are unknown provide a massive amount of data at a low cost within a short time. Such datasets are referred to as nonprobability samples. Examples are administrative data, opt-in online surveys and big data. The low cost and fast availability of nonprobability samples make them very attractive to be used for statistical purposes. However, they are often selective, and because their “sampling design” is unknown, estimators based on such samples are usually biased.

To use nonprobability samples for producing estimates for population parameters, one generally has to correct for selectivity. This can be done by means of several kinds of approaches; for an overview see, e.g., [1]. Examples of such approaches are combining estimates for a target variable based on a probability sample with estimates from a nonprobability sample (see [2], [3]), and pseudo-weight approaches (see, e.g., [1], [4]). In this paper we focus on how iterative proportional fitting (IPF) can play a beneficial role in these two kinds of approaches when one aims to estimate a contingency table of proportions of a categorical target variable over domains in the population.

Section 2 sketches an approach for combining estimates for a target variable from a probability sample with estimates from a nonprobability sample, and a pseudo-weight approach, which can be used in combination with IPF. Section 3 provides results, and Section 4 ends this paper by drawing some conclusions.

2. METHODOLOGY

We suppose there is a target population of N units with a categorical target variable y with C categories, which is observed in a nonprobability sample NPS. We assume that besides NPS also a probability sample PS is available, which – for simplicity – is assumed to be drawn by simple random sampling. We assume that the units in the population are divided into K domains, where the (known) size of domain k is N_k . We aim to estimate the proportion of units Z_{kc} for each category c of y per domain k in the population.

2.1. Combining estimates based on NPS with estimates based on PS

In this section we assume that observations on y are also available from PS, regardless of whether the same unit is present in one or both samples. In [3], an estimator of the form

$$\hat{D}_{kc} = W_{kc} \hat{Z}_{kc}^{(PS)} + (1 - W_{kc}) \hat{Z}_{kc}^{(NPS)} \quad (1)$$

is proposed for Z_{kc} . In (1) $\hat{Z}_{kc}^{(PS)}$ is the direct estimator for Z_{kc} based on PS, $\hat{Z}_{kc}^{(NPS)}$ the corresponding estimator based on NPS, and W_{kc} a weight between zero and one. If we could compute the Mean Squared Error (MSE) of both $\hat{Z}_{kc}^{(PS)}$ and $\hat{Z}_{kc}^{(NPS)}$, the weight for which the MSE of $\hat{D}_{kc}$ is minimum would be given by (see [5]):

$$W_{kc} = \frac{MSE(\hat{Z}_{kc}^{(PS)})}{MSE(\hat{Z}_{kc}^{(PS)}) + MSE(\hat{Z}_{kc}^{(NPS)})} \quad (2)$$

The main challenge in constructing the weight W_{kc} is that the bias $b_{kc} = \hat{Z}_{kc}^{(NPS)} - Z_{kc}$ of $\hat{Z}_{kc}^{(NPS)}$ is unknown and therefore $MSE(\hat{Z}_{kc}^{(NPS)})$ cannot be estimated directly.

In [2], it is largely assumed that b_{kc} is known from another source. However, in [2] also a simple estimator for b_{kc} is suggested, namely $\hat{b}_{kc} = \hat{Z}_{kc}^{(NPS)} - \hat{Z}_{kc}^{(PS)}$. In [3], the approach proposed in [2] is extended by assuming a model for b_{kc} , where b_{kc} is distributed as a random variable with mean $E_b(b_{kc}) = \beta_c$ and variance $E_b[(b_{kc} - \beta_c)^2] = \sigma^2$. Next, *expected* MSEs (EMSEs) under the posited model are computed. The final estimator for Z_{kc} , which we will call the combined estimator, is given by (1) where the W_{kc} in (2) are computed with the MSEs replaced by EMSEs. For details we refer to [3].

In [3] also a simple approach based on IPF is proposed. In this IPF approach a 2-dimensional table is constructed where the initial internal values are given by $\hat{Z}_{kc}^{(NPS)}$ and the fixed marginal totals by $\hat{Z}_{+c}^{(PS)} = \sum_{k=1}^K \hat{Z}_{kc}^{(PS)}$ ($c = 1, \dots, C$) and $\hat{Z}_{k+}^{(PS)} = \sum_{c=1}^C \hat{Z}_{kc}^{(PS)}$ ($k = 1, \dots, K$). Next, IPF is applied to align the internal cell values with the fixed marginal totals.

2.2. Pseudo-weighting

In this section we do not assume that PS includes the target variable. Instead, we assume that in NPS and PS a common set of auxiliary variables x are observed that govern the inclusion process in NPS, and that the probability d_i to be included in PS is known for each unit i in the population.

In a pseudo-weight approach, one constructs pseudo-weights for the units in NPS. The constructed pseudo-weights can be used to obtain weighted estimates for the target variable by a design-based estimator, e.g., the Horvitz-Thompson estimator (see [5]). There are several approaches to estimate pseudo-weights. We use the approach proposed in [4], which we refer to as LSW. In LSW, a propensity model $f(x)$ is fitted with the non-overlapping part B of the row combined samples PS and NPS. We assign $z_i = 1$ if $i \in \text{NPS} \setminus \text{PS}$ and $z_i = 0$ if $i \in \text{PS} \setminus \text{NPS}$ as the dependent variable. After fitting $f(x)$, the odds $O_i = \frac{\Pr(z_i=1|i \in B, x_i)}{\Pr(z_i=0|i \in B, x_i)}$ are calculated. The pseudo-weights are given by the reciprocals of the inclusion propensities, i.e., $w_{\text{incl},i} = \frac{1}{\Pr(z_i=1|i \in B, x_i)} = 1 + \frac{d_i - 1}{O_i}$. Here we assumed that the inclusion in NPS is independent of the inclusion in PS. For details we refer to [4].

To estimate Z_{kc} for each category c of y per domain k after calculating the pseudo-weights one can use these pseudo-weights directly or one can additionally use some form of IPF. In this paper we apply IPF to align the internal cell values obtained by the direct estimator based

on NPS with fixed marginal totals given by N_k/N ($k = 1, \dots, K$) and $\sum_{i \in NPS} w_{incl,i} I(y_i = c) / \sum_{i \in NPS} w_{incl,i}$ ($c = 1, \dots, C$), where $I(y_i = c)$ equals 1 if $y_i = c$ and 0 otherwise. We refer to this method as LSW+IPF.

3. RESULTS

We first compare the IPF approach (without pseudo-weighting) to the combined estimator (see Section 2.1). In [3], a population of $N = 100,000$ units was simulated from which a PS and an NPS were repeatedly drawn. For each unit an outcome variable was generated that defines to which category c of the target variable and which domain k it belongs. Also, two auxiliary variables were generated, one of which was used to manipulate the level of selectivity in NPS. Four domain sizes ($K = 2, 4, 10, 15$), four numbers of categories ($C = 3, 5, 8, 15$), a scenario where these categories are of equal size, and a scenario where categories are of unequal sizes were simulated. Also, sample sizes per domain of 20, 100, 400 and 900 for PS and 100, 1000, 2000 and 6000 for NPS were simulated, and three different levels of selectivity (strong, medium, and weak) were introduced.

To evaluate the results, $MARMSE = \frac{1}{K} \frac{1}{C} \sum_{k=1}^K \sum_{c=1}^C \sqrt{\frac{1}{R} \sum_{r=1}^R (Z_{kc} - \hat{Z}_{ck}^{(r)})^2}$ was calculated where $R = 1000$ is the number of simulations and $\hat{Z}_{ck}^{(r)}$ is either the combined estimator or the estimator based on using IPF (see Section 2.1) in the r -th simulation. Bias was assessed by $MAB = \frac{1}{RKC} \sum_{r=1}^R \sum_{k=1}^K \sum_{c=1}^C |Z_{kc} - \hat{Z}_{ck}^{(r)}|$.

Table 1 presents the results for the scenario where the categories are of equal size. The results when the categories are of unequal size are similar. Table 1 shows that the combined estimator outperformed the IPF estimator in terms of bias most of the time, in particular when the level of selectivity of NPS was small. However, in terms of MARMSE, the combined estimator and IPF performed about equally well. Apparently, the combined estimator outperforms the IPF approach in terms of bias correction, but the IPF approach is more stable, i.e., has lower variance than the combined estimator.

Table 1. Proportion ($\times 100\%$) of IPF with lower MARMSE respectively MAB than the combined estimator

Results for MARMSE	
Selectivity	Proportion IPF best
Strong	56
Medium	48
Weak	41
Results for MAB	
Selectivity	Proportion IPF best
Strong	37
Medium	35
Weak	12

We now compare LSW+IPF to the direct estimator based on NPS and LSW (see Section 2.2). In [6], the SD2011 dataset from the R package *synpop* is used as the population. Its size is 4,867. The target variable *sport* has two categories: “Active involvement in some form of sport or exercise” and “Otherwise”. The domain variables are *age group* and *education level*. From this population $R = 5,000$ pairs of PS (of size 500) and NPS (with various sizes, see Table 2) are drawn. To evaluate the results $MSE_A = \frac{1}{R} \sum_{r=1}^R \sum_{c=1}^C \sum_{k=1}^K (Z_{kc} - \hat{Z}_{ck}^{(r)})^2$ is used, where $\hat{Z}_{ck}^{(r)}$ is estimated proportion for category c and domain k for the r -th pair of PS and NPS.

Both LSW and LSW+IPF lead to the lowest average bias for the marginal totals $\sum_k \hat{Z}_{kc}$ (results not reported here). From Table 2 below we see that the direct estimator based on NPS performs better in terms of MSE_A than using LSW, and that LSW+IPF performs the best for this dataset.

Table 2. MSE_A ($\times 1,000$) for the direct estimator, LSW and LSW+IPF.

Sample size NPS	direct	LSW	LSW+IPF
1500	1.10	40.78	0.18
2000	0.93	41.43	0.11
2500	0.84	41.51	0.08

4. MAIN FINDINGS

In this paper we found that the combined estimator proposed in [3] is better in terms of MAB than the IPF approach (also proposed in [3]), but the IPF approach is more stable as is shown by the results for MARMSE.

We also found that when one uses pseudo-weights to correct for selection bias in the estimator based on the nonprobability sample and then uses IPF to aligns the obtained estimates with domain sizes, the results for MSE_A clearly improve compared to only using pseudo-weights.

This shows that, in some situations, using IPF can be quite beneficial – either as a supplement to, or as an alternative for, more advanced techniques – when one wants to correct an estimator based on a nonprobability sample for selection bias.

5. REFERENCES

- [1] M.R. Elliott, and R. Valliant, Inference for Nonprobability Samples. *Statistical Science*, 32 (2017), 249–264.
- [2] M.N. Elliott and A. Haviland, Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology*, 33 (2007), 211–215.
- [3] S. Villalobos-Alíste, S. Scholtus and T. de Waal, Combining Probability and Non-Probability Samples on an Aggregated Level. Accepted for publication in the *Journal of Official Statistics* (2024).
- [4] A.-C. Liu, S. Scholtus and T. de Waal, Correcting Selection Bias in Big Data by Pseudo Weighting. *Journal of Survey Statistics and Methodology* (2022).
- [5] C.-E. Särndal, B. Swensson, B. and J.H. Wretman, *Model Assisted Survey Sampling* (1992). New York: Springer.
- [6] A.-C. Liu, S. Scholtus, K. van Deun and T. de Waal, Correcting Selection Bias in Contingency Tables (2024) (work in progress).

Construction of a probabilistic scoring model for food allergy profiles in Tunisia

Fadia Bougacha & Ahmed Rebai (Centre of Biotechnology of Sfax)

1. INTRODUCTION

Food allergy is a serious public health problem, affecting between 1 and 10% of citizens and between 1.9 and 5.6% of children aged 7 to 10 years (Lyons et al, 2020). Although it has been well studied in developed countries, few data are available in Low and Middle income countries, in Africa and elsewhere. In Tunisia, a northern African country, the prevalence of food allergies in Tunisia is understudied. To fill this information gap, a Federated Research Project (FRP) entitled “PARADIS” has been launched in 2017. PARADIS aims to develop a big data platform to contain data on the different types of food allergies encountered in Tunisia. The first milestone was to carry on a population survey (including children and adults) in different delegations of the city of Sfax (south eastern Tunisia) (Belmabrouk et al, 2023). The present work uses the data collected from the survey to build a probabilistic score model based on predictive analysis methods such as Decision Trees, Random Forest and Bayesian Networks to propose a score that reflects an individual's level of allergic severity and identify the types of allergies most associated with this score.

2. METHODOLOGY

We are going to work on data collected from a survey as part of PARADIS project. We worked on a database containing 509 individuals and 11 variables (status: child/adult, age, sex and allergy status about 8 food groups). In this work, various models were applied, including Bayesian networks, decision tree and Random Forest, to assess the risk of food allergies in Tunisia:

2.1. Bayesian Network

A Bayesian network is a probabilistic graphical representation of knowledge about a complex phenomenon. Building a Bayesian network involves two stages: a) parameter learning (this consists in estimating the parameters of the Bayesian network using different estimation methods and b) structure learning (this consists in determining the best structure of Bayesian network (graph of nodes and arcs).

2.2. Decision tree

Decision trees are an easy-to-use classification/prediction model that can be interpreted by experts. A decision tree is made up of nodes in a tree structure. The node at the base of the tree is called the root, and each node below the root is either a leaf or a sub-tree. A leaf is a terminal node representing the result of a classification. The root of a sub-tree is labeled with the attribute that has been chosen, and the branches are labeled with the different values that the attribute chosen for the node can take. There are several tree-building algorithms. In our project, we used the CART algorithm.

2.3. Random Forest

Random forests are a model aggregation technique proposed by Breiman in 2001. Random forests are composed, as the term “forest” implies, of a set of binary decision trees into which randomness has been introduced. These trees are distinguished from each other by the subsample of data on which they are trained. These subsamples are drawn at random from an initial data set. In a decision tree, we take a single decision tree, but in the random forest, we take several decision trees, and the final prediction is given by: if the answer is quantitative, the prediction is the mean of the observations in the terminal node, if the answer is qualitative, the prediction is the most frequent modality (majority rule) in the terminal node.

2.4. Construction of the Naïve score and the Smart score

We begin by developing a food allergy severity score using two methods:

- A Naïve score equivalent to the number of foods to which the individual is allergic, i.e. the sum of the 0/1 states of allergy to all foods; for example, the score is equal to 1 if the individual suffers from food allergy to a single product, 2 if the individual suffers from food allergy to two products, etc.
- A Smart score is defined on the basis of the probabilities calculated by the Bayesian network model. So, some food products may have higher weights than others if, for example, they are likely to be accompanied more frequently by other allergies. In particular, we multiplied the 0/1 state of each foodstuff by its marginal probability if it is a parent node, and by its conditional probability (conditional on its parents being in state 0) for a child node. After rounding off the probabilities and standardizing (by a constant multiplicative factor), we obtain:

$$\text{ScoreSmart} = 16 * \text{Fruit} + 12 * \text{Poisson} + 3 * \text{Nuts} + 6 * \text{Cereals} + 6 * \text{Milk} + 3 * \text{Vegetables} + 4 * \text{Eggs} + 2 * \text{Peanuts};$$

This score therefore varies theoretically between 2 and 52; in practice, since we have individuals allergic to at least one food, this score varied in our sample between 2 and 46.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Application of Bayesian network

We worked on a database containing 368 adults (age ≥ 18 years) suffering from food allergies. According to the Bayesian network, gender is not connected to the network, suggesting that there is no difference between males and females in terms of allergy. Age and nuts also emerge as an independent subgraph of the general graph, suggesting that for adults, age is not an important factor except for nut allergy, which is an allergy that persists into adulthood. We note that patients with a food allergy to fruit and cereals will have a food allergy to fish with a probability of 0.34, meaning that an individual suffering from an allergy to these two types of food has a one-in-three chance of also having an allergy to fish (a randomly selected allergic individual has a probability of 0.23 of having an allergy to fish). A very similar trend can be observed for patients with a food allergy to fruit

and vegetables, who will have a high probability of having a food allergy to milk (conditional probability of 0.337 vs. a marginal probability of 0.065). We worked on a database containing 141 children (age < 18) with food allergies. According to the Bayesian network, patients with a food allergy to fruit will have a food allergy to cereals, milk and eggs. We will then work on a database containing 509 individuals suffering from food allergies. According to the Bayesian network, we can see that patients with a food allergy to fruit are more likely to have a food allergy to fish, milk, cereals, eggs and peanuts. Also, patients with a food allergy to fruit and vegetables will have a food allergy to milk with a probability of 0.30, meaning that an individual suffering from an allergy to these two types of food has a one-in-three chance of also having an allergy to milk. Patients with a food allergy to nuts have a food allergy to fish and peanuts.

3.2. Application of Decision tree

We applied the decision tree to all individuals with a food allergy. We will use the variables: scores, other variables such as allergic status to various non-grouped food products, age and sex. The decision tree indicates (figure 1), as does the Bayesian network model, that fruit allergy appears as a main node in score prediction, followed by fish allergy, which is consistent with their high prevalence in the population studied.

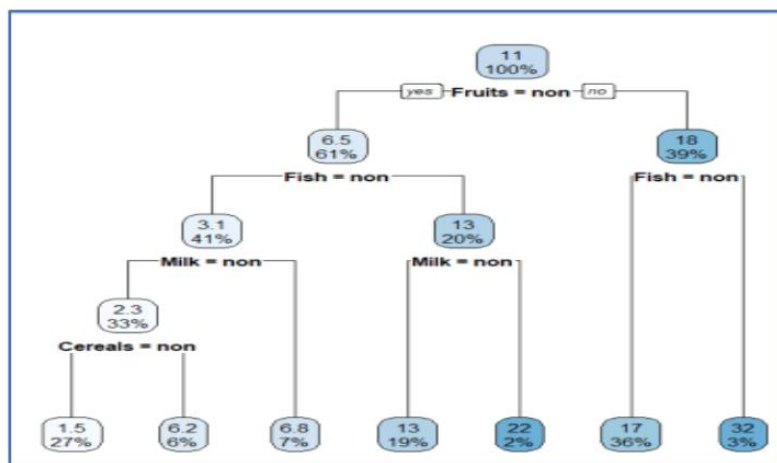


Figure 1. Decision tree based on Smart Score for all individuals

3.3. Application of Random Forest

We applied Random Forest to all individuals with at least one food allergy. The aim was to identify the variables (allergies) most associated with the severity score. In the model we have calculated (figure 2), the variables most likely to be associated with the Naïve severity score for all individuals are fruit, fish and peanut; in the model we have calculated (figure 3), the variables most likely to be associated with the Smart severity score for all individuals are fruit, fish.

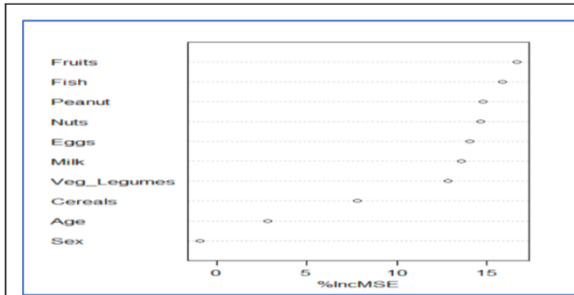


Figure 2. the Naïve severity score for all individuals

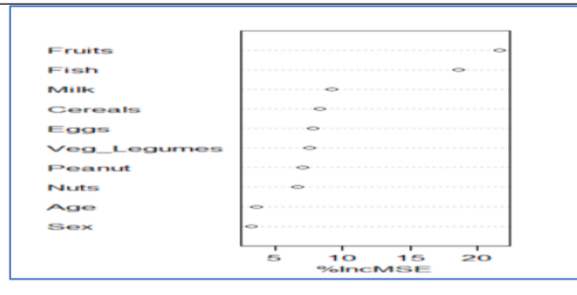


Figure 3. the Smart severity score for all individuals

4. MAIN FINDINGS

The aim of this study is to identify food allergy profiles in Tunisia. It also estimates the risk of developing an allergy via a score defined on the basis of calculated probabilities. In this context, 11 variables, defined by status (child/adult), age, gender and allergy to the 8 food groups identified in our population and collated from 509 individuals, were used as the input basis for various probabilistic models. The first model developed, based on Bayesian networks, enabled us to identify that there is no gender predominance in food allergy in adults. In contrast, a French multicenter study reported a greater predisposition of males to allergies than females in adults over 60. This result is to be expected, given that our sample contains few individuals in the over-60 age bracket, and is consistent with those found in other studies. Age is a less important factor, except for nut allergies, which supports the hypothesis that food allergies are generally permanent, except for cow's milk protein and hen's egg allergies, which disappear in children between 18 months and 2 years of age.

Bayesian networks have also enabled us to identify the profile of adult patients predisposed to developing other allergies, such as cereal and fruit allergies, which predispose to fish or milk allergies. In children, on the other hand, we find the predominance of allergens described in most other populations, with a difference in predisposition to classic food allergy to cereals, milk and eggs in the presence of an allergy to fruit.

The link between the different allergies would reflect the severity of the food allergy and can be translated into a naive score and a smart score. The latter have been used in the decision tree and Random Forest applications and show that fruit (with the highest naive score reflecting the highest marginal probability) ranks first, followed by fish, and in some cases (adults) peanuts. The results obtained in our study population are consistent with the culinary characteristics of the Mediterranean region or diet, and with certain genetic predispositions (Scott.H et al, 2020).

In this context, better identification of at-risk individuals will be of major interest in raising awareness of these allergy problems, particularly among children, and making them more attentive to preventive actions in their social and family environment. In addition, the models we have developed can be used as a decision-support tool to predict cross-allergies, i.e. allergic clinical manifestations

due to different allergens, particularly food-food allergens. Later, we can also envisage a monitoring system for industrialists and controls by administrations for good management of the risk of food allergy, such as the introduction of exhaustive and precise labelling in the agri-food sector, or the development of an allergeo-vigilance network constituting secondary prevention actions. In addition, a graphical interface of probabilistic methods for the different allergic profiles developed in Tunisia could be envisaged to predict an individual's allergic risk and offer personalized prevention of allergic reactions.

5. REFERENCES

- [1] Sabrine Belmabrouk, Rania Abdelhedi, Fadia Bougacha et al (2023),
- [2] Prevalence of self-reported food allergy in Tunisia: General trends and
- [3] probabilistic modelling
- [4] Lyons SA, Clausen M, Knulst AC et al (2020). Prevalence of Food
- [5] Sensitization and Food Allergy in Children Across Europe.
- [6] H. A Hampson (2016), Food allergy: past, present and future.
- [7] W.LOH, L.M.K. Tang (2018), The epidemiology of food allergy in the
- [8] global context.
- [9] R. Sharon Chinthrajah et al (2018), Development of a Tool Predicting
- [10] Severity of Allergic Reaction During Peanut Challenge.
- [11] Jessica Savage and Christina B. Johns (2015). Food allergy:
- [12] epidemiology and natural history.
- [13] V. Sampath, E.M. Abrams, B. Adlou, et al (2021), Food allergy across
- [14] the globe.
- [15] M. E. Pettersson, G. H. Koppelma, B. M. J. Flokstra-de Blok et al
- [16] (2018). Prediction of the severity of allergic reactions to foods.
- [17] Scott H. Sicherer et al (2020). Critical issues in food allergy: A National
- [18] Academies Consensus Report.

Cyclic Outlier Detection in AIS Data: A Direct Approach Based on Average Speed.

Francesco Sisti, Angela Pappagallo & Luca Valentino (ISTAT - Italian National Institute of Statistics)

1. INTRODUCTION

The Automatic Identification System (AIS) is an automated ship tracking system commonly used in the maritime domain for security and management purposes. AIS data are capable of accurately describing the identity of a vessel, its main characteristics (e.g. dimension, type) and dynamic information, like the geolocation at sea. AIS signals are transmitted at approximately 10-second intervals, which renders them a significant source of Big Data.

Thus, AIS represents a significant chance to enhance the accuracy and reliability of statistical data, in particular for ISTAT to improve the timeliness of the transmission of specific tables to Eurostat. These include, for instance, the data on the number of ships arriving in Italian ports, which are being requested on a quarterly basis.

A voyage is mainly characterised by two ship events: the departure event and the arrival event. Previous research has already investigated techniques for detecting ship arrivals and departures from AIS. For example, see [1], [2], and [3]. Unfortunately, certain issues affect the quality of AIS data, frequently leading to the incorrect reconstruction of voyages. In particular, in this work we investigate the problem of outliers in AIS data and the methods to handle them as explained in the next section.

2. KIND OF ANOMALIES AND REGULAR APPROACHES

Outliers in AIS positions are defined as jumps in the vessel's location of hundreds, or even thousands, of kilometres over the course of a few temporally contiguous signals (within 30 minutes), observed in data without apparent regularity.

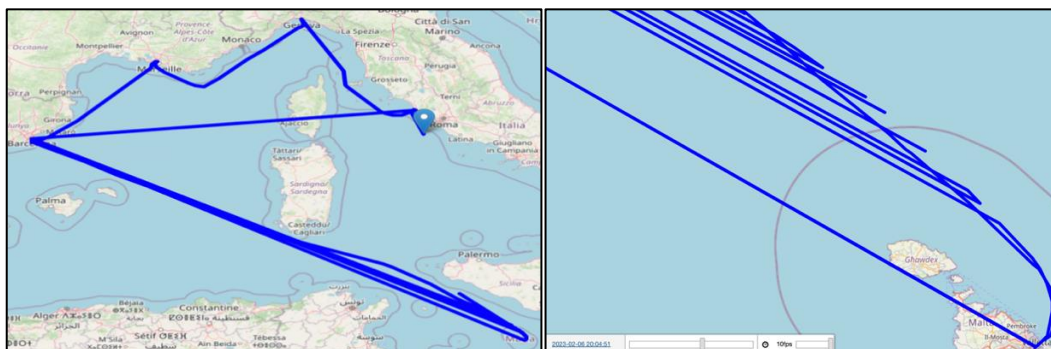


Figure 1. Left: Example of trajectories with outliers. Right: Route detail, trajectories can be heavily affected by outliers.

A comprehensive overview of the various types of anomalies in positioning data, illustrated in Figure 2, can be found in a recent review paper on the subject [4]. The outlier of interest

concerns sudden, substantial large jumps in the GPS-reported position. Examples of such outliers in real data are shown in Figure 1.

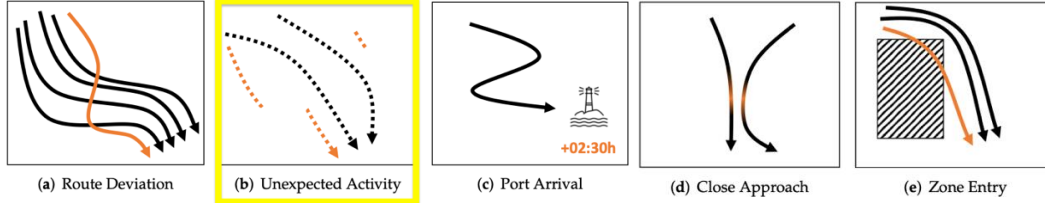


Figure 2. The five general types of AIS anomalies, summarized in [4]. The data marked in orange in each figure indicate a potential anomalous route. The type of anomaly we are interested into is highlighted in the yellow box.

A variety of approaches are presented in the literature to address this problem, with machine learning and statistical methods being the most commonly employed.

Several machine learning techniques can be used to detect outliers in GPS data, each with distinct advantages and limitations. A modified k-Nearest Neighbours (k-NN) calculates distances to flag outliers, but it requires temporal integration and is unable to effectively address clustered anomalies. t-DBSCAN, which builds on DBSCAN, incorporates time; however, adapting it to linear trajectories like ship paths is challenging and further requires careful parameter tuning. The Local Outlier Factor (LOF) is effective when applied to data with varying densities, though temporal adjustments are necessary. Lastly, the Isolation Forest algorithm can efficiently detect multiple outliers by splitting data. Nevertheless, it requires custom temporal integration and is unable to cope with varying densities.

In our outlier analysis, we also considered several traditional statistical approaches, including z-score, interquartile range, and median filtering, which are effective for univariate data. Although these could be applied separately to each coordinate (x, y) in our bivariate data, this would result in the disregard of the correlation between variables, which is of crucial importance in our case. Therefore, we excluded them. Similarly, for the detection of outliers in bivariate data, methods such as Mahalanobis Distance and confidence ellipses are inadequate in capturing the curved trajectories of ships over time.

3. DIRECT APPROACH FOR OUTLIERS' DETECTION

For the reasons mentioned, we adopted a direct approach, evaluating the average speed between pairs of data observations as the sole criterion. Although this method was relatively straightforward, it yielded some interesting observations and results.

Firstly, we created three key variables by ordering the data based on the timestamps of the recorded ship positions. The variables are:

- Δt_i which is the time interval between two consecutive timestamps at the i-th position of the ship, helping to identify periods of signal absence.
- Δs_i which is the distance between consecutive ship positions at timestamps t_i and t_{i+1} (calculated as the geodetic distance between two points on the Earth's surface, using the Haversine formula).
- $\bar{v}_i = \frac{\Delta s_i}{\Delta t_i}$ which is the average speed of the ship during Δt_i .

Accordingly, the identification of outliers is based on the following condition:

- The average speed $\bar{v}_i$ is greater than a maximum threshold: $\bar{v}_i > v_{max}$
 where v_{max} is selected on the premise that the maximum speed attainable by a ship cannot exceed a certain threshold. In our case, we set $v_{max} = 50 \text{ knots}$.

This approach reduces the problem to a univariate analysis. However, it is not possible to apply this procedure indiscriminately as outliers do not always appear individually but rather in complex groupings. In the case of a group of neighbouring outliers (see example in Figure 3), the 'average speed' allows to detect only the first outlier. In fact we observe only one jump; the subsequent outliers of the group remain unidentified due to the lack of a jump from the previous position in the raw sequence.

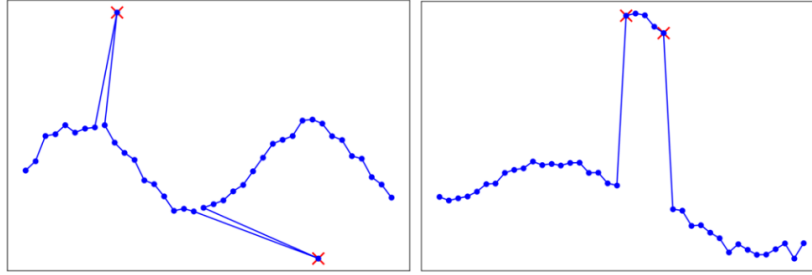


Figure 3. Left: isolated outliers in route, Right: contiguous outliers, they are only detected at the extremes (red x's) where they cause a jump in the calculated speed. These are fictitious routes to illustrate the concept

We implemented a solution involving the cyclical application of the algorithm. In the first cycle only the first outlier was removed, while in the subsequent cycle the second outlier was eliminated, and so on. Finally, the entire group was removed, as shown in Figure 4.

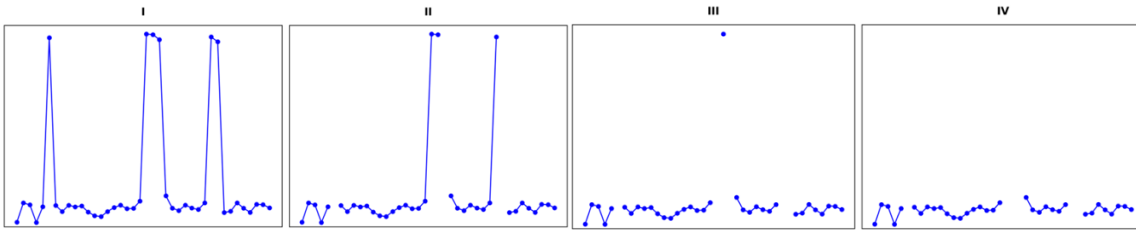


Figure 4. Illustrative diagram of the algorithm in its successive stages. The outliers are removed starting from the outermost points and moving inward. This is a fictitious route to illustrate the concept.

3.1. Key points of the Algorithm

A crucial decision that must be made by the algorithm concerns which data point to eliminate in the event of a jump.

Removing only the second point of the jump is not always appropriate. In the case of a group of outliers, this approach may be effective at the beginning of the group but not at the end (when the jump starts from an anomalous point and take to the correct trajectory). Furthermore, in the most extreme case where the initial point in the series is an outlier, the entire series would be removed. Similarly also the removal of the first point of the jump is not a consistently effective solution.

The optimal solution to this problem is to remove both points at each cycle: the point where the jump occurs and the point immediately after.

Figure 5 shows the points that the algorithm remove in the three cases. The images on the left, centre and right illustrate the effects of removing the first, second and both points, respectively. The issue is symmetrical, and all choices are possible. In all cases, for n outliers in the series, roughly $2n$ points will be removed: all n real outliers and n false outliers.

Additionally, these false outliers will be data points contiguous to the outliers, and therefore of more questionable reliability. However, the third approach brings some advantages: firstly, the performance of the algorithm are better because the number of cycles are halved (in the example 2 cycles instead of 4); the false positives removed are no longer asymmetric; finally, also in the extreme cases (when first or last point of the series are outlier) the algorithm works well.

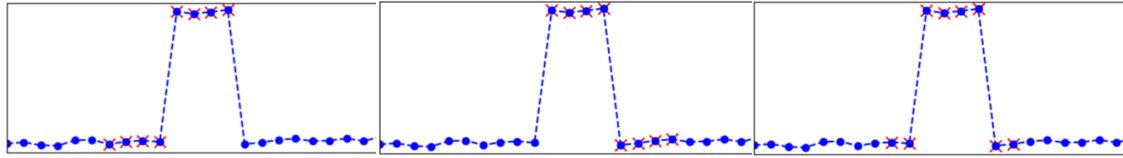


Figure 5. Approaches to outlier removal. Left: points after the jump are removed. Middle: points at the jump are removed. Right (chosen algorithm), both are removed. This is a fictitious route to illustrate the concept.

4. RESULTS AND CONCLUSIONS

We applied the algorithm to real data routes and successfully removed all outliers. Furthermore, we discovered that the number of false positives was significantly lower than that of the outliers in many instances. This was because the algorithm identified additional jumps within groups of outliers, triggering action both outside and inside the groups. This reduced the number of false positives. Figure 6 demonstrates the algorithm's effectiveness.

The method is promising and clear in its application. In the future, we will integrate the average-speed-based algorithm with machine learning models to improve outlier detection and reduce false positives.

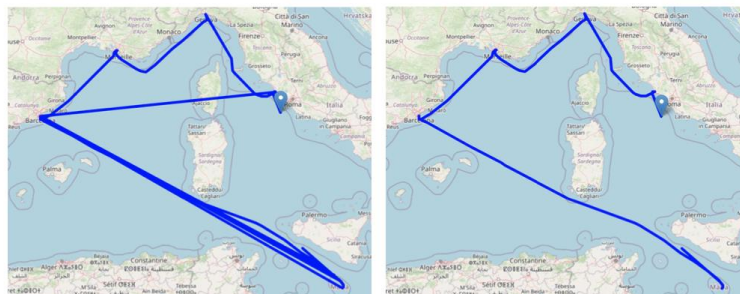


Figure 6. Left: a real route showing anomalies. Right: the same route after applying the anomaly removing algorithm.

5. REFERENCES

- [1] Port Visits Using Real-Time Shipping Data – CSO (2022) <https://www.cso.ie/en/releasesandpublications/fp/fp-pvrts/portvisitsusingreal-timeshippingdata/datasourcethodsandquality/>.
- [2] AISDAG: Daily number of vessel (2019) - <https://www.statistikbanken.dk/aisdag>
- [3] A. Pappagallo, N. Salamone, F. Sisti, M. Bruno and L. Valentino 'AIS-Driven Maritime Insights: Improving Italian Port Traffic Analysis' European conference on Quality in Official Statistics (2024).
- [4] K. Wolsing, L. Roepert, J. Bauer, and K. Wehrle. Anomaly Detection in Maritime AIS Tracks: A Review of Recent Approaches. J. Mar. Sci. Eng., 2022, 10, 112.

DATA CATALOGUE - a modern and harmonised approach to standardised data management in the NSI

Antoaneta Ilkova (National Statistical Institute, Bulgaria)

1. INTRODUCTION

The development of the Information System 'Data Catalogue' lays the foundations for the implementation of the 'Single Point of Truth' principle, introducing a modern and harmonised approach to standardised data management in the NSI. In a content and functional context, the information system is implemented as a catalogue containing information on all available structural and reference metadata in the NSI.

2. METHODOLOGY

The system is implemented through two separate modules:

- External (public) - available to the general public from anywhere in the world with internet access.
- Internal to the organisation - available only to NSI staff for the purpose of managing and administering the metadata in the catalogue.

The development of the Data Catalogue has been carried out in accordance with the National, European, and International legal framework, taking into account future legal changes in the field of storage and management of public data, including at European level (Data Governance Act, Data Act).

3. RESULTS AND PRACTICAL APPLICATION

The development of the ISDC provides the following opportunities:

- **For Users**
 - Organization of the data in the presentation part of the catalogue by thematic areas;
 - Ability to search effectively;
 - Information on timeliness, coverage and data sources;
 - Ability to export the catalogue in standardized digital formats (pdf, csv, xml).
- **For Respondents**
 - Reducing the number of re-collections of NSP survey data;
 - Reducing the administrative burden.
- **For Processes in the NSI**
 - Organization of data management and control processes;
 - Provision of a universal and standardised metadata storage and management tool;
 - Control over the organisation of data access;
 - Ability to validate and manage metadata;

- Unification of protocols and technical tools for data exchange with other public institutions and private holders.

4. MAIN FINDINGS

The Data Catalogue provides:

- a universal, standard approach for systematizing NSI's data and metadata
- strict control over access to all data stored in the system (Data Governance) and as well as ensures management and traceability of data access rights.
- opportunities for integration with new information systems within NSI
- termination of operation of the information systems containing not up-to-date data and/or data stored in other systems.

Data Dialogue: Perspectives and guidelines on the use of Mobile Phone Data for Statistics, Development and Humanitarian Initiatives

Sophie Delaporte (Flowminder)

1. INTRODUCTION

Official statistics are essential in the information systems of societies, and the digital revolution has enabled the abundance of new data types, generated as by-products of digital services held by private sector actors. The attainment of the Sustainable Development Goals (SDGs) strongly relies on governments' ability and capacity to provide services that meet the needs of its people. In order to achieve this, policies should be formulated based on timely and relevant statistics. To this end, the statistical community has been exploring the use of new data sources and tools to inform policies and programmes of governments and private sector.

2. METHODOLOGY

Among these new data sources, mobile phone data (MPD) have generated a lot of interest over the last few years. With mobile phones nearly ubiquitous worldwide, and widespread even in low-income countries, there is an expectation that data generated by users of mobile phones and maintained by mobile network operators (MNO) can help address data gaps, improve timeliness and provide disaggregated data that are needed for targeted policy-making needs.

The Mobile Phone Data (MPD) Task Team of the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD), through its member organisations, facilitates the mobilisation of this exciting form of big data by developing and disseminating methodologies and best practices for the safe, secure, consistent and responsible use of MPD, particularly by national statistical offices.

To achieve this, the UN-CEBD MPD task team has developed [six methodological guides](#), each focusing on a specific area where MPD can be effectively utilised:

- **[Displacement and Disaster Statistics](#)**: Leveraging MPD to estimate population movements before, during and after natural disasters or conflicts, and providing near real-time data for emergency response and relief efforts.
- **[Dynamic Population Mapping](#)**: Utilising MPD to create detailed and up-to-date population density maps, aiding in urban planning and resource allocation.
- **[Measuring the Information Society](#)**: Using MPD to assess access to and usage of information and communication technologies, essential for bridging the digital divide (and supplementing the data for SDG indicator 17.8.1 on Internet use).
- **[Migration Statistics](#)**: Analysing MPD to understand internal migration patterns and the scale of the IDP crisis within a country, informing migration policies and programmes.
- **[Tourism Statistics](#)**: Employing MPD to monitor tourist flows, supporting the development and promotion of the tourism industry.

- **Transport Statistics:** Applying MPD to study transportation patterns and infrastructure usage, enhancing transport planning and policy-making.

Additionally, an awareness-raising course has been launched to educate stakeholders on the potential and methodologies of MPD.

3. RESULTS AND PRACTICAL APPLICATION

In this session, the UN-CEBD MPD task team (which brings together 20 countries, 9 international and regional agencies, and 6 non-profit or private sector organisations) will present its MPD guides and discuss the value and applications of MPD in different areas of official statistics, humanitarian and development purposes, drawing on past and current initiatives in this field, alongside the importance of data literacy capacity building, communications and user engagement.

Several practical applications of MPD have been identified:

- **Health:** MPD has been utilised to monitor the spread of infectious diseases, plan healthcare services, and manage healthcare resources. For example, during the COVID-19 pandemic, MPD helped track movement patterns to understand the spread of the virus and the effect of mobility restrictions / quarantine measures.
- **Disaster Management:** In disaster-prone areas, MPD provides real-time data on population movements, enabling quicker and more efficient emergency responses. It helps identify affected areas and allocate resources accordingly.
- **Information Society:** MPD assists in measuring access to digital services and the internet, helping governments design policies to improve digital inclusion and bridge the digital divide.
- **Tourism:** By tracking the movements of tourists, MPD offers insights into tourist behaviours and preferences, aiding in the development of targeted marketing strategies and infrastructure improvements.
- **Transport:** MPD is used to analyse traffic patterns and public transport usage, informing decisions on infrastructure development and traffic management.

These applications demonstrate the versatility and value of MPD in addressing various statistical and policy-making needs.

4. MAIN FINDINGS

The discussion will bring perspectives from various actors and experts in the field, and integrate audience participation to surface the most important standards and learnings when dealing with these commercially and privacy sensitive datasets owned by the private sector. The audience will get to hear from representatives from the UN, government, including NSOs, technical service providers, and academia.

The main findings from the task team's work with MPD can be summarised as follows:

- **Enhanced data availability:** MPD provides a rich source of near-real-time data that can fill gaps in traditional statistical systems, offering more comprehensive insights if processed and analysed safely and accurately.
- **Timeliness and granularity:** MPD offers timely and detailed data that is crucial for targeted policy-making and effective response in various sectors. The high spatial and temporal resolutions can be particularly useful for addressing specific needs.
- **Methodological rigour:** The development of standardised methodologies ensures the consistent and responsible use of MPD, addressing concerns around privacy and

security. The methodological guides provide clear frameworks for collecting, processing, and analysing MPD. Additionally, partners and experts in this field are continuously investing into research and development to develop and apply the most rigorous and appropriate methods for data processing and analysis.

- **Cross-sectoral applications:** MPD has proven valuable across multiple sectors, including health, disaster management, tourism, transport, and SDG monitoring. Its versatility makes it a powerful tool for diverse applications.
- **Collaboration and integration:** Successful use of MPD relies on collaboration between governments, private sector actors, and technical service providers, fostering an integrated approach to data utilisation. The involvement of multiple stakeholders ensures that the data is used effectively and ethically.
- **Governance and Data Privacy:** Establishing robust governance frameworks is critical to ensuring the responsible use of MPD. This includes setting clear policies and regulations around data access, usage, and sharing, as well as implementing measures to protect individual privacy and maintain data security. Governance structures must be transparent and inclusive, involving all relevant stakeholders to build trust and accountability from regulators to data protection agencies and more.

In conclusion, the integration of mobile phone data into official statistics represents a significant advancement in the ability to meet the evolving needs of societies. By adopting and refining the methodologies developed by the UN-CEBD MPD Task Team, national statistical offices can enhance their capacity to provide timely, accurate, and relevant data for policy-making and program development.

5. REFERENCES

- [1] Guiding principles to maintain public trust in the use of mobile operator data for policy purposes, <https://www.cambridge.org/core/journals/data-and-policy/article/guiding-principles-to-maintain-public-trust-in-the-use-of-mobile-operator-data-for-policy-purposes/A4A2A8723151F901F5315279E76F651F#>
- [2] UN-CEBD Task Team on Mobile Phone data, <https://unstats.un.org/bigdata/task-teams/mobile-phone/index.cshtml>

Data Science in the Production of Official Statistics: Utilizing Reproducible Analytical Pipelines (RAPs) for Custom Declarations Processing at PCBS as a Case Study

Asala Khalid (Palestinian Central Bureau of Statistics, Palestine)

1. INTRODUCTION

Data science is playing an increasingly significant role in the production of official statistics by modernizing traditional methods and incorporating new data sources and technologies. It enhances the timeliness, frequency, and granularity of statistics and indicators. Since 2019, the Palestinian Central Bureau of Statistics (PCBS) has been actively working to introduce data science in Palestine through the "Data Matters" initiative. The mentoring agreement signed in 2023 between PCBS, the UK's Office for National Statistics (ONS), and the Data Science Campus marks a pivotal step toward integrating data science techniques into the production of official statistics.

The mentoring agreement aimed to automate the data processing phase of Palestinian external trade, starting with customs declarations as a major source of trade statistics with over 500,000 rows and numerous variables, processing this data typically takes five days. Previously, the data was managed manually using Excel and Access databases to generate new variables for annual export and import statistics.

The development of a Reproducible Analytical Pipeline (RAP) using R was key to automating the processing of 2021 customs declarations. This process included managing code updates via GitHub, creating user-defined functions, and automating data validation. By comparing the automated RAP method with the previous manual process, the staff became able to assess improvements in time reduction, process efficiency, and data quality.

This paper offers two primary contributions. First, it presents the integration of data science techniques in the production of official statistics. Second, it underscores the benefits of using RAP over manual processing, particularly in terms of efficiency and data quality.

2. METHODOLOGY

This paper was written based on the actual experience of building and applying RAP to custom declarations for the year 2021, comparing the results, quality, time consumed, and effort with those of the manual method, along with reapplying it to data for 2022 and 2023 after modifications and improvements to suit the nature of the data.

Additionally, international data science projects, including the Office for National Statistics (ONS) mentoring initiatives, were reviewed in terms of necessary tools and outcomes.

3. RESULTS AND PRACTICAL APPLICATION

The implementation of a Reproducible Analytical Pipeline (RAP) has significantly improved process efficiency and contributed to capacity building within the team. By automating workflows and integrating modern tools, RAP has delivered noticeable improvements in both operational speed and technical skill development.

3.1. Process efficiency

Once the automation process was implemented, its efficiency became immediately evident, delivering significant benefits:

- **Time saving:** Using the manual method with Access and Excel, custom declarations processing requires five working days. The new method needs 25 to 30 minutes, with the added benefit of being adaptable for additional validation rules, offering greater data reliability.
- **Reduction in repetitive tasks:** Creating functions within the RAP significantly reduced the effort needed to carry out recurring tasks.
- **Comprehensive process tracking:** The RAP serves as a form of documentation, with each step meticulously defined and converted into R code. It is also well-organized, incorporating validity checks, data cleaning, and storing validation outcomes in easily accessible tables for review.
- **Improving quality:** Pipelines are easier to validate than manual processes, resulting in improved quality.

3.2. Capacity building

Building the RAP promoted the use of R for automating data workflows, allowed the staff to enhance technical skills in programming. Through GitHub integration, RAP fosters collaboration, version control, and code-sharing best practices.

4. MAIN FINDINGS

Key findings reveal that automation reduced processing time to 25-30 minutes, significantly minimized repetitive tasks, and improved overall process quality through comprehensive tracking and validation using R and GitHub. The RAP not only streamlined the workflow but also enhanced data reliability by incorporating validation checks and data cleaning in a well-organized manner. These findings highlight the potential for RAP to transform the production of official statistics, offering a replicable, efficient, and high-quality approach compared to traditional methods.

5. REFERENCES

- [1] S. Schweinfest and R. Jansen, Data Science and Official Statistics: Toward a New Data Culture, Harvard Data Science Review • Issue 3.4, 2021.
- [2] Reproducible Analytical Pipelines (RAP)
<https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/>
- [3] K.Gask, Data science for official statistics, Sixth Statistical Commission for Africa, 2018, pages 4-5.
- [4] Case study: working with the Vanuatu National Statistics Office (VNSO)
<https://datasciencecampus.ons.gov.uk/making-an-international-impact-with-data-science-mentoring/#Vanuatu>

Detecting Change Points of Expectations and Judgments of Price Dynamics within the Italian Population

Rosaria Simone (Department of Political Science, University of Naples Federico II – Italy), Carmela Cappelli (Department of Humanities, University of Naples Federico II – Italy) & Francesca Di Iorio (Department of Political Science, University of Naples Federico II – Italy)

1. INTRODUCTION

For the analysis of ordered data, CUB models [1,2] entail the estimation of two main latent components of the rating process: feeling and uncertainty, parameterized with a two-component mixture of Binomial and uniform distributions.

The contribution aims at presenting how change point detection of temporal series of estimated feeling and uncertainty parameters can be pursued to identify if and to what extent Italian people modified their perception and judgments of price dynamics from 1994 to 2019 (see also [3]). To this goal, Atheoretical Regression Trees (ART, [4,5,6]) are applied considering the series of monthly response distributions to the following questions, issued by the Italian National Statistical Institute (ISTAT) within the consumers' confidence survey¹:

1-(Judgments): How do you think the price level changed over the previous 12 months?

2-(Expectations): How do you think the price level will change over the next 12 months?

Responses are collected over a scale with $m=5$ categories (1 = “fall”, 2 = “stay about the same”, 3 = “rise slightly”, 4 = “rise moderately”, 5 = “rise a lot”). Results indicate that ART are effective in partitioning the series into sub-intervals characterized by different levels of the estimated model parameters, allowing to study and compare over time the change points of both feeling and uncertainty.

2. METHODOLOGY

This paragraph is devoted to a brief methodological overview of mixture models with uncertainty and Atheoretical Regression Trees.

2.1. Mixture models with uncertainty

For a response R collected over m ordered categories, the class of mixture models with uncertainty [1,2] considers a discrete mixture of a shifted Binomial distribution with feeling parameter ξ in $[0,1]$ and a discrete distribution for uncertainty u_r :

$$Pr(R = r) = \pi b_r(\xi) + (1 - \pi) u_r, \quad r = 1, \dots, m.$$

¹ <https://ec.europa.eu/info/business-economy-euro/indicators-statistics/economic-databases/business-and-consumer-surveys>.

A relevant feature of the model is that the distribution u_r can be specified in different ways, in particular in such a way to account for inflated frequencies at some ‘shelter’ category (by choosing a degenerate distribution) and for heterogeneity of the distribution (in which case, u_r is the discrete uniform model and the model is called CUB). Given the possible specifications of uncertainty, it is of interest to focus on the mixing weight π of the feeling component, indicating the relative importance of the latent sentiment with respect to the discrete noise modeled with the uncertainty component. These parameters can be estimated with profile likelihood methods as pursued in [3], or by referring to the best fitting model within this class, as proposed hereafter: parameters can be promptly estimated with the R package ‘CUB’ available on CRAN. Note that for the data under examine, given the orientation of the scale, the lower the value of the feeling parameter ξ the higher the expectation- judgment that prices will increase.

2.2. ART: ATHEORETICAL REGRESSION TREES

ART is a procedure originally proposed in [6] and applied in various contexts [4,5] that exploits Least Square Regression Trees (LSRT) to locate multiple changes in a time series occurring at unknown dates. In LSRT, a node h is split into its descendants h_r and h_l to reduce heterogeneity of the target numerical variable Y ; the best split s of node h is the one maximising the difference in sum of squares:

$$\Delta SS(h|s) = SS(h) - [SS(h_l) + SS(h_r)], \quad SS(h) = \sum_{i=1}^{n(h)} (y_i - \hat{\mu}(h))^2 .$$

Since the child nodes form an exhaustive binary partition of the father, this splitting criterion searches for the split that generates the child nodes that are as far as possible in terms of squared difference between their means.

This goal is achieved by tree-regressing the time series Y_t on the time itself, represented by an artificial covariate t that is an arbitrary sequence of strictly increasing numbers. Then, in the binary tree generated by ART, the split points are interpreted as change points in the time series while the terminal nodes are homogeneous subintervals.

The procure mimics Bai and Perron’s method [7], and it is extremely efficient in terms of computation as it is based on a recursive algorithm. Moreover, the resulting hierarchical structure provides information on the order of importance of the detected changes.

3. RESULTS AND PRACTICAL APPLICATION

For illustrative purposes, we focus on feeling and uncertainty of expectation of price levels. Figure 1 displays the feeling series of expectation and judgments of price levels: red vertical lines highlight the identified change points that is possible to interpret in light of political events.

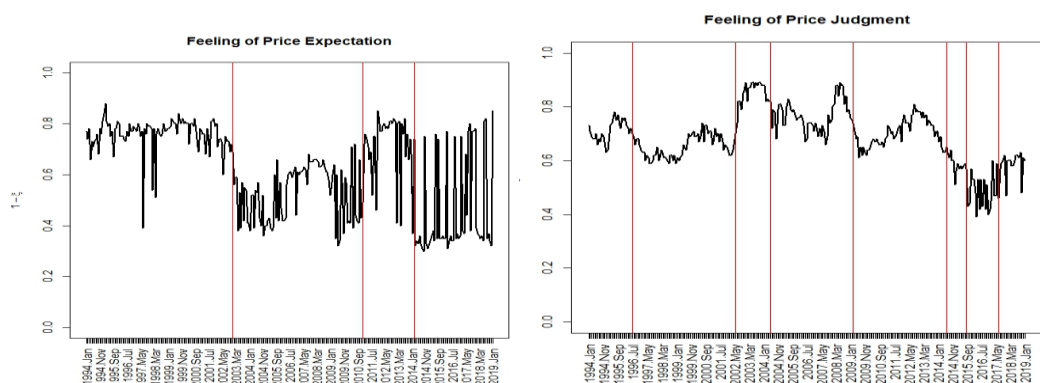


Figure 1. Estimated feeling of expectation (left) and judgment (right) of price levels

For instance, for price expectation, the first period covers the implementation of policies to reduce inflation in order to fulfill the goals of the Maastricht Treaty. The second period covers the introduction of the EURO, the 2008 worldwide financial crisis originated in the United States, as well as the outbreak of the Italian spread crisis. The third period can be read as a reaction to the news that current government was about to resign and replaced by a technocratic government composed entirely of unelected professionals that was perceived as a solution to the sovereign debt crisis. The last period is characterized by a persisting instability resulting in people fear about future price level. The series of judgments, is much more smooth yet it is concerned by some more break points than the series of expectations.

Uncertainty parameters of both expectations and judgments, instead, undergo several more break points: for the sake of illustration, Table 1 reports the average level of uncertainty weight of expectation of prices for each subperiod identified by ART, whereas the classical tree visualization is provided in Figure 2 for uncertainty of judgments.

Table 1. Break points identified by ART on the series of uncertainty weights of price expectations

Terminal node	From	TO (<u>Break date</u>)	Average π
Nodo 13	Jen -1994	Mar 1995	0.79
Nodo 49	Apr 1995	Mar 1996	0.70
Nodo 48	Apr 1996	July 1999	0.51
Nodo 51	Aug 1999	July 2002	0.73
Nodo 50	Aug 2002	Apr 2007	0.57
Nodo 7	May 2007	<u>Oct 2008</u>	0.81
Nodo 4	Nov 2008	Oct 2009	0.39
Nodo 22	Nov 2009	Feb 2011	0.51
Nodo 23	Marc 2011	Nov 2012	0.60
Nodo 20	Dec 2012	Mar 2014	0.41
Nodo 43	Apr 2014	June 2015	0.58
Nodo 42	July 2015	Jen 2019	0.44

It is worth noticing that the detected change points overall correspond to those identified using Bai and Perron's method.

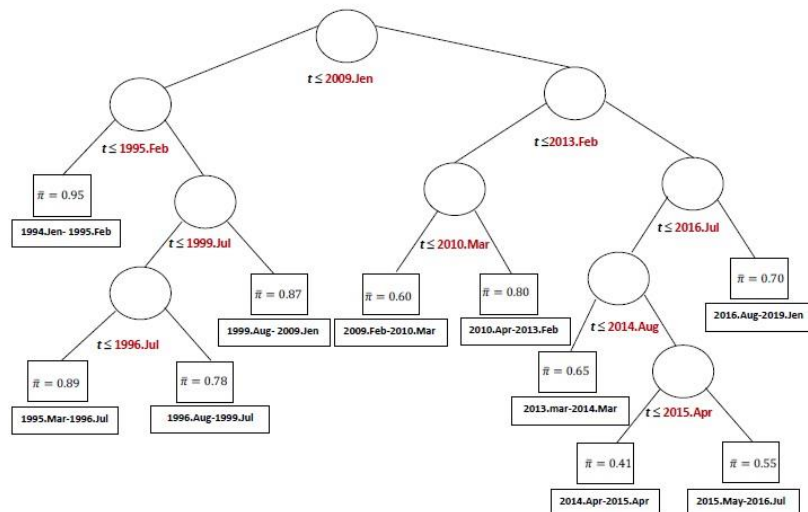


Figure 2. ART for the series of uncertainty weights related to price judgements

4. MAIN FINDINGS

For the analysis of price judgment and perception, CUB model parameters refer to two different aspects of the respondents' perception and judgment of price level, thus the study of their change points may reveal that they show different number and location of break dates providing an innovative twofold analysis of respondents' answers. The chosen framework is particularly useful for official survey data, whose results are usually released at aggregated level only. In these cases, classical methods and models cannot be applied due to the absence of individuals' covariates. The class of mixture models with uncertainty [5,6], instead, provides a useful and flexible framework to summarize the main response features and possibly their evolution over time when combining it with the setting of Atheoretical Regression Trees.

5. REFERENCES

- [1] Piccolo D. On the Moments of a Mixture of Uniform and Shifted Binomial random variables. QUADERNI DI STATISTICA (2003), 5: 85-104.
- [2] Piccolo D., Simone R. The class of CUB models: statistical foundations, inferential issues and empirical evidence. STATISTICAL METHOD AND APPLICATIONS (2019), 28:389-435.
- [3] Simone R., Corduas M., Piccolo D. Dynamic modelling of price expectations and judgments. METRON (2023), 81:323-342.
- [4] [4] Cappelli C., Maddaloni A., Di Iorio F., D'Urso P. Atheoretical Regression Trees for classifying risky financial Institution, ANNALS OF OPERATION RESEARCH (2020), doi: 10.1007/S10479-019-03406-9
- [5] Cappelli C., Penny R. N., Rea W. S., Reale M. Detecting multiple mean breaks at unknown points in official time series, "Mathematics and Computers in Simulation" (2008), Volume 78, Issues 2-3, Pages 351-356, ISSN 0378-4754.
- [6] Cappelli C., Reale M. Detecting changes in mean levels with atheoretical regression trees, in S.Co. 2005: Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione, pp. 479-484. Cleup, Padova.
- [7] [7] Bai, J., Perron, P. Computation and analysis of multiple structural change models. JOURNAL OF APPLIED ECONOMETRICS (2003), 18:1-22.

Detecting and Correcting Erroneous OCR and Manual Diary Registrations in a Modernised Household Budget Survey

Marius Runningen Larsson (Statistics Norway)

1. INTRODUCTION

The household budget survey (HBS) measures expenditures on goods and services. Respondents participate by registering their purchases in a diary over a sample period. The HBS diary has traditionally been recorded using pen and paper. Several NSIs have adopted smart survey tools to conduct the HBS. When using smart survey tools to record the HBS diary, respondents can typically choose between two different methods for registering their purchases: Optical character recognition (OCR) or manual registration. Adopting new methods for data collection can introduce new and unforeseen errors in the data. We present methods to identify erroneous diary registrations and how the registrations can be corrected. The results are summarized using quality indicators.

1.1. Data

The data consists of OCR and manual diary registrations collected during the Norwegian HBS 2022, conducted by Statistics Norway. Each respondent was assigned a sample period of 1 week. Approximately 3500 households participated in the survey recording a total of 44 000 diary registrations. Of these, 24 000 were OCR registrations and 20 000 were manual registrations.

The respondents were not given any direct feedback on the quality of their data input during or after the registration of the purchases. This was a deliberate choice by the NSI to reduce response burden. However, the respondents could inspect the results of the OCR or manual registration. If they discovered errors, they could re-take the photo or re-enter the manual entry. They could also submit the erroneous registration without further action.

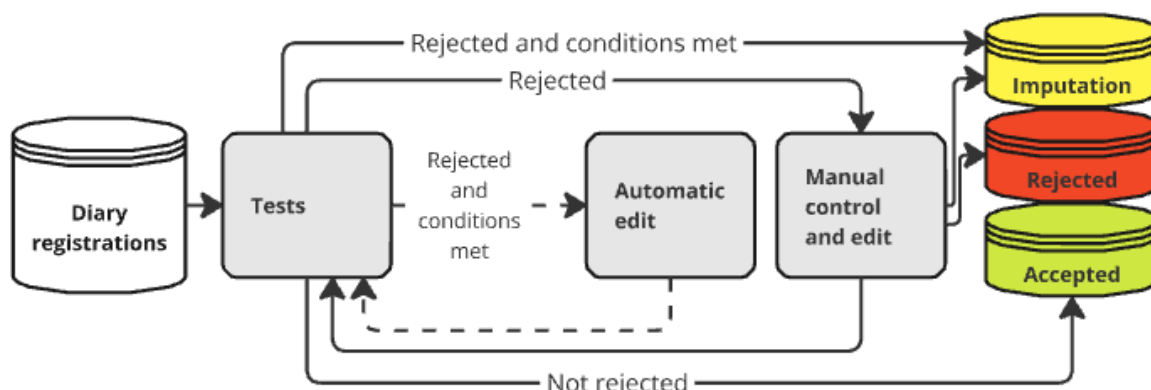
The OCR system is developed and maintained by a third-party provider. The system is satisfactory for interpreting and recognizing the basic layout and content of the receipt (e.g. the store name, total sum, item name, prices, quantity etc.), but it often fails to interpret more complex receipt logic like discounts and promotions.

2. METHODOLOGY

The diary registration quality control pipeline can be seen in Figure 1. The process is divided into three main steps: (1) We use multiple tests to identify erroneous diary registrations. The tests are set up to detect errors and deviations from what is expected. If no errors are found, or the deviations are within a threshold, the quality is considered acceptable, and the diary registration is not rejected. If one or more of the tests fail, the registration is either rejected or sent to imputation. (2) If a registration is rejected and certain conditions are met, the receipt is sent to be automatically edited. After attempting automatic edit, the edited registration is re-tested. (3) If a registration is rejected without the automatic edit conditions being met, the original registration is sent to manual control and edit.

The manual control and edit procedure is the final step of the quality control. However, the registration is tested one final time to ensure that the manual edit has been conducted correctly. The manual procedure is repeated until all tests passes.

Figure 1. Diary registration quality control pipeline



Each diary registration will be given one out of three main statuses after the quality control pipeline: Accepted, rejected or imputation.

2.1. Tests

The tests are defined to identify erroneous diary registrations. We make the following assumptions regarding the registrations: First, the registrations are assumed to be correctly registered and of acceptable quality. Second, it is assumed that the price recorded by the OCR is the price per item or per kilo/liter (unit price) and that discounts have already been subtracted. Third, we can obtain the value of the line item (what the respondent actually paid) by multiplying price by quantity. The errors can be separated into three main groups.

The first group (E1) contains errors pertaining to the OCR failing to interpret the text and layout of the receipt correctly. This can generate erroneous item prices, quantities and names. The main reasons for these errors are either because of low quality/dark images or curly/very long receipts. This kind of error can also lead to line shifts, meaning that some or all prices are either shifted up or down relative to their respective item names. The error group also includes registrations where the image submitted by the respondent to the OCR is either cut in half (e.g. only the bottom part is included omitting the top items) or multiple receipts are present in a single image.

The second group (E2) of errors pertains to manually entered purchases with insufficient information. A registration in this group will typically contain a single item with a generic item name, e.g. <food>, <groceries>, <miscellaneous>, or lists of goods and services, e.g. <bread, milk, shampoo, batteries, TV, restaurant>. Both cases must be sent to imputation.

The third group (E3) pertains to errors associated with the OCR failing to correctly interpret the receipt logic, despite having interpreted the rest of the information correctly. This is for the most part errors occurring if the receipt contains discounts and/or promotions. Some receipts could also display refunds and recycling amounts in a complicated way. This error group also includes cases where the price recorded by the OCR is the total value of the item (i.e. not the unit price), or price and quantity have swapped places.

Table 1 lists the tests conducted, which error group it is intended to identify, and which stage it is sent to in the quality control pipeline. Some tests might seem to detect only minor errors. However, small and seemingly insignificant errors might be symptoms of larger errors. For example, if the price of a single item in a registration is 0, it might indicate that all prices have

been shifted one level down or up and consequently all items have been given the wrong price.

Table 1. Tests to detect erroneous diary registrations

Test	Error group	Sent to
Sum of prices*quantities deviates more than 2% from the total sum of the receipt	E1, E3	Manual control
Sum of price*quantity is exactly twice that of the total sum	E3	Automatic edit
At least one price is negative	E1, E3	Manual control
At least one price is zero	E1, E3	Manual control
At least one item name includes a word associated with discount/refund	E3	Manual control
Purchase contains only one (multiple) item, is manually registered (OCR or manual), and the item name includes at least two decimals	E2	Imputation (manual control)
Purchase contains only one (multiple) item, and the item name equals (includes) a generic word	E2	Imputation (manual control)
Less than three characters item name after removing signs and numbers	E1, E2, E3	Manual control
Line item quantity is at least 5 times larger than price	E3	Automatic edit

Note: that we do not test for price or value outliers during this stage. This is done after assigning items to their respective COICOP group.

3. RESULTS AND PRACTICAL APPLICATION

The results are summarised using quality indicators [1] and can be found in table 2. The manual control share is the share of how many of the diary registrations that were sent to manual control. 77% of registrations sent to manual control needed some form of manual edit.

The automatic edit share is the share of how many of the diary registrations that were automatically edited. Imputation share includes both fully imputed registrations (8.6%) and partial imputations (0.6%).

The average time it took to manually control and edit includes all registrations sent to manual control, also those not needing editing. The total time usage spent on control and editing is the total amount of hours used by personnel at the NSI. The average time usage indicator is a conservative estimate. To estimate the automatic edit time saving, we multiply average time usage by number of registrations which were automatically edited.

Table 2. Quality indicators

Description	QI
Manual control share	26%
Manual edit share	20%
Automatic edit share	3%
Imputation share	9.2%
Manual control and edit average time usage per registration	62 seconds
Manual control and edit total time usage	198 hours
Automatic edit estimated time saving	22 hours

The results from the quality indicators are used by the NSI as a framework for prioritising which part of the HBS data collection needs improvements. It also allows the NSI to monitor changes over time.

We believe that by providing direct feedback to the respondents based on automatic testing at the time of registration, we can significantly reduce errors from groups E1 and E2. This does not necessarily present a trade-off between response burden and the resource burden of the NSI as this is simply intended to nudge the respondent to register the purchases according to the instructions.

Handling E3 errors does require a trade-off between response burden and the resource burden of the NSI. If the OCR system is not configured to handle discounts and complex receipt logic, either the respondent or the NSI must edit the registered purchase after the registration. While it might be tempting to give this task to the respondent, the HBS already imposes a large response burden. In addition, forcing respondents to manually edit might incur errors in the data which might be difficult to detect. Improving the OCR to better handle receipt logic and/or developing more sophisticated automatic editing might be better actions to take.

4. MAIN FINDINGS

We show that using smart survey tools to conduct the HBS can generate new types of errors in the data. Approximately 200 hours (conservative estimate) were used by the NSI to manually control and edit erroneous diary registrations. In addition, more than 9% of all registrations were imputed due to errors. This can contribute to obscure important relationships and incorrectly reduce variance at the micro level. We believe that simple and basic measures can be implemented to reduce some of the errors generated during the diary registration process. This will increase data quality and reduce the resource burden of the NSI.

5. REFERENCES

- [1] Foss, Aslaug Hurlen, and Per Ola Haugen. "Anbefalte kvalitetsindikatorer i offisiell statistikk." (2024).

Detection and correction of quarterly effects in administrative versus survey turnover data

Jetske Marcelis & Arnout van Delden, Sander Scholtus (Statistics Netherlands)

1. INTRODUCTION

From 2012-2014 Statistics Netherlands (CBS) redesigned their monthly short term statistics based on panel sample survey data (SSD). The output is in the form of monthly indices that express the growth rate of month t compared to a base period 0. The estimator is designed to accurately estimate monthly growth rates, but in the long run some bias might occur. To avoid such a bias, CBS intended to calibrate the monthly SSD indices to a series of quarterly indices. These quarterly indices are based on a 'census combined data' (CCD) system that uses the monthly survey data from the large and complex units and value added tax (VAT) data for all other units in the population.

In 2015, CBS calibrated the monthly SSD series to the quarterly CCD series of that year using a Denton method, but that did not yield good results (see Van Delden and Scholtus [1]). The reason was that the quarterly turnover levels based on the CCD data were smaller in the first quarter and larger in the fourth quarter than the quarterly levels derived from the SSD, see Figure 1 for an example. The impression of our analysts was that the yearly level of turnover from VAT is more reliable than that based on the sample survey data, but the relative distribution of VAT turnover over the four quarters within a year is less reliable, due to administrative effects in tax declarations.

The current paper aims to contribute to a better understanding of these quarterly differences between VAT and survey data. It follows up on Marcelis et al. [2] by modelling quarterly reporting differences among different groups of enterprises within the population. The work could also be useful for other National Statistical Institutes that use administrative sources to replace survey data, especially when they use VAT for turnover estimates, see for instance Lewis and Woods [3].

2. METHODOLOGY

We use a mixture model for detecting groups of enterprises that differ in their reporting behaviour. The model has been described in Marcelis et al. [2], and tests on simulated data showed that it correctly identified different groups in the population. The additional steps that we have taken compared to Marcelis et al. [2] will be explained in section 3.

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ denote the observed vectors of four quarterly turnover values for a given year of enterprise i , in the survey and VAT data respectively. The total quarterly turnovers Y_k and X_k are estimated as weighted sums over the observed values of Y_{ik} and X_{ik} for the set of units that overlap (see section 3). The corresponding vectors with relative turnovers are given by $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})'$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$, with $\sum_{k=1}^4 y_{ik} = \sum_{k=1}^4 x_{ik} = 1$. We only include records for which $y_{ik} \geq 0$ and $x_{ik} \geq 0$, which holds for nearly all of the data. We use a Dirichlet distribution which provides a flexible way to model such data. To that end we first compute a transformed value for the differences between $\mathbf{y}_i$ and $\mathbf{x}_i$, using:

$$\mathbf{d}_i = \frac{1}{4} \iota_4 - \frac{\mathbf{y}_i - \mathbf{x}_i}{4}, \quad (1)$$

where $\iota_4 = (1,1,1,1)'$. The elements of the vector $\mathbf{d}_i$ satisfy $\sum_{k=1}^4 d_{ik} = 1$ with $0 \leq d_{ik} \leq 1/2$. In the absence of systematic differences between the distributions of $\mathbf{y}_i$ and $\mathbf{x}_i$, we would expect that $E(d_{ik}) = 1/4$ for all k .

Our model for $\mathbf{d}_i$ is a mixture of Dirichlet distributions with $g = 1, \dots, G$ groups:

$$f(\mathbf{d}_i) = \prod_{g=1}^G \{\alpha_g \cdot f_{Dir}(\mathbf{d}_i; \boldsymbol{\beta}_g)\}^{z_{gi}} = \prod_{g=1}^G \left\{ \alpha_g \cdot f_{Dir} \left(\mathbf{d}_i; \kappa_g \left(\frac{1}{4} \iota_4 + \boldsymbol{\delta}_g \right) \right) \right\}^{z_{gi}}. \quad (2)$$

where α_g denotes the relative size of group g and $z_{gi} \in \{0,1\}$ indicates the group that unit i belongs to. Each group has a Dirichlet distribution with hyperparameter $\boldsymbol{\beta}_g = \kappa_g (\iota_4/4 + \boldsymbol{\delta}_g)$, where the scalar parameter κ_g determines the amount of variance within group g , while $\boldsymbol{\delta}_g = (\delta_{g1}, \delta_{g2}, \delta_{g3}, \delta_{g4})'$ describe potential systematic quarterly differences between survey and VAT data. We use the natural restriction that $\sum_{k=1}^4 \delta_{gk} = 0$. The set of model parameters is denoted by $\boldsymbol{\theta}$ and for a given group it consists of α_g , κ_g , and $\boldsymbol{\delta}_g$. Besides these parameters also the group indicators z_{gi} are unknown. The model is estimated using an Expectation (Conditional) Maximisation (E(C)M) algorithm, which iterates over an expectation and maximisation step until convergence.

In Van Delden et al. [4] we modelled quarterly effects by using a mixture model that was based on the absolute turnover values $\mathbf{Y}_i$ and $\mathbf{X}_i$. The differences in absolute turnover values between VAT and survey data made it (more) difficult to appoint units to a quarterly turnover group, therefore we were interested to model relative turnover values $\mathbf{y}_i$ and $\mathbf{x}_i$.

3. APPLICATION OF THE MODEL

We applied the model to four economic sectors (Manufacturing, Construction, Retail Trade and Job Placement) and three years (2014–2016), which all had monthly short-term statistics output based on monthly survey data at the time. Enterprises that responded to the sample survey were linked to the VAT turnover values. We estimated the model on units that responded all periods of the year for both sources, and we excluded units with a thousand-error, units that belong to an industry within the sector for which VAT is not suitable, and enterprises that changed from composition (mergers, splits and so on).

Marcelis et al. [2] tested models with different number and sorts of groups. They found that a six group model was most appropriate: groups 1 -3 and 5 all without quarterly effect and with high (group 1), medium (group 3), small (group 2) and tiny (group 5) variance; groups 4 and 6 have a quarterly effect with medium (group 4) and high variance (group 6). Note that the model is defined such that we do not impose any directions on the quarterly effects: these are freely estimated by the model, but we did experiment with restrictions on those effects, see later. The main drawback of the study by Marcelis et al. [2] is that they used only a few sets of starting values for the ECM algorithm that were manually determined. Therefore, their study could not easily be generalised to new data.

In the current study we automatically ran a range of starting values and saved the ten best solutions. Based on the estimated model of each of these ten solutions, we corrected the observed values $\mathbf{X}_i$ for those units that were assigned to a group with a quarterly effect. For those ten we computed two performance measures. The first measure is the Akaike Information Criterion (AIC), which is based on likelihood values. The model with the highest likelihood fits best with the observed data. The second measure is the root mean square of the distance (RMSD) between the corrected $\mathbf{Y}_k/\mathbf{X}_k$ ratio per quarter and the mean $\mathbf{Y}_k/\mathbf{X}_k$ value

over the four quarters. The smaller the RMSD, the more the quarterly distribution of the corrected VAT data matches that of the survey data.

Unfortunately, we found (for all years and economic sectors) that among the models with the ten best solutions, the outcomes with the smallest RMSDs often did not have the highest likelihood. Solutions with the highest likelihood sometimes resulted in very small groups with a quarterly effect. We therefore re-estimated the model, but now we added the restriction that the two groups with a quarterly effect should have $\alpha_g \geq 0.10$. We also experimented with restrictions on the size of the quarterly effects (affecting δ_g). Imposing restrictions that affect δ_g yielded better RMSD values in some cases, but they had significantly larger AIC values. Moreover the RMSD values varied more than those with restrictions on α_g . We concluded that the restrictions on α_g are very useful to avoid very small groups with a quarterly effect, but we dropped the restrictions related to δ_g because then we would impose certain quarterly effects rather than estimating them.

4. RESULTS

Figure 1 gives the result of the six-group model for Job Placement in 2014-2016 with $\alpha_g \geq 0.10$ for the groups with a quarterly effect. For all groups together (blue lines) the uncorrected proportion of VAT turnover in the first quarter was smaller than that of the survey (higher Y_k/X_k ratio) and larger in the fourth quarter (smaller Y_k/X_k ratio). The corrected values (continuous line) were clearly more straight, implying that the distribution of quarterly turnover between VAT and survey are more alike. The quarterly effects were partly captured by the first group (group 4) with quarterly effect and partly by the second one (group 6), but the extent varied from year to year. We found that the patterns were found for the appointed group as a whole but individual enterprises could show a different pattern due to the variance that is allowed within a group. The RMSD values for Job Placement clearly decreased after correction (see Table 1), in the other economic sectors the effects of the corrections were smaller. Results for the three models were rather similar.

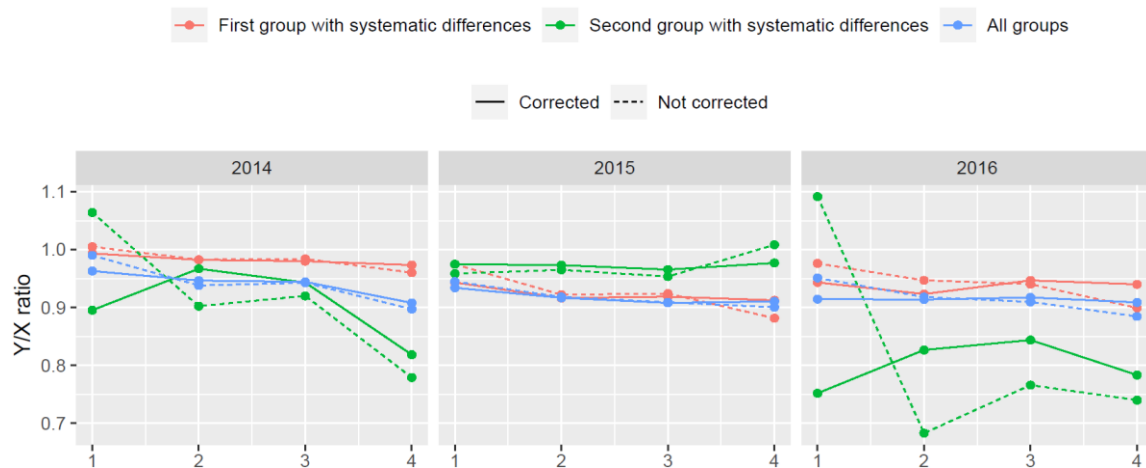


Figure 1. Uncorrected and corrected Y_k/X_k ratios for Job Placement (see text).

After estimating and selecting the best performing model (using the AIC), we corrected the original quarter-on-quarter growth rates of the CCD for the quarterly effects in the VAT data, which is also new compared to Marcelis et al. [2]. The original growth rates of the CCD were based on all data, including the large units with survey data, and including imputed values for non-response, but excluding industries for which VAT is not suitable. We corrected only the linked VAT-survey units, that comprised 10-13% of total turnover and within those only the units assigned to a group with a quarterly effect which comprised 3-7% of total turnover. The

effects of the correction were very small, see Table 2, one of the reasons was that only a small proportion of turnover is corrected. We tried to extrapolate the correction to all VAT units by modelling the probability to belong to a group with a quarterly effect as a function of the quarterly VAT turnover and background variables, but that model was not accurate enough.

Table 1. RMSD values for Job Placement for the absolute model {AM (2020), see [4]}, the relative model {RM (2022), see [2]} and the current model {RM (2024)}.

Year	Uncorrected	AM (2020)	RM (2022)	RM (2024)
2014	0.0329	0.0089	0.0134	0.0201
2015	0.0166	0.0062	0.0052	0.0105
2016	0.0237	0.0040	0.0107	0.0031

Table 2. Mean absolute difference between the quarterly growth rates of survey (SSD) and VAT (CCD) from 2014-2016 before and after correction.

	Construction	Retail	Manufacturing	Job Placement
Uncorrected	1.74	1.50	1.42	6.47
Corrected	1.69	1.49	1.41	5.93

5. DISCUSSION

We have developed mixture models that assume that the population is divided into groups that differ in their intra-annual quarterly patterns for VAT versus sample survey data. Although the results among the different model approaches were similar, we prefer to use the relative over the absolute model because the former is more parsimonious and because it avoids the level differences between the two sources. Further, we prefer to use a minimum proportion of 0.10 for groups with quarterly effects and to try a range of starting values because that can be more easily generalized to new data than Marcelis et al. [2].

The model was indeed able to identify groups with quarterly effects, but the effect of corrected VAT data on the quarter-by-quarter growth rates was very small. For the near future, CBS decided not to focus on corrections, but on analyzing the range of possible reporting differences at unit level between the two turnover sources, on how these differences relate to properties of the businesses, and on trying to identify their causes by discussing this with experts. In the long run, CBS is considering to merge the two forms of data into one production system and use VAT data where possible and survey data otherwise. Our model helps in identifying differences that can more easily be found at group level than for individual units.

ACKNOWLEDGEMENTS

We gratefully thank Femke Kessels for her contribution to the presented study. Any opinions and conclusions expressed in the current paper are those of the author(s) and do not represent the views of Statistics Netherlands.

6. REFERENCES

- [1] A. van Delden, A., and S. Scholtus, "Correspondence between survey and administrative data on quarterly turnover", discussion paper 2017-3, (2017), The Hague: Statistics Netherlands. Available at <https://www.cbs.nl/en-gb/background/2017/07/correspondence-between-survey-and-admin-data-on-quarterly-turnover>.
- [2] J. Marcelis, A. van Delden, S. Scholtus and F. Kessels, Modelling intra-annual measurement in linked administrative and survey data, Proceedings of Statistics Canada Symposium (2022).
- [3] D. Lewis and J. Woods, Issues to consider when turning to the use of administrative data: the UK experience. Paper presented at the New Techniques and Technologies for Statistics (NTTS) conference, (2013).
- [4] A. van Delden, A., S. Scholtus, and N. Ostlund, "Modelling measurement errors to enable consistency between monthly and quarterly turnover growth rates", *Proceedings of the 2018 International Methodology Symposium*, Statistics Canada. (2020)

Development and Integration of Microservices for Smart End-to-End Surveys

Joeri Minnen (hbits CV), Jerome Olsen (Destatis), Claudia De Vitiis, Fabrizio de Fausti, Francesca Inglese, Marco Dionisio Terribili, Angela Pappagallo (Istat) & Pieter Beyens (hbits CV)

1. INTRODUCTION

The Smart Survey Implementation (SSI) project focuses on developing microservices to create an end-to-end methodology for conducting Smart Surveys.

2. METHODOLOGY

The project involves defining business and functional requirements, followed by the development and testing of technical solutions aimed at collecting and integrating Smart Data.

3. RESULTS AND PRACTICAL APPLICATION

In this presentation, WP3 of the SSI project (hbits, CBS, ISTAT, Destatis) showcases the development of two key microservices designed to enhance the Household Budget Survey and Time Use Survey by leveraging Smart Data. Microservices are independent environments capable of hosting AI/ML-driven models to handle specific tasks and derive relevant variables. One microservice processes shopping receipts through scanning, while the other analyzes data from geotracking points.

The presentation will explore, from a technical perspective, how these models function and how their outputs are integrated into the MOTUS (developed by hbits) and @HBS (developed by CBS) data collection platform. It will also demonstrate how users/respondents benefit from these microservices in near real-time.

Figure 1. Generic Microservice Architecture – independent coupling

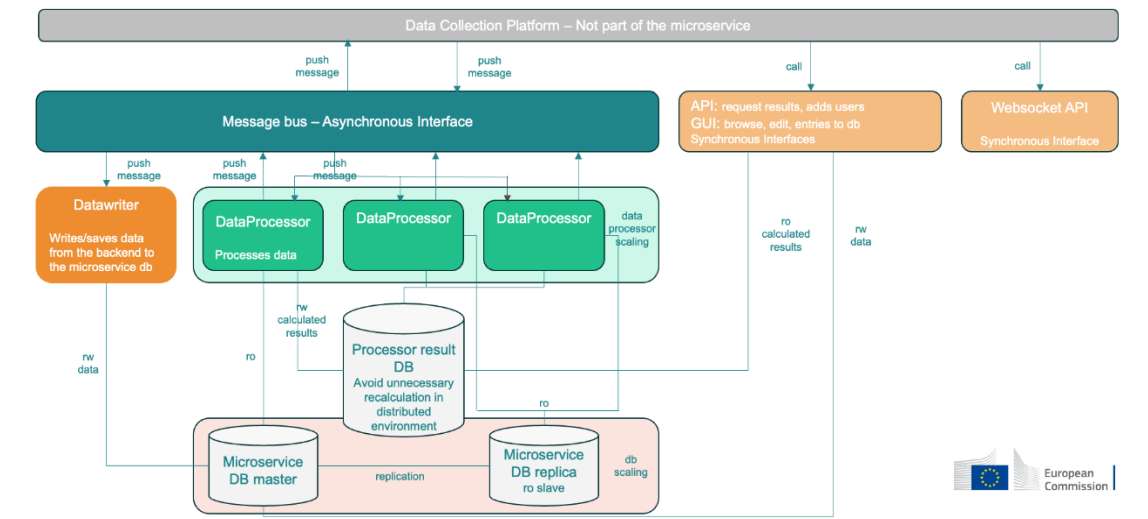


Figure 2. Receipt Scanning Management in MOTUS Back-office

The screenshot shows the MOTUS Back-office interface for Receipt Scanning Management. The main header is 'Expenses Receipt Scanning Microservice'. The left sidebar contains navigation links: Dashboard, Receipt Scanning Microservice, Overview, Flow, Builders, Respondents, Access rights, Translations, My team, and Administrator. The main content area displays 'EXPENSE TUESDAY 10TH OF SEPTEMBER 2024' with details for 'SPAR' store, 'Unknown' category, 'Unknown' location, 'be' kind of expense, 'professional expense' kind of expense, and 'expenses paid with credit card' payment method. Below this, 'EXPENSE ITEMS' are listed, including 'SPAR GEZINSBIEFSTUK', 'SPAR american prep. martini', 'PVLADELPHIA sock en f line knuden', 'Tomaat Ruby Red 500g', 'BONI aardappelhoot jes 1kg', 'BONI paprikamix 3x1 1kg', '109619 CREL ROND BROOD 0.60 Kg', '41314 BONI multivitamin PET 33x1', and 'COCA-COLA Zero Sugar 1L'. A large image of a receipt is shown on the right. The right sidebar contains a 'Navigator' section with a search bar and a list of expenses for 'Tue 10th September 2024' and 'Tue 24th September 2024'. Below this, a 'Fixed costs' section shows 'No fixed costs'.

Figure 3. Geotracking Management in MOTUS Back-office

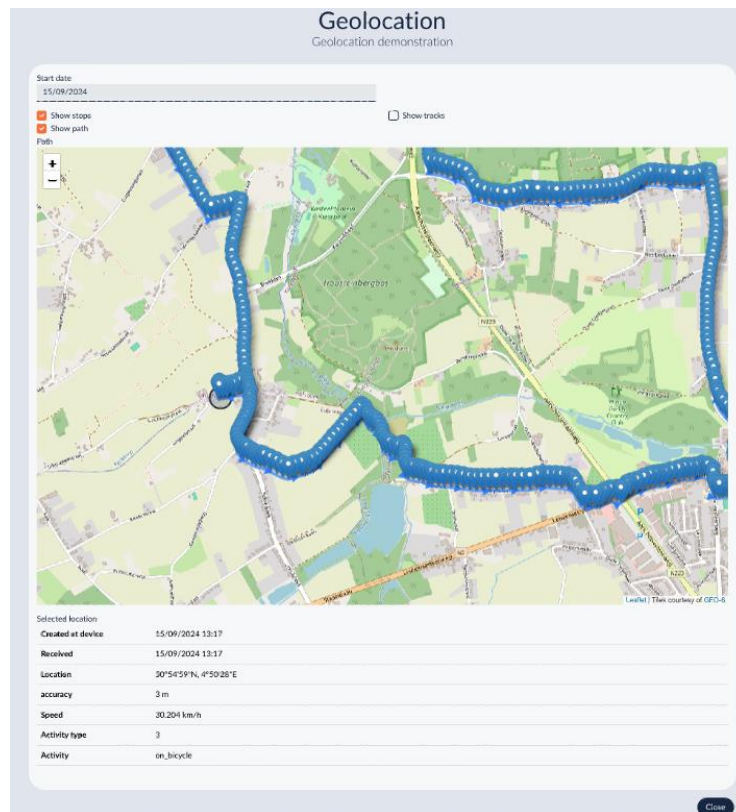


Figure 4. Geotracking in MOTUS UI/UX

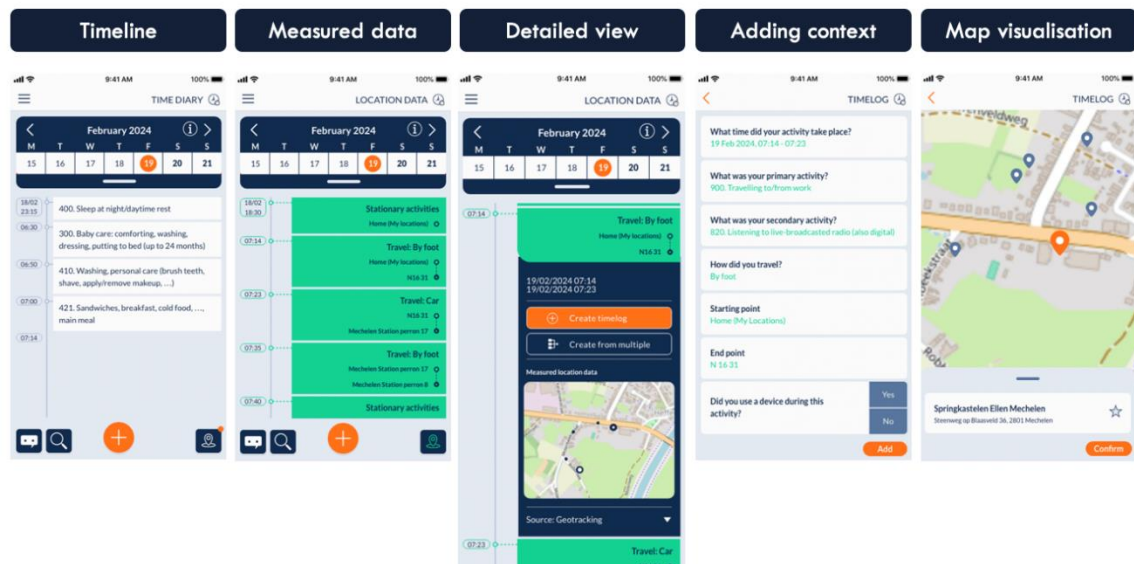
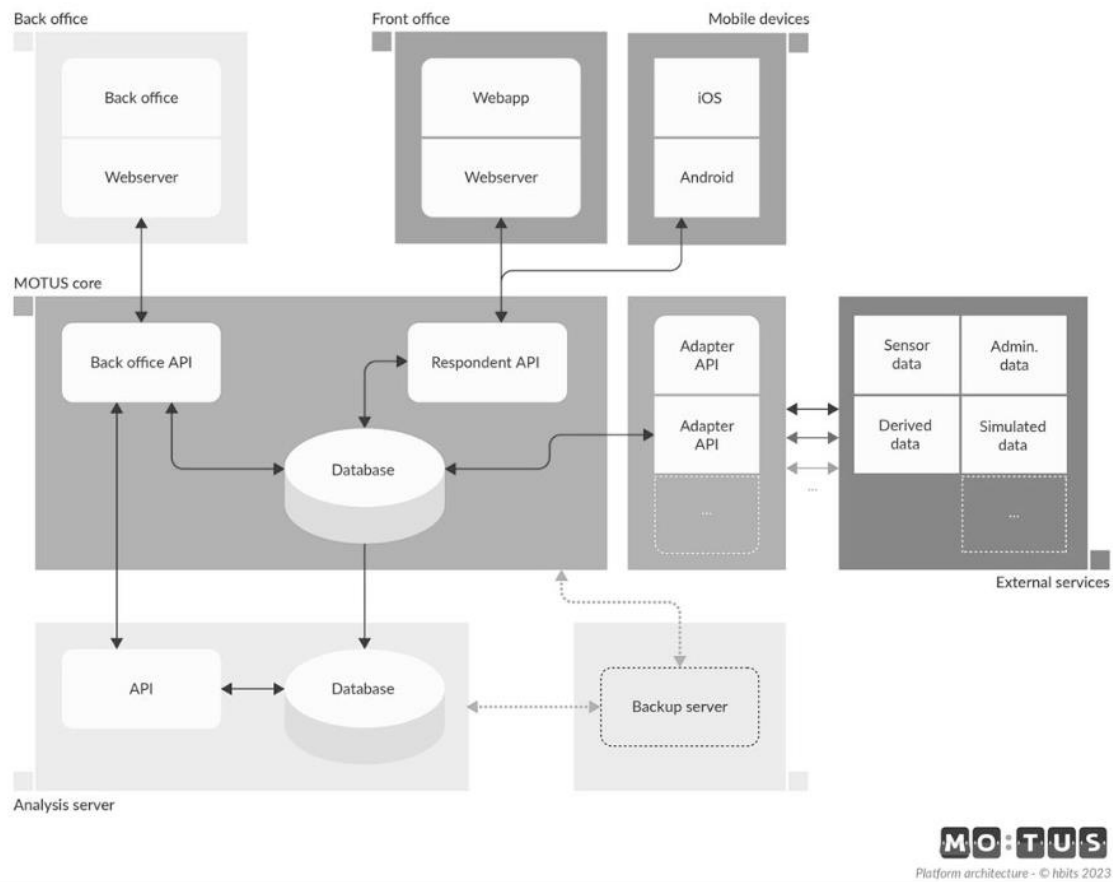


Figure 5. Platform architecture MOTUS



4. MAIN FINDINGS

A live demonstration will illustrate the full end-to-end process, from Smart Data collection to its processing and application.

Development of a Satellite Account for Culture at the Local Level - First results for Tuscany (Italy)

Alessandra Coli (University of Pisa), Sabrina Iommi & Donatella Marinari (IRPET
Regional Institute for Economic Planning of Tuscany)

1. INTRODUCTION

Culture plays a pivotal role in the economy of certain territories, particularly as a driver of tourism. Reliable estimates of the economic significance of cultural activities are therefore essential for governments in these regions.

While national and European statistics offer insights into the economic importance of the cultural sector through various sources—such as business registers, public and non-profit organization budgets, and household surveys—they often lack a cohesive framework for analysis. The satellite account emerges as a valuable methodological tool to bridge this gap, integrating diverse statistics within the structure of national accounts.

This contribution presents the preliminary findings of a research project aimed at constructing a satellite account for culture in Tuscany. The applied methodology ensures consistency with the regional accounts produced by ISTAT [1] and aligns with the definition of the cultural sector outlined in European statistics [2]. The estimates are integrated into a statistical framework consistent with the national account standards established by IRPET, Tuscany's regional institute for economic planning.

2. METHODOLOGY

Satellite accounts were officially introduced into the national accounts framework with the 1993 System of National Accounts (SNA) (§2245-2249, [3]). Their purpose is to add flexibility to the national accounting system, either by enabling in-depth analyses using concepts alternative to those of the core national accounts or by adhering to the same definitions and classifications. The latter type, which provides a more detailed focus on a specific topic of interest while remaining fully aligned with the core accounts, is referred to as an internal satellite account.

The aim of this research is to develop an internal satellite account for culture in Tuscany. Tuscany already has several statistics aligned with national accounts, including supply and use tables [4] and a satellite account for tourism [5]. The initial step in constructing the satellite account for culture is to identify and capture the flows related to the cultural sector within these existing statistical frameworks. Once this foundation is established, the analysis can be refined and expanded by incorporating additional data sources as they become available.

The strategy for constructing the satellite account has been divided into two interconnected activities: the first focuses on defining the account's framework, while the second involves creating a database to collect, integrate, and process culture-related data from various sources, ensuring alignment with national account standards.

Regarding the first activity, the process begins with defining the cultural sector and categorizing it into cultural domains (e.g., heritage, performing arts, etc.). To ensure comparability with satellite accounts for culture developed across Europe, the boundaries and domain classifications proposed by the ESSnet-Culture project [2] were adopted, albeit limited to the primary cultural domains. Next, it is essential to evaluate how the adopted classification of cultural sectors corresponds to those used in national accounts. This includes aligning with the European Classification of Economic Activities (NACE), the Classification of Public Expenditure by Function (COFOG), the Classification of Individual Consumption by Purpose (COICOP), the Classification of the Purposes of Non-Profit Institutions Serving Individual Households (COPNI), and the Classification of Products Associated with Economic Activities (CPA). Such alignment ensures that the satellite account integrates seamlessly with regional, national, and European accounts. Finally, special attention must be given to the relationships between cultural domains and the specific classifications used in Tuscany's tourism satellite account to facilitate effective integration. A fully implemented satellite account should capture culture-related monetary transactions at every stage of the economic cycle, encompassing production, income generation and distribution, and final demand for cultural goods and services. This represents a complex and ambitious endeavour that demands considerable time and resources. For now, the focus has been placed on the supply side of cultural goods and services, as data availability in this area is comparatively greater than on the demand side.

Regarding the second activity, an individual data archive has been developed, where each row represents a cultural operator, and each column corresponds to a specific variable. The archive encompasses companies, public and non-profit organizations, and individuals producing cultural goods and services in Tuscany in 2019. To identify cultural operators and gather relevant information, multiple sources were utilized, including official statistical data, administrative records from the Regional Department of Culture of Tuscany, and data collected directly through online budget publications or questionnaires sent to key cultural sector operators. The data have been processed to align with national accounting concepts. Specifically, each operator has been classified as an institutional or local unit and assigned to the appropriate institutional sector in accordance with the European System of Accounts [3].

3. RESULTS AND PRACTICAL APPLICATION

The analysis of preliminary results from the archive described in the previous section reveals that the cultural sector accounts for 1.3% of employment and 0.6% of value added in Tuscany's total regional economy (Table 1). These figures may appear to underestimate the economic significance of culture in the region, particularly when compared to other considerably higher estimates.

For example, according to Symbola [6], in 2019, the cultural sector contributed approximately 2.7% to value added when considering only the core sectors, and 5.4% when including the culturally driven sectors.

The difference between these estimates arises from the varying definitions of culture. In this study, the focus is limited to cultural sectors in the strict sense. Within this narrower framework, our estimates suggest that the value added is approximately double the figures reported by Symbola.

Table 1. Distribution of the number of enterprises, employees and value added by cultural domain, Tuscany, year 2019 – Percentage values

Cultural sectors	Cultural activities	Persons employed	Value added
Publishing & Printing	18%	19%	30%
Cine-audiovisual and musical productions	10%	8%	9%
Artistic Creations	22%	9%	6%
Performing arts	29%	17%	26%
Museum & Libraries	19%	44%	29%
Art, Dance, Music Schools	2%	2%	1%
TOTAL CULTURAL SECTORS	100%	100%	100%
<i>% cultural sector on total regional economy</i>	<i>1.7%</i>	<i>1.3%</i>	<i>0.6%</i>

Source: Our elaborations

The results also indicate that the cultural sector's contribution is significantly smaller than that of tourism-related activities: 1.3% of employment compared to 11.3% for tourism and 0.6% of value added compared to 12% for tourism. However, it is essential to recognize that Tuscany's cultural offerings, particularly those linked to its heritage, are critical drivers of tourism and likely the primary attraction for visitors.

Among cultural sectors, publishing and printing, museums and libraries, and performing arts hold the greatest economic significance, collectively accounting for 85% of the sector's value added and 80% of its employment.

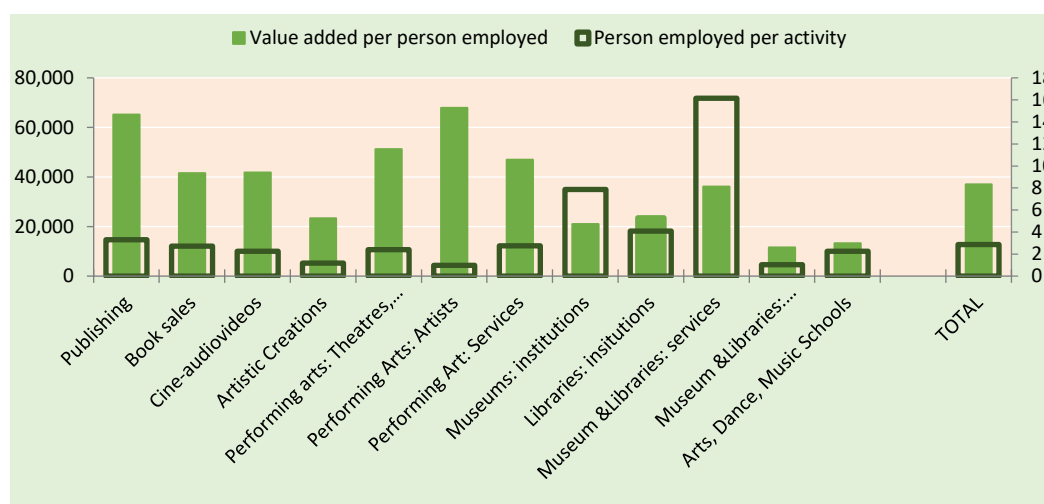


Figure 1. Value added in euros per person employed (left axis) and average number of persons employed per production unit (right axis), by cultural sector - Tuscany, year 2019. Source: Our elaborations

The archive data also facilitate the calculation of key indicators, such as the average number of persons employed per production unit and the value added per person employed across cultural sectors (Figure 1). For certain cultural sectors, the analysis delves deeper to separately examine activities with differing characteristics. For instance, within the performing arts sector, activities conducted by theatres and companies, individual artists, and enterprises providing support services are analysed independently. The highest average number of

persons employed per production unit is observed in enterprises offering support services to museums and libraries (16 employees) and in museums themselves (8 employees), compared to an average of 3 employees for the sector as a whole. In terms of value added per person employed, the highest figures are recorded in performing arts (artists) at €68,000 and publishing at €65,000, exceeding the sectoral average of €37,000. In contrast, the lowest values are seen in art, dance, and music schools (€13,000) and among professionals working in museums, libraries, and archives (€12,000).

4. MAIN FINDINGS

This research contributes to the development of satellite accounts for culture in the EU and, to our knowledge, represents the first attempt in Italy to construct a satellite account for culture based on international frameworks proposed over the past decades [2, 7, 8, 9, 10].

Compiling a satellite account involves several challenges. One of the most critical is gathering detailed information in areas that are often underrepresented in existing statistics. This requires a combination of top-down approaches (to ensure alignment with higher-level accounts) and bottom-up methods (to integrate new data specific to the cultural sector). The complexity is particularly pronounced for satellite accounts in culture, as the absence of an officially recognized international manual compounds the difficulties. These challenges are further amplified when the satellite account is developed for a sub-national area, as in this research.

One significant issue encountered thus far has been the difficulty in assigning production units to the domains and functions defined in [2]. Existing statistical classifications often fail to adequately specify the types of activities conducted within the cultural sector. As a result, it is essential to thoroughly examine the collected data and allocate production units on a case-by-case basis, especially for the most significant producers.

The research conducted so far has developed the satellite account for the supply of cultural goods and services. The next steps will focus on measuring the employment generated by the cultural sector and estimating the monetary flows associated with the demand for cultural goods and services.

5. REFERENCES

- [1] ISTAT, Conti economici territoriali, anni 2020-2022, Istat, Comunicato stampa, 21 dicembre 2023.
- [2] ESSnet-Culture project, European statistical system network on culture, final report of the ESSnet-Culture project, 2012.
- [3] United Nations, System of National Accounts, 1993.
- [4] Paniccià R. A methodology for building a multiregional Supply and Use Table for Italy: an updated and revised version, IRPET Working Paper N.12 2024.
- [5] Bocci C., Conti E. Tourism Satellite Account and Economic Analysis, a Regional Perspective, 14th Global Forum on Tourism Statistics, Venice, 23 November 2016.
- [6] Fondazione Symbola – Unioncamere, (2020) Io sono Cultura – Rapporto 2020.
- [7] LEG-Culture, Cultural statistics in the EU, Final report of the LEG on cultural statistics, Population and social conditions 3/2000/E/N1, Eurostat working papers, 2000.
- [8] UNESCO Institute for Statistics, The UNESCO framework for cultural statistics, 2009.
- [9] Eurostat, Guide to Eurostat culture statistics, Manuals and Guidelines, 2018.

[10] European Union Intellectual Property Office A Satellite Account For The European Union Creative Industries, 2019.

Development of crop pollination ecosystem service accounts in Ireland

Sam Belton, Alan Cahill, Nova Sharkey, David O'Connell & Harry Hodgins
(Central Statistics Office)

1. INTRODUCTION

Ecosystem Accounts have now been included as a new module to Regulation of the European Parliament and of the Council Regulation (EU) No 691/2011. One of these accounts, for which mandatory reporting will commence in 2026, will cover the crop pollination ecosystem service. This is defined as the contribution by wild pollinators to the production of crops, accounted for in tonnes of pollinator-dependent crops as supplied by different ecosystems providing suitable pollinator habitat. Presented here is an overview of the geospatial methods and data used to produce accounts for Ireland for the reference year 2018.

2. METHODOLOGY

These accounts were developed while adhering to guidance notes issued by an EU taskforce for developing ecosystem accounts, and results have been fed back to Eurostat to better refine the methodology. Central to estimating the contributions of ecosystems to pollinator-dependent crop provision is the quantification of pollinator habitat suitability for each hectare unit of ecosystem. In practice, this involved the incorporation of a wide range of geospatial data in a geospatial model to predict pollinator habitat suitability within a 1 km radius of pollinator-dependent crops. This corresponds to the maximum flight distance of long-distance insect pollinators.

Briefly, pollinator-dependent crops were selected from a national geospatial database before mapping landcover within the 1 km buffer zone. Next, following consultation with national experts, data sources for relevant landscape features and land management practices which could affect habitat suitability were identified and mapped. National experts also provided guidance on the assignment of appropriate 'correction factor' values for each landscape feature and area type under a particular land management practice. Finally, experts assisted with the selection of suitable floral and nesting attractiveness scores for all landcover and crop types. These geospatial data were then used alongside gridded temperate and solar irradiance data (as predictors of pollinator activity) in a biophysical model to produce habitat suitability scores for each landcover type. At this point, a crosswalk could be made between landcover and ecosystem types at level one of a new EU ecosystem typology. After assigning pollination factors (which were again informed by national experts) to each crop type, pollinator-dependent yields could be distributed from individual crop pixels to adjacent ecosystem type pixels harbouring suitable pollinator habitat. From this, the ecosystem contribution to the crop pollination service could be quantified.

2.1. Ecosystem and crop geospatial data

The source geospatial data for ecosystem types were the Corine Land Cover (CLC) Accounting Layers 2018 (25-hectare minimum mapping unit (MMU) for areal phenomena and

100 m MMU for linear phenomena) whereas the pollinator-dependent crop data source was the Irish Land Parcel Identification System (LPIS) 2018 (median polygon area = 2.7 hectare). The latter is used by the Irish Department of Agriculture, Food and The Marine (DAFM) to administer area-based payment schemes to farmers.

2.2. Geospatial data for relevant landscape features and land management

Data for relevant landscape features, such as hedgerows, rivers and treelines were sourced from a bespoke National Landcover Map (NLCM) produced by a consortium of government agencies for the reference year 2018 (MMU = 25 m). High resolution railway and road data were provided by the Irish state agency responsible for property registrations, property valuation and national mapping services (Tailte Éireann). Data for land management practices, such as whether fields are managed as traditional hey meadows, were sourced from LPIS. As a source of high nature value farmland data, the 2012 High Nature Value (HNV) farmland data from the European Environmental Agency were used.

2.3. Geospatial data for temperature and solar irradiance

Temperature (°C) and solar irradiance ($\text{kJ m}^{-2} \text{day}^{-1}$) information were sourced as gridded national data from the Irish national meteorological service (Met Éireann) and the JRC Agri4Cast Data Portal, respectively.

2.4. Crop yield data

Pollinator-dependent crop yield data were downloaded from Eurostat, after having initially been submitted by DAFM. Pollinator-dependent crops for which yield data were not available were instead imputed by deriving yield factors for similar crops (e.g., loganberry yield derived using raspberry yield factor).

2.5. Geospatial data processing

All geospatial data and biophysical modelling were performed in R (v.4.4.0) using a range of different R packages, but primarily relying on functions from 'terra' (v.1.7-78) [1] and 'sf' (1.0-16) [2]. In general, vector data (e.g., LPIS and NLCM) were rasterised to the resolution of CLC18, and calculations performed thereafter using raster data. The geospatial model used to produce the habitat suitability map yielded a habitat suitability score per pixel. This score was used to select ecosystem areas providing 'medium-to-high' suitability for pollinator habitat. A suitable threshold value was selected through consultation with national experts. The distribution of pollinator-dependent yield values from each crop pixel was done using a 'per pixel', iterative process across multiple computing cores – this was facilitated using the core R package 'parallel' (v.4.4.0) [3].

2.6. Supply and Use Table

Pollinator-dependent yield values were assigned to each crop pixel, after which – using the above iterative process – values were distributed to ecosystem pixels harbouring suitable wild pollinator habitat, as determined by the geospatial model. Values were distributed proportionally to habitat suitability scores which were 0.2 or greater. Yields per crop type could then be aggregated by ecosystem type, thereby giving the supply table for the crop pollination ecosystem service. Final pollinator dependent yields per crop were assigned to "Intermediate consumption by industry". Furthermore, because the area of crops which were not within range of a suitable wild pollinator habitat could be derived, it was possible to also model the unmet demand in tonnes of pollinator-dependent crops.

3. RESULTS AND PRACTICAL APPLICATION

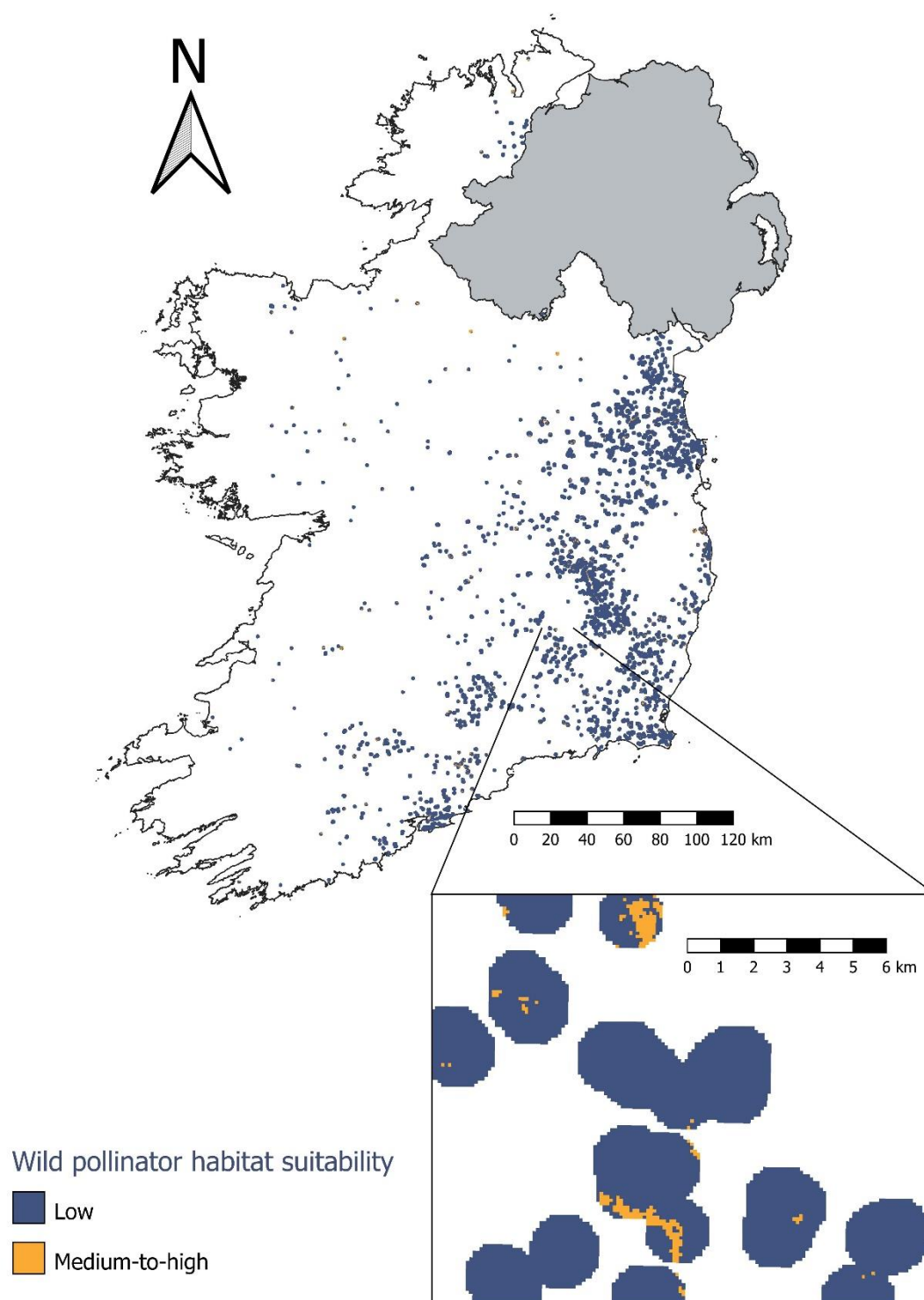


Figure 1. Habitat suitability map produced for Ireland for the reference year 2018. Pixels in orange are ecosystem types which can provide the crop pollination service.

Table 1. Supply and use table for the crop pollination ecosystem service in Ireland for 2018

Crop Category	Pulses (1.1.4)	Oil-bearing Crops (1.1.6)	Vegetables (1.1.7)	Fruits (1.1.8)	Total Yield Attributable to Wild Pollinators
Supply ('000 tonnes)					
1 Settlements & Other Artificial Areas	0	0	0	0	0
2 Croplands	0.4	0.4	0.1	3.5	4.4
3 Grasslands	0.7	0.9	0.1	4.3	6
4 Forests & Woodlands	2.7	3	0.4	2.8	8.9
5 Heathlands & Shrubs	0	0	0	0.1	0.1
6 Sparsely Vegetated Ecosystems	0	0	0	0	0
7 Inland Wetlands	0.1	0	0	0	0.1
8 Rivers & Canals	0	0	0	0	0
9 Lakes & Reservoirs	0	0	0	0	0
10 Coastal Beaches, Dunes & Associated Habitats	0	0	0	0	0
11 Marine Inlets, Transitional Waters & Wetlands	0	0	0	0	0
Total Supply	3.9	4.3	0.6	10.7	19.5
Use ('000 tonnes)					
Intermediate Consumption by Industries	3.9	4.3	0.6	10.7	19.5

Source: CSO Ireland and
DAFM

4. MAIN FINDINGS

- In Ireland for the year 2018, crops that were at least partly dependent on wild pollinators produced a yield of 84,900 tonnes. Of this yield, 19,500 tonnes (23%) were estimated to be the result of wild insect pollination (i.e., the crop pollination service provided by ecosystems).
- The ecosystems which provided the crop pollination service were – in decreasing order of their contribution – Forests & Woodlands, Grasslands, Croplands, Inland Wetlands, and Heathlands & Shrubs. These accounted for 8,900 (46%), 6,000 (31%), 4,400 (23%), 100 (1%) and 100 (1%) tonnes of the total pollinator-dependent yield, respectively.
- Of the 21,900 hectares of land growing wild pollinator-dependent crops in 2018, less than half (9,700 hectares) were within range of wild pollinator habitats. Therefore, 12,300 hectares of pollinator dependent crops were not within range of a wild pollinator habitat.
- This corresponds to 17,800 tonnes of crops which were not yielded due to an absence of suitable wild pollinator habitats.
- Altogether, with the invaluable assistance of national experts, it was possible to draw on a wide range of geospatial (including administrative) data to model the crop pollination ecosystem service in Ireland for a given reference year. This will allow for the annual production of crop pollination ecosystem service accounts in 2026 (the first official reporting year) for Ireland.

5. REFERENCES

- [1] R.J. Hijmans, terra: Spatial Data Analysis, R package version 1.7-78 (2024) (<https://CRAN.R-project.org/package=terra>).

- [2] E. Pebesma and R. Bivand, Simple Features for R: Standardized Support for Spatial Vector Data, *The R Journal* 10 (1), 439-446 (2023) <https://doi.org/10.32614/RJ-2018-009>
- [3] R Core Team, parallel (2024).

Earth Observation and AI for Construction Statistics (EO4ConStat)

Maren Köhlmann, Stefan Irrgang (Federal Statistical Office Germany, Destatis)
Carola Stolle, Frederik Stellmach (Federal Agency for Cartography and Geodesy,
Frankfurt, Germany)

1. INTRODUCTION

To support quality assurance measures for construction statistics, the project 'Earth Observation and AI for Construction Statistics' (EO4ConStat) is set up of a consortium of the German National Statistical Institute (Destatis), the Federal Agency for Cartography and Geodesy (BKG) and the German Aerospace Centre (DLR). The objective of developing a method to quality assure construction activity statistics using remote sensing data and artificial intelligence is relevant for high quality construction statistics.

Due to the German government's ambition to build new, affordable and climate appropriate housing, construction statistics are currently in the political focus. The aim of the project – which is financed by the European Commission – includes developing algorithms to detect buildings and building construction sites and possibly define new construction starts as well as works completions from earth observation data. This is to be achieved by using and adapting a permissive open license segmentation model. Furthermore, a methodology to compare these results of the change detection analysis with the data collected through the traditional statistical channels shall be contrived.

1.1. The use of remote sensing data in official statistics

Remote sensing techniques provide images of large areas of the earth's surface at relatively frequent intervals. During the last decade, technological progress has substantially influenced and improved the availability and analysis of satellite and other remote sensing data. Moreover, the demand for up-to-date data has increased. Hence, remotely sensed data is coming into play also in official statistics [1].

The construction activity statistics are intended to show the structure of construction activity in Germany, which is an important early indicator of economic development in the construction industry. Regular publications contain the monthly statistics on building permits, the annual statistics on works completions, construction progress, construction dismounts and the update of the residential building and housing stock [2]. Several political endeavours as well as a user consultation show that there is a high demand for detailed intra-year statistics of construction starts and works completions.

Optical remote sensing data cannot solve all the demands as they solely provide a view from above. Yet, with an advanced model and additional elevation data and other geo-information the idea is to develop a method that can help to quality assure and enhance construction activity statistics.

2. METHODOLOGY: CHARACTERISTICS OF REMOTE SENSING DATA

Remote sensing has become an indispensable tool in various fields, including environmental and land use monitoring as well as planning, agriculture, and disaster management. For these purposes, data can be acquired differently. In general, three types of remote sensing can be subdivided: spaceborne where the sensor is attached to a satellite

or space-missile, airborne where the sensor is attached to an airplane, airship, drone, hot-air balloon or similar and ground remote sensing where the sensor is mounted to a ship, vehicle, measuring station or any other equipment [3]. Moreover, it can be differentiated between active remote sensing instruments which operate with their own source of emission or light and passive ones, which rely on recording energy that was emitted by the sun and its reflectance from earth's surface [4]. Based on these principles there is a wide spectrum of remote sensing modalities existing that includes optical, thermal, radar, and LiDAR detection, each offering unique capabilities and characteristics for earth observation.

With these principles remote sensing enables the observation and analysis of Earth's surface and atmosphere at different spatial and temporal scales. The spatial resolution – meaning the pixel dimension of the imagery – varies between a few centimetres (so called digital orthophoto – DOP) up to multiple kilometres for satellite imagery. The temporal resolution depends on the return rate to a certain location and varies from less than a day (satellite images) up to several years (airplane flyovers for certain areas). An exemplary overview of the spatial and temporal resolution for remote sensing data for Germany is depicted in Table 1.

Table 1. Exemplary availability of remote sensing data for Germany, as of July 2022.

Type of remote sensing data	Availability / temporal frequency	Spatial resolution
Satellite imagery	Daily up to every 16 days	0,3m up to 30m
Aerial photo / Digital Orthophoto	Every 2 to 4 years (rolling updates)	0,1m up to 0,4m
3D-Modell of Earth's surface	One-off aerial survey (end of 2025)	40 points per m ²

A huge advantage is that remotely sensed data can provide georeferenced information of large areas across administrative boundaries at a small-area level. In addition, different remote sensing data sources can be combined or matched to maximize spatial and temporal resolution and adapt the data basis for specific tasks or analysis.

2.1. Data Processing and Analysis

Central to remote sensing data analysis is the extraction of meaningful information from raw sensor measurements. This process often involves a series of computational techniques, including image processing, feature extraction, and classification. Without going into detail, basically the data needs to be radiometrically calibrated, synchronised with the according position e.g. of the aircraft or satellite and then geocorrected [5].

The pre-processed remote sensing images need to be further analysed to extract the information of interest from the data. Within the last years methods of extracting information from optical data have advanced and artificial intelligence (AI) plays a major role in these developments. With segmentation techniques – which is a subfield of digital image processing and computer vision – related regions in an image can be delimited by combining neighbouring pixels according to specific homogeneity criteria. This allows to isolate specific objects or areas and thus to extract thematic raster maps within an image e.g. for navigation, urban planning, environmental monitoring and much more topics or use cases. Bearing this in mind, the question of whether construction activity can be detected through AI techniques used to analyse remote sensing data arises.

3. THE SEGMENT ANYTHING MODEL: AN OVERVIEW

In 2023 a project was introduced by Meta AI Research to build a foundation model for image segmentation [6]. A foundation model is an artificial deep learning model which is pretrained

on a very large amount of data and can be used to learn new datasets, develop more specific applications and hence perform new tasks. The Segment Anything Model (SAM) provides an approach to semantic segmentation which classifies each pixel of an image so that all objects of a certain class belong together. The SAM was trained with a large data set of 11 million licensed and privacy-preserving images of a geographically and economically diverse set of countries with which numerous masks (=classifications) were produced [6]. Hence, SAM segments objects of interest across diverse landscapes and imaging modalities. Its applicability extends to various remote sensing applications, ranging from land cover classification and change detection to object detection and monitoring. An example of how the model basically works is given in Figure 1.



Figure 1. SAM automatically segments everything in an image: (a) Putting a grid on the image, (b) Selecting certain points of interest within the image, (c) Masking the selected points according to the identified class. Source: <https://segment-anything.com>

Traditional segmentation methods need extensive human input and intervention to gain accurate results [7]. With more and more advanced AI and deep learning techniques the human input can be reduced and methods further automated. The SAM brings good generalization capabilities and is supposed to deliver accurate predictions with little further training data. This makes the model explicitly interesting for remote sensing applications, since the data acquisition and annotating process in this field is highly labour-intensive and requires expert knowledge.

Semantic and instance segmentation, in particular, play a crucial role in identifying and delineating specific objects or land cover classes within remotely sensed imagery. If the SAM performs segmentation with zero or relatively low domain-specific information this might offer an important advantage for this process [7].

4. THE USE OF SAM TO DETECT CONSTRUCTION ACTIVITY

Construction activity showcases a specific application to analyse. In general, urban areas exhibit a large variety of reflectance patterns in remote sensing data. These areas have large intra-class variations and often also low interclass variation which makes them challenging to accurately analyse [8]. Within urban areas construction sites could be detected by identifying open soil, because bare soil might not be common. However, if one building replaces another former one, the challenge is to identify the construction start and building phase from the temporally available earth observation data. In rural areas bare soil might be more frequent and therefore the aim is to differentiate between arable land / fields and open soil due to the existence of construction sites. Moreover, the frequency of remote sensing data varies largely as can be seen in Table 1. Therefore, several approaches and diverse input data will be tested within the EO4ConStat project.

To detect buildings and construction sites and possibly define new building starts as well as construction dismounts from earth observation data, the SAM will be adapted and trained accordingly with context specific imagery containing secured label information. The data

used in the project can be divided into three main categories: (1) digital orthophotos (DOP) with a spatial resolution of at least 20x20 cm, which will serve as data basis for training data for the SAM (2) house perimeter data as well as OpenStreetMap data which are used to guide the discovery of construction sites for training data generation and (3) Sentinel-2 data from the European Copernicus program, which will be used to more accurately date construction start and dismount. The DOP have a temporal frequency of two to three years, the house perimeter data is a yearly dataset and the Sentinel-2 data are available at least every six to 12 days or as a monthly cloud-free mosaic dataset.

Processing of Sentinel-2 data yields a variety of indices, which provide information on different aspects of landcover. Time series analysis of these data can provide valuable insights on change in land cover over time. Thereby it is possible to deduce on construction sites and their progress. As an example, specific changes in NDVI (Normalized Difference Vegetation Index) or NDBI (Normalized Difference Built-Up Index) can provide useful information, when monitored over time. Furthermore, new construction sites might be possible to detect through this analysis of timely more frequent data.

The project EO4ConStat is still in the first phase where certain research approaches are yet to be set up and need to be worked out in further detail. For the construction activity statistics all further details to get to know the quality of existing statistics better and to narrow down and quantify certain possibly existing gaps or errors in the data will help to quality assure the existing publications. At best, the developed methods will help in the process to further subdivide the current publications of the annual statistics into intra-yearly information.

The SAM, that is used in this research, was released in 2023 which means there is a huge potential to further adapt it with the current rapid developments in the field of AI and deep learning networks. In particular, the use of the SAM in the field of earth observation data is currently being further developed in various projects, including other National Statistical Offices (NSI). This will enhance the learning and exchange to improve the use of earth observation data for the statistical process and to quality assure existing statistics in diverse fields.

ACKNOWLEDGMENTS

This work is funded by the European Commission Grant SMP-ESS-2023-GEOS-IBA with the Grant-Nr. 101155801. Its contents are the sole responsibility of the authors listed above and do not necessarily reflect the views of the European Union.

5. REFERENCES

- [1] S. Arnold and S. Kleine, Neue Wege der Geodatennutzung: Perspektiven der Fernerkundung für die Statistik: Geplante Erprobung der Nutzung von Satellitendaten für Flächenstatistik und Ernteerhebungen, WiSta - Wirtschaft Und Statistik #5 (2017), 31–36. Online available at: https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2017/05/neue-wege-geodatennutzung-052017.pdf?__blob=publicationFile
- [2] C. Schumann and M. Schepers and A. Weigert, Eine zukunftsfähige Bautätigkeitsstatistik, WiSta - Wirtschaft Und Statistik #2 (2023), 27-38. Online available at: https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2023/02/zukunftsfaeihige-bautaetigkeitsstatistik-022023.pdf?__blob=publicationFile
- [3] L. Luo, X. Wang, H. Guo, R. Lasaponara, X. Zong, N. Masini, G. Wang, P. Shi, H. Khatteli, F. Chen, S. Tariq, J. Shao, N. Bachagha, R. Yang and Y. Yao, Airborne and spaceborne remote sensing for archaeological and cultural heritage applications: A review of the century

- (1907–2017), *Remote Sensing of Environment*, #232 (2019). Online available at: <https://doi.org/10.1016/j.rse.2019.111280>
- [4] P. Kogut, Types of Remote Sensing: Technology Changing The World, (2020). Online available at: www.eos.com/blog/types-of-remote-sensing
- [5] M.A. Warren, B.H. Taylor, M.G. Grant and J.D. Shutler, Data processing of remotely sensed airborne hyperspectral data using the Airborne Processing Library (APL): Geocorrection algorithm descriptions and spatial accuracy assessment, *Computers & Geosciences*, #64 (2023). Online available at: <https://doi.org/10.1016/j.cageo.2013.11.006>
- [6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár and R. Girshick, Segment Anything, Meta AI Research, FAIR (2023). Online Available at: <https://ai.meta.com/research/publications/segment-anything/>
- [7] L. P. Osco, Q. Wu, E.L. de Lemos, W.N. Gonçalves, A.P.M. Ramos, J. Li, and J. Marcato, The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot, *International Journal of Applied Earth Observation and Geoinformation*, #124 (2023). Online available at: <https://doi.org/10.1016/j.jag.2023.103540>
- [8] J.A. Montoya-Zegarra, J.D. Wegner, L. Ladický and K. Schindler, Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W4 (2015), 127–133. Online available at: <https://doi.org/10.5194/isprsannals-II-3-W4-127-2015>

Earth Observation to accelerate ecosystem accounting

Bruno Smets, Marcel Bucchorn (VITO), Stefano Balbi, Alessio Bulckaen (BC3 Research) Ferdinando Villa (BC3 Research), Lars Hein (Ecounting), Ioannis Kokkoris (Department of Sustainable Agriculture, University of Patras), Panayotis Dimopoulos (University of Patras) & Mathilde De Vroey, Tim Jacobs (VITO)

1. INTRODUCTION

Ecosystem Accounting is a statistical framework to quantify and integrate the value of natural ecosystems, and the services they provide, into national economic planning, accounting (System of National Accounts), decision and policy making. The framework underpins the development of Ecosystem Accounts that measure the stock, geographic extent and state of natural capital and the changes thereof.

The United Nations System for Economic Ecosystem Accounts (UN SEEA EA) offers a standardized statistical framework for Ecosystem Accounting. In Europe, this standard was adopted by the European Parliament and Council and amended to the EU Regulation No 691/2011 on environmental economic accounts, which mandates countries to report ecosystem accounts yearly, from 2026 onwards.

Given that the accounts measure the extent of natural capital and location-dependent ecosystem services, Ecosystem Accounts are inherently spatial and strongly depend on the availability of spatially explicit datasets. The emergence of dense Earth Observation (EO) data streams, like those offered by the EU Space/Copernicus programme, combined with advancements in digital technologies offer unprecedented opportunities for countries to monitor ecosystems and account systematically and efficiently for their value. The Pioneer Earth Observation apPLications for the Environment (PEOPLE) Ecosystem Accounting (PEOPLE-EA) project financed by the European Space Agency (ESA) developed and demonstrated innovative Ecosystem Accounting applications using EO data streams.

2. METHODOLOGY

First a broad review was performed to collect user requirements and to perform a desktop review on the use of earth observation for ecosystem accounting. This has resulted in a State-of-Art review [1] and Value Proposition report.

These reports have formed the base to develop in an agile way and in co-design with Early Adopters six innovative ecosystem accounting workflows making use of the potential and advantages of integrating Earth Observation datasets. The six accounts cover the entire SEEA EA typology, hence an ecosystem extent account, two ecosystem condition accounts (forest and coastal) and three ecosystem service accounts (wood provision, global climate regulation, nature-based tourism recreation). The Early Adopters are National Statistical Offices, or representatives thereof, which expressed their needs and testing of innovative solutions into their operational ecosystem accounting practises. The accounts were generated and evaluated by the Early Adopters.

Based on their experiences and further input gathered from the ecosystem accounting community during an International Workshop and evolution roadmap was defined which captured the main gaps and required research and development activities moving forward to mainstream the use of earth observation for ecosystem accounting.

3. RESULTS AND PRACTICAL APPLICATION

Habitat (EUNIS) maps for Slovakia and Peloponnese were generated through Machine Learning, which forms the backbone for natural classes in the ecosystem extent accounts according to the EU typology. The maps were evaluated through a comprehensive validation and showed a high accuracy (>90%) on the upper levels (one and two) and a good accuracy (> 60%) on the lower level (three) mainly due to some issues with wetlands. These maps were used to generate ecosystem extent reports for the year 2020.

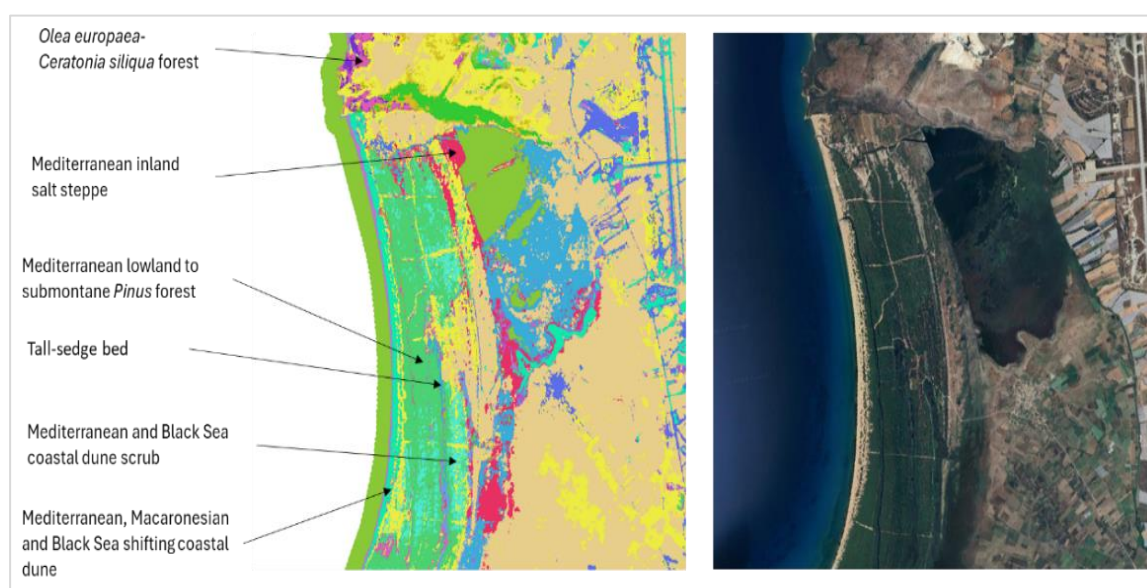


Figure 1. High-resolution EUNIS level-3 map over an area of West-Peloponnese.

An innovative system of systems [2] was setup to generate ecosystem forest condition account across full Europe, being context aware on the decisions to be taken. The accounts were generated and evaluated for the five Early Adopters (Norway, Netherlands, Slovakia, Greece and Italy), hence crossing all bio-ecological zones in Europe. The system enables statistical offices to generate the ecosystem accounts themselves in an easy way, taking always the latest available earth observation data. A coastal condition index was also developed to monitor the share of artificial impervious area cover in the coastal zone.

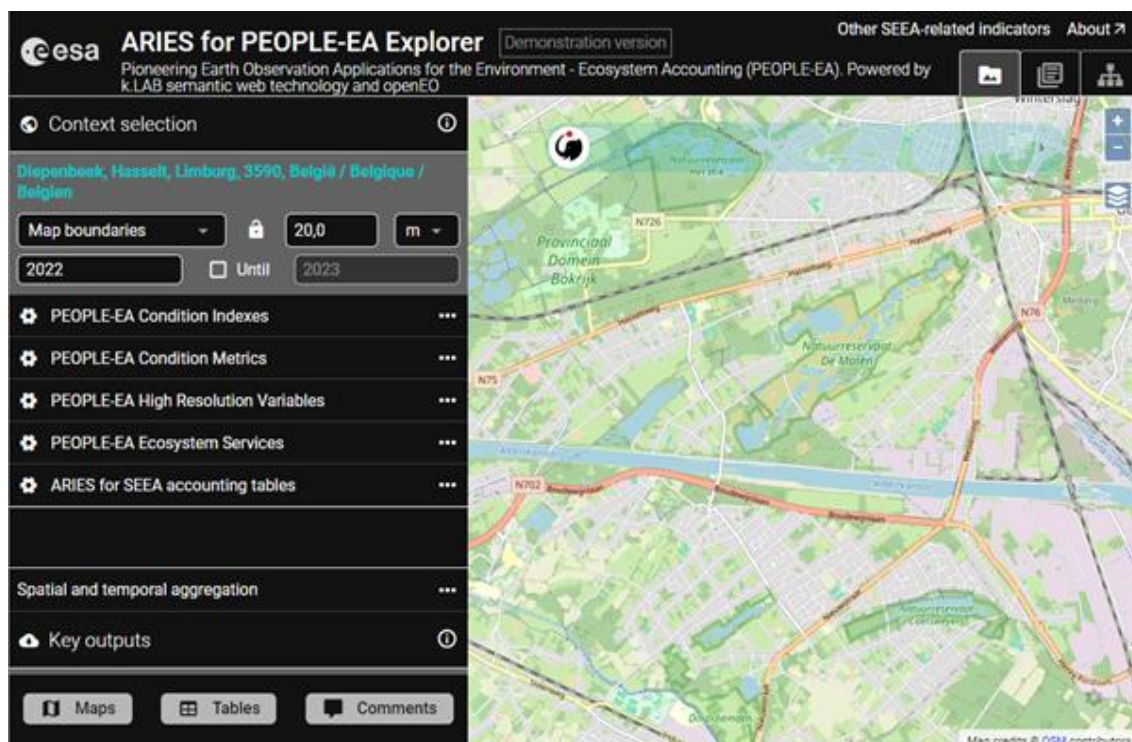


Figure 2. Forest Condition Accounting system.

On the ecosystem services side, wood provision accounts were generated through an experimental approach over Norway, Slovakia and Greece. The accounts were matching quite well with the national statistics, and provide the benefits of spatial representativeness such they can be used at regional (to local) scale.

NAME_LATN	NAI [m3 overbark]	mean_NAI_per_ha [m3 overbark]	forest_area [ha]
Oslo and Viken	4,415,176	4.1	1,076,515
Rogaland, Vestland and More og Romsdal	3,521,113	3.1	1,135,711
Trøndelag	2,819,319	2.5	1,126,883
Troms og Finnmark	1,155,342	1.3	861,044
Innlandet	5,877,166	3.0	1,959,810
Vestfold og Telemark	2,810,378	4.2	672,228
Agder	2,574,535	4.1	626,072
Nordland	1,266,325	1.9	662,814
TOTAL	24,439,354	3.0	8,121,077

Figure 3. Accounting table for wood provision in Norway.

Data limitations were found for the Global Climate Regulation Service, hence no accounts could be generated but the Earth Observation community was informed on these blocking limitations and have tailored their program to try removing such limitations.

Finally an improved recreation potential map was generated to distribute tourism statistics (overnight stays or daily visits) to the ecosystem area's with the highest potential.

4. MAIN FINDINGS

The advantages of integrating Earth Observation (EO) in ecosystem accounting (EA) was shown and (mostly) proven feasible. Some data limitations in the EO datasets were discovered and fed to the community to resolve in future.

An innovative system was setup which enables National Statistical Offices to generate ecosystem accounts in an easy manner without the need to know in depth Earth Observation techniques as well as dealing with big datasets.

A Research and Development (evolution) roadmap was created, through contributions by the EO and EA communities. This roadmap can serve as an important input for both the European Commission and the European Space Agency to prioritize future activities.

5. REFERENCES

- [1] Kokkoris, I.P., Smets, B., Hein, L., Mallinis, G., Buchhorn, M., Balbi, S., Cernecky, J., Paganini, M., Dimopoulos, P. (2024). Applying earth observation for ecosystem accounting. *Ecosystem Services*.
- [2] <https://esa-people-ea.org/en/results/aries4people-application>

Efficiently Selecting Representative Audit Samples

Lotte Mensink (EMOS Programme, Utrecht University)

1. INTRODUCTION

Audits are widely used for quality evaluation in many different fields, but they can be costly and time-consuming. When performing an audit, a sample of the population is selected for thorough examination. In performing audits within the context of official statistics, it is not uncommon that part of the population has already been audited. For example, it often occurs that the largest companies are audited by default, due to their substantial influence on the statistic of interest. Given the considerable costs associated with auditing, it would be desirable to re-use the already audited units. However, these units are not randomly chosen, and therefore form a non-probability sample, in which the inclusion probabilities of each unit are either not known or not useful. Traditionally, it is believed that inferences from non-probability samples cannot generalize beyond the sample, because they can contain serious sample selection bias (Rao, 2021).

This puts the researcher in a difficult position. On the one hand, it would be efficient to re-use the available initial audit sample, as it would save considerable resources. On the other hand, the possible non-representativeness of the initial audit sample forms a serious threat for the validity of the conclusions to be drawn from the audit sample.

To tackle this problem, Boeschoten et al. (2019) have developed an approach in which an audit sample can be selected that is representative, in which already audited units can be re-used. In this approach, also referred to as the deviance approach, sample selection viewed as a constrained minimization problem, in which the goal is to obtain the most representative sample possible, under certain audit sample size restrictions. These audit sample size restrictions include a maximum number of additional units to include in the audit sample and a maximum number of previously audited units to exclude from the audit sample.

The implementation of the deviance approach has been investigated using a simulation study, which showed that the deviance approach can be used to efficiently select a representative audit sample (Boeschoten et al., 2019). However, the deviance approach does suffer from a practical limitation. The method is focused on maximizing representativeness, instead of maximizing efficiency. Under the provided sample size restrictions, the deviance approach will find the most representative audit sample, but this does not necessarily mean that this sample is the most efficient in terms of additionally sampled units. In practice, the financial costs associated with auditing additional units are often high, and it might therefore be more desirable to find the most efficient solution, while making sure that representativeness is maintained.

To tackle this practical limitation, this study proposes to restructure the deviance approach in such a way that the number of units to include in and exclude from the sample is minimized, while making sure that representativeness is maintained. We refer to this restructured approach as the sample size approach. We believe the sample size approach fits better with the practical priorities of the researcher.

2. OBJECTIVE

First, on a practical level, the aim of this study is to introduce the sample size approach and explain how it may be applied in practice for various situations in which researchers may find themselves.

Second, on an analytical level, the aim of this study is to investigate and evaluate the performance of the sample size approach. Hereby, we intend to provide insight into how the method works, and how particular factors may affect audit sample selection using the sample size approach.

Finally, on a fundamental level, the aim of this study is to enhance insight into the relation between deviance (our measure for representativeness of the audit sample) and bias (in the estimates that the audit sample is intended to provide). The existence of a relation between these concepts forms a fundamental assumption for both the deviance approach as well as the sample size approach. Gaining more insight into the functional form of this relation is crucial for the proper application of the sample size approach.

3. THEORETICAL FRAMEWORK

This study assumes that the population that needs to be audited can be divided according to a classification variable of interest. For all units in the population, an error-prone version of the classification variable of interest is observed, denoted by Y . The true value of the classification variable of interest is referred to as W . Eventually, the researcher wants to use the audit sample to estimate the true proportions $P(W = w)$ and the measurement error of the error-prone classification $P(Y = y \mid W = w)$. Furthermore, it is assumed that one or more covariates are available for each unit, denoted by X . Finally, Z denotes the selection indicator of the audit sample, with $Z = 1$ indicating that a unit is included in the audit sample, and $Z = 0$ indicating that a unit is not included in the audit sample.

We rely on the conditional independence model, a model in which the observed classification variable of interest Y is independent of audit inclusion Z , given the covariate(s) X under consideration. If the observed classification and audit inclusion would in fact be conditionally independent, this would mean that our audit sample is representative. Hence, the deviance of the conditional independence model is used to reflect the representativeness of the audit sample. If the deviance is sufficiently small, we assume that the conditional independence model holds. This means the audit sample is sufficiently representative, in the sense that population parameters can be estimated from the audit sample without bias.

The sample size approach is formulated as a constrained minimization problem aimed at selecting a representative audit sample while minimizing the number of additional units to be included. The target function, which is the sum of the units to include in and exclude from the sample, is minimized under the constraint of a user-defined deviance threshold that should not be exceeded to ensure representativeness.

The sample size approach allows the user to choose any deviance threshold. Determining a suitable deviance threshold for an existing data set with an initial audit can be a challenging task. We assume that as deviance decreases, the sample becomes more representative, and bias decreases. However, we can only guess when it comes to which deviance thresholds yield sufficiently representative audit samples in terms of bias. Moreover, an audit sample that is considered sufficiently representative in one research context, might not be deemed sufficiently representative in another context.

The study describes a detailed approach to find out the maximum amount of bias that can be expected under different deviance threshold. Using this approach, the researcher can model the relation between bias and deviance based on the known characteristics of the data. Combining this relation with the researcher's maximum tolerable bias, the researcher can determine the most appropriate deviance threshold for a given scenario.

4. METHODS

To evaluate how the sample size approach performs, a simulation study is conducted. In this simulation study, the sample size approach, including the deviance threshold selection method, is applied to several possible situations in which researchers may find themselves. Hereby, we aim to provide insight into how the method works, and how characteristics of a data set may affect different aspects of audit sample selection. Because we want to illustrate the sample size approach for several situations in which researchers may find themselves, the factors that are varied in the simulation study are based on information that is already available to the researcher. The factors to vary are presented in Table 2. For each factor, we will investigate the main effects by varying it while fixing the other factors to the most desirable situation. Furthermore, we will take the least desirable and most desirable situation for every factor and investigate this in a full factorial design. As some of the main effects and full factorial conditions overlap, this yields 12 conditions in total. For every condition, we generate 1000 different distributions of X, Y, W and Z with corresponding data sets. We apply the sample size approach (including the method to determine the most appropriate deviance threshold based on the relation between bias and deviance) to each of these data sets.

Table 2: Conditions simulation study

Factor	Conditions	Description
Relation between Y and X	Cramér's V = 0.1	The relation between observed classification Y and covariate X
	Cramér's V = 0.2	
	Cramér's V = 0.4	
Sampling fraction	0.003	The ratio of initially audited units to units in the population
	0.01	
	0.03	
	0.1	
Maximum measurement error	50%	The maximum measurement error that the researcher expects
	33%	
	25%	

The goal of the simulation is to see whether the proposed sample size approach works as expected. Using the described deviance threshold selection method, we can obtain audit samples for any given tolerable amount of bias. The question is, however, whether these audit samples in fact contain no more bias than the tolerable amount. We can answer this question by determining the bias in the final audit samples that result from applying the sample size approach.

5. RESULTS

In the majority of the conditions, the average bias after the sample size approach was applied is below the maximum tolerable bias. When the sampling fraction is smallest, and when the maximum tolerable bias is low, average bias is sometimes higher than maximum tolerable bias. In many of the other conditions, the average bias falls well below the maximum tolerable

bias. We are also interested in seeing in how many data sets (out of the 1000 data sets per condition) the observed bias exceeds the maximum tolerable bias based on which the deviance threshold was chosen. The proportions of data sets in which maximum tolerable bias was exceeded are shown in Table 7.

Table 7: The proportion of final audit samples in which observed bias in $P(W = w)$ or $P(Y = y|W = w)$ exceeds maximum tolerable bias

Sampling fraction	Measurement error	YX relation	Bias 1%	Bias 2%	Bias 5%
0.003	Max 50%	Weak	0.338	0.220	0.061
0.003	Max 50%	Strong	0.368	0.272	0.114
0.003	Max 25%	Weak	0.304	0.190	0.057
0.003	Max 25%	Strong	0.336	0.230	0.086
0.01	Max 25%	Strong	0.230	0.110	0.030
0.03	Max 25%	Strong	0.118	0.045	0.012
0.1	Max 50%	Weak	0.056	0.022	0.004
0.1	Max 50%	Strong	0.059	0.021	0.004
0.1	Max 33%	Strong	0.067	0.023	0.002
0.1	Max 25%	Weak	0.039	0.015	0.005
0.1	Max 25%	Medium	0.035	0.011	0.005
0.1	Max 25%	Strong	0.043	0.014	0.000

These results show that there are some cases in which the sample size approach works well, and the proportion is (very close to) zero, but there are also cases in which the performance is quite bad. First, we can conclude that the method works better as we tolerate more bias. Overall, the performance is best in the conditions where the maximum tolerable bias is 5%, and worst in the conditions where the maximum tolerable bias is 1%. Furthermore, we again observe a strong effect of the sampling fraction. The method appears to work better when the sampling fraction is larger. As the final audit sample sizes were also a lot larger in these situations, this conclusion is not surprising.

Although this study mainly focused on performance in terms of bias, efficiency is another key aspect of performance. This aspect was studied more thoroughly in a pilot simulation study, which was also part of the thesis project (Mensink et al., 2023). The results of this simulation study showed that the sample size approach made it possible to obtain a representative audit sample through sampling fewer additional units than the deviance approach. Furthermore, the sample size approach outperforms the deviance approach by recognizing samples that are already representative to begin with. For representative initial audit samples, the deviance approach would suggest including more units in the sample to improve further upon representativeness, whereas the sample size approach recognizes that no additional units are needed because the sample is not selective to begin.

All in all, we conclude that the sample size approach could be a valuable addition to the auditing practice. Especially in situations when efficiency in terms of additional units to sample is prioritized, the sample size approach can provide good solutions. Further research is necessary to explore the full potential and investigate implementation in various auditing contexts.

6. CONTRIBUTION

In the context of official statistics, auditing the data used to produce a statistic is crucial to ensuring its' credibility. In many cases, part of the data has already been audited for other purposes. This study tackles the challenge of obtaining a representative sample while at the same time making optimal use of the audit sample that is already available. On a practical

level, this study contributes by introducing the sample size approach, providing researchers with a valuable tool that allows them to optimize their efficiency in auditing.

Furthermore, on an analytical level, this study contributes by evaluating the performance of the sample size approach through extensive simulation studies, offering insights into how various factors affect the application of the sample size approach.

Finally, on a fundamental level, this study contributes by gaining more insight into the theoretical foundation of the sample size approach. The study thoroughly investigates the relation between bias and deviance. It is shown that the relation exists, and guidance is provided how the researcher can take this into account in determining a deviance threshold to be able to use the sample size approach.

7. REFERENCES

- [1] Boeschoten, L., S. Scholtus, and A. van Delden (2021). A note on efficient audit sample selection. <https://arxiv.org/abs/2105.10737>.
- [2] Mensink, L., S. Scholtus, and L. Boeschoten (2023). A comparison between two approaches for efficient audit sample selection. https://github.com/lottemensink/pilot_simulation_research_archive/tree/main.
- [3] Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* 83 (1), 242–272.

Embracing Open Source: The European Commission's Journey

Miguel Diez Blanco (European Commission)

1. INTRODUCTION

This presentation will explore the European Commission's Directorate-General for Digital Services (DIGIT) and its path towards adopting open-source values and practices, as outlined in the Open Source Strategy 2020-2023.

The focus will be on a retrospective analysis of the Commission's relationship with open source, examining the most significant actions taken under the strategy and their impact on the organization.

Additionally, the presentation will discuss the future trajectory of the Commission's open-source initiatives, with a view toward continued innovation and collaboration.

Enhance statistical literacy among non-expert users: Eurostat's experience

Romina Brondino (Eurostat)

1. INTRODUCTION

One of the strategic objectives of Eurostat is to increase statistical literacy amongst users and help them turn data into knowledge. A key element in improving statistical literacy is to make official statistics easier to understand.

Eurostat offers a wide range of tools and resources that help non-expert users of statistics, such as young people, students, and teachers, better understand data and be able to use them in their studies, work, and daily life. These resources, which include explanatory tools, interactive publications, and data visualisations, are accessible for users through the dedicated section on Eurostat's website, the [Education corner](#).

The two main challenges in promoting statistical literacy are:

- how to reach out to non-advanced users and non-users;
- how to communicate on statistical literacy products and services in a way that is relevant for them.

Additionally, it is critical to engage and interact with these users to gain a better understanding of their changing needs and make sure that the portfolio of tools and materials remains targeted to their needs.

The aim of this presentation is to share Eurostat's approach and initiatives linked to meeting those challenges.

2. BUILDING PARTNERSHIPS IN STATISTICAL LITERACY

Within the framework of Eurostat's cooperation on dissemination with the national statistical offices (NSIs), Eurostat and the NSIs have established an informal group on statistical literacy to exchange practices, share experiences and mutually promote existing products and activities.

As a result of this cooperation, Eurostat has set up an [online repository](#) of statistical literacy resources developed by the NSIs in different languages. This repository, accessible through the Education corner, gives an overview of the wide range of tools and resources made available by the NSIs whilst at the same time, giving users the opportunity to access additional materials in languages other than English thus removing the language barrier.

3. STRENGTHENING PROMOTION OF STATISTICAL LITERACY PRODUCTS AND ACTIVITIES

One of the challenges of reaching non-users is finding new ways and channels to reach them.

In 2024, Eurostat launched a dedicated [statistical literacy newsletter](#). The newsletter, which currently has over 1 500 subscribers, (August 2024) includes up to date information on recently developed and existing statistical literacy products, events and statistical competitions. To increase the number of subscribers, reach a wider audience and promote the work done by national statistical offices, the newsletter also presents materials developed by the NSIs which are available in languages other than English.

Teachers represent a key target group for enhancing statistical literacy among young people. Since 2022 Eurostat has taken part in a panel with teachers organised by the European Commission's Directorate-General for Communication (DG COMM). The panel includes primary and secondary school teachers from all over Europe as well as representatives from the Europe Direct centres, the EU's information centres that answer questions about the EU. The panel meets annually, giving the Commission the opportunity to present its latest products and tools for young people and to test products in development.

Eurostat has been reaching out to various platforms to showcase available products to new audiences. Among them is DG COMM's dedicated website for teachers and students (EU Learning corner) which promotes Eurostat's materials including interactive publications and links to Eurostat's Education corner.

4. ENGAGING AND UNDERSTANDING USERS' NEEDS

Engaging with users is of key importance not only to increase awareness of products and tools relevant to them but also to gain a better understanding of their needs and to tailor products to better meet the users' needs.

The presentation will highlight one of the Eurostat tools as an example, a short animation for children developed with the involvement of a group of teachers. The video animation was developed based on relevant input and feedback from teachers on the overall concept, the storyboard and visual elements, based on their experience of teaching to pupils. At the same time, the video also responds to the need of expanding the range of Eurostat product to target new groups such as children.

Eurostat regularly carries out usability testing of its website and selected products to improve the overall user experience. The presentation will cover, an example-key findings from the usability testing run on Eurostat's interactive publications in 2024.

5. CONCLUSION

This presentation will outline Eurostat's experience in enhancing statistical literacy focusing on 3 main strands:

- Building partnerships in statistical literacy
- Strengthening promotion of statistical literacy products
- Engaging with and better understanding users' needs

Enhancing data quality of a building and dwellings register through mapping building coordinates to building polygons using a Random Forest classifier

Mathias Kästenbauer (Statistics Austria)

1. INTRODUCTION

The buildings and dwellings register (=BR), led by the National Statistics Institute, was established in 2004. It contains vast amounts of building information including, amongst others, the address, number of floors, gross and roof areas, and building coordinates (=BC). It was initially populated using data acquired during the 2001 census, with Municipalities bearing the responsibility for maintaining the BR. The BC are required to be located within a building's boundary. These boundaries, which are not part of the BR, are created by a federal institute, hereafter FI. Out of roughly 3.1 million BCs, 630,000 are located outside of the building polygons (=BP).

Due to the diversity of cases and suboptimal data quality, the automatic matching of BCs to a BP could not be achieved for a long time. The aim of this project was to establish a match using a Random Forest (=RF) classifier to reduce the number of misplaced BCs.

2. METHODOLOGY

2.1. Data sources and pre-processing

Three datasets are required: the BP and property polygon (=PP) layers provided by the FI, as well as the BC and additional information from the BR. The BP layer is updated on a three-year cycle, with different regional areas being covered each year. This results in having BP's from 2019-2022. Consequently, all instances where the BC's creation date (e.g. 2024) is more recent than the BP's creation date (e.g. 2019) were omitted.

Figure 1 shows the preprocessing of the three datasets. In step (0), all residential building coordinates (red), BPs (orange) and PPs (purple) are displayed. Step (1) shows that only BCs placed outside of BPs were kept. In step (2) the BCs were counted for each PP. Step (3) demonstrates that all instances with one BC per PP were kept, while the others omitted. During step (4), all PPs without a BC were removed, such that all BPs outside of a PP could be omitted in step (5). A one-to-many spatial join was then performed between the remaining BCs and BPs per PP. As seen in (5), not every PP must have a BP, as long as a BC is present in the PP. The final dataset of this exemplary case would consist of 5 observations – one for each BC and BP relation per PP.

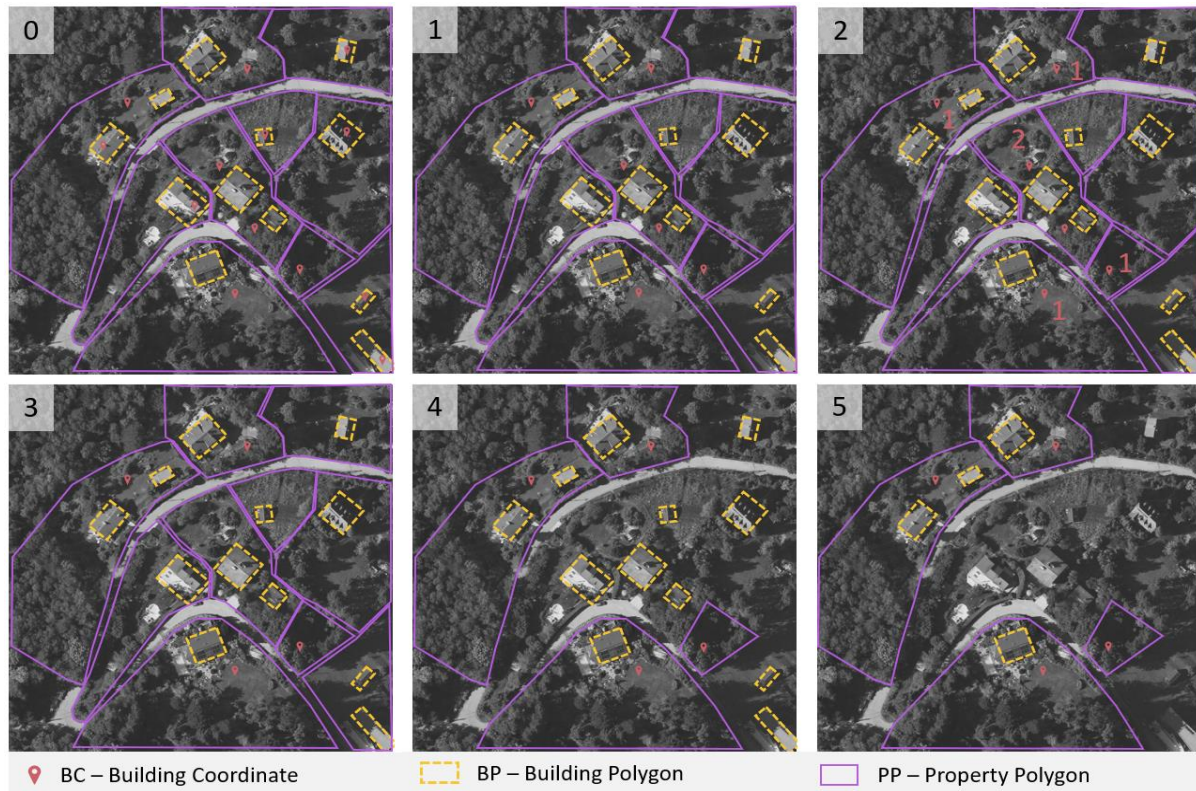


Figure 1. Illustration of the required input data preprocessing.

The spatial join holds the BC's geometry and unique IDs for the BR, BPs and PPs to preserve information from the original data sources. To match the BCs to BPs, the BCs' gross and roof area, number of floors, the BPs' geometry, creation date, median- and maximum height, and the PPs' geometry were used. The BPs' and PPs' geometry were used to extract geometric information, such as its area. The dataset holds nearly 192,000 unique BCs; however, due to the one-to-many relation, the dataset stores nearly 350,000 observations.

2.2. Features

Several features were defined to describe the relationship between the BC and BPs. The features contain geometric information, such as the distance between the BC and the BPs, the areas recorded in the BR for each BC, as well as the areas of the BP and PP. Additional information, for example, includes whether a BP overlaps a PP, as well as relational information such as the distance between the BC and one BP compared to the distance between the BC and all BPs per PP. Information on the possible relation between a BC and other BPs per PP is required since each of the BPs might be a feasible match for the BC. Consequently, the features should also express how the BC relates to other BPs.

2.3. Machine Learning modelling

The training data consists of roughly 7000 observations, which were created by randomly splitting 2% of all observations. Validation plots were made to ensure that the distribution of the training data (2%) and the feature data (98%) match. Then, a binary target (0 if the BC does not match a certain BP, otherwise 1) was manually assigned to the training data. After assigning the targets, the training data was further randomly split into 75% for the model training and 25% for the model's validation.

A RF classifier using the scikit-learn implementation in Python [1] was trained on 75% of the training data. The RF classifier consists of multiple tree-like classifiers where each tree receives an independent but identically distributed vector of inputs. Each of the trees cast a

vote. The final prediction results from the majority vote from all trees. Due to the independence of the trees, the RF classifier is robust against noise [2]. An analysis of feature importance was performed to evaluate whether certain features improve or detract from the model's accuracy. Finally, the hyperparameters were tuned using a Random Search followed by a Grid Search. The parameter ranges can be seen in Table 1.

Table 1. RF classifier parameters of the hyperparameter tuning. Names refer to the RF classifier of Python's scikit-learn package version 1.5.1. [1]

Parameters	Parameter Range					
	Random Search			Grid Search		
	Start	Stop	Step	Start	Stop	Step
n_estimators	200	800	100	650	750	3
min_samples_split	2	10	1	3	8	1
min_samples_leaf	1	4	1	1	4	1
max_depth	10	50	4	8	16	1
max_features	Sqrt	None	-	Sqrt	None	-

3. RESULTS AND PRACTICAL APPLICATION

3.1. Random Forest Model results

Table 2 presents the parameters of the best-performing RF model.

Table 2. RF classifier best hyperparameters. Names refer to the RF classifier of Python's scikit-learn package version 1.5.1. [1]

	n_estimators	min_samples_split	min_samples_leaf	max_depth	max_features	Bootstrap
RF	683	5	2	14	sqrt	True

Table 3 shows that the model has an out-of-bag (=OOB) accuracy of 0.9547 while the validation against the set-aside test data led to an accuracy of 0.9489. The model's precision, recall and F1 scores are slightly higher for class 0 than class 1. The ROC AUC scores at 0.9443.

Table 3. RF Classification metrics for both classes.

Class	Precision	Recall	F1-Score	Accuracy	OOB Accuracy	ROC AUC
0	0.9598	0.9606	0.9602			
1	0.9295	0.9280	0.9287			
Overall				0.9489	0.9547	0.9443

An analysis of the permutation feature importance showed that not all features contributed positively to the model's performance. Omitting some features resulted in a slight increase in the model's accuracy. Features representing area-related variables were found to be most important to the model's performance.

The confusion matrix (see Figure 2) for the RF model shows that the model excels in predicting both true positives and true negatives. The false positives (top right) and false negatives (bottom left) are nearly equally distributed. The True Negative Rate is slightly higher than the True Positive Rate, with 96.06% compared to 92.80%. The False Positive Rate is 3.94%.

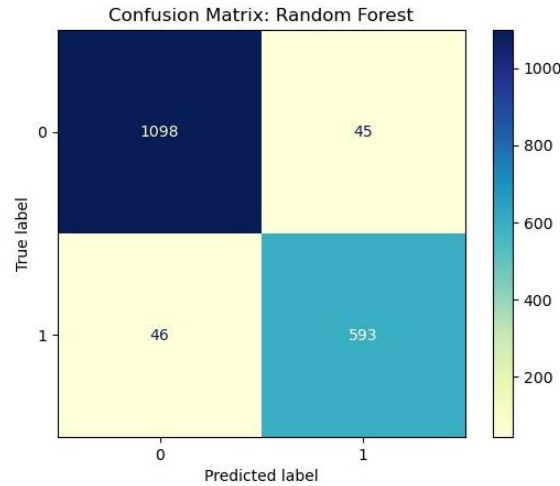


Figure 2. Random Forest Confusion Matrix of the best performing model.

3.2. Feature Data

Figure 3 shows the cumulative probability distribution of both classes, 0 in blue and 1 in orange. The model exhibits more certainty in predicting class 0 compared to class 1, indicating a stronger ability to identify cases where no match between a BC and a BP exists. The left-hand side of the cumulative distribution function shows a steep incline for both classes, suggesting few samples are assigned under terms of uncertainty. The model assigns 20% of all observations to class 0 and 25% to class 1 with a probability of less than 50%. Assignments under terms of higher certainty with probabilities that are exceeding 80% are reached for 55% and 40% of all observations for classes 0 and 1 respectively.

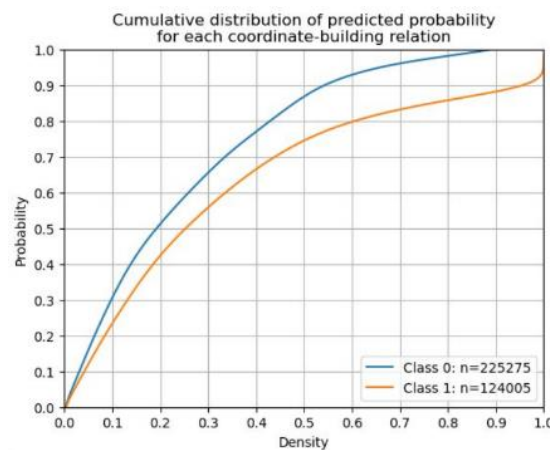


Figure 3. Random Forest Cumulative Distribution of class 0 and class 1.

About 125,000 out of roughly 192,000 BCs were assigned to a matching BP. The RF classifier could not identify a match for roughly 66,000 BCs. Nevertheless, the implementation of the RF classifier during the first iteration facilitated the automatic assignment of 20% of all misplaced BCs.

4. MAIN FINDINGS

The availability of accurate building polygon and property polygon vector datasets allows for automatic matching between the building coordinates and building polygons per property polygon using a Random Forest classifier. Especially area-related features played a crucial role in accurately classifying whether a building coordinate should be mapped to a building polygon or not.

5. REFERENCES

- [1] F. Pedregosa and G. Varoquaux and A. Gramfort and V. Michel and B. Thirion and O. Grisel and M. Blondel and P. Prettenhofer and R. Weiss and V. Dubourg and J. Vanderplas and A. Passos and D. Cournapeau and M. Brucher and M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011), 2825-2830.
- [2] L. Breiman, Random Forests, Machine Learning 45 (2001), 5-32.

Enhancing European Business Statistics: A Harmonised Microdata Linking Approach

Nikola Sunjka (European Commission, Eurostat)

1. INTRODUCTION

European business statistics are increasingly confronted with new and emerging user needs for timely and relevant data to support policy decisions. Traditional statistical practices often involve substantial delays in data delivery, reducing the efficacy of statistics for areas that require rapid political action. Moreover, there is a continuous call to reduce the administrative burden on businesses, making it imperative to explore alternative data collection methods. Micro Data Linking (MDL) solves these challenges by offering a statistical method for combining existing microdata from various sources, such as surveys, administrative records, and registers, without increasing the burden on respondents.

Eurostat has conducted several MDL projects in collaboration with Member States, demonstrating the potential for creating new insights and developing new statistics from existing datasets. Despite these successes, current MDL practices vary across Member States, with differences in readiness, technical capabilities, and methodological approaches. To maximise the benefits of MDL, a more coordinated and harmonised approach across the European Statistical System (ESS) is essential. This abstract outlines Eurostat's strategic approach to standardising MDL practices, aiming to deliver high-quality, timely, and cost-effective statistics on new and emerging policy issues at the EU level. The approach can be readily adapted to other countries, regions, or statistical domains.

2. METHODOLOGY

The project proposes a two-step approach to achieve a harmonised MDL implementation within the ESS.

Step I: The first step involves establishing a multi-annual MDL work programme that prioritises specific policy-relevant topics for MDL projects. Key focus can change on year-to-year basis and may include areas such as globalisation, enterprise productivity and performance, and innovation. For instance, globalisation can be examined by linking Structural Business Statistics (SBS) with the EuroGroups Register (EGR) to gather information on characteristics of enterprises belonging to multinational enterprise (MNE) groups. Similarly, enterprise performance can be assessed by integrating data from SBS and International Trade in Goods Statistics (ITGS).

Step II: The second step aims at further harmonisation by establishing a common set of definitions, variables, breakdowns, and populations that enable the development of centrally or jointly executed MDL syntax ('language'). This syntax will be run decentralised by participating National Statistical Authorities (NSAs) using a consistent framework across Member States. This step requires uniform naming of variables for MDL purposes, harmonising concepts and data domains, and the development of common syntax for specific MDL projects. By adopting a standardised approach, Member States and the EU can address emerging policy issues more quickly and efficiently, reducing development time for MDL projects.

The methodology addresses several challenges, including ensuring unit consistency across datasets, managing data confidentiality and privacy concerns, and aligning different statistical domains with varying definitions and content. Furthermore, the MDL strategy accommodates the diverse technological landscapes of NSAs, allowing for flexible implementation of the MDL syntax using various software tools. Through this coordinated approach, Eurostat aims to build a sustainable, re-usable MDL framework that enhances the ESS's capacity to produce more timely, coherent, and policy-relevant statistics in a cost-efficient way.

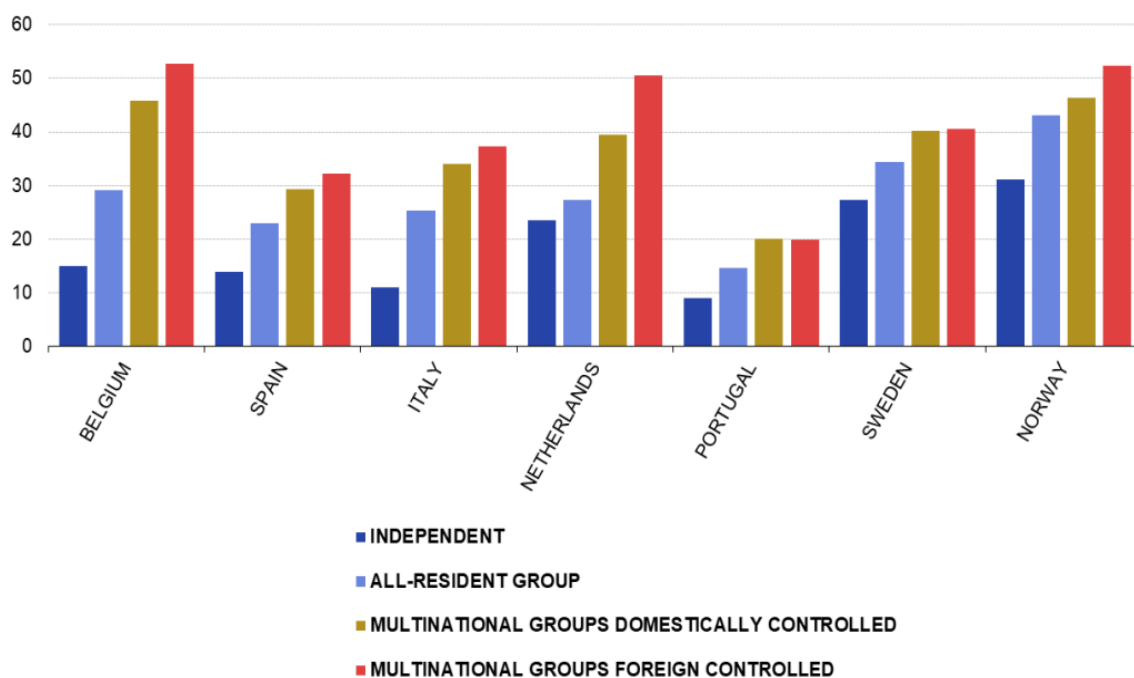
3. RESULTS AND PRACTICAL APPLICATION

The first project in the ESS MDL work programme will include a detailed proposal for data production designed to produce new statistical indicators on enterprises that import critical raw materials (CRMs) and important goods, such as solar technologies and energy storage systems. By capturing data based on origin (EU vs non-EU), enterprise size, group status, and economic activity, the data will help uncover the extent to which different sectors rely on these strategic imports. This information will provide a clearer understanding of the role CRMs and intermediate and final goods play in the production and competitiveness of European businesses.

In addition, the data production will include information on trade relations with major non-EU partners, including countries like the United States, China, and others. This aspect aims to provide a nuanced view of the European supply chain structure, particularly in terms of dependencies and trade dynamics. Another critical focus area is the analysis of trade intensity, evaluating how various levels of import and export activities influence enterprise performance, with key performance indicators such as net turnover and gross operating surplus. By assessing trade intensity, the tables will help illustrate how engagement in international trade can impact different types of enterprises, ranging from small businesses to enterprises belonging to large multinational groups.

Furthermore, the proposed data production will include a comparison of the performance of traders (importers/exporters) with non-traders, highlighting the influence of global trade on business success. It will also delve into enterprise performance by group status, examining differences between independent enterprises and enterprises belonging to national and multinational groups. The data will be compiled by a number of European NSAs beginning in 2025 and will cover data from the reference period 2021-2024.

The results of a similar successful MDL exercise, as illustrated in Figure 1, can serve as an example of the type of data that can be produced via MDL. These results indicate that enterprises belonging to MNE groups consistently offer the highest average wages across all analysed countries, often significantly surpassing those provided by independent enterprises. Enterprises belonging to domestically controlled MNE groups also tend to offer higher wages than independent enterprises, although not as high as their foreign-controlled counterparts.



Source: EuroGroups Register and SBS data

eurostat

Figure 1. Average wage and salary per employee or self-employed person in thousands of Euro, 2020

By harmonising data from diverse sources, this comprehensive data production aims to support evidence-based policymaking within the EU, providing valuable insights into the intricate dynamics of European businesses within the global market. This effort will enable policymakers to understand better the complexities of international trade and its effects on the European economy, informing strategies for enhancing the competitiveness and resilience of European enterprises.

4. MAIN FINDINGS

Applying the proposed MDL strategy is expected to enhance the depth of European business statistics. By integrating multiple data sources, such as the SBS and ITGS, and information from EGR and business registers, the MDL strategy will enable a more comprehensive analysis of enterprise activities, particularly in the context of globalisation and trade. The ability to link data across various domains will provide policymakers and researchers with richer insights into the operational characteristics of enterprises, including their import and export intensity, reliance on critical raw materials, and performance by group status and size class.

Despite its numerous benefits, MDL techniques also have some potential disadvantages. Harmonising data across different Member States and domains presents methodological challenges, such as ensuring consistency in definitions, variable breakdowns, and population coverage. Additionally, safeguarding data confidentiality while linking microdata from various sources requires careful consideration and management. Finally, if there is only a limited number of overlapping entities between statistical surveys, grossing-up techniques must be developed.

However, the overall advantages of this strategy are clear: the MDL framework aims to deliver more comprehensive European statistics with enhanced accuracy and timeliness. This improvement supports evidence-based policymaking and aligns with the ongoing goal of reducing the administrative burden on businesses, as it leverages existing data rather than relying on additional surveys.

5. REFERENCES

- [1] Eurostat. Economic indicators of enterprises belonging to multinational enterprise groups. Statistics Explained article, 2023

Enhancing housing surveys in Overseas Departments using Deep Learning and Satellite Data

Thomas Faria, Raya Berova & Clement Guillo (Insee)

1. INTRODUCTION

The Localized Buildings Directory (RIL) comprises a list of housing units pinpointed by their x, y coordinates. While this directory is comprehensive in metropolitan France, it's not the case in the Caribbean, where there is a lack of reliable sources for this registry. An annual cartographic survey is conducted in the overseas departments (DOM), where surveyors identify certain housing areas on the ground to augment the RIL. The accuracy of the RIL is crucial as it directly influences the quality of the population census and subsequent demographic estimates.

Calibrating the necessary workload in each area presents challenges, especially with rapid new housing developments, including in informal zones. These changes can quickly overwhelm the workload allocated to surveyors. This challenge underscores the importance of using earth observation data. Through satellite imagery, labeled using INSEE's data, deep learning segmentation algorithms are developed to detect housing. By comparing algorithm outputs at different times, areas requiring more intensive surveying efforts in a given year can be identified.

The application of these methods is particularly pertinent in light of the current situation in Mayotte, where conducting a complete census is not feasible this year due to risk factors. Similarly, it is beneficial for newly established institutes such as the National Institute of Saint-Martin, which lack a housing database essential for assessing urban expansion. In both cases, population estimates for each area could be derived from the number of pixels identified as housing on satellite images.

The project utilizes images data from Pleiades (0,5m resolution) and Sentinel 2 (10m resolution) satellites. A primary concern is the availability of these data, particularly their acquisition frequency. Image labeling before algorithm training is a critical step, though it faces challenges due to the imperfections of current data sources. Historical RIL provides precise geolocations but does not reflect the actual area occupied by housing. On the other hand, IGN's topographic database (BDTOPO) offers housing polygons that align with the concept of surface area but are difficult to date accurately.

A method for practical implementation is in the process of development. This method could be integrated prior to the cartographic survey, thereby enhancing the survey's organization and efficiency.

2. METHODOLOGY

A pipeline structured around five key phases was developed to enable deep learning models to effectively detect housing structures from satellite imagery and make the results available to users.

The first phase involves the collection and processing of satellite images from the Pleiades and Sentinel 2 satellites. Sentinel 2 data are openly available and free to access, whereas for Pleiades data, we established a partnership with Dinamis to acquire (freely) satellite imagery for specific regions. To ensure optimal training, the raw images undergo preprocessing steps such as splitting them into smaller tiles and removing any visual obstructions, such as clouds. This preprocessing is essential to prepare the images for further analysis and to ensure that the deep learning models receive clean, usable data.

Next, we move on to data labeling. In this phase, the processed images are annotated, assigning housing labels to specific regions based on the NSI's existing datasets (see Figure 1). This annotation is crucial, as it provides the ground truth needed to train our segmentation model. However, we acknowledge that the database used has imperfections, which results in some inaccuracies in the annotations. The main challenge is obtaining annotations that correspond to a specific date, aligning closely with the date of our satellite images.



Figure 1. Annotations from BDTOPO

After labelling the data, the next step involves fine-tuning a deep learning model that has demonstrated excellent performance in segmentation tasks: Segformer [1]. Segformer builds on recent advances in machine learning by applying transformer architectures, originally developed for natural language processing, to image segmentation. Unlike traditional convolutional neural networks (CNNs), Segformer treats images as two-dimensional sequences and uses transformer-based structures to process them. This approach, inspired by the work of Vaswani et al. [2] and extended by Xie et al. [1], enhances the model's efficiency and performance, making it more powerful than previous CNN-based segmentation models. Fine-tuning such a model with satellite images requires GPU(s).

Then, inference is performed on new satellite images, generating segmentation masks that identify housing areas. A key aspect of our approach is comparing these masks across images taken at different times for the same geographical locations. This allows us to detect changes over time, such as new housing developments or expansions. The ability to track these changes is crucial for updating and organizing housing surveys and urban planning.

To facilitate the dissemination of our results to our teams, we developed a tiny web application that allows users to visualize housing predictions and monitor urban growth. The application is accessible at <https://cratt.lab.sspcloud.fr/>.

3. RESULTS AND MAIN FINDINGS

The results obtained are extremely promising and demonstrate a potential for large-scale housing detection. The segmentation model achieved an Intersection over Union (IOU) score exceeding 0.75, reflecting its high accuracy in identifying housing structures from satellite images. This high level of precision ensures that even small-scale constructions are detected and mapped, making the model particularly useful for tracking both individual building developments and broader urban expansion.

In terms of prediction quality, the model produced detailed outputs that aligned closely with known urban expansion trends. For instance, in areas experiencing significant urban sprawl over several years the model was able to capture these developments consistently. Furthermore, its ability to detect individual buildings highlights its effectiveness not only for regional monitoring but also for pinpointing new constructions at the level of individual plots (see Figure 2 slide 34).



Figure 2. Predictions for 2023 on the background of the 2020 (left) and on the background of the 2023 (right)

This experimentation has demonstrated the potential of satellite imagery as a valuable tool for enhancing official statistics, particularly in supporting census surveys. The integration of deep learning models has already proven beneficial in lighten the workload. However, several challenges remain that need to be addressed for wider adoption in production.

One key challenge is securing partnerships to regularly access dated Pleiades images, as reliable and timely data is essential. Additionally, the project highlights the need for specialized human expertise, such as data scientists and data engineers, which might be relatively rare within official statistics institutes. Lastly, the technical demands of this approach, including the requirement for advanced IT resources like GPUs, pose logistical hurdles that must be overcome for sustained scalability.

Despite these challenges, the success of this experiment highlights the promising role of satellite data and machine learning in improving traditional data collection methods in official statistics.

4. REFERENCES

- [1] Vaswani & al (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.
<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

- [2] Xie al (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. <http://arxiv.org/abs/2105.1520334>

Estimating CO₂ emissions in Italy with local regression models

Antonio Balzanella, Gianmarco Borrata, Raffaele Mattera & Rosanna Verde
(Department of Mathematics and Physics, University of Campania “Luigi Vanvitelli”, Italy)

1. INTRODUCTION

The accurate estimation of CO₂ emissions is critical for addressing climate change, particularly in locations where direct measurements are unavailable or unreliable. Many areas lack comprehensive monitoring systems, which results in significant gaps in the local understanding of CO₂ emissions. These gaps hinder effective policymaking and the implementation of mitigation strategies. Therefore, the need for reliable CO₂ estimation methods in unobserved locations is paramount for a complete and accurate global assessment of emissions. To address these challenges, policymakers and organizations worldwide have developed various methodologies and tools for estimating CO₂ emissions. One prominent example is the Emissions Database for Global Atmospheric Research (EDGAR), developed by the Joint Research Centre (JRC) of the European Commission. EDGAR integrates emission data from various sources, including energy use, industrial activities, and transportation, to provide a global estimate of greenhouse gases, including CO₂. This system helps fill gaps in regions with limited or no direct measurements, enabling policymakers to track emissions and identify priority areas for intervention. To address this issue, we build on previous studies [1, 2] by adopting a data augmentation approach that leverages commonly used covariates—such as economic activity, transportation, energy consumption, and population density—that are known to correlate with CO₂ emissions (e.g. see [3] for a review on the subject). Indeed, as economies grow, industrialization and consumption typically increase, leading to greater energy use and, consequently, higher emissions. Also population growth and density drive up energy demand and, consequently, emissions. Therefore, energy consumption is a critical determinant, directly linked to CO₂ emissions [4]. Following the same reasoning, variables related with industrial output—typically proxied by the share of the economy involved in manufacturing and heavy industries—also significantly influences emissions. Indeed, locations with high levels of industrial activity tend to have greater energy consumption, especially from carbon-intensive sources [5]. Finally, it is important to mention transportation sector (number of vehicles, trains and train stations, airplanes traffic, etc.) which is one of the largest contributors to CO₂ emissions worldwide. All these covariates allow for the prediction of emissions in locations where direct observations are missing. However, for locations with unknown emissions, a reliable model is not readily available due to the absence of direct measurements. While one could employ the parameters of a global model to estimate emissions in these areas, such an approach is often unsatisfactory given the complexity of the problem. Global models assume homogeneity across regions and fail to capture local variations in emission patterns, which are influenced by specific regional and environmental factors. Therefore, in this paper we investigate if more accurate predictions of CO₂ emissions in locations without observed data can be obtained by estimating cluster-specific models which account for the peculiarities of the clusters while estimating emissions.

2. DATA AND METHODOLOGY

We aim to predict the per capita CO₂ emissions of Italian municipalities. Therefore, we collect data on CO₂ emissions and related explanatory for 7,920 municipalities across Italy. The dependent variable is CO₂ emissions per capita, while the independent variables consist of estimated population, the number of buildings, farms, fuel stations, and trains. With the aim of estimating local rather than global models, we propose to employ the clusterwise regression [6] for estimating k ($c=1,...,k$) different models for distinct k groups within the data. We assume that the k groups are unknown and are identified jointly with parameters of interest with the statistical model. The clusterwise approach accounts for heterogeneity in the relationships between CO₂ emission patterns and the covariates across different areas. Each cluster-specific model is tailored to the unique characteristics of each group, leading to more accurate and localized predictions. This method overcomes the drawbacks of applying a single global model, providing a more nuanced understanding of emissions in locations where data is sparse or unreliable. To successfully implement this modelling approach, we have to deal with a challenge related with the matching of the observation in the most similar cluster. For this aim, we consider k -nearest neighbours approach with different values of k , but other approaches can be also considered. Indeed, for locations where CO₂ data is unavailable but covariates are observed, we first assign each data point to the most similar cluster based on covariate patterns. The model specific to that cluster is then used to estimate CO₂ emissions for the unknown point. It is interesting to mention that the proposed approach aligns with the idea in semisupervised learning of leveraging both labeled and unlabeled data to improve predictions, in contexts where labeled data, that is CO₂ emissions in our setting, is limited or unavailable. This method provides an innovative solution for CO₂ estimation, enhancing the precision of emission forecasts in regions with scarce or unreliable data, which is essential for global environmental monitoring and policy formulation. We compare the predictions obtained with the clusterwise regression model with those of the multiple linear regression, employing a global rather than local model. To evaluate the accuracy of the predictions, we divide the dataset into a training set (75%) and a testing set (25%). We evaluate both the fitting and predictive accuracy of the models, considering both Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) losses.

3. MAIN RESULTS

Before estimating the clusterwise regression, we need to choose the optimal number of clusters. Figure 1 shows the Total Residual Sum of Squares obtained for different number of clusters k . We select the number of clusters according to the elbow criterion, leading to the choice of $k=5$. We then evaluate the differences in terms of fitting between the global model (Multiple Linear Regression, MLR) and the $k=5$ local models, estimated with the clusterwise approach. Table 1 shows the obtained results, considering adjusted R^2 , RMSE and MAE computed on training data. The best model is highlighted in bold in Table 1. While the R^2 associated with the global model is equal to 0.22, the goodness of fit of the models within the k clusters is found to be between 0.86 and 0.99. Also the RMSE and MAE are dramatically lower for the clusterwise model compared to the global model estimated with MLR.

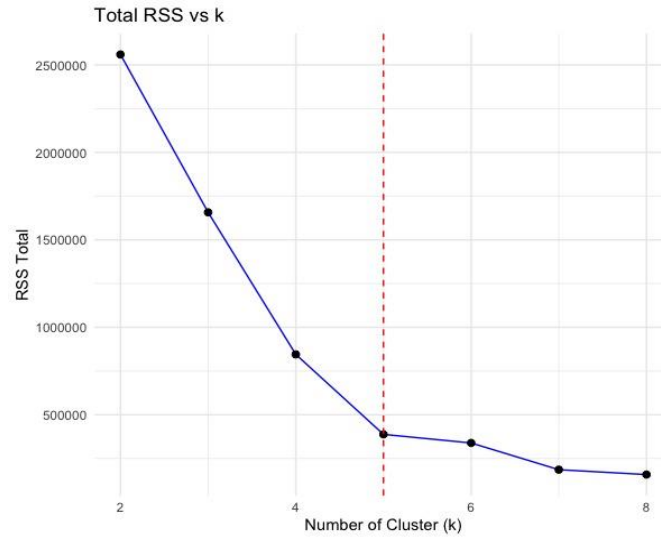


Figure 3. Selection of the number of clusters k with the Elbow method.

Table 1. Prediction accuracy of the considered models in the training dataset.

Model	Cluster	Adjusted R^2	RMSE	MAE
MLR	-	0.22	37.39	7.82
CLWR	1	0.99	6.47	4.55
	2	0.99	3.76	2.92
	3	0.86	21.39	6.98
	4	0.88	19.00	6.87
	5	0.91	15.60	6.51

The largest benefit in terms of fitting is achieved for the cluster 2. Table 2 shows the results obtained for the predictions task in the testing set. We first find that the predictive accuracy associated with the clusterwise approach is considerably larger than the global model obtained with the MLR. We evaluate both k -NN with $k=1$ and $k=5$ and find that considering larger number of k in the matching task leads to the largest improvements.

Table 2. Prediction accuracy of the considered models in the testing dataset.

Model	Matching	RMSE	MAE
MLR	-	33.47	8.22
CLWR	k -NN ($k=1$)	30.16	7.56
	k -NN ($k=5$)	29.42	6.75

We provide evidence on the usefulness of the proposed approach considering data from Italy. In particular, we train and test the accuracy of the predictions obtained with model using data with fully available observations. We then provide a map to visualize estimated CO₂ emissions patterns also including locations with unknown emissions. Extensions to clusterwise regression can also be considered. For example, research may adopt an ensemble approach, combining predictions from multiple cluster-specific models to provide a more robust estimate of emissions. This dual approach of clusterwise regression and ensemble modeling offers flexibility and enhances predictive accuracy, particularly in locations with diverse environmental and socio-economic conditions. Alternatively, methods exploiting spatial clustering can also be adopted in this context.

4. REFERENCES

- [1] Ahn, D. Y., Goldberg, D. L., Coombes, T., Kleiman, G., & Anenberg, S. C. (2023). CO₂ emissions from C40 cities: citywide emission inventories and comparisons with global gridded emission datasets. *Environmental Research Letters*, 18(3), 034032.
- [2] Moran, D., Pichler, P. P., Zheng, H., Muri, H., Klenner, J., Kramel, D., ... & Gurney, K. R. (2021). Estimating CO₂ emissions for 108,000 European cities. *Earth System Science Data*, 2022, 1-23.
- [3] Mardani, A., Streimikiene, D., Cavallaro, F., Loganathan, N., & Khoshnoudi, M. (2019). Carbon dioxide (CO₂) emissions and economic growth: A systematic review of two decades of research from 1995 to 2017. *Science of the Total Environment*, 649, 31-49.
- [4] Aller, C., Ductor, L., & Grechyna, D. (2021). Robust determinants of CO₂ emissions. *Energy Economics*, 96, 105154.
- [5] Adebayo, T. S., & Beton Kalmaz, D. (2021). Determinants of CO₂ emissions: Empirical evidence from Egypt. *Environmental and Ecological Statistics*, 28(2), 239-262.
- [6] Park, Y. W., Jiang, Y., Klabjan, D., & Williams, L. (2017). Algorithms for generalized clusterwise linear regression. *INFORMS Journal on Computing*, 29(2), 301-317.

Evaluating an ESCO skill classifier

Gabriele Marconi & Inès Baer (ADEM, Luxembourg)

1. INTRODUCTION

Data on job advertisements have become increasingly available over the last decade. These data are considered the main source of information on skill demand across economies [1] [2]. Therefore, statisticians, data scientists and researchers across academia, public and private institutions and international organisations developed a variety of skill classifiers. A skill classifier, in this paper, is defined as an algorithm that automatically reads job ads and identifies the skills thereby contained. In this way, it is possible to produce statistics on how many job ads require certain skills in given economies or data sources. Skill classifiers usually classify skills according to existing taxonomies. In this paper, we focus on the most granular level of the ESCO skill taxonomy, which is widely used in this field.

Skill data can be used for many applications, for example for labour market analysis or matching algorithms [3] [4]. For these applications, it is necessary to evaluate the classifiers. However, skill classifier evaluations are uncommon in practice and in the literature. A recent literature review on skill classifiers [2] dedicate only a short section to classifier evaluation, with only one referenced paper [3]. This paper evaluates only based on “recall-oriented” metrics that do not consider negatives in the evaluation, something we are typically not willing to do because it is concerning to have classifiers finding skills that are not actually in the ads. This scarce attention to evaluation can be surprising at first, but it is due to some challenges that are objectively difficult to overcome. To understand why, consider that the most granular level of the ESCO skill taxonomy [6], a common target for skill classifiers, contains around 14000 different skills. This has two important implications:

- An ESCO skill classifier is composed by a large multiplicity of single-skill classifiers (around 14000), and it is nearly impossible to evaluate them all and present the evaluation results for each skill in an understandable way
- The skills are narrowly defined, creating class imbalance. Most skills are not found at all, and the most common skills are found only in a small percentage of ads. This makes it difficult to evaluate a single skill, because even reading a few hundred of ads (something which is very time consuming) can yield only few occurrences of the skill.

This paper presents a method to evaluate the skill classifiers used by the Luxembourgish public employment service (ADEM) while overcoming the problems of multiplicity of skill classifiers and of class imbalance. The results of the evaluation conducted in this way are also presented.

2. METHODOLOGY

2.1. Skill extraction at ADEM

The way a classifier works is irrelevant to the evaluation methodology presented in this paper, but it is useful to briefly describe it for the sake of context. ADEM has two classifiers in place, both using methods common in the literature [2]:

- The first classifier associates a list of tags (or keywords/labels) to each skill and operates a fuzzy search for these keywords in the text of the ads. The list of tags is based on the “alternative labels” provided by ESCO, and modified based on the developers’ experience with previous skill extraction and based on potential equivalent tags identified through artificial intelligence. The skill data extracted through this algorithm have been published by ADEM after evaluation.
- The second classifier breaks an ad into sentences and produces the embedding for each sentence based on a S-BERT model [7]. Then it produces the embedding for each skill based on the labels and descriptions produced by ESCO. The skill is considered a requirement in the ad if its embedding is sufficiently close to the embedding of a sentence. The skill data extracted through this algorithm have not yet been evaluated and published.

2.2. Defining the objective of the evaluation

We would like to know how accurate the classifier is for a “typical” skill. At this point, it must be noted that most ESCO skills are not found by the classifier (so the “typical” skill is never found). This does not mean that these skills are irrelevant – we still want to know if they are not found because they are not requested in job ads or because our classifier is inadequate. However, surely our definition of “typical” skill must cover the skills that are empirically most interesting, because they are found in many ads.

The reasoning in the following paragraph leads us to define three key indicators for our evaluation:

- The average accuracy of the algorithm in finding the 102 most-often-found skills (hereafter, “top 100 skills”) in the dataset
- The average accuracy for the other skills in the top 300 (the 103rd to 304th)
- The average accuracy of the algorithm to find any other ESCO skill

The choice of evaluating on 102 skills (instead of 100) per stratum was made to simplify data annotation. The design of the data annotation process goes beyond this abstract, but it offers many lessons and it will be discussed at NTTS2025.

2.3. How to evaluate a skill classifier

We aim at having some performance information on all the top 100 skills, on about 100 of the following 200 skills, and on about 100 of the remaining skills. In this way, we will be able to produce the most robust accuracy statistics for the top 100 skills, slightly less robust for the top 300, and again slightly less robust for the full population of skills. Therefore, we randomly select our “skill sample”, made of 102+102+102=306 skills.

We would like to evaluate every skill based on at least 3 ads. This ensures that, for every skill included in the analysis, we can draw some practically useful conclusions (even though they will lack statistical power). For example, if we get 0/3 for a skill, we can already suspect that something is wrong.

To do this, we use a stratified matrix sampling. We draw a random “ad sample” of 102 ads, and for each ad, we randomly select 9 skills among those that we want to evaluate. This yields 918 ad-skill matches, allowing us to fulfil the need for 3 ads for each of the 306 different skills.

However, this still leaves open the problem of class imbalance. Out of 918 ad-skill matches, we only have 19 positive predictions. To solve this problem, we include in the evaluation sample all the skills extracted by the algorithm from the 102 ads in the ad sample. In other words, we stratify to include 100% of the predicted skills in the ad sample. This increases the number of ad-skill pairs that we need to annotate to 1618, by adding 700 predicted cases. For each of this ad-skill pairs, we needed to read the ad and annotate if the skill was required in the ad or not.

Once the annotation was finished, we calculated the normalised confusion matrix for each skill. This matrix contains the estimated proportion of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). The calculation of this proportion needs to account for the fact that positive predictions are oversampled. To illustrate this, we will show the example of the calculation of the proportion of true positives for a skill s (tp_s) below.

In the random sample of 3 ads, we always have some negative predictions for every given skill. These ones cannot be true positives by definition, so the proportion of true positives among these negative-predicted cases (tp_0) is 0. The predicted-positive cases in the random and stratified sample will be divided between true positives and false positives, allowing to calculate the proportion of true positives among positive predictions (tp_1). To get an estimate of tp , it is necessary to weight tp_0 and tp_1 by, respectively, the known proportion of negative predictions for skill s in the dataset (tn^*+fn^*) and that of positive predictions (tp^*+fp^* , where the star indicates that this is a population value). Therefore:

$$tp = (tn^* + fn^*) \cdot tp_0 + (tp^* + fp^*) \cdot tp_1$$

By adopting a similar procedure, tn , fp and fn can be calculated for each skill and average across any given set of skills. This leads to an “average confusion matrix” for the top 100, top 300 and for the full set of skills. In turn, that makes it possible to calculate an “average accuracy” besides any other common performance indicator (e.g. f1-score).

3. RESULTS AND PRACTICAL APPLICATION

The results of the evaluation are reported in Table, which reports the normalised confusion matrix for the top100, other skills in the top 300, all other skills.

Table 1. Evaluation results

Top 100	Skill not required	Skill required
Prediction is negative	$tn = 81.6\%$	$fn = 12.0\%$
Prediction is positive	$fp = 0.8\%$	$tp = 5.7\%$
Other skills in Top 300		
Prediction is negative	$tn = 85.4\%$	$fn = 13.1\%$
Prediction is positive	$fp = 1.2\%$	$tp = 0.3\%$
All other skills		
Prediction is negative	$tn = 94.7\%$	$fn = 5.2\%$
Prediction is positive	$fp = 0.02\%$	$tp = 0.04\%$

The main result is that the accuracy of the model is overall satisfactory. For example, among the top 100 skills, the average accuracy is 87.3%. This compares well to the accuracy of a simple model that always predicts the most common class (82.4%), which is often considered a good benchmark of performance [8]. This had a major practical implication: publishing the data.

Another important result is that false positives, something that worried ADEM when setting up the classifier, are a minor problem (0.8% on average among the top 100). In contrast, false negatives, i.e. not finding skills that are actually required, is the main issue faced by the text matching classifier. This result also had major practical implications, since it stimulated ADEM to experiment with a second classifier (based on S-BERT). This classifier extracts skills even in absence of a direct match with a tag in the text, which can be both a strength and a risk (the results from its evaluation may be available for NTT2025).

Beyond the top 100 skills, the ability of our skill classifier to capture true positives decreases substantially. Additional insight by type of skill will be discussed at NTTS2025.

4. MAIN FINDINGS

We found that evaluation of skill classifiers is too rare in the literature. We attribute this to two problems: the multiplicity of skill classifiers; the class imbalance in the data. We developed a method to classify our own ESCO-skill classifier while dealing with these problems, through a stratified matrix sampling of skills and ads.

We find a satisfactory average accuracy of 87.3% across the top 100 skills, which led ADEM to publish the extracted skills as open data. However, we also find that our evaluated algorithm is too conservative, missing to find many skills that were actually required in the job ads.

5. REFERENCES

- [1] Cedefop (2019). Online job vacancies and skills analysis: a Cedefop pan-European approach. Publication Office of the European Union.
- [2] Khaouja, I., et al. (2021). A survey on skill identification from online job ads. IEEE Access, doi:10.1109/ACCESS.2021.3106120
- [3] Luxembourg Employment Agency (2024). Job insights, <https://jobinsights.lu>
- [4] Naya, V., et al. (2022). Designing labor market recommender systems: the importance of job seeker preferences and competition, HAL Open Science, <https://inria.hal.science/hal-03540319>
- [5] Bholá, A., et al. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification, doi:10.18653/v1/2020.coling-main.513
- [6] European Commission. (2018). ESCO implementation manual. Publication Office of the European Union.
- [7] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, <https://arxiv.org/abs/1908.10084>
- [8] Monarch, R. (2021). Human-in-the-loop machine learning. Manning.

Evaluating the Completeness of Business Databases: A Comparison with Official Records Using Web Scraping Techniques

Xinhui Huang & Josep Domenech (Universitat Politècnica de València, Spain)

The Orbis database, a global corporate information repository managed by Bureau van Dijk, is a widely used resource for accessing company data across multiple countries. Similarly, one of its regional counterparts, the Sistema de Análisis de Balances Ibéricos (SABI), provides extensive company information for Spain and Portugal. This study evaluates the comprehensiveness of the SABI database by comparing it to the BORME, the official gazette for business registrations in Spain. Employing web scraping techniques, we identified companies listed in BORME that are absent from SABI. Our findings indicate that SABI covers only 38.3% of companies established between 2010 and 2023, with notable underrepresentation of younger firms, smaller enterprises, and certain sectors. Additionally, we observed a survivorship bias, with dissolved companies progressively less likely to be retained in the database over time. These results underscore critical biases in SABI, suggesting that researchers should exercise caution when using this database for economic and business research

1. INTRODUCTION

Bureau van Dijk's Orbis database serves as a comprehensive global resource for corporate information, offering detailed data on millions of companies worldwide. This extensive dataset is widely used by institutions and researchers worldwide as it facilitates comparative business studies and economic research across various countries. Within this broader framework, the Sistema de Análisis de Balances Ibéricos (SABI) focuses specifically on companies in Spain and Portugal. Like Orbis, SABI is widely used in academic research and business analysis [3, 4, 5], playing a crucial role in the regional economic and financial landscape. Despite its extensive use, concerns about its completeness and representativeness persist. Previous studies have highlighted potential biases and gaps in business databases similar to SABI [2, 1]. For instance, Kalemli-Ozcan et al. [2] discuss issues related to the representativeness of firm-level databases like Orbis, noting that they may not fully capture the population of firms, especially smaller ones.

This paper assesses these concerns by comparing SABI with the Boletín Oficial del Registro Mercantil (BORME), Spain's official business registry. BORME provides a comprehensive record of all registered companies in Spain, making it a valuable benchmark for assessing the coverage of SABI. Using web scraping techniques, we constructed a dataset from BORME and identified companies that are missing from SABI. Our goal is to understand the biases and gaps in SABI to improve the quality of economic analyses that rely on such databases.

2. METHODOLOGY

Data for companies founded between 2010 and 2023 was collected from two data sources: SABI and BORME. SABI is a commercial database that provides data in a tabular format, regularly updated, including its coverage. In its version 145.00 (update 293), it includes 1,911,775 companies in Spain, with 502,123 of these established between 2010 and 2023.

On the other hand, BORME data is less conveniently structured, as it consists of multiple publications in PDF format, accessible via the official website. These publications were downloaded and converted to text using web scraping and crawling techniques. Each publication contains a list of registry entries, with each entry representing a company, identified by its legal name. The entries document specific events, such as company establishment, capital changes, or bankruptcy filings. Changes in company names were tracked to match the most recent name listed in SABI.

For this dataset, around 100,000 publications were downloaded, resulting in the processing of 9,956,791 registry entries from 3,051,505 companies. After filtering out companies not established between 2010 and 2023, 2,917,784 entries related to 1,298,056 companies were retained. Each registry entry was converted into a tabular format and grouped by company, with legal name changes accounted for. Finally, data from SABI and BORME were merged to create the dataset used for analysis.

3. RESULTS

3.1. Coverage by Year of Establishment

SABI covers approximately 38.3% of companies established between 2010 and 2023. Coverage varies significantly by year of establishment, with younger companies being underrepresented. Nearly all companies aged 0-1 years are missing from SABI, and coverage improves for companies older than five years, stabilizing at around 50%.

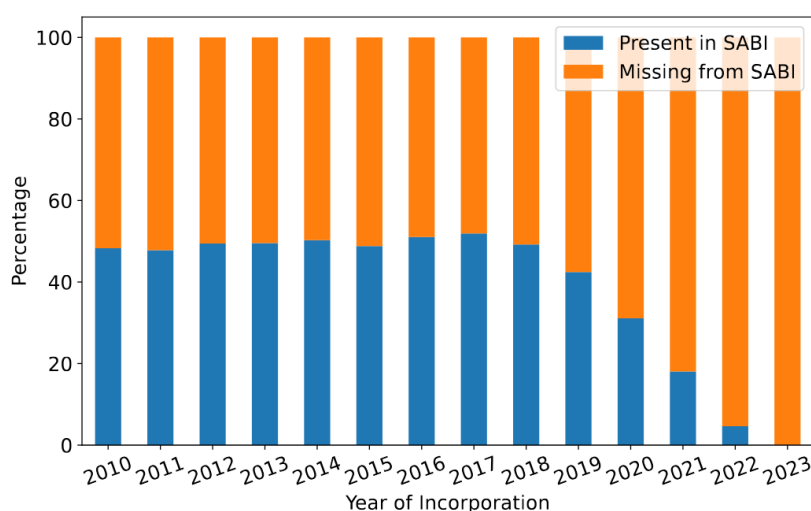


Figure 1. Coverage of established firms by year of incorporation

3.2. Survivorship Bias

We observed a survivorship bias in SABI's data. Among companies dissolved between 2020 and 2023, about 40% are present in SABI, similar to overall coverage. However, for companies dissolved more than four years ago, the likelihood of being included in SABI decreases by approximately 3.7 percentage points per year, indicating that dissolved companies are progressively less likely to be retained in the database over time.

3.3. Geographical and Sectoral Coverage

Geographical coverage varies across provinces, ranging from 65% in Lugo to 18% in Guadalajara, without a clear pattern related to population size or location. Sectoral coverage also varies, with certain sectors like Public Administration and Manufacturing having higher coverage, while others like Electricity and Real Estate are underrepresented.

3.4. Coverage by Legal Form and Initial Capital

SABI tends to favor larger companies. For limited liability companies (Sociedad Limitada), only 37.3% are included, while 50.5% of public limited companies (Sociedad Anónima) are covered. Additionally, companies present in SABI generally have higher initial capital compared to those absent, indicating a bias towards larger firms. SABI covers approximately 38.3% of companies established between 2010 and 2023. Coverage varies significantly by year of establishment, with younger companies being underrepresented. Nearly all companies aged 0-1 years are missing from SABI, and coverage improves for companies older than five years, stabilizing at around 50%.

4. MAIN FINDINGS

Our study reveals significant gaps and biases in the SABI database. Younger firms, smaller enterprises, and certain sectors are underrepresented. There is also a survivorship bias, with dissolved companies less likely to remain in the database over time. These biases can impact the representativeness of analyses based on SABI data, suggesting that researchers should use caution and consider supplementary data sources or adjust methodologies when using SABI for economic research.

ACKNOWLEDGEMENTS

This work was supported by the Generalitat Valenciana under Grant CIAICO/2023/272; the Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033) and ERDF/EU under Grant PID2023-152106OB-I00; and the Ministerio de Educación y Formación Profesional.

5. REFERENCES

- [1] Bajgar, M., Berlingierii, G., Calligarisi, S., Criscuoloi, C., and Timmisim, J. (2020). Coverage and representativeness of Orbis data. OECD Science, Technology and Industry Working Papers, 2020/06.
- [2] Kalemli-Ozcan, S., Sørensen, B. E., Villegas-Sanchez, C., Volosovych, V., and Yesiltas, S. (2024). How to construct nationally representative firm-level data from the Orbis global database: New facts on SMEs and aggregate implications for industry concentration. American Economic Journal: Macroeconomics, 16(2), 353–374.

- [3] Martínez-Matute, M. and Urtasun, A. (2022). Uncertainty and firms' labour decisions. Evidence from European countries. *Applied Economics*, 25(1), 220–241.
- [4] Rizov, M., Vecchi, M., and Domenech, J. (2022). Going online: Forecasting the impact of websites on productivity and market structure. *Technological Forecasting and Social Change*, 184, 121959.
- [5] Sánchez-Infante Hernández, J. P., Yañez-Araque, B., and Moreno-García, J. (2020). Moderating effect of firm size on the influence of corporate social responsibility in the economic performance of micro-, small-and medium-sized enterprises. *Technological Forecasting and Social Change*, 151, 119774.

Experiences from experiments with smart surveys in the Smart Survey Implementation project

Maren Fritz, Florian Keusch (University of Mannheim, Monica Perez (Istat), Nina Berg (SSB) & Patrick Lusyne (Statistics Belgium)

1. INTRODUCTION

The Smart Survey Implementation (SSI) project investigates how to establish an end-to-end methodology for conducting smart surveys in official statistics. Smart surveys combine surveys with smart elements from sensors, for example, the use of the smartphone camera for receipt scanning in a household budget survey.

2. METHODOLOGY

In 2024, fieldwork experiments were conducted in Norway, Belgium, and Germany to test various options in how to design and field a smart survey as part of national household budget surveys and time use surveys. In this presentation, experiences and results from three large field tests are shown, concentrating on the effect of design features on recruitment and response rates to smart surveys. For example, the use of different platforms from which the data collection app could be accessed and the use of CATI interviewers was tested in Norway. In Belgium and Germany different recruitment protocols were tested, including the use of different invitation letters that focus on individual features of the survey and secondary modes.

3. RESULTS AND PRACTICAL APPLICATION

We find that recruitment rates vary across countries, and that the differences between within-country experimental conditions are smaller.

4. MAIN FINDINGS

The presentation will also focus on the more practical experiences of conducting a smart survey, including country-specific experiences in designing and organising the fieldwork for smart surveys.

Farm register for agricultural sample surveys: toward a systematic integration of administrative sources

Damien Babet (French Ministry of Agriculture)

A farm register is necessary as a sampling frame to draw random samples for agricultural surveys, but administrative registration of farms is often lacking compared to firms in other industries. In previous decades, our farm register was only updated every decade with the agricultural census. In the last ten years, the increase in quality and availability of administrative data allowed for much more frequent updates, associated with a heavy workload for data preparation and integration. The resulting register has known issues, including difficulties in detecting cessation, creation, and correctly identifying statistical units when there are multiple administrative definitions.

We describe the ongoing project of building a new software architecture with automated administrative data integration. The project includes:

- Automated API updating of contact information from new national databases
- Integrated pipeline for data qualification, formatting and matching
- Priority rules to resolve conflicting information between sources
- Database historisation to allow for analysis and correction
- Integration of feedback from survey field managers

The project will go soon in software development phase. We describe the challenges and lessons of translating official statistics objectives and constraints into general and detailed specifications for an IT project.

Finding the usual environment of tourists in Mobile Phone Data: a case study applied to the same-day visits

Barbara Dattilo, Mascia Di Torrice & Mariangela Sabato (Italian National Institute of Statistics, Istat)

1. INTRODUCTION

Tourism statistics constitute a privileged application area for using the so-called *big data*. In particular, acquiring data from mobile phone providers can reveal enormous potential for information on tourist movements. Among the various experimental initiatives conducted by European states for the use of mobile phone data (MPD), the Italian National Statistical Office (Istat) started a collaboration agreement with Vodafone in 2019 [2], which continued in 2023 also with Motion Analytics [1]. The events generated by the Vodafone network monitoring systems can identify a continuous flow of data between the smartphone and the network, with enormous territorial and temporal detail. Vodafone data - non-personal and anonymized in full compliance with current legislation on privacy - are high-density event-based data generated by the mobile radio network (30 billion records per day referenced in time and space). This work explores the use of MPD as a potential auxiliary source supporting the official statistics on domestic same-day visits (SDV). As is known, the International Recommendations on Tourism Statistics [3] define tourism as an activity carried out by the individual outside his/her Usual Environment (UE). The concept of UE in traditional surveys based on direct interviews is operationalized through specific questions that delimit the space and frequency of SDVs. This is difficult to implement in an algorithm developed to extract information from mobile phone data. This work represents the first attempt to measure UE using algorithms based on mobile phone data. The experimentation concerned the domestic SDVs in the Italian regions of Tuscany and Latium, during the third quarter of 2022 and 2023.

2. METHODOLOGY

The testing activities lasted from September 2023 to April 2024. The domestic SDV extraction algorithm was created starting from the joint reflections and experiments of the working group composed of Istat (with thematic, methodological, and IT skills) and Vodafone/Motion Analytics human resources.

The Istat "Trips and Holidays" sample survey (THS) provides the official estimates on domestic SDV in compliance with international definitions [6], which define a SDV as a visit without overnight stays made outside the usual environment and originating from the place of usual residence. The agreements among European Member States led to specific rules for identifying the UE: the crossing of administrative borders, the duration of the visit (at least 3 hours at the destination), and the frequency of the visit: less than once per week. Instead, the use of mobile phone data for estimating the SDV and for comparison with the survey data required ad-hoc definitions and metrics. In particular, for a mobile phone user, the usual residence is set in the prevalent night-time municipality, i.e. the municipality in which the user has spent the most nights in the 12 months of reference, where therefore the so-called "Telephone Residence" is located. To define the prevalent night-time municipality, the night-time presence within the municipality was set for at least six hours, namely between 8 pm and 8 am the following morning. The spatial perimeter of the telephone user's UE, in addition to

the prevalent night-time municipality, also includes all those municipalities identified by combining several months and the number of hours in which the telephone SIM has been seen in a certain municipality in 12 months. Therefore, having fixed the number of months, the municipality belongs to the user's UE if the hours threshold defined for that particular number of months has been exceeded. After several iterations, two monthly thresholds were agreed upon, at 4 and 6 months, and, for each of them, an hourly threshold, identifying two possible scenarios as reported below:

1. "4m-25h": the UE includes the municipalities where the user has spent at least 4 months and a number of hours greater than the number of months multiplied by 25 (hours), plus the municipality of habitual residence;

2. "6m-0h": the UE includes whether those municipalities where the user has spent at least 6 months (continuously) or respectively 3 months and at least 125 hours, or 4 months and at least 100 hours, or 5 months and at least 75 hours, plus the municipality of habitual residence. All the visits within the UE have not to be considered tourists, therefore they are not counted as SDV. Scenarios that include a greater number of municipalities in the UE determine the contraction of the SDV estimate because the set of municipalities where the user is not a same-day visitor increases. In this sense, the 6m-0h UE definition is more "inclusive" than the "4m-25h" definition, including a greater number of municipalities where the user can move, without being a same-day visitor. Consequently, by adopting a 6m-0h UE, a lower number of same-day visitors is expected in comparison with a 4m-25h UE.

To limit the number of possible errors in the attribution of municipalities belonging or not to the UE, a new attribute called "buffer" has also been introduced. This attribute adds to the set of UE municipalities obtained as stated above, all the municipalities bordering a municipality of that set. The longitudinal observation of the data over the 12 months allows us to estimate the municipality of Telephone Residence and define different tests related to the different definitions of UE.

In this work, we compare the results of two more significant tests, from the one that includes more events (SDV) to the least inclusive one. The first, defined as *Test Only-home*, does not consider the UE because it only counts SDVs outside the municipality of Telephone Residence. The second, called *Test UE of the type 6m-0h*, counts visits external to the UE, therefore excluding visits outside both the municipality of Telephone Residence and all the neighboring ones, and municipalities where the user has been frequently in the last year (at least once a month for 6 months, or at least 125 hours in total for 3 months).

Furthermore, the metric was enriched with the additional parameter of the frequency of visits in the month, to adhere even more to international definitions. Only SDVs carried out 1-3 times a month in the same location have been included. The same-day visitor is a resident of Italy (Italian or foreign), and the place where the SDV begins/ends coincides with the municipality of Telephone Residence, where he/she spends the night before and the night after the event.

3. RESULTS AND PRACTICAL APPLICATION

If we estimate the SDV regarding the Only-home case (telephone residence only), without the frequency parameter (test T1 in Figure 1), Vodafone's estimates are much higher than those provided by the THS. The introduction of the concept of UE reduces them by 72% (test T1_buffer_not UE) while considering only visits with a frequency of 1-3 times a month reduces the estimates by 79% (test T1 buffer 1/3t). The joint application of the two criteria further increases the precision of the estimates and 82% of SDVs are excluded from the count. In this way, in Latium, the number of events drops from over 50 to just over 9 million, and in Tuscany from 46 to approximately 8 million (test T2 buffer 1/3t not UE in Figure 1). However, even in the best-case scenario developed, which better defines the events collected by Vodafone by

"reducing" them and making them conceptually and definitively consistent with the SDVs detected by the THS, the mobile phone estimates remain on average approximately 10 times higher than those provided by the survey.

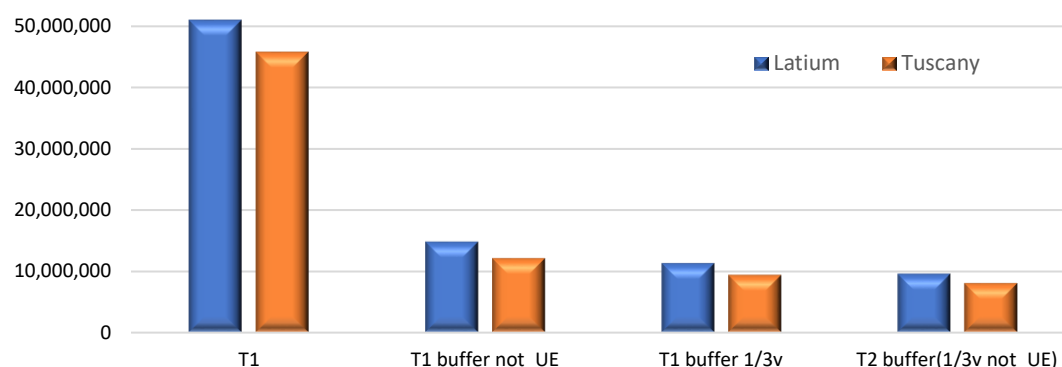


Figure 1. Different tests for defining the UE to estimate domestic SDV towards Tuscany and Latium. Third quarter 2023. (Source: Istat elaboration, acknowledgments: Vodafone, Motion Analytics)

Further comparisons between Istat data and MPD concern the regions of origin of same-day visitors in Latium and Tuscany in the two quarters under examination (Figure 2). The regions of origin detected by THS are significantly lower in number for both Tuscany and Latium, while the Vodafone data covers practically the entire list of Italian regions, including, albeit with residual shares, regions that are decidedly far away to originate a short-range visit (with minimal variations in ranking between 2022 and 2023). Moreover, as expected due to the nature of the visit, in both sources the region of destination occupies the first position in the ranking of the regions of origin. However, in the Vodafone distribution, Latium absorbs almost all of the flows directed to the region, whereas the Istat distribution highlights significant shares of SDVs also in the second region of origin (Campania). On the contrary, SDVs to Tuscany are mainly absorbed by flows from the region itself, both for Vodafone and Istat estimates; furthermore, the first two regions of origin coincide among the sources in 2022 (Tuscany and Emilia-Romagna), but not in 2023, when Istat reports Latium in second place and Vodafone Emilia-Romagna. Note in the Vodafone ranking the category "Non Definito" ("Undefined"), is constantly in top positions, albeit with residual shares. This is the result of the bottom-up anonymization process used by Vodafone, which should be better transformed into a top-down method. The inclusion of regions of origin very far from those of destination (e.g. Latium and Tuscany), should also be explored in greater depth to ascertain their plausibility.

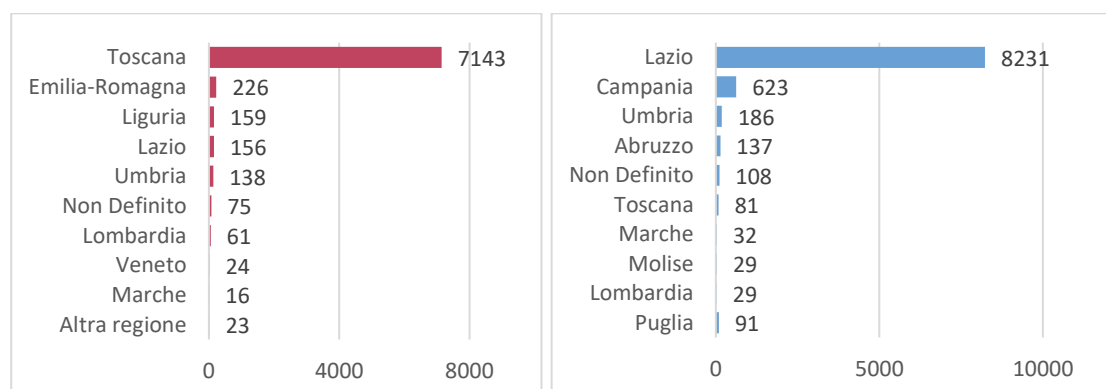


Figure 2. Regions of provenience of same-day visitors in Tuscany and Latium. Third quarter 2023 (Acknowledgments: Vodafone, Motion Analytics)

An important innovation of this experiment is the possibility of separately counting SDVs directed outside of one's UE, but originating and returning in a place other than one's Telephone Residence. This has allowed us to exclude from SDVs a significant share of those carried out during a trip (Same-day visits during a trip – SDVDT). These represent respectively 20% and 33% of SDVs in Latium and Tuscany in the third quarter of 2023 (Figure 3).

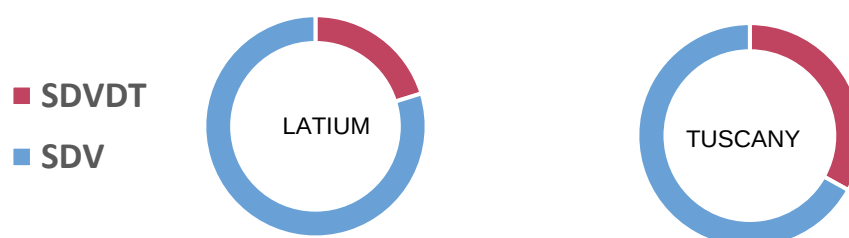


Figure 3. Same-day visits (SDV) e Same-day visits during a trip (SDVDT) in Tuscany and Latium. Third quarter 2023. (Source: Istat elaboration, acknowledgments: Vodafone, Motion Analytics)

4. CONCLUSIONS

Mobile data confirm their potential as an auxiliary source for traditional surveys; the fruitful collaboration between Istat and Vodafone has its strengths in the sharing of algorithms for the definition and analysis of data. The concept of UE applied to both Italian and foreign tourists, has proven to be fundamental for more correctly distinguishing tourists present in the territory and constitutes an innovative result for official statistics.

The Trips and Holidays survey, as mentioned, is the official source for providing Eurostat macro data on domestic SDVs (every three years) and on outbound SDVs (annually), but may be affected by the limitations related to the sample size and the memory effect for the estimate of this phenomenon.

The mobile source for the estimate of SDVs (and of SDVDTs as a new product) is interesting for the potential ability to provide timely and more granular data territorially, to overcome the aforementioned limitations of the sample survey. However, mobile phone data cannot provide the qualitative information currently required by Regulation 692/2011 for tourism statistics [4], [5], so they are necessarily an auxiliary source.

A possible use of mobile phone data as an auxiliary source is as an input for models aimed at disaggregating the estimates produced by the THS survey for the destination regions at a sub-regional level. The small size of the amount estimated by THS is a critical factor for this use. On the contrary, the estimates provided by the mobile source could constitute the reference amount for the national and sub-national SDVs levels to produce the other qualitative information based on the estimates provided by the survey.

5. REFERENCES

- [1] AA.VV. (2022), Una nuova metodologia di stima del turismo con i dati di telefonia mobile, poster per la 15° Conferenza Nazionale di Statistica, Rome, 3-4 luglio 2024. Available at https://www.istat.it/storage/15-Conferenza-nazionale-statistica/poster/09_03_Righi_POSTER.pdf

- [2] AA.VV. (2022), Exploring mobile network data for tourism statistics: the collaboration between Istat and Vodafone Business Italia, in *Rivista di Statistica Ufficiale* 3/2022, pp. 43-76. Available at <https://www.istat.it/produzione-editoriale/rivista-di-statistica-ufficiale-32022/>
- [3] World Tourism Organization. International Recommendations on Tourism Statistics 2008 (IRTS 2008), ST/ESA/STAT/SER.M/83/Rev.1, Madrid and New York. 2010 Available at https://unstats.un.org/unsd/publication/Seriesm/SeriesM_83rev1e.pdf
- [4] European Commission. 2019. Commission Delegated Regulation (UE) 2019/1681 of 1 August 2019, amending Regulation (UE) N. 692/2011 of the European Parliament and of the Council concerning European statistics on tourism, as regards the transmission deadlines and adaptation of Annexes I, and II.
- [5] European Parliament and the Council of the European Union. 2011. UE Regulation N. 692/2011 of the European Parliament and of the Council of 6 July 2011, aggiornato dal Reg. delegato UE n. 2019/1681. Available at <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF>
- [6] Eurostat. 2015. "Methodological manual for tourism statistics. Version 3.1 - 2014 edition". Manuals and Guidelines. Luxembourg: Publications Office of the European Union

Flash estimates of income and poverty indicators: an empirical study predicting labour market transitions

Andrea Fasulo, Aura Leulescu, Adriano Di Guglielmo, Mihaela Agafitei (Eurostat)

1. INTRODUCTION

Timely statistics on social issues, particularly income poverty and inequality, are a priority for policymakers and the European Statistical System. A strategy has therefore been put in place with two main objectives: (1) improving the timeliness of data from the primary income source, the European Union statistics on income and living conditions (EU-SILC); and (2) producing model-based [flash estimates of income and poverty indicators](#) (FE).

EU-SILC are essential for assisting EU policymakers in tackling poverty and social exclusion. Collaborative efforts between national statistical institutes (NSIs) and Eurostat have effectively shortened the publication delays for official statistics. However, significant events such as the 2008 financial crisis, the 2020 health crisis (i.e. COVID-19), and the recent inflation crisis have heightened the need for even more timely statistics on income distribution.

Since 2017, Eurostat, in collaboration with NSIs, has produced experimental statistics such as the flash estimates (FE), which offer insights into the distributional effects of labour and social policies just 6 months after the reference period. The FE (see Eurostat, 2024) rely on microsimulation modelling to adjust microdata from the previous year, considering the complex interactions between two key phases of the nowcasting process: (i) developments in the labour market; and (ii) the impacts of social policies. The EU-SILC dataset is integrated with current auxiliary data sources to estimate market income. This integration of primary data from EU-SILC with newer auxiliary data enables more accurate and up-to-date analyses of income and living conditions. The effects of social policies are simulated using EUROMOD (Sutherland and Figari, 2013), the EU's tax-benefit microsimulation model.

This paper focuses on the initial stage of the nowcasting process, with the primary objective of determining the most effective methodology for modelling the impact of labour market changes on income across various distribution points. The goal is to identify the individuals that are most likely to experience a change in their employment status, whether transitioning from employment to unemployment or vice versa. To achieve this, the paper evaluates different selection methods, including random selection, logistic regression and random forest predictions.

The quality assessment of the methods is carried out via a repeated k-fold cross validation (CV) for different European countries using the EU Labour Force Survey (EU-LFS) data. Cross validation is a machine learning technique that evaluates a model's performance by splitting the dataset into multiple subsets or folds. Once the dataset is divided into k equal-sized folds, each fold is used once as the test set, while the remaining k-1 folds are used for training. This process is repeated many times, and the results are averaged to provide a reliable estimate of the model's performance. The same model is reproduced in EU-SILC to estimate the impact of the labour market transitions on household income.

In brief, this paper aims to compare the above-mentioned methods investigating the empirical property in terms of bias and mean square error (MSE) in order to integrate the best methodology into the standard process of the FE, with the objective of improving their quality.

2. EMPIRICAL STUDY

An empirical study aims to evaluate the goodness-of-fit of the methods and their ability to capture the profile of individuals undergoing transitions into/out of employment. Successful methods would enable a more accurate estimate of income inequality and changes in poverty.

2.1. Repeated K-Fold Cross Validation

A Repeated k-fold (CV) study has been performed and empirical properties of the implemented methods, in terms of bias, mean square error, are computed.

The CV (See Picard and Cook, 1984) is a resampling procedure used to evaluate the performance of a method. It involves dividing the dataset into training and testing sets, training studying the model on the training set, and validating it on the test set. The CV helps to ensure that the model can generalize well to new, unseen data. There are several different CV extensions. In the Empirical Study developed in the work a Repeated K-fold CV is applied. Instead of performing the CV once, it is repeated multiple times, resulting in a higher number of folds.

The basic flow of repeated k-fold cross-validation consists of the following steps:

1. Divide the dataset into k folds.
2. For each repetition:
 - a. Use k-1 folds for training the model.
 - b. Test the model on the remaining fold.
 - c. Calculate the performance metric for the current fold.
3. Repeat steps 2a–c for all folds and all repetitions.
4. Compute the mean performance metric across all repetitions and folds.

The CV is carried out using longitudinal data from the 2022 EU-LFS and related for computational issues only to a sub-set of countries. The countries were randomly selected considering the size of the population to have small, medium and large states, which are specifically: Austria, Germany, Denmark, Estonia, Greece, Spain, France, Italy, Latvia, the Netherlands and Portugal. The samples have been divided in 5 k-folds of approximately equal size for 10 replications. For each replication, 4 over 5 k-folds are considered as training set and the remaining 1 over 5 as test set.

2.2. Simulation study

The simulation study is developed for the countries specified in the previous Section using the CV above described. For each country, the CV is run at stratum level where the strata are identified by gender and two age classes (16-29; 30-64).

The variable of interest is the labour transition probability measuring the likelihoods to move from unemployment to employment. From here on, out for the sake of brevity we will call the transition from unemployment to employment *Newemp*.

The exercise compares the performance for the Random Selection (RS), Logit regression and Random Forest (RF) methods. The Logit and RF model are fitted with the same covariates pattern composed by age, sex, education level, region of residence, duration of employment, student flag and degree of urbanization.

The comparison is presented for two level of details that are, from the finer to the most aggregated: (i) at unit level, (ii) estimating the total at strata level. In detail, the relevant variables used for *Newemp* are duration of unemployment, educational level and region.

A comparison of the performances of the methods for the target variable can be made in terms of Average Relative Mean Absolute Error (ARMAE).

$$ARMAE_{cml} = \frac{1}{S} \sum_{s=1}^S \frac{1}{K} \sum_{k=1}^K \frac{1}{R} \sum_{r=1}^R \frac{|\hat{p}_{lkr} - y_l|}{\hat{p}_{lkr}}$$

And in terms of Average Relative Root Mean Squared Error (ARRMSE).

$$ARRMSE_{cml} = \frac{1}{S} \sum_{s=1}^S \frac{1}{K} \sum_{k=1}^K \frac{1}{R} \frac{\sqrt{\sum_{r=1}^R (\hat{p}_{lkr} - y_l)^2}}{\hat{p}_{lkr}}$$

where

- c is the country indicator;
- m is the method (sampling or logit);
- l is the level of detail;
- s is the strata indicator;
- k is the folder indicator from 1 to $K=5$;
- r is the replication indicator from 1 to $R=10$;
- $\hat{p}_{skr}$ is the predicted probability of the k^{th} folder in the r^{th} replication in the l^{th} level of detail;
- y_l is true value of the l^{th} level of detail.

3. RESULTS

This section reports the results of the CV aimed to assess the accuracy of the different methods used to impute the Labour Transition probabilities.

Tables 1 and 2 show the results of the empirical study in terms of ARMAE and ARRMSE for the variable *Newemp* respectively at unit level and at strata level.

As shown in table 1, the logistic model outperforms the RF and RS methods for all the assessment indicators apart from in Estonia where RS shows better results. Overall, across countries, the use of the logit model compared to the RS leads to an average improvement of about 50% in the ARRMSE results. Overall across countries the use of the Logit model compared to the RS leads to an average improvement of about 50% in the ARRMSE indicator.

Table 1. ARRMSE and ARMAE for the NEWEMP variable at Unit Level

	ARMAE			ARRMSE		
	RS	Logit	RF	RS	Logit	RF
Average	1.17	1.06	1.17	1.81	1.25	1.38
AT	0.92	0.79	0.87	1.32	0.88	0.97
DE	1.08	0.93	1.02	1.57	1.05	1.16
DK	0.69	0.70	0.77	1.03	0.76	0.84
EE	0.81	1.26	1.39	1.22	1.44	1.58
EL	1.67	1.37	1.51	3.23	2.05	2.26
ES	1.29	1.13	1.24	1.95	1.3	1.43
FR	1.30	1.10	1.21	1.96	1.28	1.41

IT	1.33	1.20	1.32	2.03	1.37	1.51
LT	1.61	1.28	1.41	2.35	1.44	1.58
NL	0.82	0.72	0.79	1.23	0.82	0.90
PT	1.30	1.21	1.33	1.98	1.37	1.51

Table 2 shows that RS and logit have similar results in terms of ARMAE, with the former almost always showing slightly better results. On the other hand, ARRMSSE shows better results for the logit model with a gain in terms of RMSE of about 30% compared to RS.

Table 2. ARRMSSE and ARMAE for the NEWEMP variable at Strata level

	ARMAE			ARRMSSE		
	RS	Logit	RF	RS	Logit	RF
Average	0.12	0.15	0.18	0.23	0.18	0.21
AT	0.09	0.09	0.11	0.16	0.11	0.13
DE	0.09	0.09	0.11	0.17	0.11	0.13
DK	0.15	0.14	0.17	0.24	0.18	0.21
EE	0.19	0.47	0.56	0.58	0.56	0.57
EL	0.16	0.13	0.16	0.26	0.17	0.20
ES	0.05	0.06	0.07	0.08	0.06	0.07
FR	0.07	0.07	0.08	0.12	0.08	0.09
IT	0.10	0.12	0.14	0.18	0.12	0.15
LT	0.25	0.31	0.37	0.41	0.39	0.41
NL	0.07	0.08	0.10	0.13	0.09	0.11
PT	0.13	0.12	0.14	0.23	0.16	0.19

4. CONCLUSIONS AND FUTURE DEVELOPMENTS

An empirical study based on the CV shows that the use of a Logistic model, instead of a Random selection of the individuals or a Random Forest model, leads an important gain in terms of accuracy of the estimates for the Labour Market Transition. Furthermore, the simulation shows how this improvement is present both to identify the true ones at unit level and when more aggregated estimates are produced. A future development is to try to build a Bootstrap variance estimator based on the Logit predictor.

5. REFERENCES

- [1] Eurostat, Flash estimates of income inequalities and poverty indicators for 2023 (FE 2023), Methodological note (2024) [Methodological note Flash Estimates 2023 \(europa.eu\)](https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&plugin=1)
- [2] Picard, R.R., Cook, R.D., Cross-Validation of Regression Models, Journal of the American Statistical Association (1984), 79(387), 575–583. <https://doi.org/10.2307/2288403>

Fostering Collaboration: Communication Strategies in the AIML4OS Project

Orietta Luzi, Gerarda Grippo & Mauro Bruno (Istat)

1. ABSTRACT

Communication plays a pivotal role within the AIML4OS project by acting as a transversal work package (WP), supporting and enhancing the objectives and results of all other WPs. Its core function lies in amplifying the consortium's research efforts and ensuring the dissemination of findings both internally and externally.

As a first step, the communication WP (WP2) focuses on raising awareness within the consortium about the challenges and opportunities related to AI and ML research. By fostering a unified understanding of these issues across all consortium members, WP2 enables an effective and consistent communication strategy. This, in turn, strengthens outreach efforts not only within the statistical community but also towards citizens, media, and a wide range of stakeholders.

A critical aspect of WP2's mission is to highlight the achievements of participating countries and to leverage their collective knowledge. This approach helps build trust in innovative statistical products and facilitates the transition from experimentation to production within national statistical institutes.

WP2 promotes these objectives through a cross-media strategy that integrates both online and offline environments. Key communication tools include a project dashboard, social media channels, newsletters, internal and external events, and regular consortium meetings. These platforms foster collaboration, enhance interconnections, and support the visibility of the project's outputs.

Moreover, the project's communication strategy is distinctly characterized by a dedicated logo, ensuring that all project-related outputs are easily recognizable and associated with the initiative's goals. This branding not only reinforces project identity but also ensures consistency in messaging across diverse platforms and audiences.

Fostering the collaboration between Eurostat and the Academia

Tiziana Laureti (University of Tuscia), Dario Buono, Cristiano Tessitore (Eurostat),
Ilaria Benedetti (University of Tuscia)
& Hionia Vlachou (GOPA Luxembourg)

1. INTRODUCTION

This paper outlines Eurostat's strategic vision¹ to enhance collaboration with academia, with the aim of advancing statistical methods, fostering innovation, and promoting research. As the statistical authority of the European Union, Eurostat plays a crucial role in producing high-quality statistics that underpin evidence-based policymaking, economic analysis, and societal understanding across Europe. However, the increasing complexity of data ecosystems, rapid development of new technologies and requirements for innovative data resources have created a pressing requirement for closer collaboration between Eurostat and academic institutions.

The amendment of Regulation (EC) No 223/2009, while strengthening the legal environment and the professional independence of statistical authorities and of Eurostat, encourages the development of structured collaborations with academic and research institutions.

Eurostat has long recognized the importance of collaborating with academia, and several initiatives have already laid the groundwork for this partnership, see among the others the European Master in Official Statistics (EMOS), European Statistical Training Programme (ESTP) or other initiatives focusing on training, research, and data-sharing to foster a collaborative environment between Eurostat, academic institutions, and National Statistical Institutes (NSIs).

Therefore, this paper suggests a vision based on four pillars to enhance Eurostat's strategy for collaborating with academia aiming to improve statistical methods, fostering innovation, ensuring high-quality data and promoting communication. This involves education and skills development, fostering innovative methodologies, and engaging in research partnerships. The strategy aligns with the European Statistical Programme and emphasizes cooperation with NSIs, encouraging knowledge exchange, and ensuring statistical literacy among data producers and users.

This partnership is designed to harness the innovation and methodological advancements from academic research to support Eurostat's mission, ensuring that European statistics remain relevant, accurate, and capable of addressing evolving challenges.

Through a structured framework, this paper addresses critical needs in the areas of statistical education, methodological innovation, research, and communication, all of which are pivotal for supporting evidence-based policymaking and understanding key trends in Europe.

¹ The paper is being currently reviewed by Eurostat's senior management; an approval of the strategy is expected in the forthcoming weeks.

2. THE REGULATORY FRAMEWORK

The need for closer collaboration between Eurostat and academia is not only driven by practical demands but is also enshrined in European regulations and strategic documents. The legal foundation for this cooperation is established in Regulation (EC) No 223/2009, which governs the development, production, and dissemination of European statistics. In addition, the European Statistics Code of Practice (ES CoP) provides a framework of principles that guide the quality and integrity of European statistics. Further support for this collaboration comes from the European Statistical Governance Advisory Board (ESGAB), which, in its recent annual reports, has recommended that Eurostat strengthen its strategic engagement with academia. ESGAB's 2021 and 2022 reports highlight the importance of formalizing these partnerships to foster innovation and ensure that the ESS attracts highly skilled professionals. ESGAB's recommendations are aligned with the goals set out in the Eurostat Strategic Plan 2020-2024 and the European Statistical Programme (2021-2027), both of which prioritize cooperation with academic institutions to enhance the quality, timeliness, and relevance of European statistics. Moreover, Regulation (EU) 2021/690, which establishes the Single Market Programme, further reinforces the need for high-quality, comparable, and reliable European statistics to support policymaking and research across the EU.

3. RESULTS AND PRACTICAL APPLICATION

Four pillars are identified providing the fundamental areas supporting the achievement of the above strategic goals for fostering the collaboration with academia and describing the two-way interaction between Eurostat and academia. Our approach based on four pillars is illustrated in Fig.1.

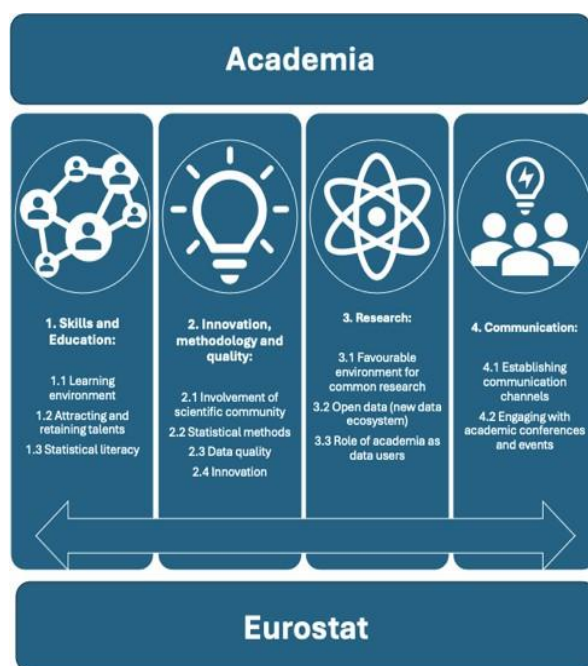


Figure 1. Approach based on four pillars

Pillar 1, named “Skills and education” focuses on three key goals: creating a robust learning environment to drive innovation and train statisticians, attracting, and retaining top talents from all Member States, and enhancing statistical literacy among data producers and users for improved data quality and decision-making.

Pillar 2, named “Innovation, methodology and quality”, groups four strategic goals aiming at promoting knowledge sharing by involving the scientific community, harnessing academic expertise, and strengthening methodological rigor in the ESS by exploring innovative methods with a view to continuous improvement.

Pillar 3, “Research”, includes three strategic goals aiming at creating a favourable environment for common research with academia, encouraging a comprehensive framework that promotes collaboration, open data, and academic involvement as data users.

Pillar 4, “Communication”, includes two objectives aimed at fostering effective communication between Eurostat and academia, encouraging collaboration and information exchange.

Taking as a basis the four pillars of the strategy, different actions are proposed for each pillar for fostering the collaboration between Eurostat and academia.

Considering the Pillar “Skills and education”, for example, in support of the objective of developing a learning environment, Eurostat will collaborate with universities to support Ph.D. and postdoctoral programs related to official statistics. This involves co-supervision of research projects, joint funding initiatives, and providing access to Eurostat data. In support of the objective of attracting and retaining talents, Eurostat will promote interinstitutional mobility and strengthen recruitment strategies and expand the talent pool by attracting candidates from diverse backgrounds and experiences. With the aim of improving statistical literacy, Eurostat will collaborate with academia to provide training courses in statistics to public sector employees to foster statistical skills.

Regarding the Pillar “Innovation, methodology and quality”, in support of the objective of enabling stronger involvement of the scientific community, Eurostat will set up a task force or working group(s) focused on facilitating communication, joint initiatives and knowledge exchange between Eurostat and academia. This task force for developing collaboration with academia will serve as central information hub where Eurostat staff and academic researchers can exchange ideas, share insights, and explore potential areas of collaboration.

In support of the objective of encouraging the exploration of innovative statistical methods, Eurostat will identify priority areas by determining which domains could benefit most from improvements and innovative methodological approaches. Academics will help in addressing data challenges, new and innovative methods including machine learning and artificial intelligence techniques. Eurostat will establish programs that allow academics to spend sabbaticals or short-term visits at Eurostat, working closely with its experts.

In order to achieve the goal of maintaining and enhancing the quality of data, Eurostat will support academic research projects on topics such as advanced data fusion techniques to integrate heterogeneous data sources to create unified datasets, quality assessment frameworks, data imputation techniques. Fusion techniques will include statistical methods, machine learning algorithms, and domain-specific knowledge to combine information effectively.

Concerning the Pillar “Research” to achieve the objective of creating a favourable environment for common research, Eurostat will provide access to research infrastructure, data repositories, and analytical tools to support joint research efforts. Eurostat will organize or participate in workshops, conferences, and seminars with academic institutions in addition to existing ones. These events provide a platform for researchers to share their findings, discuss methodologies, explore potential collaborations and may also represent an opportunity for promoting EMOS emphasising the benefits that academic institutions get by joining EMOS.

With the aim of encouraging the development of tools and methodologies to facilitate open data use, Eurostat will stimulate the development of tools and methodologies that facilitate the processing, analysis, and visualization of open data. The focus will be on the harmonization

of metadata standards, ensuring that data from different sources, including NSIs, adhere to common standards.

To achieve the objective of fostering the role of academia as data users, Eurostat will collaborate with academic journals and publications to publish articles or special issues showcasing the use of European statistics in research. This can demonstrate the real-world applications and impact of the data.

Turning to the Pillar “Communication” , in support of the objective of fostering communication channels, Eurostat will create a dedicated section on the Eurostat website highlighting collaborations with academia. Provide detailed information about ongoing and past collaborations, objectives, and outcomes.

Eurostat will participate in academic conferences and present findings from collaborative projects and actions, arrange students group visits and organise “job-day” or sponsor career fairs hosted by universities or academic institutions. During the “job-day” students (involved in tertiary education and bachelor’s degree programs) will be informed about Eurostat educational initiatives in the field of statistics (e.g. EMOS program) and career opportunities at Eurostat.

4. REFERENCES

- [1] Regulation (EC) No 223/2009 of 11 March 2009
- [2] ESDG Annual Report, 2021 Luxembourg: Publications Office of the European Union, 2021
- [3] ESDG Annual Report, 2022 Luxembourg: Publications Office of the European Union, 2021

From NACE 08 to Nace 2025: Retraining an ML model in production using Large Language Models

Thomas Faria & Julien Pramil (Insee)

1. INTRODUCTION

The French company registry, SIRENE, lists all companies in France and assigns them a unique identifier, the Siren number, for use by public institutions. As part of the registration process, companies must provide a description of their economic activity, which is then classified into an industry using the French classification of activities (NACE Rev. 2)

In our institution, to automate this classification we deployed our first machine learning (ML) model into production [1] in late 2022, using the FastText algorithm [2][3]. Despite its simplicity, the model delivered high quality results, significantly reducing the workload of the manual classification team. Since its deployment, only a minor retraining was required to address a recurring classification error.

The model's performance has been stable, allowing us to develop a monitoring dashboard that continuously tracks its effectiveness over time. The model's reliability also enabled public agents to annotate and label data continuously, gradually building a robust gold-standard test dataset for model evaluation.

However, the recent revision of the NACE (Statistical Classification of Economic Activities in the European Community) nomenclature to NACE 2025 presents a new challenge. This revision not only requires the model to classify future enterprises using NACE 2025 codes but also necessitates the reclassification of over 14 million registered companies in the SIRENE database, previously coded in NACE 2008. Addressing this twofold challenge — forecasting new classifications and re-coding historical data — demands significant retraining of the model and adaptation to the new nomenclature.

2. METHODOLOGY

Manually recoding the entire stock of companies from NACE 2008 to NACE 2025 is impossible given the short timeline. To handle this, we developed a three-step methodology:

1. **Nomenclature Analysis:** Since the transition from NACE 2008 to NACE 2025 is not bijective (i.e., one-to-one mapping), nomenclature experts first analysed the new classification scheme. They identified both unequivocal codes, where a direct mapping is possible, and multivocal codes, where multiple potential classifications exist.

2. **Annotation Campaign:** To address multivocal codes and refine the model, we launched an annotation campaign. In this phase, public agents labeled data according to the NACE 2025 nomenclature. This process helped create a high-quality gold standard test set for future model evaluation.

3. **Leveraging Large Language Models:** To accelerate the reclassification, we used a large language model to automatically recode the dataset according to NACE 2025, generating a new training set.

4. Retraining an ML model: The retraining process involved training an ML model on this new NACE 2025-labeled dataset and evaluating its performance using the gold standard test set developed during the annotation campaign.

In the third step, we focused on handling multivocal codes, where the classification was ambiguous and used a RAG approach. A *Retrieval-Augmented Generation* (RAG) is a methodology that combines two key elements: *retrieval* of relevant information from a large database and *generation* of a final output using a (large) language model. The retrieval component ensures that the language model has access to factual and context-specific knowledge, while the generation step allows it to construct responses or predictions that are coherent and contextually appropriate. In our case, this means that the RAG model can retrieve relevant documentation, such as explanatory notes from the NACE 2025 nomenclature, and use this information to generate more informed and accurate predictions for classifying ambiguous (multivocal) codes.

Two distinct approaches were implemented:

- **Naïve RAG Approach:** A Retrieval-Augmented Generation (RAG) pipeline was designed, using the explanatory notes of the NACE 2025 nomenclature. All these notes have been embedded into a vector database. Then, a similarity search is made between the activity description of the company and all the documents contained in the vector database. The K most similar/relevant documents are then given to the LLM as context so that it can propose the right NACE code to the company.
- **Manual RAG Approach:** In this approach, we assumed that the NACE 2008 code was accurate and used it to filter the set of possible NACE 2025 codes. We do not use a vector database to retrieve the most relevant documents/codes, since we know *a priori* the list possible NACE2025 codes given the NACE 2008 code. Hence, this list of possible code is given as context to the LLM that can determine the most appropriate NACE 2025 code. By doing so we reduce the sources errors coming from the stage of retrieval, providing a more targeted reclassification.

3. RESULTS AND MAIN FINDINGS

Our evaluation showed that both the *naïve* and *manual RAG approaches* yielded promising results. However, in our specific case, the *manual RAG approach* provided better performance, largely because it restricted the selection to a prefiltered list of potential NACE 2025 codes based on the company's existing NACE 2008 code. This targeted reclassification resulted in over 90% accuracy when evaluated on the test set labeled by nomenclature experts. Importantly, when the model made errors, they almost always occurred at the final level of the nomenclature hierarchy. This means that the mistakes were relatively minor and had a limited impact on the broader classification, thus minimizing any significant misclassification.

Taking advantage of this method, we successfully reclassified the entire stock of over 14 million companies in a relatively short period, while maintaining a high level of accuracy. Additional quality checks are still required on the dataset before officially informing companies of their new NACE 2025 codes. However, the reclassified dataset has already been used to train a new machine learning model for future classifications under the NACE 2025 nomenclature. This new model demonstrated performance comparable to the earlier NACE 2008-based model, providing similar levels of accuracy and reliability in predictions.

While the *naïve RAG approach* delivered slightly lower accuracy compared to the manual approach, we still plan to explore its potential for full classification in an experimental context. Despite its lower accuracy, the naïve approach offers valuable insights and may be useful in

cases where fewer constraints are needed or in situations requiring greater flexibility in code selection. That said, given the strong performance of our existing, lightweight ML model, we opted not to transition to a more resource-intensive and potentially less stable LLM-based model. Ongoing experiments are being conducted to assess the feasibility and performance of the LLM approaches, and the final results of these experiments will be shared during the session.

4. REFERENCES

- [1] R. Avouac, T. Faria and F. Comte, A cloud native data science platform for official statistics, 2024.
- [2] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, *Bag of Tricks for Efficient Text Classification*, 2016.
- [3] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching Word Vectors with Subword Information, 2017
- [4] P. Lewis and al, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020.

From web to data: Quality and Methodological Considerations

Alexander Kowarik (Statistics Austria), Magdalena Six & Piet Daas (Statistics Netherlands)

1. INTRODUCTION

In the rapidly evolving landscape of data acquisition, traditional sources of data collection, such as surveys and censuses, are increasingly supplemented by alternative data sources like web-scraping. This presentation outlines the methodologies and quality guidelines developed for using web-scraped data in official statistics production, based on the work in the project “Web Intelligence network”, co-funded by the European Commission.

Web-scraping has emerged as a critical tool for capturing information from the internet, especially when traditional data sources are incomplete or unavailable (see, e.g., [3] Maślankowski, J. et al., 2023). However, using web data for official statistics successfully presents challenges related to data quality, methodological issues and legal and ethical considerations. The guidelines developed aim to ensure that web-scraping methods conform to the principles of the European Statistics Code of Practice. The methodologies discussed highlight the role of selective scraping, which targets specific sub-populations relevant to the statistical needs of National Statistical Institutes (NSIs), allowing for more controlled and targeted data acquisition.

2. METHODOLOGICAL FRAMEWORK

The scraping process is divided into several phases, starting with source identification, followed by source selection, and data extraction (see, e.g., [2] Six M. et al., 2023). The methodology emphasizes the importance of maintaining transparency throughout the scraping process.

In selective scraping (see, e.g., [1] ten Bosch, 2024), a priori knowledge, such as business registers, is used to target specific entities, enhancing the efficiency of data collection. For example, selective scraping has been applied to the Online Job Advertisements (OJA) use case, where targeted websites are selected based on predefined criteria, including stability, popularity, and coverage. This allows for a refined, representative data set tailored to the needs of official statistics.

3. QUALITY CONSIDERATIONS

The quality of web-scraped data is crucial for its integration into official statistics. WP4 developed a comprehensive set of guidelines for assessing the quality of scraped data, focusing on several key aspects:

Coverage and Representativeness: Ensuring that the scraped data covers the relevant population is essential. Under-/over coverage needs to be measured and considered in the estimation.

Comparability: Structural changes in web data sources, such as website redesigns, must be monitored continuously to maintain comparability over time.

Measurement Errors: Scrapers should be regularly updated to ensure the accuracy of the data collected, and measurement errors should be assessed through manual reviews or cross-validation with other data sources.

Deduplication errors: In many cases when using web data, a specific unit, e.g., a job advertisement, a hotel or an apartment, could be present in multiple sources. There this error category deserves special attention.

Classification errors: Often the needed information, especially the official statistics classifications, e.g., the economic activity of a company, the category of a job, etc. are not directly available from web source. Errors in deriving these classifications must be assessed by appropriate means.

3.1. Technical Infrastructure

The establishment of a centralized Web Intelligence Platform (WIP) for scraping, processing, and storing web data was a key outcome of the WP4 activities. The platform is designed to reduce the maintenance burden on individual NSIs by centralizing the scraping processes and ensuring that data is securely processed and stored. The guidelines recommend the use of open-source, cloud-agnostic platforms that can be adapted to the specific needs of different statistical offices, allowing them to bring the platform to their data rather than uploading sensitive information to a shared infrastructure.

4. MAIN FINDINGS

Web-scraping presents a valuable opportunity to enhance official statistics by providing timely, detailed data from online sources. However, ensuring the quality, comparability, and legal compliance of web-scraped data is essential for its effective use. The guidelines developed in WP4 provide a robust framework for acquiring, processing, and using web-scraped data in a manner that aligns with the principles of official statistics, while also addressing the specific challenges posed by web data sources. The ongoing development of centralized infrastructures, such as the Web Intelligence Platform, will play a crucial role in facilitating the sustainable integration of web-scraped data into statistical production processes.

5. REFERENCES

- [1] Ten Bosch, O., Kowarik, A., Quaresma, S., Salgado, D., van Delden, A. (2024). Statistical scraping: informed plough begets finer crops. Paper for the Q2024 conference, Estoril, Portugal. Link: https://www.researchgate.net/publication/380532685_Statistical_scraping_informed_plough_begets_finer_crops, page 61.
- [2] Six, M., Kowarik, A., Gussenbauer, J. (2023). Landscaping of Websites for Webscraping with Focus on Selection Modes. Draft version of the report for WP4 of the ESSnet WIN. Oct. 2023.
- [3] Maślankowski, J., Daas, P. (2023). Current Challenges and Possible Solutions for the Use of Web Data as a Source for Official Statistics. Presentation at the Methodology of Statistical

Research 2023 (MET2023) conference, 3-5 July, Warsaw, Poland. Link:
<https://met2023.stat.gov.pl/Content/Presentations/Sesja%2016.4%20MET2023.pdf>

From web to data: Selective scraping and WIN.The hackathon

Olav ten Bosch (Statistics Netherlands)

1. INTRODUCTION

Web data is becoming increasingly important for official statistics. In combination with traditional inputs web data may add up-to-date information to the existing data portfolio, it may improve or speed-up statistical processes or it may creating new opportunities for (experimental) indicators. There are successful examples in price statistics on web shops, enterprise statistics on business websites and social statistics observing social media. However, there are also challenges: web data may be volatile, may contain multiple types of biases and quality may depend on many factors, such as the business value of the data for the content provider.

In the European ESSnet project Trusted Smart Statistics – Web Intelligence Network (TSS-WIN) [1] multiple ways of using web data in official statistics have been explored. Apart from major use cases such as online job advertisements (OJAs), online business enterprise characteristics (OBECs) a work package was dedicated to exploring other use cases, such as characteristics of the real estate market, construction activities, online prices of certain product categories, indices in tourism statistics. Another major use case is business register quality enhancement. They all had their successes and challenges such as volatile inputs, deduplication challenges, mapping web data onto statistical concepts and operational challenges, but also helped further shape the relatively new thinking about *selective scraping versus bulk scraping*, first explained in [2].

2. THE CONCEPT OF SELECTIVE SCRAPING

In many cases the use of web data starts with collecting initial samples of data from one or more websites that seem valuable for the statistical use case at hand. The websites are studied, some data is collected and analysed and, if found to be valuable, data collection is expanded to cover the web domains needed for the quality desired. This may involve gathering raw data from various online sources and subsequently linking it to the relevant statistical context. This can be a labour-intensive process, especially as new data sources emerge and the volume of data increases. We call this method bulk-scraping.

Statistical¹ or selective scraping, on the other hand, leverages the existing knowledge base of statistical offices. By querying the web with specific identifiers, names, categories, or statistical definitions, already contained in the registers maintained by the statistical organisation, it ensures that the collected data can be directly linked to the statistical context. This approach is akin to conducting an automated survey on the vast expanse of the web, where the results, though potentially messy, can be connected to the relevant statistical units. By applying such a targeted approach to data collection, data volumes are limited, data in registers are

¹ In [2] the term “statistical scraping” is used. However “selective scraping” seems to become more popular and also adheres to the well-known “selective editing” in editing in official statistics. Hence, in the sequel of this abstract we will use the term selective scraping.

maximally used and statistical agencies can reduce the risk of representation errors and calculate accuracy and reliability of their findings. Figure 1 sketches the principle.

It is important to note here that bulk scraping certainly has its value in certain statistical use cases. Even stronger, in cases where the statistical population yet has to be discovered it may be the only option. However, in other cases a selective scraping methodology may complement or in some cases replace bulk scraping methods.

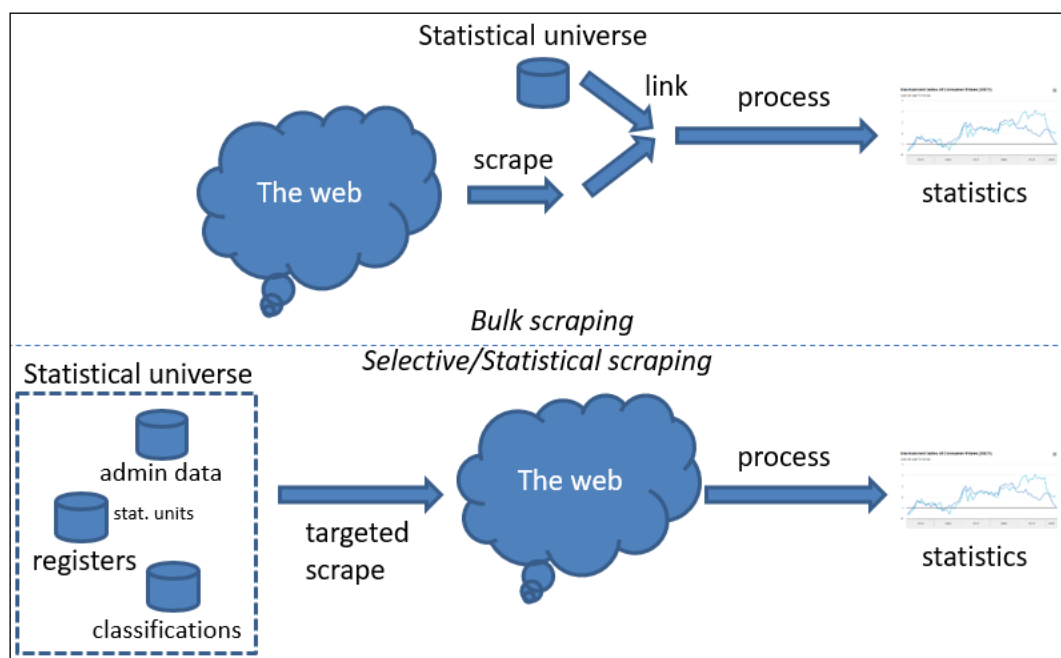


Figure 1. Selective scraping versus bulk scraping

3. BUSINESS REGISTER ENHANCEMENTS FROM WEB DATA

One notably example of selective scraping is improving the business register with web data, one of the use cases studied in the WIN project. Statistical Business Registers (SBRs) are indispensable assets for National Statistical Institutes (NSIs). They provide detailed information about enterprises, serving as the foundation for numerous statistical analyses and surveys.

Online data offers a wealth of insights into the activities of enterprises. Websites, media advertisements, product listings, customer interactions, Wikipedia articles, job postings can all be usual information to better classify the statistical unit or to check or complete the information available in the SBR.

Using selective scraping, the first step in this process is to identify the URLs associated with each enterprise. This can be challenging as URLs are often not readily available or may be unreliable. To address this, NSIs can utilize search engines using information from the SBR and select the most relevant matches based on factors such as the enterprise's name, contact information, and tax identification number.

Once URLs are identified, the next phase involves extracting statistical variables from the associated web contents. This typically involves scraping text and applying natural language processing (NLP) and machine learning techniques to interpret the content. Common use cases include deriving economic activity (NACE), ecommerce, social media use, enterprise relationships, sustainability practices, and / or job vacancies. Of course, this is a simplification,

in reality the relationships between legal units in the SBR and websites or other web data may be more complex. A single enterprise may operate multiple websites for different customer segments, while a website may be used by multiple legal units within an enterprise group. Nevertheless, generally speaking selected scraping for enhancing business registers has proven to be a valuable concept.

4. OTHER EXAMPLES

Other domains where the selective scraping concept has been applied or can be applied are 1) price statistics (the basket approach applied to the web), 2) Tourism statistics (scraping a selected sample of hotels in certain categories) and 3) job vacancies (checking whether a selected sample of enterprises from the business register have job vacancies, how many and possibly what kind of).

5. WIN.THE HACKATHON

Building on the selective scraping concepts explained above, the WIN project issued a hackathon on OBEC indicators in 2024. Target indicators were ecommerce and social media use, where social media use was specialized into 6 types of social media. A set of URLs, geographically spread over 4 countries (Poland, Austria, Germany and the Netherlands) was derived. A subset was manually labeled and a community of data scientists was asked to come up with the best open source software to automatically derive these indicators from websites. The results were scored against the labeled set, which was secret to the participating teams.

In this “From web to data” session the teams that were closest to the labeled results will present their results and experiences and the innovative way they came to their solution.

6. REFERENCES

- [1] WIN, the Web Intelligence Network project, <https://cros.ec.europa.eu/book-page/web-intelligence-network-project-overview>
- [2] Ten Bosch, O., Kowarik, A., Quaresma, S., Salgado, D., van Delden, A. (2024). Statistical scraping: informed plough begets finer crops. Paper for the Q2024 conference, Estoril, Portugal. Link:
- [3] https://www.researchgate.net/publication/380532685_Statistical_scraping_informed_plough_begets_finer_crops

From web to data: Winning Hackathon with Offline AI for E-Commerce Site Detection Strategies

Riccardo Corradini & Rita Lima (Italian National Institute, Istat)¹

Instructions: This paper demonstrates how AI-powered automation can efficiently analyze URLs, providing actionable insights into e-commerce trends. The approach emphasizes the need for scalability, efficiency, and effective resource management in modern data analysis tasks. Future research directions include exploring more models that are sophisticated or expanding the range of analysis to different types of online platforms.

1. INTRODUCTION

This paper proposes a methodological procedure that utilizes Generative AI models, served by `llama-server`, to analyze URLs and detect e-commerce activities and social media presence. This approach was applied in the 2025 Hackathon Competition (2025 HC)² to analyse and extract information from big and open data sources, developing a prototype useful for addressing specific EU policy issues announced at the beginning of the hackathon. The proposed approach is hybrid and technically feasible, offering several benefits: a) By running `llama-server` locally, we avoid sending sensitive data to remote servers, reducing the risk of data breaches and ensuring regulatory compliance; b) By monitoring online activities, we gain real-time insights into societal trends, including changes in work patterns and consumer behavior during events like the COVID-19 pandemic. Similar approaches can be applied by National Statistical Institutes (NSIs) to URL classification, using diverse data sources and adhering to ethical standards. As digital landscapes evolve, NSIs are increasingly exploring new data sources and methods for producing official statistics. Furthermore, relying solely on traditional data collection methods may lead to incomplete understanding of digital engagement and potential biases in statistical analysis³. Integrating online activities into data collection methodologies allows NSIs to gather more comprehensive and timely statistics. A key challenge for NSIs is adapting their strategies to effectively leverage new data sources and methods, while addressing the challenges of producing official statistics in a rapidly changing digital landscape. By processing data locally using offline AI, NSIs can significantly reduce risks related to data privacy and system latency, while also gaining better control over sensitive information and minimizing exposure to cyber threats.

¹ The paper is the result of a joint research of the authors. As far, it concerns this version of the work, Sects. 1 and 3 should be attributed to R. Lima, Sect. 2 to R. Corradini.

² <https://cros.ec.europa.eu/system/files/2024-08/Hackathon%202025%20-%20Event%20description.pdf>

³ https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/05/digital-transformation-of-national-statistical-offices_17c00b5b/ee4b1b85-en.pdf

2. METHODOLOGY

The proposed procedure to tackle the specific 2025HC task emphasizes the creation of an offline URL classification procedure that respects netiquette principles while effectively answering the specific EU policy issues announced at the beginning of the 2025HC. The procedure consists of a hybrid approach that mixes the analysis of URL Structure, hardcoding by Python script with BeautifulSoup language [1], and the offline Artificial Intelligence (AI) to speed and automate URL analysis and classification procedure.

2.1. URL Classification Challenges

Detecting non-e-commerce sites is relatively straightforward, as these websites often contain specific keywords such as "barber," "dentist," "lawyer," or "notary." These terms can be easily identified and used to classify the website as non-e-commerce. However, detecting e-commerce sites is a more complex task. While words like "credit card" and "payment" may indicate an e-commerce site, their frequency and distribution within the source code can be difficult to interpret. Moreover, the complexity of modern web development means that the source code can be heavily obfuscated, making it challenging to accurately detect e-commerce sites based on keyword frequency alone. Parallel analysis does not necessarily help in this case, as the sheer volume of data and the nuances of website design can lead to false positives or false negatives.

2.2. Methodological Steps

Here is a synthetic list of the methodological steps involved in the procedure:

1. URL Processing Script:

- Input Handling: Loads URLs from a CSV file, allowing for structured data input.
- Asynchronous Processing: Utilizes ``asyncio`` to process each URL concurrently, enhancing efficiency and reducing processing time.
- AI Integration: Communicates with ``llama-server`` to analyze each URL, extracting insights such as e-commerce status and social media handles.
- Progress Tracking: Implements ``tqdm`` for real-time progress monitoring, improving user experience.

2. System Monitoring:

- GPU Usage Check: Monitors GPU resources to ensure optimal performance and prevent overload.
- Server Availability: Includes checks (commented-out) to verify ``llama-server`` status before processing, ensuring smooth operation.

3. Result Management:

- Saves analysis results in CSV format, including fields for e-commerce detection, social media presence, and AI response data.

The following link <https://github.com/freesoftwdreamer/Web-Intelligence> is a reference of our methodological procedure used to tackle the 2025HC task.

2.3. Main Findings

The procedure employs asynchronous processing to efficiently handle large datasets, ensuring scalability and performance, and addresses challenges such as exploring new data sources and innovative methods for producing official statistics while upholding ethical standards.

Here is a detailed exploration of some benefits and rationale behind the procedure:

- AI systems process vast amounts of data in real-time, providing valuable insights and reducing the time and effort required for the task. This leads to improved operational efficiency (Reduced Errors and Improved Operational Efficiency);
- Given the dynamic nature of frequent changes in URL structures, LLaMA's training on vast datasets allows it to adapt to new patterns and threats effectively [4]. This adaptability is crucial for maintaining accurate detection rates over time (Reduced Latency and Network Issues);
- Implementing LLaMA models offline respects netiquette rules by avoiding unnecessary network requests that could infringe on user privacy or violate terms of service. Offline models can analyze URLs directly without needing real-time internet access, which is particularly important in sensitive environments (Cost Reduction).

3. CONCLUSIONS

This paper presents a hybrid procedure with potential applications in various fields related to producing official statistical information. For instance, it offers a specialized approach to understanding text, synthesizing large amounts of information, and generating responses that mimic the data production phases of the Generic Statistical Business Process Model. Additionally, regarding privacy considerations and offline AI usage, this procedure proposes a solution to mitigate potential issues related to data privacy and system latency. The paper has a broader purpose: to explore how NSIs can effectively utilize the vast amounts of data accessible through URLs, including e-commerce sites and social media platforms. As online activities continue to grow, surveys on internet usage, e-commerce, and digital skills are conducted regularly to capture the evolving behavior of individuals and households.

To harness this digital landscape effectively, we suggest the following key strategies:

a. Importance of Digital Transformation

- Adaptation to Digitalization: NSIs should adapt their methodologies to leverage digital technologies, which can enhance data collection and analysis processes. The COVID-19 pandemic highlighted the need for NSIs to integrate alternative data sources, such as mobile networks and social media interactions, particularly when traditional data collection methods were disrupted.
- Comprehensive Digital Strategy: Developing a clear digital strategy is essential for NSIs. This strategy should align with the NSI's core objectives and incorporate innovative data collection methods that can be scaled over time. Pilot projects can help test new approaches before broader implementation.

b. Leveraging Online Data Sources

- Utilizing E-commerce and Social Media Data: E-commerce platforms and social media are rich sources of consumer behavior data. NSIs could analyze trends in online shopping and social commerce, which are increasingly influencing economic indicators. For example, understanding how platforms like Facebook and Instagram facilitate direct sales can provide insights into consumer spending patterns.
- Engaging with Social Media Analytics: Social media platforms serve as valuable tools for gauging public sentiment and consumer preferences. By analysing engagement metrics, NSIs can gain insights into societal trends that may not be captured through traditional surveys.

c. Capacity Building

- Training and Development: Investing in training programs for staff on digital tools and data analytics is crucial. E-learning platforms can offer cost-effective solutions for building capacity within NSIs, enabling staff to stay updated on the latest methodologies in data collection and analysis.
- Partnership with Private Tech Companies: Collaborating with technology companies can facilitate access to advanced analytical tools and methodologies, enhancing the NSI's ability to process large datasets from online sources effectively.
- Addressing Data Privacy: As NSIs expand their data sources, they must navigate ethical issues related to privacy and data sensitivity. Establishing robust regulatory frameworks will be necessary to protect individual privacy while harnessing the power of big data.

By focusing on these areas, NSIs can significantly enhance their capabilities in understanding and utilizing online activities, ultimately leading to more accurate and timely statistical outputs that reflect the realities of a digital economy. Moreover, by adopting similar approaches, they can leverage diverse data sources while addressing the growing demand for new statistics. These approaches should prioritize transparency in how data is collected, processed, and classified, uphold strict ethical standards, and adhere to principles of netiquette when engaging with online platforms. In summary, understanding online activities is vital for NSIs as it strengthens their ability to gather relevant data, inform evidence-based policy decisions, and promote transparency and accountability in governance.

4. REFERENCES

- [1] Python Libraries: ``asyncio``, ``tqdm``, and ``requests`` for HTTP communication.
- [2] ``llama-server`` Project: For providing the AI model used in URL analysis.
- [3] <https://github.com/unclecode/crawl4ai> for web crawling and source code structure download.
- [4] <https://github.com/ggerganov/llama.cpp> for the AI model used in URL analysis.
- [5] <https://chat.deepseek.com/> for helpful discussions and hints on code development.
- [6] <https://chatgpt.com/> for providing valuable insights and suggestions during the coding process.

Functional Clusterwise LDQ Regression for Assessing Poverty Risk Across European States

Gianmarco Borrata (University of Naples "Federico II"), Rosanna Verde, Antonio Balzanella (Università della Campania Luigi Vanvitelli) & Francisco Carvalho (Universidade Federal de Pernambuco)

This work deals with a clusterwise algorithm based on a new regression method for distributional data. The innovativeness of the method consists of a new transformation of the distributional data, which allows of mapping density functions into a Hilbert space, that is a logarithmic transformation of the Derivative Quantile functions (LDQ). This distributional processing approach allows for an in-depth analysis of these functions while enabling a return to the original quantile functions through an inverse transformation. The LDQ functions are a functional data representation of distributional data. According to functional data representation, a B-spline smoothing at points corresponding to the quantiles of the distributions is provided. Another contribution of such work is a clusterwise functional regression method, which estimates the response variable by partitioning the set of objects into K clusters, optimizing the fit of local regression models within each cluster. Preliminary results based on Eurostat data have confirmed the effectiveness of the proposed method. By analyzing socioeconomic factors and degrees of urbanization, we identified distinct clusters that reveal the varying impacts of these socioeconomic variables on poverty risk in cities, towns, and rural areas of European states. This application highlights the effectiveness of our approach in handling distributional data and provides valuable insights for policymakers aiming to address poverty in diverse contexts.

1. INTRODUCTION

In recent years, there has been a growing interest in developing regression methods tailored to handle data represented in the form of distributions (Bock H.H.- Diday E., 2000). Traditional functional regression techniques may not fully capture the nuances inherent in such data structures. Various methods have been proposed to address the regression modeling analysis for interval-valued data. A common idea is to convert the interval-valued variable into classical single-valued variables and then apply traditional regression analysis methods; for instance, the center method (Billard L., Diday E., 2000), the Min–Max method (Billard L., Diday E., 2002), the center and range method (Neto E.A.L., de Carvalho F., 2008), etc. On this basis, different models were constructed, where the coefficients were limited to being nonnegative. Regression models for distributional variables has been even proposed by Irpino-Verde (2014) and Brito-Dias (2015).

Our research introduces an innovative approach to characterizing a probability distribution based on its minimum and the outcome derived from the logarithmic transformation of the derivative of its quantile function (LDQ). This transformation, previously utilized in (Petersen, A., & Müller, H. G., 2016) to map density functions to a Hilbert space of functions under the assumption that density functions share a common support interval, has been adapted in our method to remove constraints on the coefficients. The main problem with the LDQ transformation is that it loses information about the location parameters of the density distribution due to the derivative operation, for which, the derivatives of two quantile functions

that differ by a constant term are equal. For this reason, we introduced minimum regression to reconstruct the position of the distribution.

We establish two linear regression models related to the minimum and LDQ functions, respectively. For the i th observation, the predictions of minimum value $\hat{y}_i^m$ and the LDQ function $\hat{y}_i^l(t)$ can be obtained, respectively, by:

$$\hat{y}_i^m = \hat{\beta}_0^m + \sum_{j=1}^p \hat{\beta}_j^m x_{ij}^m \quad \text{and} \quad \hat{y}_i^l(t) = \hat{\beta}_0^l(t) + \sum_{j=1}^p \hat{\beta}_j^l(t) x_{ij}^l(t)$$

where $\hat{\beta}_j^m$ is the coefficients' vector to estimate in the minimum regression and $\hat{\beta}_j^l(t)$ is the functional coefficients' vector to estimate in the LDQ functional regression. Finally, we can determine the predictions of quantile distribution $\hat{y}_i^Q(t)$ by the predicted result of $\hat{y}_i^m$ and $\hat{y}_i^l(t)$ through the inverse mapping of the LDQ.

Traditionally, regression methods involve fitting a single functional (statistical) model to a data sample. This estimated model is then used to describe or summarize the dataset and, potentially, to make predictions for new observations. While a single-model approach can be effective for some real-life datasets, it may not always result in accurate regression modeling. This is because the data sample might consist of several unknown subgroups, each of which follows its own local functional model between response and explanatory variables. Therefore, to accurately describe all relationships within the dataset, each smaller data group must be individually fitted with its own regression model. Clusterwise regression (CR) encompasses a series of methods that aim to partition large heterogeneous data sets into smaller, homogeneous groups (clusters) of observations while simultaneously fitting a regression model to each group (Späth, H., 1979). This approach simplifies the understanding of regression relationships in large heterogeneous data sets. By creating clusters, CR on LDQ functions improve the estimation of linear models that effectively summarize data relationships and the prediction power with respect to an unique regression model as also demonstrate on real data.

2. THE FUNCTIONAL CLUSTERWISE REGRESSION MODEL

The LDQ functions are treated as data function representations. Let's X_j (for $j=1, \dots, p$) be distributional variables, which assume as values probability density distributions or empirical ones based on observed data distributions. The elements of each distributional variable X_j (for $j=1, \dots, p$) are transformed into LDQ functions and represented as functional data (Ramsay, J. O., & Silverman, B. W., 2002) by considering a B-splines smoothing at the points corresponding to the quantile of the distributions.

Unlike traditional linear models, the functional linear model can capture dynamic characteristics, allowing for a more nuanced analysis of the direction and intensity of influencing factors (Xu, Y., et al., 2023) in the several cluster.

Let us consider an extension of Clusterwise regression method for distributional data on the LDQ functions using a functional representation of the data. The CRM-LDQ combines the dynamic clustering algorithm (Diday, E., & Simon, J. C., 1976) with the OLS-LDQ regression method for distributional-valued data.

For a fixed number K of clusters, it seeks the better partition $P_k = \{C_1, \dots, C_K\}$ and the best fitting models $\hat{y}_k$ for each cluster C_k , by minimizing the $SSE_{OLS-LDQ}$:

$$SSE_{OLS-LDQ}(P_k, \hat{y}_k \mid \beta_{ldq}(t), \beta_m) = \sum_{k=1}^K \sum_{e_i \in P_k} \left[|\epsilon_{i(k)}^l(t)|^2 + (\epsilon_{i(k)}^m)^2 \right]$$

where:

- ϵ_i^m is the i-th residual error of the minimum values;
- $\epsilon_i^l(t)$ is the i-th residual error function of the LDQ function.

After setting the number K of clusters, the algorithm is performed by alternating the following two steps, until convergence to a stationary value:

Step 1 - Representation step (best fitting):

The local regression models are estimated by minimizing the objective function $SSE(P_k, \widehat{y}_k^*)$ on the parameters $\beta_{(k)}^l(t)$ and $\beta_{(k)}^m$ ($1 \leq k \leq K$). The OLS estimations are provided as solutions of the two independent systems.

Step 2 - Assignment step (partitioning P_k):

The optimal clusters P_k which minimize the criterion $SSE(P_k, \widehat{y}_k^*)$ are obtained according to the following assignment rule:

$$P_k = \left\{ e_i \in E : |\widehat{\epsilon_{i(k)}^l(t)}|^2 + (\widehat{\epsilon_{i(k)}^m})^2 = \min_{h=1}^K \left[|\widehat{\epsilon_{i(h)}^l(t)}|^2 + (\widehat{\epsilon_{i(h)}^m})^2 \right] \right\}$$

Thus, the observation e_i is assigned to cluster P_k if the sum-of-squared errors are minimal for this cluster regression model.

In order to evaluate the fitting of the LDQ model we use the following indices:

- $\Omega = \frac{\{\sum_{i=1}^n [(\widehat{y}_i - \bar{y})^2 + s_{\widehat{y}_i}^2]\}}{\{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2 + s_{\bar{y}_i}^2\}}$ proposed by Dias, S., & Brito, P.(2015);
- $Pseudo R^2 = \min \left[\max \left[0; 1 - \frac{\sum_{i=1}^n d_W^2(y_i^Q(t), \bar{y}_i^Q(t))}{\sum_{i=1}^n d_W^2(y_i^Q(t), \bar{y}_i^Q(t))} \right]; 1 \right]$

proposed by Irpino, A., Verde, R., 2015

3. RESULTS

Our analysis employed Functional Clusterwise LDQ Regression to assess poverty risk across European Countries from distributional data, utilizing data sourced from Eurostat. The key variables used in the study are described in Table 1. The results, presented in Tables 2, show the model's strong performance in predicting poverty risk within these different contexts. The Functional Clusterwise LDQ Regression model exhibits high predictive accuracy and reliability across all regions. The consistently high value of the GoF indexes (Pseudo R^2 and W), coupled with low RMSE values, show the efficacy of the model in capturing the complex relationships between socio-economic factors and poverty risk. The variation in performance metrics across clusters and regions indicates that different predictors impact poverty risk differently depending on the degree of urbanisation. These results validate the utility of our approach in handling complex distributional data and highlight its potential for informing policymakers. By considering the differences in the influence of socioeconomic variables in different subgroups (clusters) on poverty risk, targeted interventions can be outlined to effectively mitigate poverty across Europe.

Table 1. Variables' description

Label	Variable
Y: Poverty	Persons at risk of poverty or social exclusion by degree of urbanisation
X1: Educational Attainment	Population by educational attainment level and degree of urbanisation
X2: Median Income	Share of people having income greater or equal to median national thresholds by degree of urbanisation
X3: Employment Rate	Employment Rate by degree of urbanisation

Table 2. Performance Measure of Regression Clusterwise on:

	European Cities			European Towns and Suburbans			European Rural Zones		
	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3
Pseudo-R2	0.91	0.86	0.88	0.90	0.91	0.87	0.88	0.90	0.91
RMSE	1,43	1.53	1.50	1.,45	1.40	1.50	1.50	1.43	1.38
Ω	0.90	0.84	0.87	0.91	0.92	0.86	0.88	0.90	0.93
R2 LDQ	0.88	0.84	0.86	0.87	0.88	0.85	0.86	0.89	0.90

4. REFERENCES

- [1] Bock H.H.- Diday E., Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, (2000).
- [2] Billard L., Diday E., Regression Analysis for Interval-Valued Data, in: Proceedings of the Seventh Conference of the International Federation of Classification Societies, Springer, 369–374, (2000).
- [3] Billard L., Diday E., Symbolic regression analysis, in: Classification, Clustering, and Data Analysis, Springer, 281–288, (2002).
- [4] Neto E.A.L., de Carvalho F., Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Stat. Data Anal. 52(3) (2008) 1500–1515.
- [5] Petersen, A., & Müller, H. G. Functional data analysis for density functions by transformation to a Hilbert space, (2016).
- [6] Späth, H. Algorithmus 39. Klassenweise lineare Regression. Computing, 22, 367-373. (1979).
- [7] Ramsay, J. O., & Silverman, B. W. (Eds.). Applied functional data analysis: methods and case studies. New York, NY: Springer New York. (2002).
- [8] Xu, Y., You, T., Wen, Y., Ning, J., Xiao, Y., & Shen, H.. Air quality research based on B-Spline functional linear model: a case study of Fujian province, China. *Applied Sciences*, 13(20), 11206. (2023).
- [9] Diday, E., & Simon, J. C. Clustering analysis. In Digital pattern recognition (pp. 47-94). Berlin, Heidelberg: Springer Berlin Heidelberg. (1976)
- [10] Dias, S., & Brito, P. Linear regression model with histogram-valued variables. Statistical Analysis and Data Mining: The ASA Data Science Journal, 8(2), 75-113. (2015)
- [11] Irpino, A., & Verde, R. Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. Advances in Data Analysis and Classification, 9(1), 81-106. (2015)

Future of Data Collection: Automated Data-Based Reporting

Maarja Männik (Statistics Estonia)

1. INTRODUCTION

1.1. Current situation

Currently, business entities are obligated to fill in statistical questionnaire forms, in which they need to aggregate data into different variables and where data for past periods are asked. The required information could be stored in different information systems and the data might not have been originally collected or retained in the level of detail that is asked for. Often, filling in a questionnaire requires the cooperation of several departments. This could be very time-consuming process, and it could also affect the quality of the data.

Various state agencies in Estonia request more than 400 reports (including statistical questionnaires) from enterprises.^[1] It is also complicated by the fact that there are many different standards in use, which means the same data is asked in different ways. As the reporting burden increases over time, it takes more and more people and money to fulfil all obligations. It also leads to the issue that enterprises have less time to focus on their core business activities and objectives.

1.2. The future of data collection: automated data-based reporting

To reduce the administrative burden on business entities, an alternative way of data collection is provided. The idea behind automated data-based reporting is that enterprises send their detailed data in a standardised and structured format directly from their enterprise resource planning (ERP) systems to different state agencies, where an automatic process is triggered that validates the data, aggregates it into needed variables, and submits the related online forms for them. Thus, the burden of aggregating the data is shifted from enterprises to state agencies. This also ensures that all needed variables are calculated in a unified way, which increases data quality. To make this solution to work, state agencies have agreed on a unified taxonomy and business processes towards clients.

This solution requires initial mapping of data by ERP developers but after that, there will be no or a little additional effort for accountants/data entrants to send the data regularly to state agencies. Currently, it might take several hours to fill in questionnaires but with data-based reporting, this could be replaced by a manual or automated 'Send' button in the ERP systems.

1.3. Background

Automated data-based reporting is one of the building blocks in the Real-Time Economy (RTE) project in Estonia. The project is led by the Ministry of Economic Affairs and Communications, and various state agencies are cooperating to apply the new method of data collection. In addition, private sector partners are included in working groups to obtain their valuable feedback.

The assessment of the economic impact of RTE solutions (data-based reporting, e-invoices, e-receipts, etc.) shows that these solutions could save the business sector more than 200

million euros a year. It is possible to save about 14.10 million working hours a year, which equals the full-time work of 7,000 people.[\[2\]](#)

1.4. Main challenges

- One of the main challenges in this project is the number of different questionnaires that need to be analysed thoroughly to create a logical and optimal taxonomy for automated workflow.
 - The questionnaires with large samples, and the ones including data that are commonly stored in ERP systems, are prioritised.
- Creating a unified taxonomy with various state agencies is a time-consuming and complex task.
 - Several working groups have been created to make the process as efficient as possible.
- Questionnaires include estimated/predictive questions, which cannot be automated.
- Questionnaires include questions which require data from different ERP systems.
 - Questionnaires should be designed in a way that they include similar types of data that could be retrieved from one information system.
- Samples do not support the automation of data collection.
 - Currently, many samples are taken at the end of the year based on past business results (e.g. turnover), but for automation to work, enterprises must know in advance who will need their data, and in what level of detail it will be asked. This means that new methodology for taking samples must be developed.
 - The samples are not constantly the same each year and every company is not required to provide same amount of data about the same field (e.g. job cost and payroll data). This means a lot of additional complexity and cost for ERP systems. The questionnaires and samples should be harmonised within the same field throughout various questionnaires.
- Some questionnaires (or questions) appear at rare intervals, and this does not support regular automated processes.
 - The same amount of data should be asked regularly.

2. METHODOLOGY

2.1. Taxonomy

Several state agencies have been working together to create a central unified taxonomy for automated data exchange, including the classifications and data fields, which will cover the data required for the various questionnaires and reports. If several agencies require similar information, they need to come to an agreement on classifications, ensuring that everyone's needs are considered. Classifications must be generated in a way that there are no duplicates.

In data-based reporting taxonomy, classifications are not managed as versions. Instead, all classifications must be floating. This reduces the development needs for both, the ERP systems, and the state agencies, as with versions all the mapping must be done again each time changes are made to the content.

For creating the taxonomy, questionnaires' variables are broken down into raw data for analysis, and after that the taxonomy needed for automated data collection is created. This is one of the most time-consuming steps in the process.

2.2. The process – how does it work?

After the taxonomy is created, ERP system developers need to map the data in their systems with the taxonomy and with technical tags to create a standardised data file. Once the

necessary data mapping has been completed, the process of sending and processing the data is as follows:

- 1) Data file is sent to Statistics Estonia from the ERP system over the secure data exchange layer X-Road.
- 2) Data file is validated by Statistics Estonia (structural and content checks).
- 3) A validation report is provided for the client.
- 4) If a file passes validation checks and all needed data sections are received, a related questionnaire form is submitted automatically in the electronic data submission environment.
- 5) Data is stored in a statistical activity database at the exact same aggregation level as the data collected by a questionnaire. This means that for the statistical teams there is no difference, from which source the data is collected.

2.3. Data file format

The data files are in XBRL GL based XML file format. It is an open standard that supports presenting data in a detailed level. The standard allows to present data in a dynamic way, which reduces duplicated information.

2.4. Data file content

The data is structured by logical data sections inside the data files. Data sections are grouped in logical data sets of similar content from the viewpoint of ERP systems. For example, labour cost data is in one section, accommodation data in other sections, etc.

One data section might be equal to one questionnaire, but it could also be the case that one questionnaire needs several data sections, or that one section could be input for several questionnaires.

2.5. Interval

Currently, the agreed interval for the submission of data to Statistics Estonia (and to other institutions) is once a month as this fulfils the data needs for producing official statistics.

3. RESULTS AND PRACTICAL APPLICATION

Currently, one ERP system has developed the new solution, and one enterprise is sending their salary data to Statistics Estonia through that system. Several other software enterprises are also currently developing the solution to offer the service to their clients.

The internal automated workflow in Statistics Estonia is fully designed, developed, and working, but we are also continuously testing and improving the process.

4. MAIN FINDINGS

As the automated data-based reporting method is currently being implemented, we cannot yet provide enough data for comparison or for conclusions to be drawn.

Since detailed data will be received straight from the ERP systems, we can assume an increase in data quality. As the data will be aggregated into different variables by Statistics Estonia, there will be assurance that it is done in a unified way.

There is currently one enterprise that is sending their data from their ERP system to Statistics Estonia and already this one example has shown that the method is working well. The feedback from the enterprise's representative was that they would never wish to go back to the traditional way of filling in questionnaires as it is far too time-consuming and complicated.

The centre of gravity of aggregating the data in different complex variables will shift from enterprises towards Statistics Estonia (and other state agencies).

Business entities will have more time to focus on their core business. They could also create additional internal digital analytical solutions based on the implementation of data-based reporting methodology.

5. REFERENCES

- [1] The information in this section is excerpted from the website of the Ministry of Economic Affairs and Communications, available at <https://realtimeeconomy-bsr.eu/data-based-reporting>.
- [2] The information in this section is excerpted from the article of the Ministry of Economic Affairs and Communications titled "The impact of real-time economy", available at <https://realtimeeconomy-bsr.eu/impact-real-time-economy>.

Gemma 2 Large Language Models for Aspect-Based Sentiment Analysis on Italian Texts: A Use Case for Tourism Statistics

Francesco Ortame, Mauro Bruno, Elena Catanese & Francesco Pugliese (Italian National Institute of Statistics, Istat)

1. INTRODUCTION

Aspect-Based Sentiment Analysis (ABSA) is an advanced Natural Language Processing (NLP) technique that aims to identify specific aspects of a given text and the sentiments associated with them. ABSA represents a promising opportunity for National Statistical Offices (NSOs) to extract more detailed insights from textual data compared to traditional text classification techniques. For instance, the Italian National Institute of Statistics (Istat) conducts a yearly survey titled “*Viaggi e Vacanze in Italia e all’Estero*” (*Trips and Holidays in Italy and Abroad*), which gathers data on the travel habits of Italian residents, such as the number of trips, destinations, reasons for travel, types of accommodation, and duration of stays. This survey is carried out in compliance with European regulations for tourism statistics (Reg. 692/2011). The inclusion of unstructured data sources, such as online accommodation reviews, could enrich this survey by providing more detailed information regarding tourist preferences and satisfaction levels. This additional data could even be exploited to rank tourism activities by geographic area based on information systematically extracted from reviews.

Recent Large Language Models (LLMs) have shown remarkable capabilities in understanding and generating human-like text. Pre-trained LLMs offer significant advantages in a variety of NLP tasks, including ABSA, due to their deep contextual understanding and versatility without requiring extensive retraining or fine-tuning. However, their high computational requirements pose challenges for large-scale inference in production environments. A potential solution is to use LLMs to train smaller, more efficient models. For ABSA, LLMs can be used to automatically label textual training datasets, particularly appealing for non-English texts. Additionally, the rise of open-source and open-weight LLMs provides opportunities for customization, fine-tuning, and local deployment—addressing concerns about data privacy and control in Official Statistics (OS).

Although LLM-based approaches to automating ABSA have shown promising results [1], our study presents a novel contribution by focusing on the Italian language and predefined aspect categories. We work in the framework of the 2018 ABSITA campaign by EVALITA [2], classifying Italian accommodation reviews from *Booking.com* based on predefined aspects and sentiment polarities. Our approach prioritizes generalizability by avoiding LLM fine-tuning for specific aspects, making it adaptable to various topics in tourism statistics and OS more broadly.

For this study, we employ Google’s open-weight Gemma 2 models [3]—specifically the 2-billion and 9-billion parameters variants—to examine the performance of different model sizes. Additionally, we explore the impact of slight variations in prompting strategies, comparing results with and without detailed aspect descriptions in the prompts.

2. METHODOLOGY

In this section, we introduce the data used to validate our approach and the models implemented for Aspect-Based Sentiment Analysis.

2.1. Data

We utilized the test set from the EVALITA 2018 ABSITA task, which consists of 2,718 sentences in Italian from hotel reviews. Each sentence is annotated with aspect categories and polarities. We only used the test set for evaluation, to have a term of comparison and because our approach does not require training on task-specific data.

There is a predefined set of aspects, namely: *cleanliness*, *comfort*, *services*, *staff*, *value for money*, *wi-fi*, and *position*. The polarities can be *positive*, *negative*, *neutral*, and *mixed* (if both positive and negative sentiments are expressed towards the same aspect). Here is an example of a sentence from the dataset:

“Ottima la disponibilità del personale, e la struttura della stanza”
(*Great staff availability and room layout*)

The aspects identified in this sentence and the associated polarities are: (*staff*, *positive*), and (*comfort*, *positive*).

2.2. Models

We evaluated two models from Google’s recently released Gemma 2 family of open-weights LLMs: Gemma 2 (2B), a 2 billion parameter model, and Gemma 2 (9B), a 9 billion parameter model. The Gemma models are based on the research and technology used to create Google’s commercial Gemini models. They are relatively lightweight open models designed to be efficient while still showing strong language understanding and generation capabilities.

The models are based on a decoder-only Transformer architecture, widely adopted for LLMs due to its capacity to capture contextual relationships in natural language. The core innovation of Transformers is the multi-head self-attention mechanism, which allows the model to weigh the importance of different words in a sentence based on their relevance to one another [4]. This is critical in ABSA, as sentiment analysis depends on understanding the relationships between words to properly assign aspect categories and corresponding polarities. For example, in the sentence “*Ottima la disponibilità del personale*” (*Great availability of the staff*), the model needs to recognize the positive sentiment conveyed by “*Ottima*” (*Great*) and relate it to the correct aspect category *staff*.

2.3. Prompting

We used the following (simplified) prompting template:

“You are an Aspect-Based Sentiment Analysis (ABSA) assistant for Italian texts. Your task is to analyse accommodation reviews in two steps:

1. Aspect Category Detection (ACD)
Identify one or more aspect categories from this list: ...

2. Aspect Category Polarity (ACP)
For each identified category, assign a sentiment polarity: ...

Output Format
Return the result using this format without any extra text:

(Aspect Category 1, Polarity 1) (Aspect Category 2, Polarity 2)

Examples...”

We provided the model with two examples, one with multiple aspects identified, and one with a single identified aspect. Furthermore, we implemented two slightly different versions of this prompt: a standard version, that only includes the names of the aspect categories, and a descriptive version, that includes a detailed description of each category.

3. RESULTS AND PRACTICAL APPLICATION

Here, we show the comparative performance of the models we tested, both with standard and descriptive prompting. Performance metrics are listed in Table 1.

Table 1. Model performance

Task	Metric (micro)	Gemma 2 (2B)		Gemma 2 (9B)	
		Standard	Descriptive	Standard	Descriptive
Aspect Category Detection (ACD)	Precision	0.7090	0.6988	0.7055	0.7182
	Recall	0.6801	0.6970	0.7455	0.7683
	F1-score	0.6943	0.6979	0.7250	0.7424
Aspect Category Polarity (ACP)	Precision	0.6697	0.6741	0.7048	0.7172
	Recall	0.6514	0.6666	0.7037	0.7221
	F1-score	0.6604	0.6703	0.7042	0.7179

As expected, Gemma 2 (9B) consistently outperforms the smaller Gemma 2 (2B) model (except for Precision in the ACD task with a standard prompt). Moreover, the advantage of including detailed aspect descriptions in the prompt, albeit small, seems consistent, reducing the risk of aspect misclassification. The Gemma 2 (9B) model tends to balance Precision and Recall more effectively, particularly with the descriptive prompt. For instance, in the ACD task, Recall for Gemma 2 (9B) increases from 0.7455 (standard) to 0.7683 (descriptive), with Precision also increasing from 0.7055 to 0.7182, reflecting a well-rounded improvement, while in the smaller Gemma 2 (2B) model, there is a slight trade-off between Precision and Recall in some cases, particularly in the ACD task, where Recall improves with the descriptive prompt, but Precision decreases marginally.

Another key observation relates to inference speed: the Gemma 2 (2B) model processes approximately 23 texts per second, while Gemma 2 (9B) processes around 10 texts per second, on average. Both models were tested on an NVIDIA A100 GPU (80GB vRAM), with batch sizes of 256 and 64, respectively.

There are several ways in which such a pipeline can be integrated into the production process of tourism statistics: 1) Begin by sampling a relatively small amount of textual data related to tourism, like X posts and accommodation reviews. Utilize larger LLMs to automatically label this training set. Once labelled, the training set can be validated and subsequently used to train more efficient models, such as BERT-based models or LSTMs. These models can then be employed to make inferences on vast amounts of tourism data. 2) Alternatively, one can

opt for lighter LLMs to directly extract predictions from large datasets. While this approach demands significantly more computational power compared to traditional models, it can yield superior results, especially for complex tasks like aspect-based text classification where domain-specific training sets are scarce.

4. MAIN FINDINGS

These preliminary results make way for a more extensive exploration of using LLMs to integrate and enrich tourism statistics. While the relatively smaller LLMs do not achieve the performance levels of models specifically fine-tuned for this specific domain [5], their adaptability and generalization capabilities position them as promising candidates for labelling unseen data for which training sets are not yet available. Furthermore, larger LLMs may demonstrate significantly enhanced performance compared to those analysed in this study.

Future research could benefit from the development of aspect-based indicators that extend beyond accommodation to encompass a wider range of tourism activities, including museums, restaurants, and historical sites. This broader focus could provide a more comprehensive understanding of tourist preferences and satisfaction across various dimensions of tourism statistics.

5. REFERENCES

- [1] Simmering, Paul F., and Paavo Huoviala. "Large language models for aspect-based sentiment analysis." *arXiv preprint arXiv:2310.18025* (2023).
- [2] Basile, Pierpaolo, et al. "Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA)." *EVALITA Evaluation of NLP and Speech Tools for Italian*. CEUR, 2018. 1-10.
- [3] Mesnard, T., et al. "Gemma: Open models based on gemini research and technology." *arXiv preprint arXiv:2403.08295* (2024).
- [4] Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).
- [5] Cimino, Andrea, Lorenzo De Mattei, and Felice Dell'Orletta. "Multi-task learning in deep neural networks at evalita 2018." *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* (2018): 86-95.

Harmonised statistics on income and living conditions in Europe, innovations and development over time

Estefania Alaminos, Alessandro Albano, Emilio di Meglio, Agata Kaczmarek-Firth
(European Commission, DG Eurostat); Ruzhdie Bici (Sogeti)

1. INTRODUCTION

The EU statistics on income and living conditions (EU-SILC) data collection covers data on income, poverty and living conditions in European countries, providing a comprehensive picture of the social and economic situation of individuals and households across Europe. To monitor social inclusion, inform policy decisions and compare social conditions across countries, it is crucial that all participating countries collect data in a harmonised manner. This can be achieved through a robust legal basis.

[Regulation \(EU\) 2019/1700](#) establishes a common framework for collecting European statistics related to persons and households. It aims to improve the quality, comparability and timeliness of European statistics, by integrating social surveys with administrative data and other sources. This integration is expected to reduce the burden on respondents, and improve data quality and timeliness. However, challenges remain, including ensuring data access while complying with data protection and confidentiality rules, and improving data integration and analysis.

This article presents the latest developments in countries that carry out the EU-SILC survey, from a traditional survey-based framework to its current role in a comprehensive, multisource data collection process. In addition, it shows how countries are enhancing the survey by using administrative data for information on income and introducing future plans.

2. CHANGES IN DATA COLLECTION

2.1. Data collection in EU-SILC

EU-SILC consists of two types of data: (i) cross-sectional data which provides an overview of income, poverty, social exclusion and living conditions at a specific point in time or over a certain period; and (ii) longitudinal data which follows changes at individual level over time, observed at regular intervals over a 4-year period.

Data collection for EU-SILC is a decentralised process, meaning that each country is responsible for collecting its own data, and using its own survey design, sampling frame and data collection methods.

2.2. Recent innovations at national level

When carrying out an EU-SILC survey, each country is responsible for: i) collecting data from a representative sample of households; and ii) data processing, always using Eurostat methodology [1] for harmonising output.

Innovation in data collection methods can have a positive impact on response rates [2]. In recent years, countries have been moving from traditional data collection methods towards

more innovative methods¹ (Figure 1). During the COVID-19 pandemic, due to fieldwork limitations, some countries introduced the computer assisted telephone interview (CATI) alternated with other methods (such as computer assisted web interview (CAWI)), which had a mixed impact on data quality. While some countries saw improvements in data quality thanks to the use of administrative data, others saw a decline in quality due to the changes in data collection methods. In addition, response rates were negatively affected, as many countries faced challenges in conducting surveys during the pandemic, resulting in lower response rates.

During the 2023 data collection, CAWIs are used by several countries (Austria, Denmark, Estonia, Finland, Germany, Lithuania, Luxembourg, Netherlands, Spain and Sweden) [3]. Norway collected 2023 SILC data using only CATIs.

Some of the countries that use mixed-mode approaches reported that they encountered challenges when using these methods but achieved high response rates, ranging from 70% in Austria to 80% in Luxembourg. In Finland, the use of CAWIs has led to significant increases in response rates, especially among younger respondents, and improved data quality. In addition, some countries leveraged administrative data to reduce response burden and further improve data quality, highlighting the effectiveness of these approaches in improving survey outcomes.

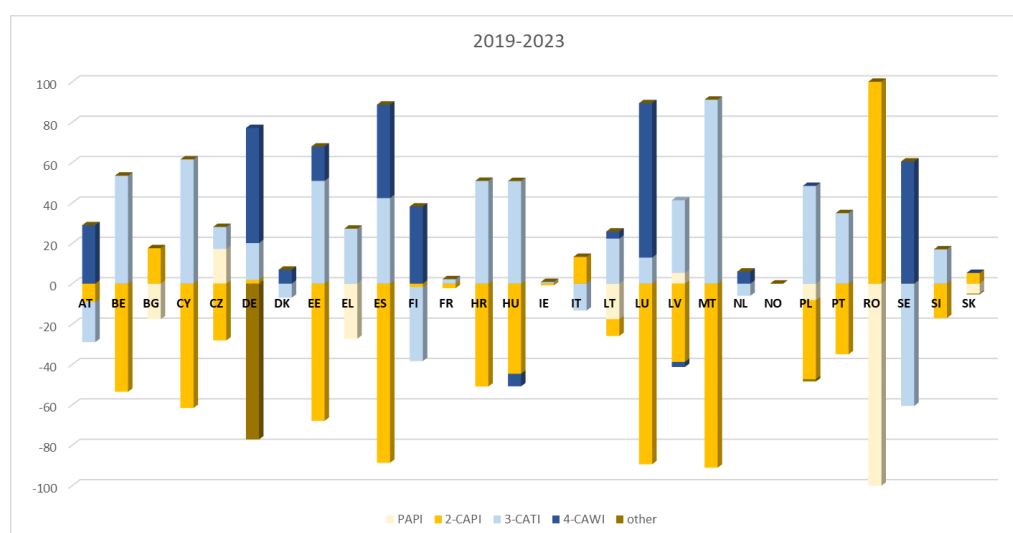


Figure 1. Differences in data collection methods followed by EU-SILC national teams, 2019-2023

Note: Eurostat's own compilation from the information reported by countries through the EU-SILC variable PB270 'interviewing mode used (person)'. The results are the differences in the proportions that each method is used.

¹ There are different methods used when collecting data for EU-SILC: i) PAPI (paper and pencil personal interview); ii) CAPI (computer assisted personal interview); iii) CATI (computer assisted telephone interview); iv) CAWI (computer assisted web interview); and v) self-administered (respondents complete the questionnaire themselves).

2.3. Use of data from administrative records in collection of income data

Administrative data are widely used by selected respondent countries² (Denmark, Finland, the Netherlands, Slovenia, Sweden and Norway) even for non-income variables. Using administrative data has become a useful practice in EU-SILC, especially for collecting income data (Austria, Belgium, Croatia, Cyprus, Denmark, Estonia, Finland, France, Ireland, Italy, Latvia, Lithuania, Malta, Netherlands, Norway, Slovenia, Spain and Sweden) [4].

Regarding income data, some countries reported using administrative data when conducting their EU-SILC surveys to reduce response burden and improve data quality. Specifically, Estonia used administrative data to support its mixed-mode approach, while Spain used a mixed-mode approach to improve the quality of data on employee income. Similarly, Portugal also used administrative data to improve data quality, particularly on employee income, and achieved a high response rate of 80%. It is very interesting to see that many countries have developed algorithms to integrate administrative data with data from other sources.

3. THE FUTURE OF EU-SILC

EU-SILC faces different challenges for the future:

- integration with other data sources: the survey component will be integrated with other data sources, such as administrative data and big data, to improve the quality and timeliness of statistics;
- increased use of administrative data: EU-SILC will make greater use of administrative data, such as tax and social security records, to reduce response burden and improve data quality;
- development of new modules: new modules will be developed to collect data on different relevant topics such as energy poverty, transport poverty, and mental health and well-being;
- improvement of data quality: efforts will be made to improve the quality of EU-SILC data – which includes strengthening IT data validation tools – and editing techniques.
- increased focus on policy relevance: EU-SILC is designed to remain policy-relevant, with a focus on providing data that can be used to shape and inform policymaking.

4. MAIN FINDINGS

EU-SILC is the primary source of data for policymakers on income, poverty and social exclusion. Initially, data were collected mostly through surveys. Over time, administrative sources have increasingly been used. Imputation and small-area estimation have also been increasingly used. Register data are an important data source used in EU-SILC in a growing number of countries. However, it is still unavailable in 14 countries.

Eurostat has encouraged national SILC teams to:

- increase the use of administrative data to improve data quality and reduce response burden;
- improve data collection methods; and
- strengthen data validation procedures to ensure the quality of EU-SILC data.

² 'Selected respondent model' means a mode of sampling based on individuals, where the household to which the selected respondent belongs is the sample household and the selected respondent is the sample person. In this model only selected respondents are subject to interviews at individual level for the collection of data. [1]

Also, there have been many innovations on the users' side of EU-SILC. The most notable is the development of EUROMOD [5], a tax-benefit microsimulation model that allows users to simulate EU-SILC data based on certain scenarios. EUROMOD is already used in EU-SILC production to extrapolate gross income from net. In future, machine learning algorithms should enable data collection and processing to become further automatised in order to further reduce the response burden and increase efficiency.

5. REFERENCES

- [1] Eurostat - [Methodology - Income and living conditions - Eurostat \(europa.eu\)](#): EU Statistics on Income and Living Conditions (EU-SILC) methodology.
- [2] Aguilera, A., Fernández-Esquinas, M., & Santa-Marina, L., New data collection methods for innovation studies: An overview and a research agenda, *Technological Forecasting and Social Change*, 168, 2021. doi: [10.1016/j.techfore.2021.120392](#)
- [3] Eurostat – [EU-SILC metadata](#)
- [4] Eurostat – [EU-SILC national quality reports](#)
- [5] [European Commission – EUROMOD - Tax-benefit microsimulation model for the European Union.](#)

Harnessing Machine Learning for Superior GDP Forecasting: A Comparative Study of Machine Learning versus Econometric Models

Bogdan Oancea (National Institute of Statistics and University of Bucharest)

1. INTRODUCTION

Forecasting GDP is vital for economic planning and policy-making, providing critical insights for government and financial institutions. Traditional GDP forecasting methods, which often depend on statistical models and historical data, can struggle with the intricacies and nonlinearities inherent in modern economies, especially during periods of rapid change. In this context, machine learning techniques have become valuable tools for economic forecasting, offering significant improvements over traditional statistical methods (Hull, 2021). Machine learning algorithms, capable of processing large datasets and uncovering complex patterns, have been shown to enhance GDP prediction accuracy compared to conventional autoregressive models (Chu & Qureshi, 2023).

Previous studies have demonstrated the superior performance of various machine learning models, including Random Forests, Support Vector Machines, and Neural Networks, in economic forecasting (Tamara et al., 2020; Flannery, 2023). These models not only outperform traditional methods but also provide more reliable and timely predictions. For instance, Random Forests have shown excellent individual performance in GDP nowcasting, and ensemble approaches have further improved accuracy. Deep learning approaches, such as LSTM networks, have also proven effective, capturing long-term dependencies in data better than traditional time series models (Kurihara & Fukushima, 2019).

This paper proposes the use of LSTM networks for GDP forecasting in a univariate setting and compares their performance with the SARIMA econometric model. The results demonstrate the superiority of the LSTM network in forecasting accuracy, underscoring the benefits of machine learning techniques in handling complex economic data.

2. METHODOLOGY

The study utilizes quarterly GDP data for Romania from Q1 1995 to Q4 2023, sourced from the National Statistics Institute of Romania. This dataset, provided in national currency and adjusted to 2020 average prices, offers a comprehensive view of economic performance over nearly three decades. Descriptive statistics highlight the pronounced seasonality and upward trend in the data (see also figure 1), confirming its non-stationary nature (ADF statistics is 0.31 with p-value=0.97). Table 1 shows the descriptive statistics for our data series. Traditional statistical methods often require stationarity for accurate predictions, a limitation not faced by many machine learning models, including LSTM networks.

Table 1. Descriptive statistics of the GDP

Statistics	GDP
Mean	199681.0552
Standard Error	6114.13206
Median	192018.65
Standard Deviation	65851.21759
Kurtosis	-0.483138316
Skewness	0.389917727
Range	274466.4
Minimum	84817.9
Maximum	359284.3

LSTM networks, a type of Recurrent Neural Network designed to capture temporal dependencies, are particularly suited for time series forecasting (Hochreiter & Schmidhuber, 1997). Unlike traditional RNNs, which suffer from the vanishing gradient problem, LSTMs can maintain and update information over long periods due to their unique architecture, making them ideal for sequential data such as GDP (Graves et al., 2009).

The LSTM model used in this study includes two LSTM layers and a final dense layer, optimized using regularization techniques to prevent overfitting. We used the dropout of the recurrent neurons in the second LSTM layer technique as well as a L2 kernel regularization in the final dense layer. The activation function of the LSTM neurons was ReLU and the optimization algorithm was *adam*.

The study compares the LSTM network's performance with the SARIMA model, an extension of ARIMA that includes seasonal components (Nontapa et al., 2020). The dataset was split into training and testing subsets, with the last four quarters reserved for testing. Hyperparameter tuning was performed using grid search, and the models were implemented using Python libraries *Keras* (Chollet, 2015) and *TensorFlow* (Abadi et al., 2016) for the LSTM network, and *statsmodels* (Seabold & Perktold, 2010) for SARIMA. The hyperparameters used in the grid search procedure for the neural network were the number of training epochs, the number of neurons in the LSTM layers, the batch size, the L2 constant for the regularization in the dense layer and the percent of dropout neurons in the LSTM layer. For SARIMA we used p, d, q, P, D, Q , while s was kept fixed at the value of 4.

3. RESULTS

The LSTM network's performance was evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The LSTM network consistently outperformed the SARIMA model across all metrics, demonstrating its superior capability in capturing complex patterns and trends in the data. The study's findings align with existing research, confirming the advantages of machine learning methods over traditional econometric techniques for GDP forecasting.

In figure 1 we show the training and testing, actual versus predicted, values while in table 2 we show the values of the performance metrics.

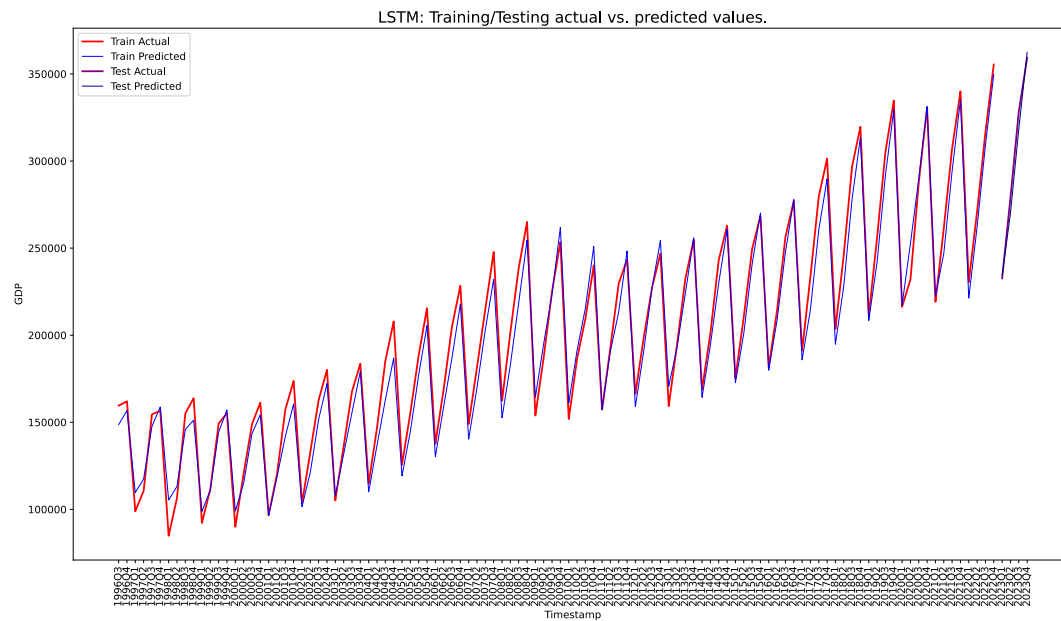


Figure 1. Training/testing actual versus predicted values

Table 2. Performance metrics for LSTM and SARIMA models

Performance metrics	LSTM	SARIMA
MAPE	1.96%	2.56%
MAE	5955	7144
MSE	48670983	63999869

4. MAIN FINDINGS

Machine learning, particularly LSTM networks, offers significant advantages for economic forecasting, providing enhanced accuracy and the ability to handle complex, nonlinear data. This study shows that LSTM networks outperform traditional econometric models like SARIMA in GDP forecasting, highlighting the potential of machine learning techniques to improve the precision and reliability of economic predictions. The results support the growing body of evidence that machine learning approaches are highly effective in economic analysis and policymaking, offering robust tools for better-informed decisions.

5. REFERENCES

- [1] Hull, I. (2021). Machine Learning and Economics. In: Machine Learning for Economics and Finance in TensorFlow 2. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-6373-0_2.
- [2] Chu, B., & Qureshi, S. (2023). Comparing Out-of-Sample Performance of Machine Learning Methods to Forecast U.S. GDP Growth. *Computational Economics*, 62, 1567–1609 (2023).
- [3] Tamara, N., Muchisha, N. D., Andriansyah, A., & Soleh, A. M. (2020). Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms. Munich Personal RePEc Archive, MPRA Paper No. 105235.
- [4] Flannery, J. (2023). Recurrent Neural Networks for flash GDP estimates in Ireland: A Comparison with Traditional Econometric Methods, ICT.44 Retrieved June 1, 2024, from <https://arc.cct.ie/ict/44>.
- [5] Kurihara, Y., & Fukushima, A. (2019), AR Model or Machine Learning for Forecasting GDP and Consumer Price for G7 Countries, *Applied Economics and Finance*, 6(3), 1-6.
- [6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [7] Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), 855-868.
- [8] Nontapa, C., Kesamoon, C., Kaewhawong, N., Intrapai boon, P. (2020). A New Time Series Forecasting Using Decomposition Method with SARIMAX Model. In: Yang, H., Pasupa, K., Leung, A.C.S., Kwok, J.T., Chan, J.H., King, I. (eds) *Neural Information Processing. ICONIP 2020. Communications in Computer and Information Science*, vol 1333. Springer, Cham. https://doi.org/10.1007/978-3-030-63823-8_84.
- [9] Chollet, F. (2015). Keras: The Python Deep Learning library. <https://keras.io>.
- [10] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX symposium on operating systems design and implementation*, 265-283.
- [11] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In van der Walt, S. & Jarrod Millman, J. (Eds.) *Proceedings of the 9th Python in Science Conference*, 92 – 96.

Households and their dwellings as communities in administrative data

Helle Visk (Statistics Estonia)

1. INTRODUCTION

The statistics of households and families is vital part of census. In a register-based census, households and families have to be constructed solely based on administrative data. The definition of a household is different from surveys: rather than viewing household as a housekeeping unit, we follow household-dwelling concept: household consists of *all* people living in same dwelling.

Each household may contain one or more families, or no families at all. Here, family stands for nuclear family in a strict sense: it is either a couple with or without children or a lone parent with children. If a child has a partner or their own children in the household, they are no longer considered a child but form their own family.

In the presence of nationwide Population Register (PR) it is natural to use its data on place of residence to assign people to households. If this data lacks accuracy on address level, it may result in biases in household and family statistics. For example, in Estonian register-based pilot census in 2016, PR-based families were skewed towards lone parents: 41% of families were of lone parents, whereas there were only 24% lone parent families in 2011 census [1]. The reason behind this bias is that family members are dispersed between multiple dwellings in PR.

We describe a novel method for constructing households that was used in Estonian 2021 census and improved register-based household and family statistics.

2. METHODOLOGY

2.1. Reuniting families means reassigning dwellings

The key to better statistics is to restore families whose members are divided between multiple addresses. In nuclear families, people are connected by parent-child relationship or partnership. Links between children and their parents are well covered in Estonian PR. Finding partners is more complicated but administrative data offers clues: being married to someone or co-owning property suggest possible partnership, while alimony disputes or divorce make partnership unlikely.

Reconstructing families is closely tied with assigning dwellings for each household. If we have decided that people *A* and *B* from different registered addresses most likely live together, we must also decide which dwelling they inhabit. Possible home candidates include either person's registered addresses or properties. Situation is obscured further by other family members that *A* and *B* may have. For example, they might live by parents of one of the spouses. If one partner coparents children with some third person, children's addresses may also need to be changed. Given example was about partners that were reunited despite their different addresses, but sometimes shared relationship with a dwelling (as place of residence, co-ownership, or some other link) is the only hint that people are partners.

As constructing families and households and assigning dwellings for them are closely intertwined, we approach both problems simultaneously.

2.2. Network of people and places

We consider people and places as nodes of a network, or in mathematical terms, a graph. Edges are relationships between people (e.g., parenthood, marriage) or between people and places (registered place of residence, electricity contract). Basically, we are constructing a network of all people and places. Our goal is to divide that graph into subgraphs, each including household members and their dwelling.

We assume that household members are strongly connected, and they are also strongly connected with their dwelling. This objective aligns well with definition for a community in a network: “group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups in the network” [2]. Finding communities in a network is a well-studied problem and there are many algorithms implemented in various software packages.

Some connections are stronger than others. For example, it is highly probable that married couples share the same dwelling, but most adult people do not live at their parents’ house. For that reason, we consider a weighted graph: for each edge, we attach a numerical value representing the connection strength.

2.3. Algorithm to find households and dwellings

To build the graph of people and places we used data from 17 registers. Altogether, there were 24 types of connections between people (child-parent connections, marriages, divorces, alimony disputes, sharing vehicle, buying prescription drugs for another person, family and subsistence benefits, etc.) and 23 types of connections between people and places (postal addresses and registered places of residence from various sources, real estate, electricity contracts, workplaces, schools, kindergartens, health care facilities, pharmacies, etc.). Linking data was straightforward because in Estonia, people, addresses and businesses have unique identifiers that are used across different sources.

Additionally, we used data from Estonian Labour Force Survey and Estonian Social Survey (part of European Union Statistics on Income and Living Conditions survey, EU-SILC for short) to model edge weights. Links between people were assigned weights representing probability of living in same household (modelled with logistic regression). Links between people and places were assigned weights representing probability of living on that address (random forest). Covariates were different types of connections. For links between people and places, the explanatory variables included municipality level data (e.g., municipality of visited pharmacy) and geographical distances from kindergarten, school, and work.

The graph of all people and places consisted of 5.2 million nodes and 7.8 million edges. Next, community detection was applied. For efficiency, this was performed in two steps. First, Louvain method [3], known for speed and scalability, was used to break the initial graph to smaller subgraphs of up to 5000 nodes. After that, the Infomap algorithm [4] was applied to each subgraph recursively, until the resulting communities were small enough or there was no significant improvement in modularity.

The resulting communities were very close to households, albeit a bit skewed towards larger households. In postprocessing phase, some of the least connected communities were broken into smaller ones according to heuristics. For example, if there were multiple adult siblings living with parents, the oldest one would form their own household.

The final step was to ensure that each household would have exactly one place of usual residence (more details in [5]). People in institutional households and homeless people were identified from other sources, not via the graph-based process.

3. RESULTS

After applying the graph-based algorithm, most people (75%) were left on their officially registered address in PR. By reallocating people to addresses, some people ended up in different municipalities. As a result, population count increased by at least 2% in 27% of the municipalities and decreased by more than 2% in 29% of the municipalities. The losses were greatest in small islands and other popular summer destinations, growth was largest in Russian speaking small towns in Northern Estonia.

To assess quality of the graph-based households and dwellings we compared the results with EU-SILC, the source for annual household statistics. Survey data was collected 1–5 months after census moment 31 December 2021.

Assigned places of residence were not more accurate than PR. When comparing graph-based dwellings with EU-SILC, there was 96.6% (standard error (SE) 0.3%) coincidence on municipality level and 79.3% (SE 0.8%) on dwelling level. PR dwellings matched with EU-SILC in 97.2% (SE 0.3%) of people on municipality level and 83.7% (SE 0.7%) on dwelling level. Such comparisons probably overestimate accuracy in population, as it is easier to reach respondents that have correct data in PR. Also, EU-SILC is a panel survey, there is an overlap between respondents of consecutive years: some of the people were part of training data.

Comparing housekeeping unit households from survey with household-dwelling units from registers is not straightforward. Majority of households are the same under both definitions, but some household-dwelling units include multiple housekeeping units. In Estonian population and housing census 2011 (PHC 2011), there were 7% less households and even 20% less one person households with household-dwelling definition [6]. In Table 1, column EU-SILC* corrects the share of each household type to match household-dwelling definition according to patterns observed in PHC 2011.

Graph-based households are great improvement over PR: share of couples and lone parents is consistent with EU-SILC*. Still, multiperson non-family households and lone fathers with adult children are likely overestimated and one person households underestimated. Also, in graph-based households there were less multifamily households and couples without resident children or with adult children, and more lone mothers and couples with young children.

Table 1. Household type in PR, EU-SILC and 2021 census.

Household type ¹			EU-SILC	(95% CI)	EU-SILC*	Census (graph)	PR
No families	One person		44.6	(43.4 – 45.8)	39.2	37.0	40.5
	Multiperson		0.5	(0.4 – 0.7)	0.9	3.2	3.5
One family	Couples	Without resident children	21.4	(20.3 – 22.4)	22.6	21.1	13.0
		At least one resident child under 25	22.3	(21.4 – 23.2)	24.1	25.5	15.6
		Youngest resident son/daughter 25 or older	2.2	(1.7 – 2.7)	2.6	1.8	4.1
	Lone fathers	At least one resident child under 25	0.7	(0.5 – 1.0)	0.7	0.7	2.1
		Youngest resident son/daughter 25 or older	0.2	(0.1 – 0.3)	0.3	0.6	1.2
	Lone mothers	At least one resident child under 25	4.5	(3.9 – 5.0)	4.6	6.1	9.4
		Youngest resident son/daughter 25 or older	2.8	(2.3 – 3.2)	3.2	3.0	6.0

¹ Family households may include additional members that are not part of family nuclei.

Household type ¹	EU-SILC (95% CI)	EU-SILC*	Census (graph)	PR
Multiple families	0.9 (0.5 – 1.3)	1.8	1.1	4.5

4. CONCLUSION

Inaccuracies in place of residence data in PR makes it unsuitable to form households for census. But even then, it is possible to achieve feasible household and family statistics by using additional administrative data. Household and its dwelling can be modelled as a community in network of people and places. Then, finding households reduces to community detection. In Estonian 2021 census, statistics of graph-based households was reasonably close to survey results and significant improvement over PR households.

5. REFERENCES

- [1] Statistics Estonia, Estonia's first register-based pilot census (2017). <https://www.stat.ee/sites/default/files/2020-07/Results%20of%20the%20first%20pilot%20census%202016%20%281%29.pdf>
- [2] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Communities in Networks, Not. Am. Math. Soc., vol. 56, no. 9 (2009), 1082–1097.
- [3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech., vol. 2008, no. 10 (2008), P10008, doi: 10.1088/1742-5468/2008/10/P10008.
- [4] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proceedings of the National Academy of Sciences, vol. 105, no. 4 (2008), 1118–1123, doi: 10.1073/pnas.0706851105.
- [5] Statistics Estonia, Estonian 2021 census: determining households and dwellings from registers using graphs, Geneva: United Nations Economic Commission for Europe (2023) https://unece.org/sites/default/files/2023-09/DOC%2010_2314617E.pdf
- [6] K. Kütt, Households and families in the register-based census, Quarterly Bulletin of Statistics Estonia, vol. 1 (2015), 148–155.

How far would you walk to the supermarket? Estimating access to shops across the United Kingdom

Joni Karanka, Davide Deflorio & Ahmet Aydin (Office for National Statistics, UK)

1. INTRODUCTION

The years 2022 and 2023 were characterised in the UK by remarkably high inflation. In February 2022, the Consumer Price Index rose above 5% for the first time since the 1990s, reaching a high of 9.6% in November 2022. One key concern of such sudden and high inflation relative to recent periods was its impact on different demographics.

There was considerable public debate on how inequitable the impact of inflation could be, such as whether essential items and services (such as food and heating) experienced more inflation than the rest of the basket, potentially impacting those with the lowest incomes.

As part of research carried out in ONS we investigated the impact of inflation on those with the lowest incomes, such as analysing the inflation of a basket of low-cost essential items. Would consumers with the least income experience less choice and potentially fewer retail outlets from which to buy goods from? A lack of access to shops would reduce consumer choice, potentially making them more vulnerable to inflation: if the price of goods raises, they would not be able to substitute them for cheaper items, or to consider retailers that might stock cheaper items.

Here we present evidence on the geospatial distribution of supermarkets within Britain in relation to the households that can access them. The key finding is that low-income households living in rural areas have the lowest access to supermarkets close to them, putting them at greater risk during periods of high inflation.

2. METHODOLOGY

The aims were to estimate the distribution of and accessibility to specific supermarkets and small retailers for several types of households. Initially we did the analysis by identifying the location of each shop of each retailer which could then be mapped onto an ONS geographic classification known as Lower layer Super Output Areas (LSOAs). LSOAs divide the UK into 42,619 geographic areas of comparable population (which makes rural LSOA larger in area than urban LSOAs). For each LSOA we obtained demographic information, such as its income decile and whether it is a rural or urban settlement. This allowed us to analyse the coverage of shops depending on these characteristics of individual LSOAs. A specific shop was deemed accessible to all people living in the same LSOA where the shop is located.

Given the large area of some LSOAs, the approach mentioned above can lead to biases on our estimates to retailers' accessibility in rural areas. Therefore, we developed methods to estimate the distance to shops for households on a point-to-point basis, providing a finer grained and more accurate estimate access to shops. We calculated 'as the crow-flies' distance between each dwelling and each shop within the same postcode area (e.g., CF for Cardiff area) to estimate the proportion of residential dwellings that would have a supermarket

within 1.5km; a distance that would be appropriate to reach on foot if a vehicle was not available.

2.1. Data Sources

LSOAs codes, income decile and LSOAs classification in Rural or Urban were obtained from the publications available on the respective government official websites of England and Wales [1],[2], Scotland [3],[4], and Northern Ireland [5],[6]. We used the overall grocery market share obtained from Kantar website,[7] and retailers' name and coordinates were obtained from Geolytix for the last quarter of 2022[8]. Dwelling coordinates and type (e.g., residential) are held within ONS and not publicly available.

3. RESULTS AND PRACTICAL APPLICATION

Regarding the distribution of shops in the UK, we first analysed their location based on which LSOAs (Lower layer Super Output Areas) they belong to. Figure 1 shows the distribution of shops based on the income of the LSOA they are sited in. This gives an idea of the overall target market of the different retailers.

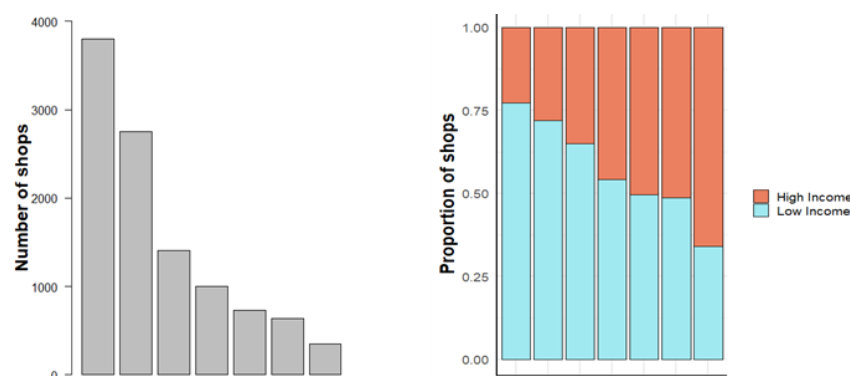


Figure 1. Left. Total number of shops for each retailer. Right. Proportion of shops for each retailer in low- (light blue) and high-income areas (red).

Considering the distribution of these outlets in urban and rural areas, Figure 2 shows the distribution of the retailers in rural and urban areas. It is hardly surprising that most outlets are in urban areas in which most of the UK population (just over 80%) lives. On the right-hand side of Figure 2 we can also see the distribution of shops across income deciles in urban and rural areas. Note that the lowest income areas are prevalently urban and the highest income areas have an increasing proportion of rural areas.

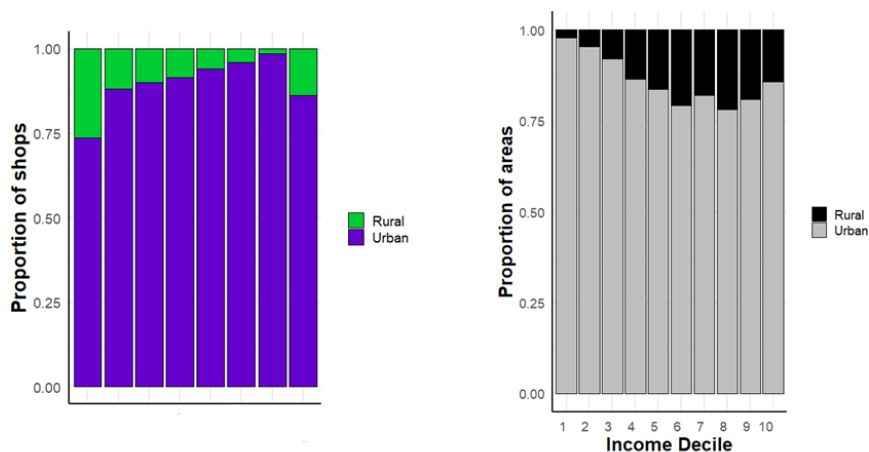


Figure 2. Left. Proportion of shops for each retailer in urban and rural areas. Right. Proportion of urban and rural areas in each income decile.

Shop accessibility was calculated as the proportion of households (i.e., dwellings), in low- and high-income areas and in rural and urban settlements, within 1.5 kilometres of a store. Figure 3 shows the percentage of dwellings that have access to at least one shop from any of the seven retailers, displayed by income and urban/rural areas.

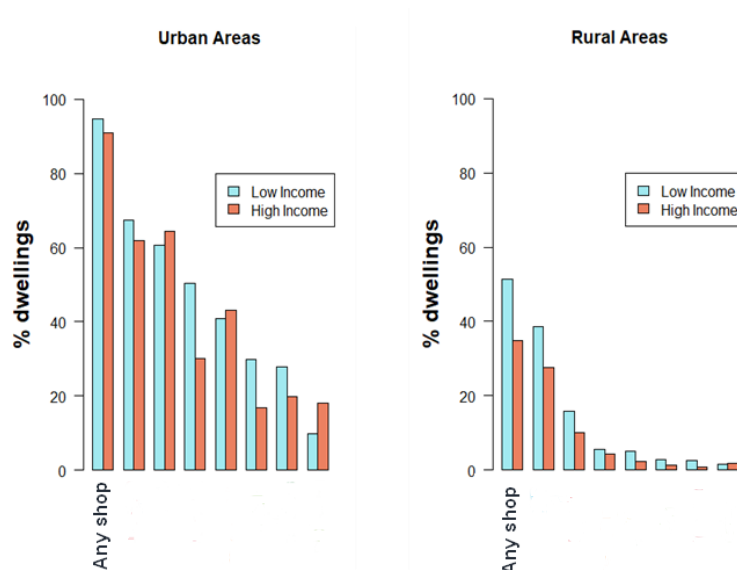


Figure 3. Percentage of dwellings with access to each of the seven retailers in urban areas (left panel) and rural areas (right panel).

In urban areas (Figure 3, left panel), there is high accessibility for those living in both low- and high- income areas with 94.7% and 91% of dwellings that can access at least one shop of the seven retailers, respectively. In rural areas (Figure 3, right panel), accessibility is significantly lower than urban areas with only 51.4% and 34.8% of dwellings within 1.5 kilometres from at least one shop of the seven retailers.

4. MAIN FINDINGS

We show that access to various shops depends more on whether a household lives in urban or rural areas than on their income. This warrants further research on how rural households experience inflation, as they are a smaller group with limited access to supermarkets that could face additional challenges during inflationary periods.

5. REFERENCES

- [1] MHCLG. 2020. Indices of Deprivation 2019 <https://www.gov.uk/government/statistics/indices-of-deprivation-2019-income-and-employment-domains-combined-for-england-and-wales>
- [2] DEFRA. 2021. 2011 Rural Urban Classification lookup tables for all geographies <https://www.gov.uk/government/statistics/2011-rural-urban-classification-lookup-tables-for-all-geographies>
- [3] Scottish Government. 2020. Scottish Index of Multiple Deprivation. <https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020>
- [4] Scottish Government. 2019. 2011 Data Zone Lookup <https://statistics.gov.scot/data/data-zone-lookup>
- [5] NISRA. 2017. Northern Ireland Multiple Deprivation Measure 2017 <https://www.nisra.gov.uk/statistics/deprivation/northern-ireland-multiple-deprivation-measure-2017-nimdm2017>
- [6] NISRA. 2017. Urban- Rural Geography Documents (2015) <https://www.nisra.gov.uk/publications/urban-rural-geography-documents-2015>
- [7] KANTAR. 2023. Great Britain Grocery Market Share. <https://www.kantarworldpanel.com/grocery-market-share/great-britain/snapshot/22.01.23>
- [8] Geolytix. 2022. <https://geolytix.com/blog/supermarket-retail-points-v22>

How many diary days? Smart surveys can help to reduce burden of data collection for household statistics

Danielle Remmerswaal (Utrecht University & Statistics Netherlands), Barry Schouten (Statistics Netherlands) & Bella Struminskaya (Utrecht University)

1. INTRODUCTION

Traditionally, diary studies are used to capture time use behaviour, such as traveling, that is otherwise prone to recall errors, telescoping and other measurement errors when asked retrospectively. However, the traditional diary modes, web and paper, pose a high burden for respondents, resulting in nonresponse and measurement errors such as rounding errors, and underreporting. Given these limitations and the technological advances of smartphones and other smart devices, growing attention is given to smart surveys. For example, smartphone-based travel apps can make use of geolocation sensors present on the smartphone to construct a travel diary. Another example is a smartphone-based budget app that can make use of the camera to scan receipts. In both examples (see e.g. [1], [2], [3], [4] for use cases) the manual entry of travel or expenditure data is (partially) replaced by sensors on a smart device.

Currently, the common approach in travel surveys is to survey one to three days in order to minimize response burden, despite recommendations from researchers to collect data for a longer period in order to capture within-person variability (see e.g. [5], [6], [7]). In smart surveys the response burden is different. The respondent has to install an app on a device where data can be collected (passively) and does not have to manually enter all the data. Therefore, increasing the number of diary days contributes less to response burden compared to traditional diary surveys. This reduction in burden opens up the possibilities to prolong the study period.

Our Research aim is to answer “How many days should we collect data for a smartphone-based diary study in official statistics?” by considering the response burden and the variability of statistics between and within persons. If additional days of one person provide additional information and the burden of asking one more day is low, why would we not prolong the study duration?

2. METHODOLOGY

2.1. Variability of statistics

We measure the variability of a statistic within a person, and thus between days. We compare the variation across 7 days with the variation across 2, 3, 4, 5, and 6 days for different statistics. One measure to guide the choice on the number of diary days is reliability. By calculating the reliability of multiple-day data we can assess which proportion of the variance can be attributed to within-person differences. We use the same idea as the authors [8] who calculate the reliability of 7-day accelerometer data with the reliability formula based on work by McGraw and Wong [9].

2.2. Motivation for case studies

We include datasets of three different topics on which National Statistical Institutes (NSIs) collect data: household budget, travel behaviour, and physical activity. The three studies are chosen because they each report on daily behaviour of persons. Making the survey ‘smart’ (see [10] for an elaboration of the meaning of smart statistics for official statistics) changes the amount of input necessary by the respondent, and thus the response burden. In a physical activity study, accelerometers attached to the body can measure the duration and intensity of physical activity. The measurements are collected passively without respondent interaction necessary. An app-based travel study automatically compiles a travel diary, consisting of a series of movements and stops, by passively collecting geolocation data via the smartphone. In a travel app AI-ML predictions (travel mode/motive prediction) can be made, but a respondent still needs to validate the movements and stops of the travel diary. In a household budget app participants are able to manually enter expenses in a user-friendly app interface, and they are able to (actively) upload receipts of their expenses which are automatically (using Optical Character Recognition (OCR)) compiled.

The three studies have different levels of response burden based on i) whether the nature of the data collection is active or passive, and ii) how much respondent interaction is necessary to validate the data. Based on this the case studies can be ordered from relative low to high burden. Relatively lowest burden for physical activity study, middle burden for a travel app, and the relatively highest burden for a budget study app.

There are however potential challenges regarding the willingness to participation in smart surveys related to coverage errors, nonparticipation, privacy and ethics. To read more on these challenges we refer to [6], [7], and [8].

2.3. Case studies

For each data source we calculate two statistics based on the frequency and duration of the studied behaviour.

Table 1. Description of analysed data sources

Topic	Data source	Year	Data collection mode	Data collection duration
Physical activity	LISS panel	2021	Hip-based physical activity tracker ‘Activpal’	1 week
Travel	Statistics Netherlands	2018 + 2022	Smartphone-based app using passive geolocation sensors to construct diary	1 week
Budget	Statistics Netherlands	2022	- Web option - Smartphone app using OCR	1 week (web) 2 weeks (app)

3. RESULTS AND PRACTICAL APPLICATION

Reliability is calculated for each statistic. The increase in reliability is the greatest for the addition of the first days, after that the level of increasement declines. The change in reliability is different when the days are not consecutive but random (one can imagine that weekdays are more alike to each other than a weekday with a weekend day).

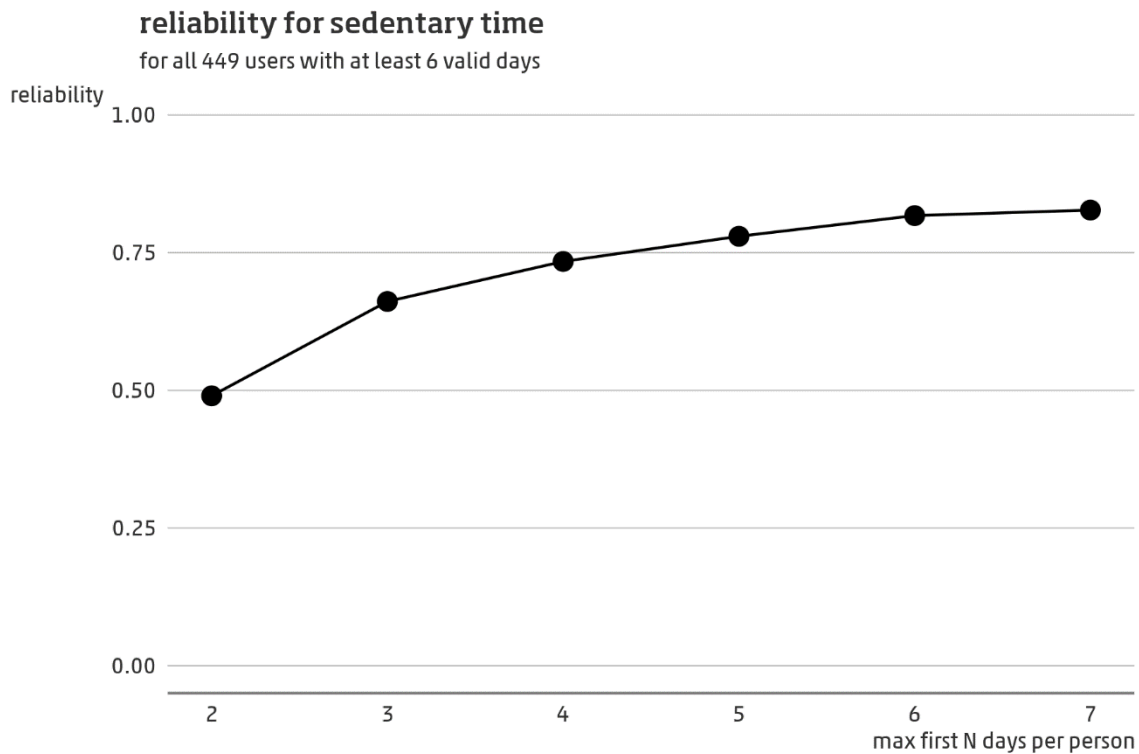


Figure 1. Reliability for the first d ($d: 2, 3, 4, 5, 6, 7$) consecutive days for sedentary time (time spend sitting in a day).

4. MAIN FINDINGS

By combining the results of the reliability analysis based on the ‘variability of statistics’ with burden (measured by dropout), we can make recommendations on the study duration for studies with different levels of burden.

Our conclusions and recommendations have relevance across a wide range of research areas concerning studies on (time-use) behaviour, studies considering the use of smartphone apps, and studies using (smartphone) sensors.

5. REFERENCES

- [1] C. Winkler, A. Meister, and K. W. Axhausen, “The TimeUse+ data set: 4 weeks of time use and expenditure data based on GPS tracks,” *Transportation*, Aug. 2024, doi: 10.1007/s11116-024-10517-1.
- [2] D. McCool, P. Lugtig, O. Mussmann, and B. Schouten, “An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges,” *J. Off. Stat.*, vol. 37, no. 1, pp. 149–170, Mar. 2021, doi: 10.2478/jos-2021-0007.
- [3] A. Jäckle, M. P. Couper, J. Burton, and C. Lessof, “Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases,” *Surv. Res. Methods*, pp. 23–44, 2019, doi:10.18148/srm/2019.v13i1.7297.
- [4] R. Kraakman, A. Luiten, V. Toepoel, and M. Kompier, “The Feasibility of Using Consumer-Level Activity Trackers in Population Monitoring of Physical Activity: Comparing Representativeness and Measurement Quality With Self-Report and a Professional Research-

- Grade Accelerometer,” *Surv. Res. Methods*, vol. 17, pp. 429–445 Pages, Dec. 2023, doi: 10.18148/SRM/2023.V17I4.7919.
- [5] P. R. Stopher and Y. Zhang, “Repetitiveness of Daily Travel,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2230, no. 1, pp. 75–84, Jan. 2011, doi: 10.3141/2230-09.
- [6] R. Schlich and K. W. Axhausen, “Habitual travel behaviour: Evidence from a six-week travel diary,” *Transportation*, 2003.
- [7] P. R. Stopher, K. Kockelman, S. P. Greaves, and E. Clifford, “Reducing Burden and Sample Sizes in Multiday Household Travel Surveys,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2064, no. 1, pp. 12–18, Jan. 2008, doi: 10.3141/2064-03.
- [8] T. Scheers, R. Philippaerts, and J. Lefevre, “Variability in physical activity patterns as measured by the SenseWear Armband: how many days are needed?,” *Eur. J. Appl. Physiol.*, vol. 112, no. 5, pp. 1653–1662, May 2012, doi: 10.1007/s00421-011-2131-9.
- [9] K. O. McGraw and S. P. Wong, “Forming inferences about some intraclass correlation coefficients,” *Psychol. Methods*, vol. 1, no. 1, pp. 30–46, Mar. 1996, doi: 10.1037/1082-989X.1.1.30.
- [10] F. Ricciato, A. Wirthmann, and M. Hahn, “Trusted Smart Statistics: How new data will change official statistics,” *Data Policy*, vol. 2, p. e7, 2020, doi: 10.1017/dap.2020.7.
- [11] B. Struminskaya, P. Lugtig, F. Keusch, and J. K. Höhne, “Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges,” *Soc. Sci. Comput. Rev.*, p. 089443932097995, Dec. 2020, doi: 10.1177/0894439320979951.
- [12] F. Keusch, S. Bähr, G.-C. Haas, F. Kreuter, M. Trappmann, and S. Eckman, “Non-Participation in Smartphone Data Collection Using Research Apps,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 185, no. Supplement_2, pp. S225–S245, Dec. 2022, doi: 10.1111/rssa.12827.
- [13] B. Struminskaya, V. Toepoel, P. Lugtig, M. Haan, A. Luiten, and B. Schouten, “Understanding Willingness to Share Smartphone-Sensor Data,” *Public Opin. Q.*, vol. 84, no. 3, pp. 725–759, May 2021, doi: 10.1093/poq/nfaa044.

“I don’t know what I did last summer” – measurement error and memory effect in tourism statistics

Christophe Demunter (Eurostat) & Koen Hooyberghs (Statistics Belgium)

1. INTRODUCTION

Tourism statistics generally comprise two main components: statistics on tourism supply, often operationalised through data on capacity and occupancy of tourist accommodation, and statistics on tourism demand, often operationalised through data on trips made by a population and the characteristics of those trips. This two-fold approach to tourism statistics is also the cornerstone of the European statistics on tourism as laid down in Regulation (EU) 692/2011 [1].

In most cases, these statistics are collected via business and household surveys. The accuracy and reliability of (probability) survey-based data can be affected by sampling error and non-sampling error. Non-sampling error can occur in accommodation statistics in the form of coverage error (for instance service providers not being included in the statistical business register), measurement error (for instance a service provider may underreport to tax authorities and therefore also underreport in a survey for official statistics) or non-response error. Non-sampling error in household surveys on tourism demand mainly takes the form of non-response error (unit non-response as well as item non-response) and measurement error.

This abstract focuses on a particular type of measurement error that impacts on the accuracy of tourism demand statistics: the recall bias or memory effect. The recall bias leads to underreporting of the number of trips made, and underestimation of the expenditure made during that are reported. This can cause significant underestimation of the tourism volumes and the sector’s contribution to countries’ economies.

The NTTS contribution will make a quantitative assessment of the recall bias in tourism surveys and explore solutions to deal with this type of measurement error.

2. METHODOLOGY

In the above introduction, some examples of non-sampling error were listed. The impact on the accuracy will be a cumulative function of the different types of error combined. While it is difficult to assess to what extent coverage error (for instance resident population not entirely overlapping with the register population from which samples are drawn, or frequent travellers who may be difficult to ‘catch’ for the survey or non-travellers assuming their empty reply would be useless) or non-response error (selected sample elements not providing data) have a systematic element, measurement error due to recall bias or memory effect is expected to cause a systematic downward bias in the estimated volume of tourism trips or the estimated amount of tourism expenditure. It is the combination of uncertainties that can jeopardise the reliability of the statistics, and the trust users put into them. Based on available metadata [2], for instance, unit non-response rates in tourism demand surveys tend to be high, exceeding 50% in one third of the countries and being lower than 20% in only a few countries.

2.1. In-depth analysis of surveys estimates the memory effect at 20% or more

Quantifying the level of recall bias is not straightforward using the meta-information of a production round or survey cycle. However, some studies made use of looking into the number of trips in subsequent survey waves concerning the same reference period.

A French study from 2010 [3] looked at the trips reported during interviews spanning 13 weeks to cover the three months prior to the 13 weeks' survey period. Looking at the distribution of the number of reported trips for the reference quarter, by survey week of reporting, the number of reported trips declined strongly as the distance (in time) between the end of the reference period and the survey moment increased (see Figure 1). After correcting for the error caused by underreporting of trips due to recall bias, the total number of trips increased by 20.7%.

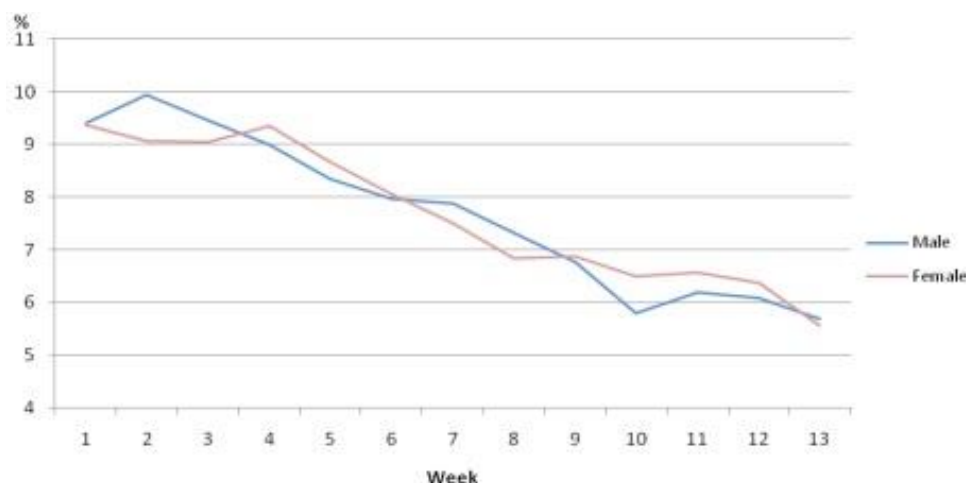


Figure 1. Distribution of reported trips, according to the number of weeks elapsed after the reference period

Source: S. Roux et al. [3]

Similar conclusions were obtained in a Spanish study [4] that analysed the responses in different monthly survey waves with a three months' reference period. The methodology of the survey implied that each reference month was covered by three subsequent survey waves, for instance the reference month June 2006 was covered by the July survey wave spanning the quarter April-May-June, by the August survey wave spanning May-June-July and the September survey wave spanning June-July-August. In other words, respondents were reporting June trips respectively one month, two months and three months after the reference month. In the July wave, 13.0 million trips were reported for June, in the August wave this dropped to 11.7 million trips, another month later only 8.8 million trips were recalled. All other sampling and survey conditions being equal, the memory effect was the only explanation for the drop. The gaps in the second and third survey month following the reference month, as compared with the base level of the first survey month, correspond to a memory effect of 15% underreporting (not yet considering a possible memory effect already in the first survey month as well).

2.2. Alternative data sources reveal weaknesses in surveys attributable to the memory effect.

The recall bias does not only affect the estimate of the number of tourism trips or night but, in addition, the expenditure made during those trips. This will be non-observed expenditure for non-reported trips as well as underreported expenditure for trips that the respondent did recall. A recent project in the Nordic countries exploring the use of payment transaction data to compile tourism statistics [5] found that the tourism effects were 30% higher using this source compared to the traditional surveys.

In 2016, Eurostat explored the feasibility of using mobile network operator (MNO) data for tourism statistics, based on a data cooperation with a Belgian MNO [6]. The results included a comparison of the distribution of trips by duration using this new data source and using official tourism statistics obtained from surveys. Two memory effects were striking in the results (see Figure 2). Firstly, the respondent's memory tends to be bias to cut-off reference values such as “one week”, “a ten days’ trips”, “a fortnight” or “quinze jours”, while the MNO data showed less pronounced peaks for these durations. Secondly, the survey-based official statistics performed particularly poor in recording very short trips of one overnight stay.

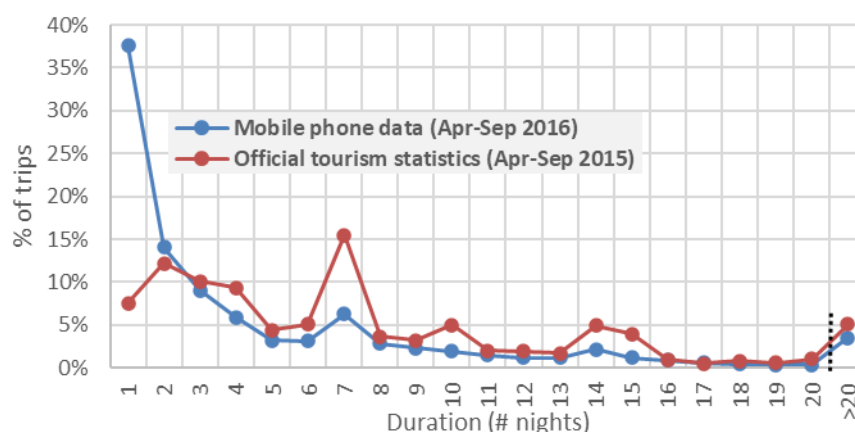


Figure 2. Distribution of outbound trips from Belgium to other EU countries, by duration, mobile network operator data versus survey-based official statistics

Source: G. Seynaeve & C. Demunter [6]

The same study found generally higher estimates using the MNO data compared to the survey data, but the deviation appeared to be bigger for nearby destinations that can easily be reached for a short trip (and easily be forgotten?). Figure 3 shows the ratio for the number of trips estimated from MNO data versus survey data (for eleven European countries for which the survey had enough observations). While relatively high ratios were observed for countries close to Belgium (the Netherlands, Luxembourg, Germany), lower ratios were found for trips to Croatia, Greece, Portugal, Italy, Austria).

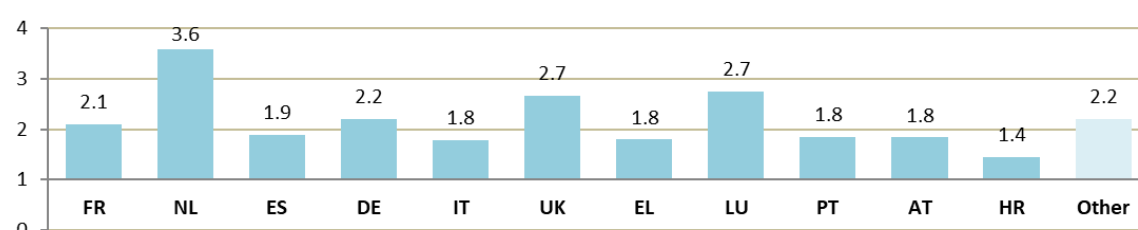


Figure 3. Ratio of the estimated number of tourism trips using mobile network operator data versus survey-based official statistics, by country of destination

Source: G. Seynaeve & C. Demunter [6]

2.3. Comparing metadata across countries to find methodological factors impacting on the recall bias

After having detected that its residents had a low number of tourism trips per person, as compared with neighbouring countries, the country in question analysed the national metadata available from Eurostat's website to test the hypothesis that countries with a monthly survey

would have a higher proportion of short trips with 1 to 3 overnight stays (compared with all trips of 1 or more overnight stays) than countries with a quarterly survey.

The presentation at the NTTS2025 will further examine the impact of different methodologies on the presumed recall bias, not only looking at the periodicity of the survey tools, but considering in the analysis the length of the reference period. Some first results are included in the next section of this abstract.

3. RESULTS AND PRACTICAL APPLICATION

The comparison of monthly surveys with quarterly surveys, as discussed in section 2.3 above, showed an important difference in the proportion of short trips: while countries with quarterly surveys observed on average 48% of trips with a duration of 1-3 nights among the total number of observed trips of 1 night or more, this share increase to 61% on average in countries using a monthly survey. The country for which the analysis was initially performed, recorded 41.5% of short trips in its quarterly survey, to reach a 61% level the number of short trips would need to be more than doubled (and the total number of trips grossed up with around 50%).

The current study as well as the findings reported in sections 2.1 and 2.2 used the assessment and quantification of the recall bias as input for models to estimate more accurate statistics on tourism demand. In other words, the practical application lies not only in an assessment of the level of underestimation but also in the use of this information as input for reweighing or recalibrating survey results.

4. MAIN FINDINGS

The work done so far confirms earlier findings that the memory effect can be a very significant type of non-sampling error in tourism demand surveys. To ensure high quality statistics, quantifying the error and correcting for it, is essential.

The underestimation does not only affect the primary statistics, but also derived products such as tourism satellite accounts (TSA) or the travel item of the balance of payments. Since TSA make use of data on expenditure to estimate tourism ratios for the economic activity in tourism industries. An increase of the expenditure measured in surveys by 20% could impact the estimate of the tourism direct gross value added by 7-12% and thus possibly revise the share of tourism in the economy upwards.

The findings on the memory effect are wake-up call to adjust the survey results and to optimise the survey methodologies, but can also be a trigger to rethink tourism statistics towards relying on sources that are less subject to the respondents' cognitive limitations, for instance payment transaction data or mobile network operator data used as a main source for measuring tourism flows or tourism expenditure, or as an auxiliary source to assess the accuracy of surveys.

5. REFERENCES

- [1] Regulation (EU) No 692/2011 concerning European statistics on tourism, [OJ L 192, 22.07.2011, p. 17-32.](#)
- [2] Eurostat, metadata for "Trips of EU residents - annual data (tour_dem)" [\[link\]](#)
- [3] S. Roux, J. Armoogum, J.-L. Madre and F. Potier, *Sampling strategies and correction of measurement errors for tourism travel surveys*, Paper presented at the 10th International Forum on Tourism Statistics (Belem, Portugal, 22-23 November 2010). (*)

- [4] T. Guardia and S. García, *Memory effect in the Spanish domestic and outbound tourism survey*, Paper presented at the 8th International Forum on Tourism Statistics (Paris, France, 19-21 November 2008). (*)
- [5] M. Wiberg, *A joint model for Nordic European countries to compile tourism statistics based on payment transaction data*. Presentation made at the IAOS-ISI Conference 2024 (Mexico City, Mexico, 15-17 May 2024). [link](#)
- [6] G. Seynaeve, C. Demunter et al., *When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics*, Paper presented at the 14th Global Forum on Tourism Statistics (Venice, Italy, 23-25 November 2016). (*)
- [7] (*) conference papers no longer publicly available through the event website, but available from the abstract authors upon simple request.

Implementation of a new coding service at Insee

Yves Coltier & Julien Guay (Insee)

1. INTRODUCTION

For public statistics, nomenclatures facilitate the organization of economic and social information.

Codification consists of classifying labels provided in a survey into nomenclatures. It can concern activities, geographical locations, professions, etc.

Until the beginning of 2024, all automatic coding processes for household surveys in France were based on the Sicore¹ system.

Sicore is an expert system for which all the components of the algorithm (normalization rules, reference indexes, logical rules) must be specified in full by one or more nomenclature experts.

Coding with Sicore takes place in 3 parts:

- Standardization of the label to be coded to simplify its handling;
- Finding the standardized label within a reference label index according to a decision tree questioning groups of consecutive characters called "atoms" that make up the label;
- Application of logical rules taking into account categorical ancillary variables if the label alone is not sufficient to code in the nomenclature.

However, several limitations appear at a time when the Internet collection method allows an increase in sample volumes and therefore a larger volume of labels to be coded:

- The proper functioning of Sicore is based on the ability to extract expert knowledge and formalize it in the form of rules and over time, it becomes difficult to add new rules without creating side effects;
- The coding is only deterministic and does not return multiple echoes. It is not possible to make a trade-off between efficiency and reliability and therefore to control the volume of labels sent in manual recovery;
- over time, Sicore has become a black box, very difficult to maintain in operational condition and to upgrade (for example, to support a new nomenclature).

That's why Sicore is being replaced by a mixed codification system, involving three tools:

- a list collection service;
- a deterministic codification service;
- a probabilistic codification service based on supervised learning methods.

In addition, the technology used for the deployment of this service is still innovative at INSEE since we use the Kubernetes environment which is a portable, extensible, open source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation.

¹ Système Informatique de Codage de Réponse aux Enquêtes (in French)

2. METHODOLOGY

2.1. The list collection service

The list collection service (also called "Suggester") is an autocomplete tool that allows for label suggestions while typing a label when answering a question.

It involves parameters (list of non-significant words, list of synonyms, etc.) and lists of labels which are developed by nomenclature experts.

It takes the form of a javascript component.

2.2. The deterministic codification service

The deterministic coding tool allows the conversion of a label, possibly associated with additional variables, into a code of the target nomenclature.

For each nomenclature, the possible labels and set of variables leading to a code are listed in a "rule file".

As a first step, it allows the coding of labels entered on a list. As a second step, it can also be used on labels entered in clear text by the respondent or the investigator.

It is a web service developed in Python language.

2.3. The probabilistic codification service

The supervised learning model "statistically" predicts categories of the nomenclature for each label entered in plain text and provides associated "probabilities" for each category.

The service therefore returns for each label and additional input variables the code that emerges as the best prediction, according to a confidence score calculated from the probabilities, and also returns the X best predictions with the associated probabilities, X being an adjustable parameter when calling the service.

It is a web service developed in python, serving a Natural Language Processing model implemented with fastText². The model is stored in a MLFlow model store. This architecture allows model updates without any code modification.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Global service performances

The new coding service is still being deployed but has been used already in production for several surveys.

The following table shows the breakdown of service usage for the Information and communication technologies survey to code French classification of professions.

² FastText is a free library from Facebook AI Research for learning word embeddings and word classifications.

Table 1. Profession and socio-professional category (PCS) codification in the information and communication technologies survey

	Returned questionnaires	Uses of Suggester	Deterministic codification	Probabilistic Codification*
Paper	7,144		1,664	1,768
Phone	2,472	1,933	1,947	157
Internet	8,840	5,805	5,905	862
Total	18,096	7,738	9,516	2,787

* : Above a confidence score, the first prediction of the probabilistic coding is considered correct.

The coding service was able to code more than 65% of occupational labels, taking into account empty labels (2,601 labels).

The confidence level set for the probabilistic codification was quite high (0.6) due to the large amount of manual coding capacity available over the period.

3.2. Probabilistic codification service performances: the exemple of PROFS model

PROFS model is a NLP model trained on several surveys to code the salaried profession. The convergence rate (i.e. the convergence rate between manual annotations and the model predictions used as a proxy for accuracy) is around 62%.

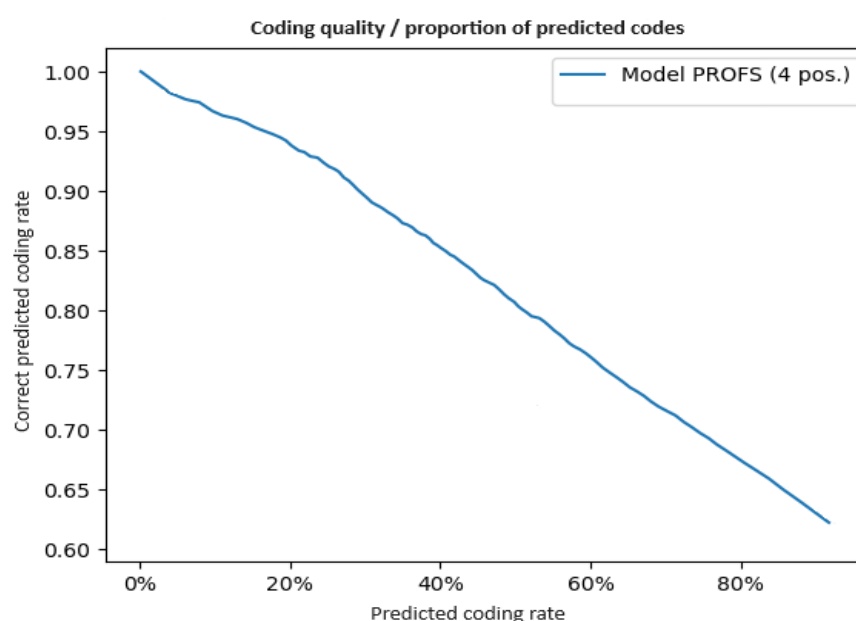


Figure 1. the coding quality of PROFS model

The figure 1 shows that if the PROFS model predicts 60% of labels then 75% of labels should be coded correctly, under the assumption that 100% of manually coded labels are correct.

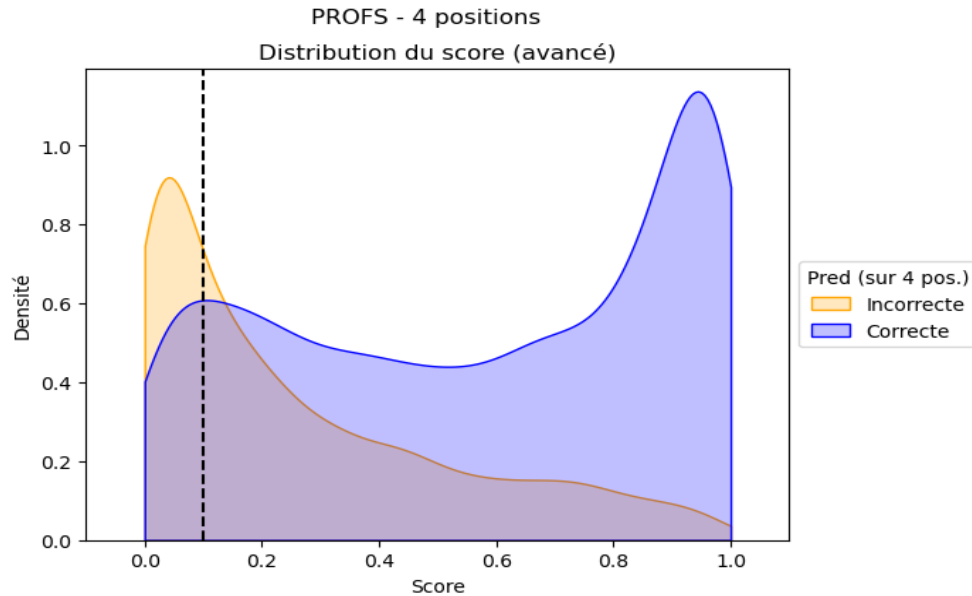


Figure 2. the confidence score distribution of PROFS model

The figure 2 shows that if we set a confidence threshold of 0.15, for example, then all predictions coded by the model will have at least a 50% chance of being correct (and this probability increases as the score rises).

4. MAIN FINDINGS

The new coding service, and in particular probabilistic coding, provides a balance between the volume of labels to be coded manually and the expected coding quality.

There are not necessarily any gains in terms of human resources, as constant monitoring of the service and regular re-training of the NLP models are required, which implies an additional annotation workload.

However, it does allow us to respond to the increase of sample size, made possible by the growing use of the Internet as a data collection method.

5. REFERENCES

- [1] T. Leroy, [Quelques limites de l'algorithme implémenté dans l'outil Sicore](#) (JMS 2022).
- [2] T. Leroy, L. Malherbe and T. Seimandi, [Application de techniques de machine learning pour coder les professions dans la nomenclature des professions et catégories socio-professionnelles 2020](#) (JMS 2022)

Improving AI Language Models via Specialised Training Corpora for Official Statistics

Cristiano Tessitore, Marius Felecan & Jean-Marc Museux (Eurostat)

1. INTRODUCTION

Generative AI systems and especially Large Language Models (LLMs) have proven to be a transformative force in various sectors, demonstrating substantial utility. Forecasts by industry experts (Marr, 2024) indicate that LLMs will continue to integrate with robotics and task automation, while regulatory bodies are developing frameworks to ensure their ethical use. One of the noticeable use cases for official statistics is the capability LLMs provide to interact with information systems, in particular open data systems. However, many of the public domain models may show unwanted biases, hallucination and still poor performance for handling data which are incompatible with official statistics standards. There are many ways to improve their performance and reliability for instance using fine tuning or Reinforcement Learning from Human Feedback (RLHF). In most situations the presence of high-quality data based on Official Statistics corpora (see next paragraph) is instrumental to augment the performance of the model. In this paper we explore the need for building an official statistics corpus and a set of specialized model benchmarks.

2. LLM MODEL TRAINING

To build a successful model, two critical components are the model architecture itself and the training data (so called *corpus*, the Latin word for “body”). The model's architecture defines how the system processes information, learns patterns, and makes predictions, determining its capabilities. Equally important is the training data, which serves as the foundation for learning. A proper *corpus* ensures the model learns effectively, reducing biases and improving accuracy.

Large language models require extensive training to develop their ability to understand and produce human-like text. By processing huge amounts of data, LLMs learn language patterns, enabling them to understand context, answer questions, and generate coherent and relevant content. Continuous training, particularly with high-quality datasets, refines their capacity to provide good answers. Effective training also helps to mitigate biases and errors, ensuring LLMs are reliable and trustworthy in practical applications.

Training AI models, especially large language models (LLMs), involves various learning methods as highlighted by Buono et al. (2024). Supervised learning relies on labelled datasets to train models for accurate predictions, though it is difficult to scale due to the scarcity of labelled data. Self-supervised learning (SSL) addresses this challenge by enabling models to learn from unlabeled data, generating pseudo-labels in the process. Unsupervised learning allows models to learn patterns from raw data without explicit labels. Reinforcement learning, particularly with human feedback (RLHF), is used to align models with human values but is resource-intensive, though newer techniques like Direct Preference Optimization (Rafailov, 2023) are increasing efficiency. Fine-tuning pre-trained models is common practice, allowing customization for specific domains. Parameter Efficient Fine-Tuning (PEFT) methods like Low-

Rank Adaptation (LoRA) (Hu, 2021) reduce the cost of model adaptation and help prevent forgetting.

3. SPECIALISING LLM MODELS

Recent developments like Zhou (2023) and Microsoft (2024) have demonstrated that smaller, open-source LLM models can achieve performance levels comparable to those of significantly larger models. This breakthrough defies the traditional approach of scaling up model sizes to improve functionality. The new “compact” models, which are more resource-efficient, can deliver results with reduced computational power and lower energy requirements, making them more accessible for a wide range of applications. Such advancements could democratise AI, allowing researchers, smaller institutions, and developers to deploy powerful AI tools without needing access to vast resources. The role of open-source collaboration in driving these innovations is especially noteworthy, as it fosters a more inclusive and collaborative environment for future AI development. By focusing on optimization and efficient algorithms rather than sheer size, these new models open the door to broader AI adoption and utilisation across various sectors.

Specialising a small language model (SLM) through fine-tuning allows it to perform optimally on specific tasks or domains by adapting its pre-trained parameters using task-relevant data. Fine-tuning starts with a pre-trained model and involves updating its weights on a smaller, more specialised dataset, making the model capable of handling domain-specific queries with greater accuracy (see previous paragraph).

Fine-tuning LLMs offers several key benefits across multiple categories. In terms of performance, fine-tuning helps reduce hallucinations, improve consistency, and minimise the inclusion of unwanted information. From a privacy perspective, fine-tuning can be performed on-premises, which prevents data leakage and helps avoid security breaches. Cost efficiency is another significant advantage, as fine-tuning allows for lower costs per request, enhanced transparency, and greater control over resource allocation. Additionally, reliability is improved, offering control over uptime, reduced latency, and more effective moderation. These benefits underscore the value of fine-tuning in creating more secure, efficient, and reliable models.

4. BENCHMARKING THE MODELS

Benchmarks are key to checking out LLMs. These tests provide a common way to measure and compare models. Benchmarks can measure models' precision and reliability; by testing LLMs in a structured way, researchers can find both the model's strong and weak points. This process is vital for moving the field forward, making sure LLMs can be trusted and giving users in many different areas confidence that the models will give them accurate information.

A benchmark metric for language models should offer a quantitative measure to evaluate different characteristics such as performance, accuracy, and effectiveness of the model in Natural Language Processing tasks. It typically involves comparing the model's output against a set of predefined standards or references across various tasks such as text generation, translation, summarization, or question answering.

Specialised benchmarks can be used as comprehensive evaluation tools for assessing the proficiency of current models in official statistical knowledge. By providing standardised metrics, they facilitate the fine-tuning of models to enhance their performance in statistical tasks.

Emphasising metrics in benchmarking ensures that the models meet the highest standards of accuracy, reliability, and relevance to a specific domain application. This rigorous evaluation process not only fosters transparency and trust in AI applications but also guides iterative improvements, ultimately leading to more effective and accountable tools that serve the public interest.

5. IMPROVING LLM FOR OFFICIAL STATISTICS USE CASES: THE ROLE OF STATISTICAL AUTHORITIES

By merging qualitative and quantitative aspects of statistical knowledge, Statistical Authorities can influence the development of LLMs and to tune them to manage complex statistical problems. This is expected to not only improve the analytical capabilities of LLMs but also bridges the gap between data (not necessarily provided by the model) and the rich and quite formal semantic assets, mainly metadata that accompany the data. This could lead to nuanced interpretations necessary for informed decision-making in fields like policy-making, economic forecasting, and social research.

Openness and collaboration with Universities and research institutes, industry experts, private sector data holders, start-ups, and providers of innovative ICT solutions, is seen as an accelerator for endeavour. Such partnerships would ensure that the Generative AI training datasets would be comprehensive, diverse, and reflective of real-world scenarios where statistical intelligence is applied.

These collaborations also foster innovation by incorporating a wide range of expertise, ensuring that the training corpora remains relevant and valuable for future advancements in AI and statistical analysis. This is crucial as it ensures that LLMs continue to evolve and improve, keeping pace with new developments and emerging trends in data science and statistics. By maintaining a dynamic and updated dataset, (National) Statistical Authorities, academic institutions, industry experts, and other data agencies can drive forward the potential for LLMs to make significant contributions to official statistics as well as, providing deeper, more nuanced insights into statistical data. This not only enhances the utility of LLMs in practical applications but also supports the broader goal of using AI to generate actionable intelligence that can lead to better decision-making and improved outcomes in numerous domains.

Further, National Statistical Authorities can significantly enhance the semantic richness of the training corpora through the promotion of multilingualism. By incorporating a diverse range of languages spoken within a country, these institutions can ensure that the dataset accurately reflects the linguistic diversity of its population. This approach not only improves the language model's ability to understand, generate and translate text in multiple languages but also helps in preserving less used languages. This would result in developing more inclusive and performant LLMs with regards to Official Statistics.

6. CONCLUSIONS

The optimization of Large Language Models through specialised training corpora is essential for improving accuracy, reliability, and reducing biases. Recent advancements in smaller, efficient models have shown that optimization can outperform complex ones, making AI more accessible and reducing resource demands.

Benchmarking remains crucial for assessing model performance. Standardised evaluation metrics enable consistent testing, ensuring models meet accuracy and reliability standards, particularly for official statistical applications.

(National) Statistical Authorities may play a key role in enhancing LLMs by building specific training corpora and sectorial benchmarks, leveraging collaborations with academic, industry, and private sector partners. These partnerships will ensure comprehensive, real-world training data, leading to better handling of complex statistical tasks. NSAs can also support multilingualism, ensuring LLMs reflect linguistic diversity and inclusivity.

By improving training, fostering collaboration, and setting benchmarks, LLMs will become more effective, trustworthy, and widely accessible, driving advancements in official statistics.

7. REFERENCES

- [1] Buono, D., Felecan, M., Tessitore, C., An introduction to large language models and their relevance for statistical offices : 2024 edition, Publications Office of the European Union. <https://data.europa.eu/doi/10.2785/716217>, 2024.
- [2] Hu, Edward J., et al, LoRA: Low-Rank Adaptation of Large Language Models, ArXiv, /abs/2106.09685, 2021.
- [3] Marr, B., The Future Of Generative AI: 6 Predictions Everyone Should Know About, Forbes contributor, Mar 5, 2024 - <https://www.forbes.com/sites/bernardmarr/2024/03/05/the-future-of-generative-ai-6-predictions-everyone-should-know-about/>
- [4] Microsoft, Tiny but mighty: The Phi-3 small language models with big potential, 2024, <https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/>
- [5] Rafailov, Rafael, et al, Direct Preference Optimization: Your Language Model Is Secretly a Reward Model, ArXiv, /abs/2305.18290, 2023.
- [6] Zhou, Zhengping, et al, Mini-Giants: Small Language Models and Open Source Win-Win, arXiv preprint arXiv:2307.08189, 2023.

Improving estimates of short-stay accommodation occupancy in the EU using online platform data

Simon Bley (Eurostat)

1. INTRODUCTION

The proposed presentation outlines methodologies estimating a minimum threshold for the undercoverage of short-stay accommodation occupancy estimates within the European Union (EU), focusing on the integration of data from online platforms. Traditional tourism statistics (TOUR), collected under Regulation 692/2011, show significant undercoverage for short-term rentals categorized under NACE 55.2, specifically due to threshold limitations and incomplete business registers. The CETOUR project, which compiles occupancy data from four major online booking platforms, offers an alternative, complementary data source for short-stay accommodation. The analysis we suggest presenting at NTTS 2025 suggests that the actual occupancy of short-stay accommodation is significantly higher than currently recorded in TOUR statistics.

2. METHODOLOGY

Leveraging the fact that Regulation 692/2011 allows member states to limit the scope of observation to all tourist accommodation establishments having 10 or more bed places, two estimation methods are proposed to quantify the discrepancies:

2.1. Method A

The first method applies to countries where a threshold (≥ 10 bed places) is applied to the sampling frame. In these countries, the sum of nights spent in dwellings with less than 10 bed places according to CETOUR and the nights spent in short-term rentals according to TOUR are combined, resulting in significant adjustments:

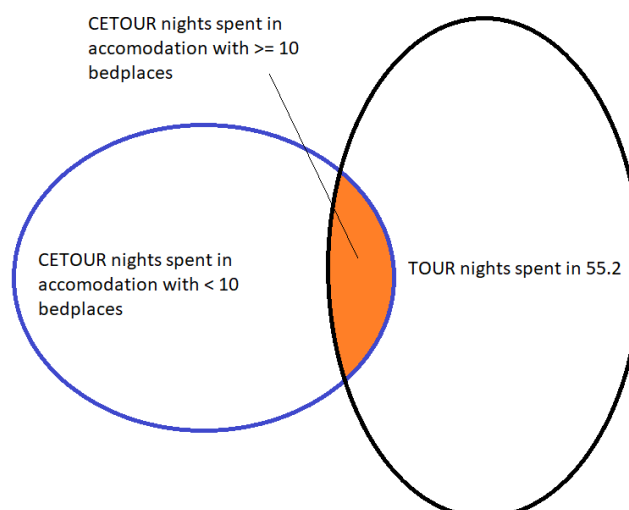


Figure 1. Schematic illustration of method A

2.2. Method B

The second method is used for countries without a threshold. In this case, the greater of the two occupancy values (TOUR or CETOUR) is taken as the minimum estimate, accounting for undercoverage in some regions:

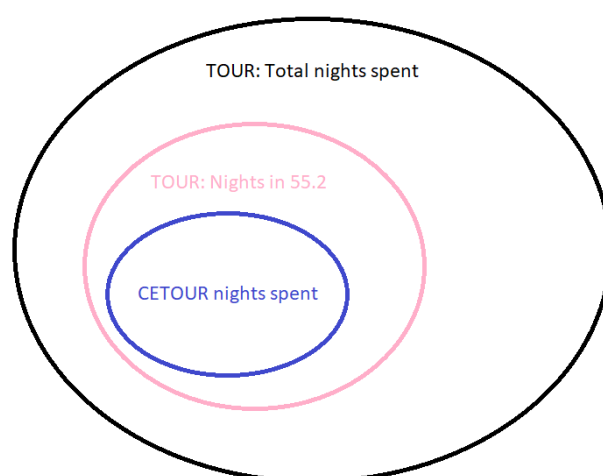


Figure 2. Schematic illustration of method B

3. RESULTS AND PRACTICAL APPLICATION

By applying these two basic methods, we can show that the true occupancy of short-term rentals in the EU in 2022 was at least 869 million, 32% more than the 658 million guest nights recorded through TOUR.

In country group A (the countries applying a threshold), including Germany, the Netherlands, Poland, Portugal, Czechia, Denmark, Sweden, Finland and Bulgaria, the undercoverage was at least 124.9 million guest nights, 57.9 % more than the reported TOUR total:

Table 2. Estimated undercoverage in Group A

Country	CETOUR, <10	CETOUR, >=10	TOUR, >=10	Estimate
Germany	40 547 034	1 934 621	91 692 940	132 239 974
Netherlands	8 225 889	560 984	51 057 588	59 283 477
Poland	24 943 054	1 405 399	33 574 428	58 517 482
Portugal	30 284 062	2 910 949	12 714 304	42 998 366
Czechia	5 804 331	901 249	9 012 454	14 816 785
Denmark	4 384 691	391 841	7 218 559	11 603 250
Sweden	4 539 025	450 231	6 108 212	10 647 237
Finland	3 106 642	250 106	2 355 138	5 461 780
Bulgaria	3 069 061	161 425	1 987 164	5 056 225

In country group B, for 10 countries (France, Spain, Greece, Hungary, Romania, Estonia, Latvia, Luxembourg, Malta and Cyprus), the CETOUR totals exceeded the TOUR totals in 2022, meaning that the undercoverage was at least 86.1 million guest nights:

Table 3. Estimated undercoverage in Group B

Country	CETOUR	TOUR, NACE = 55.2	Difference
France	137 166 518	96 342 814	40 823 704
Spain	117 954 021	86 715 327	31 238 694
Greece	31 621 497	29 005 077	2 616 420
Hungary	7 578 355	5 070 574	2 507 781
Romania	5 732 807	4 171 585	1 561 222
Estonia	1 286 201	1 261 551	24 650
Latvia	1 179 187	708 966	470 221
Luxembourg	322 201	249 587	72 614
Malta	2 822 783	187 155	2 635 628
Cyprus	4 135 742	0	4 135 742

4. MAIN FINDINGS

The two basic methods presented show that the true occupancy of short-term rentals in the EU in 2022 was at least 869 million, 32% more than the 658 million guest nights recorded through TOUR. Since (a) we were only able to apply method A to a subset of 9 countries, representing about a quarter of short-term rental nights; and (b) CETOUR itself does not capture short-term rentals booked outside of the four platforms, we can be sure that the true value is even higher.

Given these significant insights gained from applying these two methods, it might be beneficial for the ESS to engage in further discussions on refining the data collection and estimation techniques applied to TOUR statistics. This could help align the different datasets more closely and enhance the overall accuracy. For Eurostat, a potential next step could be to apply these estimation methods to historical data, to better understand long-term trends in short-stay occupancy.

Independent Address Identification Search Engine for National Statistical Institute Using ElasticSearch

Raya Berova, Lino Galiana & Julien Pramil (Insee)

1. INTRODUCTION

In order for survey interviewers to reach individuals, it is essential to accurately identify and geolocate their addresses. Address data, used in a wide range of statistical processes—from census to surveys—is often difficult to process due to inconsistencies, variations in input, and the volume of records.

In matters of address search, many rely on established services like Google Maps or OpenStreetMap. However, these platforms often pose limitations in terms of data control and reliability. Creating a custom address identification search engine provides complete control over the data, addressing concerns about data source transparency and monthly data updates.

A solution employing ElasticSearch (ES), a powerful software used to create and configure search engines, is here proposed to build an independent process for identifying address data for the National Statistical Institute (NSI). Moreover, ES enables text-based address search and supports the storage of geometric objects, considering the spatial aspect of addresses. This approach optimizes both processing time and accuracy by employing a two-step strategy: an initial strict search for precise address identification, followed by a flexible matching phase for addresses not identified in the first step, which accounts for spelling errors and variations in the input.

2. METHODOLOGY

The methodology is based on the implementation of ES to handle large-scale address datasets. Data is organized into JSON **documents**, which represent individual entities. To enable efficient searching, the search engine employs inverted **indices**—a data structure that links terms to the documents where they appear. Its distributed architecture supports the rapid search and analysis of vast data volumes, providing near real-time performance [1].

2.1. Address data indices

Two key indices were developed: one for complete addresses and one for street names.

Table 1. Example address index

idAddress	Number	Suffix	Full Address	idStreet	Postal Code
A1	10		10 Boulevard Royal Crescent	V1	W1B4AA
A2	25		25 Road Kingsway Court	V2	W1B4AA
A3	55	bis	55 bis Street Michael's Mount	V3	SW36PE

Table 2. Example street index

idStreet	Type of Street	Name of Street	Postal Code
V1	Boulevard	Royal Crescent	W1B4AA
V2	Road	Kingsway Court	W1B4AA
V3	Street	Michael's Mount	SW36PE

In reality, other variables are present in the indices, including geolocation variables. However, these are not useful for conducting text-based address search.

To enable effective searches, it is crucial that the data in the indices is processed in the same way as the addresses being identified. For this reason, the same filters are applied, called **analyzers**, to both the indices and the inputs to the search engine. These include:

- Lowercasing
- Accents handling
- Punctuation handling
- Use of synonym lists to normalize spellings

2.2. Searching process

To search for an address using this engine, the input must include a complete address string, for example, "10 bd Royal Crescent", along with its postal code string, such as "W1B4AA". Multiple queries will then be executed:

1. **Exact Match Query:** A strict query is performed within the town to match the exact address string. The query returns addresses from the index where the *"full address"* field matches exactly with the address provided. It bypasses tokenization, which is applied by default in ES, to ensure precision. This step allows a quick identification of addresses with perfect spelling after applying the analyzers, resulting in a time-saving process. If the address is not identified at this stage, the process moves on to the next step.
2. **Street-Level Query:** This step involves a more complex query with fuzzy matching to identify the street within the town. A fuzzy matching is an efficient approximate string matching technique [2]. Correcting a string with a fuzziness level of 1 means adding a letter, removing a letter, replacing a letter, or swapping two letters. In order for a street to be returned in the search results, at least one of the following conditions must be met:
 - Matching the whole street name with a fuzziness level of 1
 - Matching one token of the street type or street name with a fuzziness level of 1
 - Matching one n-gram of the street name (n = 3, 4 and 5)

Each time a condition is met, the street's **score** increases. Additionally, weights have been applied to these conditions, referred to as “**boosts**”. Here, the first condition carries the most weight. The street with the highest score will be chosen. In addition, an external verification outside of Elasticsearch is performed to validate the selected street. If the address provided contains less than 10% of the trigrams from the reference street, the street is considered not correct.

3. **Number and suffix Query:** If the street is identified, the algorithm searches for addresses within the street (*idStreet*). To be returned, the address numbers in the index must match exactly the number extracted from the address to identify. This extraction is performed using a custom-developed **RegEx**. Additionally, a function has been implemented to check the provided address for suffix, and if any are found, the reference addresses with the same suffix will be boosted.

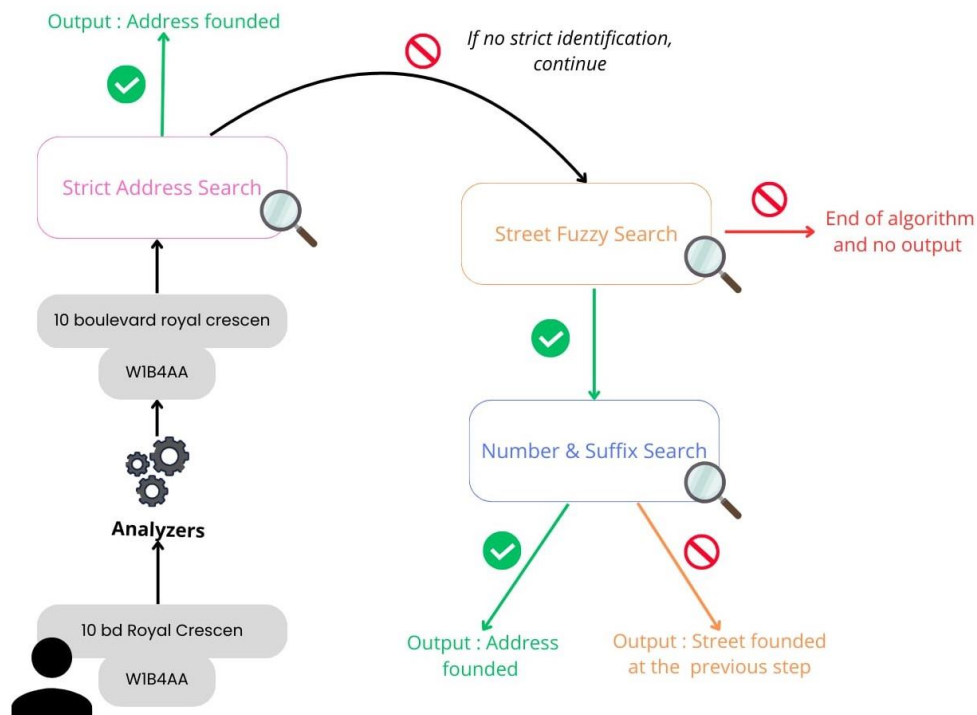


Figure 1. Overview diagram of the full identification process

3. RESULTS AND PRACTICAL APPLICATION

The ES indices were populated with addresses derived from multiple clean and reliable sources from the NSI. In total, approximately 27 million addresses and 3 million streets were indexed.

To evaluate the search engine’s performance, addresses from a survey conducted by the NSI were collected, where participants manually provided their addresses. These addresses contain noise: spelling errors, acronyms... Using the national individual code for each participant, the survey data was matched with our national individual-address database, which serves as the ground truth by providing the correct *idAddress*. This annotated dataset, consisting of 100,000 addresses across the country, was used to provide quality metrics on the search engine.

These addresses went through the identification process:

- The address was identified 86.0% of the time with an accuracy of 0.94.
- The street was identified 99.5% of the time with an accuracy of 0.94.
- 0.5% of the addresses were not found, at any level.

A test set of 100,000 clean addresses, sampled from the data in the address index, was created to evaluate the quality of its own data identification: the address was identified 100.0% of the time with an accuracy of 1 (all with the strict identification step).

To assess the efficiency of the actual identification process, a comparison was made between two different pipelines:

- P1: The actual full pipeline including strict address search, street name search, and then number and suffix search.
- P2: A pipeline that starts with the street name and then the number and suffix search.

To improve processing times, an additional step has been added to P2, creating the P1 process: a strict address search without **tokenization/ngrams**—which is time consuming—, only on addresses in the town required. This step resolves a significant portion of the addresses, around 30% in practical cases. If the database is clean, like administrative addresses, this rate can rise to 100%. With the same test set of 100,000 clean addresses as before, it took **one-sixth the time** to identify them using this strict identification step compared to doing it without, with the same rate of address/street identification and accuracy. Adding this step does not increase processing time if very few addresses are identified, but it saves time if many addresses are found there.

4. MAIN FINDINGS

The primary findings indicate that Elasticsearch is a highly effective tool for optimizing address identification processes. This is much more effective than strict searches in a SQL-style database, which lack of analyzers and do not account for spelling errors *and acronyms*.

The P1 structure has proven its effectiveness through the evaluation of the quality of identification in an annotated database containing noisy addresses, as well as by comparing its processing time performance with the P2 pipeline.

This solution is scalable and can be adapted to other country datasets by taking into account the specificities of its national addresses with the boosts. This search engine represents a significant advancement in address data processing, offering data control, independency and reliability for official statistics.

5. REFERENCES

- [1] Kathare, N., Reddy, O.V., & Prabhu, D.V. (2022). A Comprehensive Study of Elastic Search. Journal of Research in Science and Engineering.
[https://doi.org/10.53469/jrse.2022.04\(11\).07](https://doi.org/10.53469/jrse.2022.04(11).07)
- [2] Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. (2003). Robust and efficient fuzzy match for online data cleaning.
<https://courses.cs.washington.edu/courses/cse591d/03sp/chaudhuri.pdf>

ineapir: an R package to query statistical data and metadata from National Statistics Institute (INE Spain)

David Crespo Rocas (INE)

1. INTRODUCTION

The Spanish National Statistics Institute (INE Spain) stores the statistical information it produces in the INEbase system in open electronic formats. Besides, this information can be accessed via web services using an Application Programming Interface (API). To facilitate the access to the information that the API can provide, an R package has been developed since R is widely used among the users who consume statistical data. The `ineapir` package [1] provides a set of functions to get open data and metadata through API calls. It is available to install and has a website with detailed documentation and use cases.

1.1. The INE API web service

The API web service accesses the statistical information available in INEbase through URL requests, which have the following structure:

`https://servicios.ine.es/wstempus/js/{language}/{function}/{input}[?parameters]`

Fields that appear in curly braces, { }, are required, while fields that appear in square brackets, [], are optional. Below is a brief description of the different fields:

- {language}: output language (ES for Spanish and EN for English).
-
- {function}: functions already implemented in the system to make different types of queries.
-
- {input}: identifier code of the object queried by the implemented function.
-
- [?parameters]: optional parameters.

2. DEVELOPING THE R PACKAGE

A standard procedure has been followed for the development of the package with the help of the `devtools` package [2]. In terms of internal structure, `ineapir` contains 21 functions that are organized in 5 files depending on the information they request (see Figure 1):

1. The file `data_tables.R` contains the function to get data from a statistical table.
2. The file `data_series.R` contains functions to get data from time series.
3. The file `metadata_tables.R` contains functions to get metadata from statistical tables.
4. The file `metadata_series.R` contains functions to get metadata from time series.

5. The file `metadata.R` contains functions to obtain information from other objects.

Besides, two main guidelines were considered when developing the R package:

- All functions and features of the API service must be present in the package.
- Minimum number of external dependencies.

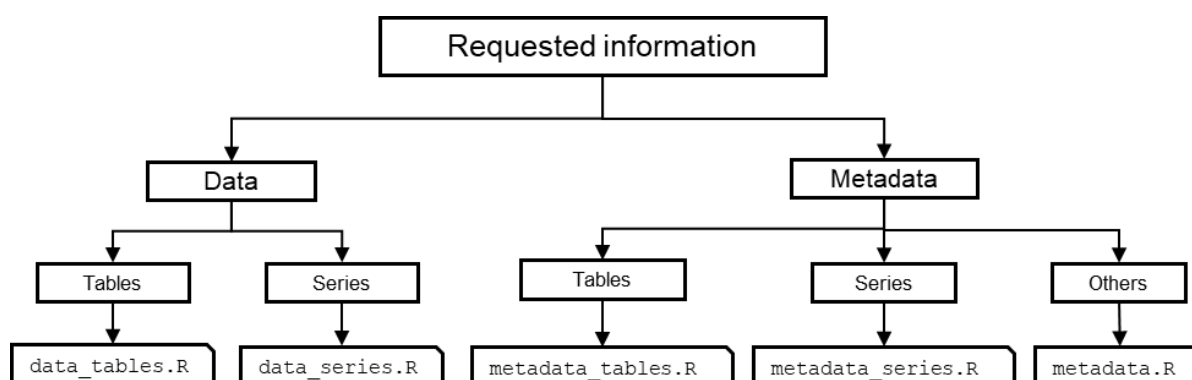


Figure 1. The functions of the package are organized internally in 5 files depending on whether they query data from tables or series, or whether they query metadata from tables, series, or other objects.

2.1. Inclusion of functions and features of the API service

The R package incorporates the functions of the API service and allows the user to query all the information related to tables, time series, and statistical operations. A time series is a set of observations measured at certain moments in time and ordered chronologically, which is defined by a set of variable-value pairs. Furthermore, each series has several immutable characteristics such as a unique identifier, periodicity, unit, name, among others. Time series belong to statistical operations and do not necessarily have the same periodicity. However, the statistical tables belonging to an operation can only contain time series of the same periodicity. In addition to the functions, the optional parameters of the API service have also been included in the package (see Table 1), allowing the user to perform more precise queries.

Table 1. Correspondence between API service parameters and package function arguments.

API parameter	Package argument	Description	Use
nult	nlast	Query n last periods	Filter
tv	filter	Filter data	Filter
date	dateStart dateEnd	Initial and last date of the query	Filter
geo	geo	Geographical disaggregation	Filter
det	det	Output level of detail	Output format
tip	tip	Friendly output and metadata	Output format
page	page	Paginated query	Output format

2.2. Minimum number of external dependencies

For better maintenance, the number of external dependencies on other R packages are kept to a minimum. Thus, `ineapir` has only two dependencies: the `httr` package [3], which is used to handle the API requests, and the `jsonlite` package [4], which is used to convert the json file retrieved by the API service to an R data frame. By default, `ineapir` makes GET requests. However, in case the URL reaches a length limit in the number of characters, the requests are made with a POST method.

3. MAIN FEATURES

The core function of the package is to obtain data from statistical tables. To perform this type of queries it is necessary to previously know the table's identifier code (there is an article on the package website about this topic). In addition to getting all the data of a table simply by using this code, there exists the possibility to refine the query with additional arguments.

3.1. Filtering data

The first option to filter data is by period. In this case it is necessary to specify the `nlast` argument with the number of last periods to request. Besides, there is the possibility of filtering data based on the variables and values that define the time series of the table (there are articles on the package website about this topic). To do this it is necessary to specify the `filter` argument, which is a `list()` of variable-value pairs. Each element of the filter is sent as additional parameter to the API request and the response is already filtered. To filter it is necessary to know the codes of the variables and values. However, it is possible to use certain shortcuts instead of variable codes, which enables the feature of filter by string value names. Although, for better performance is recommended to filter by codes. Furthermore, we can also include regular expressions as values when using variable shortcuts.

3.2. Input arguments validation

By default, the arguments specified in a function are validated. However, this validation can be turned off with `validate = FALSE`. This is useful especially when filtering, since turning off validation means making fewer API calls to check the filter. If the validation is activated and the filter contains invalid variables or values, the execution will stop, and an error message will appear. For better performance, it is recommended to use codes in the filter instead of shortcuts and disable validation if the filter is valid.

3.3. Friendly output

By default, the API response contains metadata information encoded with internal codes. It is possible to retrieve this information in a more friendly output with the equivalent names using the argument `tip` set to 'A' for friendly output (e.g., epoch to human-readable dates).

3.4. Obtaining data and metadata at the same time

When the data of a table is requested, it is possible to obtain the metadata as well using the argument `tip = 'M'`. The metadata consists of information about the variables and values that define each time series. To extract this information into additional columns the arguments `metanames = TRUE` and `metacodes = TRUE` can be used. The first one will create additional columns for each variable with the names of the values and the last one will create additional columns for each variable with the identification codes of the values.

4. CONCLUSIONS

The `ineapir` package provides users with a direct way to obtain statistical information disseminated by the Spanish National Statistics Institute and process it in their developments in R. It contains functionalities such as facilitating data filtering even without knowing internal codes of variables and values or having metadata information next to the data.

4.1. Future work

The next step would be to develop improvements to better handle API web service version updates. The goal is to allow access to the new changes introduced by a new version and at the same time prevent software developed by users with previous versions from being affected by an update.

5. REFERENCES

- [1] Crespo D (2024). `ineapir`: Obtaining data published by the National Statistics Institute. <https://es-ine.github.io/ineapir>, <https://github.com/es-ine/ineapir>.
- [2] Wickham H, Hester J, Chang W, Bryan J (2022). `devtools`: Tools to Make Developing R Packages Easier. <https://devtools.r-lib.org/>, <https://github.com/r-lib/devtools>.
- [3] Wickham H, `httr`: Tools for Working with URLs and HTTP (2023), <https://httr.r-lib.org/>, <https://github.com/r-lib/httr>.
- [4] Jeroen Ooms, The `jsonlite` Package: A Practical and Consistent Mapping Between JSON Data and R Objects (2014), arXiv:1403.2805.

Integration of MNO data with survey data to produce commuting statistics

Tiziana Tuoto, Erika Cerasti, Loredana Di Consiglio, Danila Filipponi,
Tiziana Pichiorri (Istat) & Li-Chun Zhang (SSB)

Use of Mobile network data (MNO) integrated with surveys are exploited to produce commuting statistics, traditionally estimated with census data. Availability of MNO data allows for the derivation of recurrent flows between home and work/study locations. These flows serve as a valuable information and can be used as the primary source for producing official statistics by adjusting MNO data coverage with a quasi randomisation approach. On the other hand superpopulation model such as spatial interaction models can describe the movement across different places. The methods are applied to estimate commuter flows between municipalities of an Italian region.

1. INTRODUCTION

In this work different approaches for integrating Mobile network operator (MNO) data with survey data to produce official statistics are explored.

The study is developed under the MNO MINDS ESSnet.

Traditionally, decennial population censuses have been the basis for gathering commuter data, proving flows of individuals across municipalities. However, for more timely information, some NSIs have shifted towards annual census surveys with reduced sample sizes. This transition presents various challenges in deriving accurate commuting statistics.

Availability of MNO data allows for the derivation of recurrent flows between home and work/study locations. These flows serve as a valuable information and can be used as the primary source for producing official statistics by adjusting MNO data coverage with a quasi randomisation (QR) approach.

In the superpopulation framework, aggregate MNO data serve as covariates, complementing target variables obtained from census sample surveys.

Spatial interaction models are widely used in transportation planning and urban studies to describe the movement of people, goods or information across different places.

These models describe the relationship between the origin and destination of flows within a geographic area accounting for factors such as distance, attractiveness, and connectivity of the different locations, also including spatial auto-regressive components.

2. METHODS

2.1. Superpopulation models

Let Y denote an $n \times n$ square matrix of origin-destination flows from each of the n origin zones to each of the n destination zones as shown in Equation (1). The elements on the main diagonal of the matrix represent intra-zonal flows.

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{bmatrix} \quad (1)$$

We can produce a $N \times 1$ stacked vector $y = \text{vec}(Y)$ of these flows by origin-centric ordering with $N = n^2$ and where the first n elements reflect flows from origin zone 1 ($i = 1$) to all n destinations, and the last n elements reflect flows from origin zone n ($i = n$) to destinations $1, \dots, n$

The classical interaction model (LeSage and Pace, 2008) is:

$$\mathbf{y} = \alpha + \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (2)$$

where:

- $\mathbf{H} = \mathbf{H} \otimes \mathbf{I}_n$ is an (N, Q) matrix of Q origin-specific variables that characterize the ability of the origin zones to produce flows,
- $\mathbf{G} = \mathbf{I}_n \otimes \mathbf{G}^d$ is an (N, R) matrix of R destination-specific variables that represent the attractiveness of the destination zones,
- $\mathbf{D}$ is an (N, S) matrix of S factors influencing the flow, such as distance, transportation infrastructure, information between origin and destination zones (including MNO data),
- $\boldsymbol{\epsilon}$ is an $(N, 1)$ vector of errors with $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_N)$.

LeSage and Pace (2008) extend the non-spatial model (2) by introducing a spatial autoregressive component, filtering for spatial dependence associated with destination and origin regions:

$$(\mathbf{I}_N - \rho \mathbf{W}_d)(\mathbf{I}_N - \lambda \mathbf{W}_o)\mathbf{y} = \alpha + \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + \mathbf{d}\vartheta + \mathbf{x}\xi + \mathbf{e} \quad (3)$$

Here, if we denote $\mathbf{W}$ as a row-standardized spatial weight matrix, the $N \times N$

matrix $\mathbf{W}_d = \mathbf{W} \otimes \mathbf{I}_n$ describes the spatial connectivity between regions.

To account for possible bias small area models can be applied (Fay 1979).

For making inferences about the small area means, Y_{ij} we assume that direct estimators of the commuting flow from the origin i to the destination

j , $\hat{y}_{ij}$ are available. Suppose that out of n areas of destination only m areas

are observed in the sample. As in the small area framework, $\hat{y}_{ij} = y_{ij} + e_{ij}$,

$i = 1, \dots, n$ and $j = 1, \dots, m$ where e_{ij} are the sampling errors. We further assume that the areas obey the population model (3), where MNO data serve as auxiliary data $\mathbf{x}$.

A simplification can be achieved if the dependence is introduced only in the error term u :

$$\hat{\mathbf{y}} = \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + \mathbf{d}\theta + \mathbf{x}\xi + (\mathbf{I} - \rho_j\mathbf{W}_j)^{-1}\mathbf{u} + \mathbf{e} \quad (4)$$

If $\mathbf{W}_j = \mathbf{W}_o$, then the model can be defined the Origin Spatial Error Model (O-SEM). For $\mathbf{W}_j = \mathbf{W}_d$, we have the Destination Spatial Error Model (D-SEM). Finally, if $\mathbf{W}_j = \mathbf{W}_w = \mathbf{W}_o \cdot \mathbf{W}_d$, we have the Origin-Destination Spatial Error Model (OD-SEM).

2.2. QR approach

Assuming that among the movements detected by MNO, the movements due to commuting by work and study m_{ij} can be identified, estimates can then be obtained by adjusting these counts by their coverage in a QR-approach.

A measure of coverage at broader area can be given by

$$\hat{Y}_{i+|z}/m_{i+|z} \quad (5)$$

where $\hat{Y}_{i+|z}$ are the direct estimates of commuters originating from area i in the socio-economic group z and $m_{i+|z}$ the corresponding MNO count. Then

$$\hat{Y}_{ij|z}^M = \hat{Y}_{i+|z} \frac{m_{ij|z}}{m_{i+|z}} \quad (6)$$

and the target estimate based on MNO is

$$\hat{Y}_{ij}^M = \sum_z \hat{Y}_{ij|z}^m \quad (7)$$

Alternatively, the coverage adjusting factor(s) can be estimated through a survey where the use of mobile phone data is recorded.

3. RESULTS AND CONCLUDING REMARKS

Results of the application of the previous methodology for estimating the commuting flows between municipalities in an Italian region will be presented at the conference. The application is based on real data coming from:

- the Italian census survey, used as direct estimates in the super-population modeling and in the adjustment factors in the QR approach;
- CDRs from an Italian mobile operator, used as auxiliary variables in the super-population modeling and primary source of information in the QR approach;

- the Italian survey on Daily Life Aspects, used as alternative in adjustment factors in the QR approach.

The findings will be compared and commented. In addition, a proposal for combining the estimates derived by different approaches to produce robust results will be introduced and discussed.

ACKNOWLEDGMENTS:

This work was co-funded by the European Commission Project “MNO-MINDS”
- 101132744 — 2022-IT-TSS-METH-TOO.

4. REFERENCES

- [1] ESSNet MON-MINDS (2024) Deliverable 3.1 Preliminary Report on Methodology, Work Package 3 *Methodologies and open source tools for integrating MNO and non-MNO data sources*
- [2] Fay, R.E. and Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 85, 398-409.
- [3] LeSage, J. and Kelley Pace. R (2008). Spatial econometric methods for modeling origin-destination flows. In *Journal of regional science*, Volume 48, Issue 5, 941-967

Integration of Mobile Network Operator (MNO) data with further data sources: How to treat errors and biases

Gloria Deetjen & Maurice Brandt (Destatis)

1. INTRODUCTION

In the digital age, a large variety of new data sources have emerged and so has the need to provide statistics that comprise advantages of new data sources on the one hand and that produce high quality and reliable outcomes in a world of fast data on the other hand. Mobile Network Operator (MNO) data are amongst the most favourable types of new data sources because of their high spatial resolution as well as their timely availability. The integration of MNO data with traditional or further new data sources is a promising way to benefit statistics as various experimental studies have shown. However, the implementation of integrated MNO data into regular statistical production requires profound solutions and innovative methodologies because of potential errors and quality issues which can mostly arise from either the data itself (data configuration), the difference between the statistical unit and the observation unit, or from the way the data is used (target or auxiliary). In addition, more insights on socio-demographic characteristics of mobile phone users and device usage are required for quality improvements and for better application.

The ESSnet MNO-MINDS project “Mobile Network Operator – Methods for Integrating New Data Sources” addresses these needs by proposing a reference frame for methods regarding the combination of MNO and non-MNO data for official statistics production. The non-MNO data sources selected result from a landscape analysis which studies different data sources and examines each source for its data access, availability and quality as well as for its relevance to concrete application scenarios. This process is presented in the dedicated conference paper as part of the same ESSnet MNO-MINDS session at NTTSS2025 conference. Further, a questionnaire will enable a better usage of MNO data itself by improving quality and addressing several sources of biases, e.g. multiple device usage, mismatches between the device and user, and diverse usage patterns across different demographic groups. This paper aims to present intermediate results of methodological developments and a questionnaire design to address these aspects. Results have been achieved by collaboration of NSIs from Austria, France, Germany, Italy, Norway, Portugal, Romania, Spain, Sweden, and the Netherlands. Altogether, the final project outcomes aim to provide a full package of solutions and concrete starting points for NSIs and further stakeholders in the European Statistical System (ESS).

2. HOW TO IDENTIFY ERRORS AND FIND THE RIGHT METHODS

In most applications of methodology, MNO data is not used alone but together with further data sources. So, after carefully selecting the most promising data sources to be integrated with MNO data for a concrete statistic or application scenario, the next step is to think of errors and sources of bias. Because of the wide range of potential applications, one should identify possible error sources for each specific scenario. For this, a total error analysis is helpful and comprises three elements: First, is the statistical unit which is derived from the target statistic different from the observation unit? This can often be the case, whenever one is not interested

in signals but e.g. in the presence or movement of people. Then, at which level is the available MNO data aggregated? This naturally affects the choice of appropriate methods and since many National Statistical Institutes (NSIs) have access to macro MNO data, it represents a limitation that needs to be addressed. Last, is MNO data used as a target or auxiliary variable? Depending on how it is used, different aspects may or may not be an issue. Most issues to be considered are (non-)representativity and user ambiguity.

After the total error analysis for each specific scenario, suitable methods need to be selected to deal with the identified errors. In general, relevant methodologies can be categorised into three approaches: Randomisation, quasi-randomisation, and super-population modelling (Zhang, 2019). Behind these broad categories, there is a range of relevant methods, e.g. statistical calibration or origin-destination estimation. Further, software and open source tools support application of methodology.

But how can such application scenarios look like? It seems that for many NSIs applications regarding the following themes are of high relevance: Tourism, Population, Mobility, and Environment. The project outcomes shall respect the different thematic focuses that different countries may have within the ESS. Methods are developed having such scenarios in mind. In more detail, the project aims to focus on inbound tourism, domestic mobility (e.g. commuting and tourism), de facto population, and on the effect of car traffic and pollution.

3. ADDRESSING BIAS OF MNO DATA WITH A DEDICATED SURVEY

But not only the integration of data sources with MNO data is affected by biases and errors, also the quality of MNO data itself contains shortcomings due to multiple reasons (Saidani et al., 2022). One approach to improve representativity of MNO data is by a traditional survey. Because MNO signals do not equal persons, several biases need to be adjusted for: Device-User-Ambiguity refers to situations where the device user is not the same person as the contract holder (e.g. business contracts). Next, individuals may carry multiple devices and SIM cards, and user behaviour might differ between users or user groups. As most NSIs have access to anonymised and aggregated MNO data, there is lack of information on socio-demographic information on mobile phone users, which is required e.g. for extrapolation.

Considering the relevant application scenarios from the landscape analysis, the methodological developments, and the need to address the abovementioned issues, the survey contains questions regarding the mobile phones and SIM cards in use (e.g. number of SIM cards in phone 1, 2, 3), contractual relationships (e.g. relation to the contract owner), and on user demographics and lifestyle (Kowarik et al., 2024). The last aspect collects general socio-demographic information as well as responses about device usage in every-day life and during travels. Insights on contractual relationships may give an idea on market shares among e.g. demographic groups.

4. MAIN FINDINGS

To treat errors and bias of MNO data or MNO data together with further data sources, one first needs to identify the potential errors with the total error analysis and consider three elements: What does MNO data represent in the scenario, how MNO data is used, and at what aggregation level it is available. Depending on the outcome, a randomisation, quasi-randomisation, or super-population model approach can be followed to handle the errors. With the help of a traditional survey, necessary quality improvements of MNO data can be achieved: For example, extrapolation could be improved with more detailed information on mobile phone users and device usage.

ACKNOWLEDGMENTS

This work was co-funded by the European Commission Project “MNO-MINDS” - 101132744 — 2022-IT-TSS-METH-TOO.

5. REFERENCES

- [1] Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3:103-113. DOI:10.1080/24754269.2019.1666241
- [2] Saidani, Y., Bohnensteffen S., Hadam S. (2022). Qualität von Mobilfunkdaten – Projekterfahrungen und Anwendungsfälle aus der amtlichen Statistik. *WISTA – Wirtschaft und Statistik*, 74 (5), 55–67.
- [3] Kowarik, A., Tuoto, T., di Consiglio, L., Deetjen, G., Brandt, M., Kamali, R., Gylling, K., Faricelli, C., Piombo, S., Pichiorri, T. (2024). Enhancing the Quality of Mobile Network Operator Data with a traditional survey with the right questions. *European Conference on Quality in Official Statistics 2024*.

Inveus. Remote access to data space for researchers

Jose Jabier Zurikarai, Jesus Nieto (EUSTAT- Basque Statistical Institute), Zunbeltz Izaola, Belen Muñiz (Mondragon University), Marivi Garcia, Javier San Vicente (EUSTAT- Basque Statistical Institute) & Begoña Laibarra (Software Quality Systems, SQS)

1. INTRODUCTION

Statistical institutes have access to a large volume of data, both personal and non-personal, and from both public and private sources. This data can assist in the development of a more competitive economy, but various legislations limit its sharing by external agents on the part of the institutes. There is an interesting case which is the use of data for Scientific Research. Official Statistics receive requests from the Researchers to access data. If those data are protected, researchers can use blind computers and disclosure control techniques to preserve confidentiality. But those solutions are in-site and that is an obstacle to promote scientific research. Thus, the development of data space that will enable remote access to the researchers is the aim of this project.

2. METHODOLOGY

2.1. Actors

There have been three institutions involved in INVEUS and they are playing complementary roles to develop a Statistical Data Space for Researchers:

1. EUSTAT - Basque statistical institute

Promoter of the Data Space

Service and data provider

Data Space Designer

Data Space Governance Authority

2. Mondragon University

Researching agency

Data Space Designer

Services Designer

Technology-tester

3. SQS. Software Quality Systems

Provider and developer of the data space infrastructure

Application and services developer

Data Space Designer

2.2. The model and the roles

The data space is a technological solution offered to the researchers. Through a web environment, the researcher can access and consult the data catalogues published by Eustat, as well as the various data processing services. Hence, the Data Space works as a bridge between the researchers and the statistical institute.

1. Researcher:

Upon registering, the researcher gains access to a data catalog and a services catalog, which allow them to explore the potential of incorporating statistical data into their research. The catalog describes the main characteristics of the provided data. Additionally, the offered services include basic data management activities, such as uploading their own data, data merging, and tabulations, as well as some techniques based on Machine Learning.

According to the configuration detailed above, the process defined for the researcher begins when the researchers select which data and services they wish to use and submit a request that is executed in a sandbox available within the Eustat environment. The results undergo a semi-automated filtering process that ensures complete confidentiality, as they are neither visible nor accessible until the results have been reviewed.

Consults the **data catalogue** to know what datasets are available and which variables can work with

Decides which of the **services** is the needed one: merge, tabulation...

Makes a **request** and **retrieves** data

...But doesn't have direct access to data and doesn't see any confidential information

2. Data Space as a service for data processing.

Data Space technology guarantees sovereignty, confidentiality, and security by leveraging decentralized control, robust identity and access management, secure data processing environments, and encrypted data exchanges. These mechanisms ensure that data owners retain full control over their data while also enabling secure and confidential data collaboration across organizations. In this proof of concept, an architecture based on the IDS-RAM has been defined, and certified components such as the Clearing House and the Identity Provider have been used. These components facilitate the deployment of a Data Space aligned with the policies defined by the promoter. Since the statistical data does not leave Eustat's servers and there is a defined review process prior to sharing, its confidentiality is ensured. However, it is not only the statistical data that is protected; the data shared by the researcher also benefits from the security provided by the connectors that facilitate the exchange. Thus, all this infrastructure has been configured as a set of services that Eustat offers to researchers to use statistical data in their research.

It uses Connectors and Digital Security Controls as a service.

In order to build the data sharing infrastructure it includes IDS core elements: Broker, Clearing House, Digital Certificates

The data space shows to the researches the catalogue. This catalogue is maintained by the Statistical Institute. It also sends the request to the statistical institute and uploads the results for the researcher.

3. Statistical institute

It develops and publishes the data and service catalogue which are shown in the data space.

It has its own connectors to get the request from the data space and send the results.

When receives a request, it operates with the data selected by the researcher and makes the calculations needed as a service. Then it has to validate the results and make sure that the results do not break the confidentiality rules. In the case of a breakdown of the rules, the researcher will get an advice saying that the results cannot be released.

3. RESULTS AND PRACTICAL APPLICATION

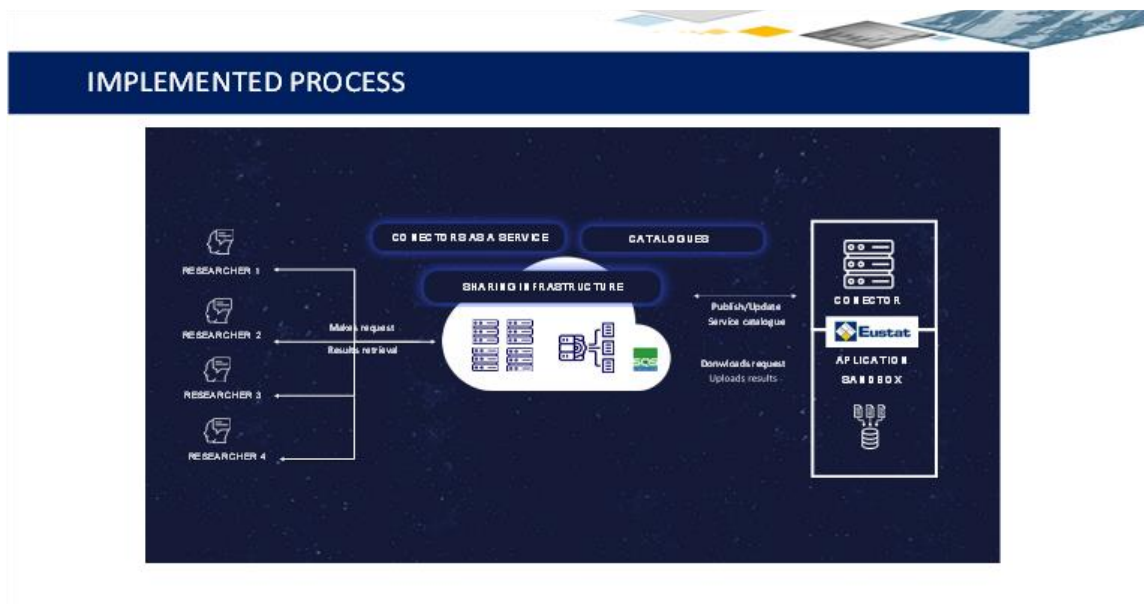


Figure 1. the process

The figure shows the actors and their roles

4. MAIN FINDINGS

We made a Proof of concept to validate the whole process and it worked.

Technically is feasible, but there are some decisions to be taken. The Data Space can automatically register a researcher but there is no standard to certify who is a researcher.

During the Proof of concept, simple data and services were tested. We have to scale them to real situations where the boundaries of the system can be tested intensively.

5. REFERENCES

- [1] https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-White-Paper-IDSA-Rule-Book.pdf
- [2] <https://docs.internationaldataspaces.org/knowledge-base/ids-ram-4.0>

JDemetra+ 3.2.4: New (R) tools for (high-frequency) time series analysis

Tanguy Barthelemy (Insee)

1. INTRODUCTION

The latest version of JDemetra+¹, 3.2.4 to be released in July 2024, completes the advances in seasonal adjustment, in particular for high frequency. A state-space framework gives access to Basic Structural Models, offering a seasonal adjustment procedure with explicit decomposition and time varying trading day correction.

This open source software ([al](#)) is unique in its combination of very fast java routines, a graphical user interface and a family of R packages. The graphical interface offers a structured and visual feedback, suitable for refined analysis and training, whereas R tools allow the user to mix the capabilities of JDemetra+ with the versatility of the R world, be it for mathematical functions or data wrangling.

The first R package allowing to use JDemetra+ core algorithms **{RJDemetra}** R package was created in 2018 ([Alain Quartier-la-Tente and Baeyens](#)) and since then the software has been rapidly expanding: algorithms have been extended and tools upgraded. In version 3 **{RJDemetra}** R package has been replaced with a family of R packages covering each a very specific perimeter. This modular structure makes the functions more readable, fosters evolution and encourages the user to enhance the functions according to its own needs.

The tedious task of generating calendar regressors for quarterly, monthly but also high frequency data is now possible using simple R functions of the R package **{rjd3toolkit}** ([Palate et al. 2024](#)), for the first time in R. This kind of feature is of high importance for seasonal adjustment practitioners and even more important is the possibility to refresh estimation when new data is available, which is covered in greater detail in the methods section.

Our paper aims at describing and illustrating the new capabilities of JDemetra+ 3.2.4 as well as the R packages allowing to access them. In the remainder of this abstract we highlight three categories of features which seem to be much sought-after by users: seasonal adjustment of high frequency data, revision analysis and tools for building seasonal adjustment production processes entirely in R.

2. METHODOLOGY

Seasonal adjustment of High-frequency data

Infra-monthly economic time series have become increasingly popular in official statistics in recent years, more and more users ask for timely weekly and even daily indicators of economic developments. Many of those indicators display seasonal behavior and, thus, are in need for seasonal adjustment. JDemetra+ seasonal adjustment algorithms have been augmented to meet this need, offering an enhanced reg-ARIMA pre-treatment model and extended versions

¹ JDemetra+ is an open source software for time series analysis. It has been officially recommended by Eurostat to the European Statistical System members since 2015.

of the ARIMA model-based, STL and X-11 seasonal adjustment approaches able to deal with multiple and non-integer periodicities common in high frequency data, as described in [Webel and Smyk 2023](#). These extensions are accessible through the **{rjd3highfreq}** ([Palate 2024](#)), **{rjd3stl}** ([Palate 2024](#)) and **{rjd3x11plus}** ([Palate 2024](#)) R packages as well as through the graphical user-interface. Some key features like extended X-11 to any (fractional) periodicity or fractional airline model are unique to JD+. Fractional periodicities are tackled using a Talyor approximation for the backshift operator $B^{s+\alpha} \cong (1 - \alpha)B^s + \alpha B^{s+1}$, where α is the decimal part of the periodicity. A short illustration is provided in the Results section.

Revision analysis

Revision analysis provides important information on the efficiency of preliminary estimates, allowing to identify potential issues and/or improvements that could be made in the compilation process. The R package **{rjd3revisions}** ([Lemasson and Barthelemy 2024](#)) provides a tool to automatically perform a battery of relevant tests on revisions and create a visual report including both the main results and their interpretation. This tool can perform analysis on different types of revision intervals and on different vintage views.

The **{rjd3revisions}** R package performs revision analysis and offers both detailed and summary output including the generation of a visual report. It is composed of a selection of parametric tests which enable the users to detect potential bias (both mean and regression bias) and other sources of inefficiency in preliminary estimates. What we mean by inefficiency in preliminary estimates is whether further revisions are predictable in some way.

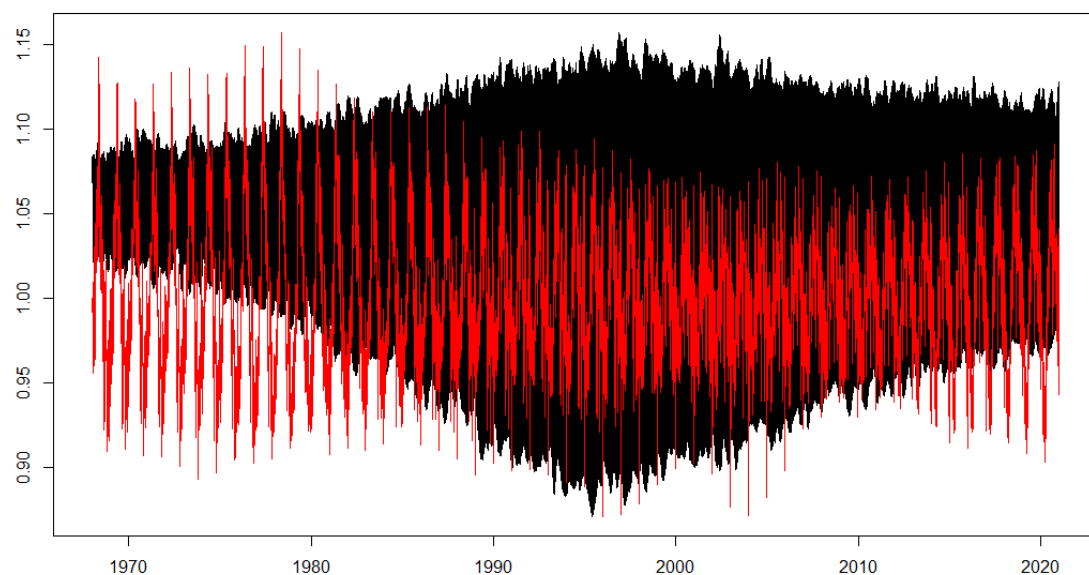
Mass production of seasonally adjusted data in R

R has become ubiquitous in official statistics and the demand for its use in production of seasonally adjusted data is growing fast. JDemetra+ offers the speed and the pre-specified refresh policies recommended by Eurostat Guidelines on SA ([Eurostat 2015](#)). A wide range of “partial-concurrent adjustment” options, in which parameters and re-estimated and/ or re-identified gradually have been long available in JDemetra+. But, until now, in the versions 2 family, these options were linked to updating a workspace (specific data structure) via the graphical user interface or more probably via the cruncher (a batch module). It was quite a liability for full production in R as explained in [Smyk and Tchang 2021](#). Revisions policies are now even more customizable if implemented in R, as time spans on which options are applied can be chosen by the user. Before version 3.0, the user could chose between re-identifying outliers on the whole series span or on the last year of the data (this is the widely applied “partial concurrent last outliers policy”), now the period is customizable, which really makes sense when progressively remodelling the impact of the covid crisis. It is now possible to fully manage workspaces and data providers with the **{rjd3workspace}** ([Palate et al. 2024](#)) and **{rjd3providers}** ([Palate and Piovani 2024](#)) R packages.

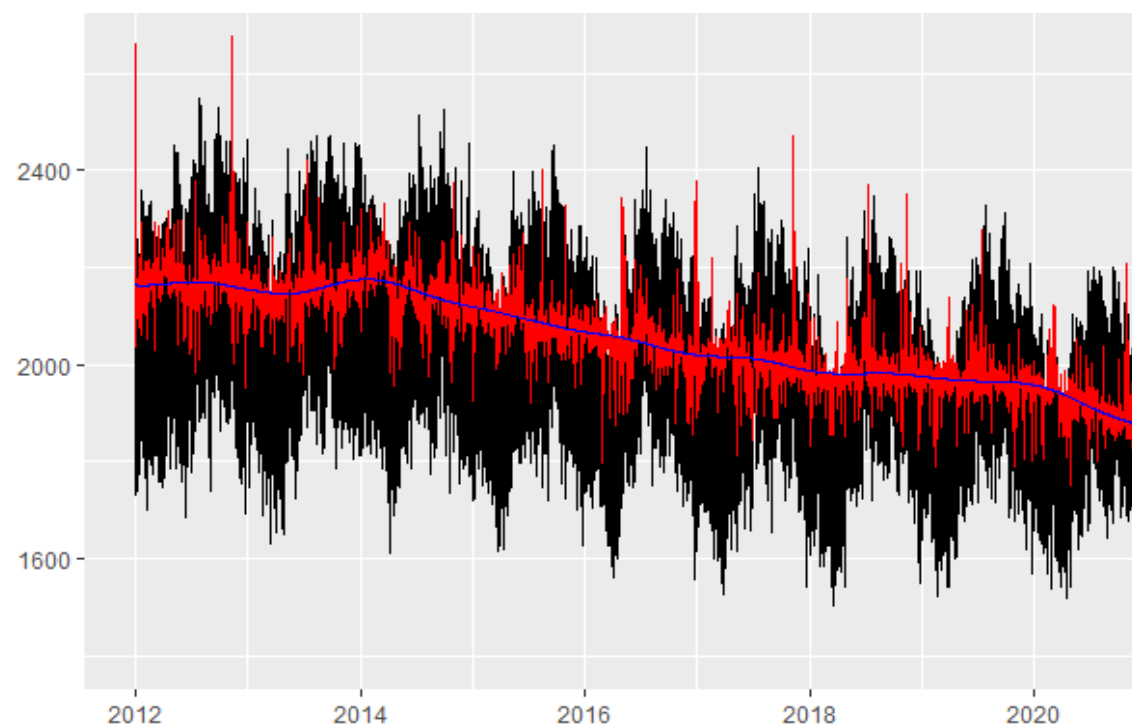
3. RESULTS AND PRACTICAL APPLICATION

Seasonal adjustment of French daily births series

We consider the series of daily french births from 1968 to 2020. Spectral analysis shows that two periodicities $p_1 = 7$ and $p_2 = 365.25$ are present. The series is first linearized: outliers are detected and calendar effects removed with the following fractional airline model: $(1 - B)(1 - B^7)(1 - B^{365.25})(Y_t - \sum \alpha_i X_{it}) = (1 - \theta_1 B)(1 - \theta_2 B^7)(1 - \theta_3 B^{365.25})\epsilon_t, \dots \epsilon_t \sim \text{NID}(0, \sigma_\epsilon^2)$ with $1 - B^{365.25} = (1 - 0.75B^{365} - 0.25B^{366})$. Then a decomposition is performed with extended X-11, using modified filters with the Taylor approximation, which avoids imputing data.



French daily births: estimated seasonal factors , $p=7$ (black) and $p=365.25$ (red)



French daily births: raw (black), trend (blue), seasonally adjusted (red)

4. MAIN FINDINGS

This abstract is meant to give a hint on the new developments of an open source time series software, already widely used in official statistics throughout Europe and the world. JDemetra+ provides new solutions to some critical issues, which could benefit many time series analysts. Our forthcoming paper will describe in greater detail those innovations, providing detailed explanations on algorithm's extensions as well as R code snippets.

Further developments are under way in several domains, among them a better calibration of filters and more adapted tests to high-frequency data.

5. REFERENCES

- [1] Anna Smyk et al. “JD+ Documentation.” <https://jdemetra-new-documentation.netlify.app/>.
- [2] Alain Quartier-la-Tente, Jean Palate, Anna Michalek, and Raf Baeyens. “RJDemetra: Interface to ‘JDemetra+’ Seasonal Adjustment Software. R Package Version 0.1.6.” <https://CRAN.R-project.org/package=RJDemetra>.
- [3] Eurostat. 2015. “ESS Guidelines on Seasonal Adjustment.” Eurostat Methodologies; Working Papers, European Commission. <https://doi.org/10.2785/317290>.
- [4] Lemasson, Corentin, and Tanguy Barthelemy. 2024. “Rjd3revisions: Revision Analysis with ‘JDemetra+ 3.x’.” 2024. <https://github.com/rjdverse/rjd3revisions>.
- [5] Palate, Jean. 2024. “Rjd3highfreq: Seasonal Adjustment of High Frequency Data with ‘JDemetra+ 3.x’.” 2024. <https://github.com/rjdverse/rjd3highfreq>.
- [6] “Rjd3stl: R Interface to ‘JDemetra+ 3.x’ Time Series Analysis Software.” 2024. <https://github.com/rjdverse/rjd3stl>.
- [7] “Rjd3x11plus: Interface to ‘JDemetra+ 3.x’ Time Series Analysis Software.” 2024. <https://github.com/rjdverse/rjd3x11plus>.
- [8] Palate, Jean, and Alessandro Piovani. 2024. “Rjd3providers: Interface to ‘JDemetra+ 3.x’ Time Series Analysis Software.” 2024. <https://github.com/rjdverse/rjd3providers>.
- [9] Palate, Jean, Alain Quartier-la-Tente, Tanguy Barthelemy, and Anna Smyk. 2024a. “Rjd3toolkit: Utility Functions Around ‘JDemetra+ 3.0’.” 2024. <https://github.com/rjdverse/rjd3toolkit>.
- [10] “Rjd3workspace: Interface to ‘JDemetra+ 3.x’ Time Series Analysis Software.” 2024. <https://github.com/rjdverse/rjd3workspace>.
- [11] Smyk, A, and A Tchang. 2021. “R Tools for Jdemetra+, Seasonal Adjustment Made Easier.” Institut National de la Statistique et des Etudes Economiques. <https://www.insee.fr/en/statistiques/5019812>.
- [12] Webel, K, and A Smyk. 2023. “Towards Seasonal Adjustment of Infra-Monthly Time Series for JDemetra+.” *Bundesbank Discussion Paper*. <https://www.bundesbank.de/en/publications/research/discussion-papers/towards-seasonal-adjustment-of-infra-monthly-time-series-with-jdemetra--915460>.

Joint Estimation of Common and Specific Dynamic Factors from Vectors of Time Series: Application to Electricity Prices in Several Markets in the EU

Carolina García-Martos (Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid), Julio Rodríguez (Departamento de Análisis Económico: Economía Cuantitativa, Universidad Autónoma de Madrid) & María Jesús Sánchez (Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid)

Abstract: Dynamic Factor Models (DFM) and Dynamic Factor Analysis (DFA) are dimensionality reduction techniques useful for dealing with vectors of series. Recent surveys on these topics are Peña (2021) and Poncela et al. (2021). In some cases the focus is not just on reducing the number of parameters to estimate but also on the interesting issue of separating common and specific dynamic components.

This type of models can be easily expressed under state-space (SS) formulation and then estimation is carried out by direct maximization of the log-likelihood or using the EM (Expectation-Maximization) algorithm (Dempster et al., 1977).

The common part of the original m series in $\mathbf{y}_t$ is captured by the r unobserved common factors, $\mathbf{f}_t$ (r -dimensional vector of series) and the specific ones, $\boldsymbol{\varepsilon}_t$, m -dimensional vector of series, do not incorporate any common information. Thus, in the case of correct specification of the multivariate model for the r common factors $\mathbf{f}_t$, then we should not expect any correlation between adjacent elements of $\boldsymbol{\varepsilon}_t$. This assumption is imposed in the estimation procedure making the variance covariance matrix of the specific factors, $\mathbf{S}$, diagonal by construction.

This could be mathematically convenient but may be suspect in practice when dealing with complex real data, since it is very restrictive for the specific factors, as it imposes that they cannot evolve over time according to different univariate models. In fact, when dealing with some real data and applications, there is sometimes a need of a subsequent step after estimating the parameters involved in the model, just to capture any remaining correlation along each specific factor, as in Peña and Poncela (2004 and 2006) or in Alonso et al. (2011).

In this work, a joint estimation procedure that overcomes this problem is proposed.

Then, a complete simulation Monte Carlo study is carried out and the conclusions are clear: Not only the dynamics of the specific components are better estimated when using the joint estimation procedure, but also the dynamics of the common factors could be biased when using the 2-step estimation procedure (Quasi-Maximum Likelihood, QML).

Finally, the proposed joint estimation procedure (Maximum Likelihood estimation) is applied to properly estimating the common and specific components of electricity prices in several markets in the UE (The Netherlands, Belgium, Germany, Denmark, Finland, Italy, Spain and Portugal and France). Results for nowcasting and forecasting are also included.

1. INTRODUCTION

When using Dynamic Factor Analysis (DFA), each series in the observed m -dimensional vector of time series $\mathbf{y}_t$, i.e., each $y_{j,t}$, where $j = 1, \dots, m$ and t is the temporal index, is decomposed into common and specific part respectively, according to the observation equation:

$$y_{j,t} = \sum_{k=1}^r \omega_{jk} f_{k,t} + \varepsilon_{j,t}$$

Where the r common factors are $f_{k,t}$ with $k = 1, \dots, r$. The m specific ones are $\varepsilon_{j,t}$ with $j = 1, \dots, m$.

With respect to the Dynamic Factor Model by Stock and Watson (2002), where both the number of series under study, m , as well as the length of the data, n , are huge, Bai (2003) stated that although some correlation is present in the specific factors (univariate dynamic structure for each), consistent estimates are obtained for the parameters involved in the DFM.

DFA is usually expressed and estimated using SS methods. As far as asymptotic distribution of estimates of the parameters of a SS model is concerned, this subject is studied extensively in Caines (1988, Chapters 7 and 8). The consistency and asymptotic normality of the estimators is established under general conditions.

However, for the computation of consistent estimates, not only for the parameters of the state-space model under study, but also consistent estimates of the state variables, additional conditions involving the dimension of the vector of series, m , or the variances in the diagonal of matrix $\mathbf{S}$ must be satisfied.

Thus, in DFA, when using the 2-step (or subsequent) estimation procedure described above, when the conditions for consistent estimation of the latent variables do not hold, then the estimated specific factors are badly estimated and as a consequence, the estimates of the parameters of each univariate model are systematically biased.

In this paper a joint estimation procedure that overcomes this problem is introduced. The complete model, that includes not only the decomposition of the vector of series into common and specific components and the parameters of the VARIMA model for the evolution of the unobserved common factors over

time, but also the univariate modelling of the specific factors is written under its state-space formulation and jointly estimated.

The rest of this summary is organized as follows: In Section 2 the complete formulation of the model is introduced. In Section 3 the Monte Carlo experiment developed is summarized and in Section 4 the real dataset used in the paper is briefly described. Section 5 concludes.

2. THE MODEL

The complete formulation of the model that relates the observed vector of series with the unobserved common factors following a VARIMA model is given by:

$$\begin{aligned} \mathbf{y}_t &= \Omega \mathbf{F}_t + \varepsilon_t, \\ \mathbf{F}_t &= \Psi \mathbf{F}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathbf{N}_r(0, \mathbf{I}_r) \\ \varepsilon_{j,t} &= \sum_{i=1}^{p_e} \phi_{i,j} \varepsilon_{j,t-i} + a_{j,t}, \quad a_{j,t} \sim NID(0, \sigma_{aj}^2), \end{aligned}$$

where the notation is as in Alonso et al. (2011) and in García-Martos et al. (2011). In the full paper complete derivations for the differences between the QML estimation and ML estimation are obtained. This justifies the bias that appears in the QML (or 2-step) estimation.

3. SIMULATION STUDY

The models considered consists of $m = 4$ observed series, generated by $r = 2$ common factors that follow a VARI(1,1) model, such that:

$$(I - B)(I - \Phi_1 B)\mathbf{f}_t = \mathbf{w}_t, \quad \mathbf{Q} = E(\mathbf{w}_t \mathbf{w}_t') = \mathbf{I},$$

and different values have been considered for matrix parameter Φ_1 : For models called A, the AR coefficients of the 2 common factors are respectively 0.6 and 0.7. For models called B, the AR coefficients are 0.2 and 0.3, respectively.

The loads for the first factor are all equal to 0.5 both for A and B models. And equal to the column vector (0, -0.2, 0.1 and 0.3)^t for the second factor, as well for models A and B.

The specific factors are univariate AR(1) processes (with no dynamic cross-correlation). We consider the cases: I) the AR coefficient of the specific components is 0.5, II) the AR coefficient of the specific components is 0.8 and III) the AR coefficient of the specific components is -0.5.

Thus, the Simulation experiment includes results for estimation and forecasting for Models A.I, A.II, A.III, B.I, B.II and B.III. Also, the cases: A.0 and B.0 are included as benchmarks (specific factors with no dynamic structure and same structure for the common factors as in all the A and B models respectively).

For illustration purposes the estimation results for the parameters of the specific components of the first and second series in $\mathbf{y}_t$, for the case A.II when the length of the time series is $n=1000$ is shown in Figure 1.

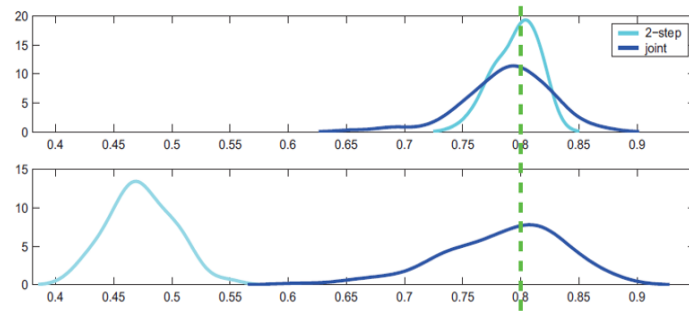


Figure 1. Densities of estimated autoregressive coefficients for the specific factors of the first (top) and second (bottom) series in y_t , models A.II, $n=1000$. Similar results are obtained for the third and fourth specific factors. True values are in green.

Estimation results for the AR parameter of the first common factor (true value = 0.6) in the cases A.0, A.I, A.II and A.III, when $n=1000$ is shown in Figure 2 for the 2-step estimation procedure.

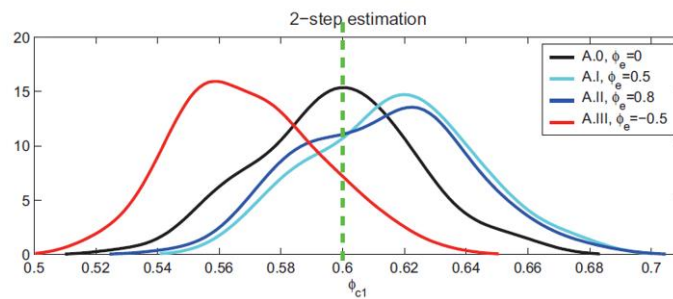


Figure 2. Densities of the estimates of the common dynamics for simulated cases in the case of performing the 2-step estimation method. Results are provided for the AR parameter of the first common factor for $n=1000$, for specifications A.0, A.I, A.II and A.III. True values are in green.

Joint estimation with the new full model here proposed -joint estimation of the common and specific dynamics- provides clearly better results (unbiased) as shown in Figure 3.

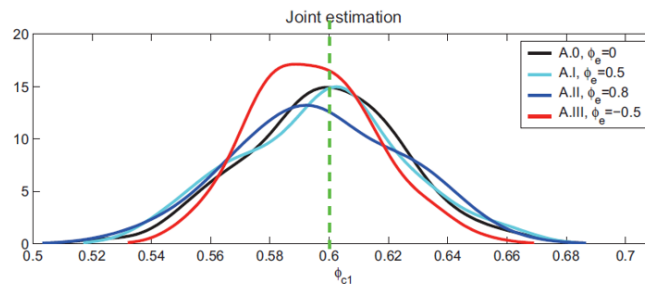
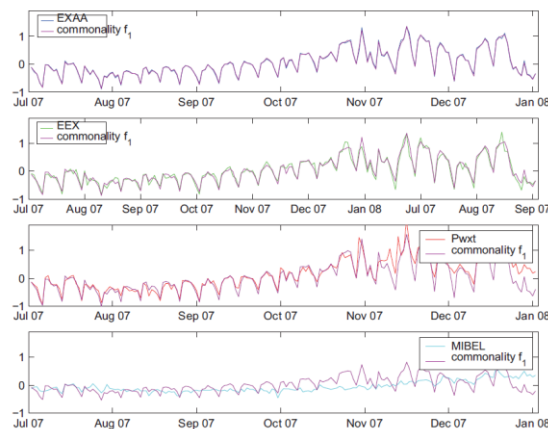


Figure 3. Densities of the estimates of the common dynamics for simulated cases in the case of performing the joint estimation method, the improvement are clear compared to Figure 2. Results are again provided for the AR parameter of the first common factor for $n=1000$, for specifications A.0, A.I, A.II and A.III. True values are in green.

4. REAL DATA EXAMPLE

Our methodology is illustrated with the series of electricity prices in some of the most important markets in the EU, for which we model and interpret the estimation results and compute forecasts.

Partial results (just a couple of years from a longer dataset used) are shown in Figure 3, where the log-series of electricity prices for some of the markets under study as well as the part of it explained by the first common factor extracted and estimated using the joint estimation procedure are depicted.



5. MAIN FINDINGS

A general Dynamic Factor Model in which the specific factors can evolve over time is introduced. A general procedure for the joint estimation of all the parameters involved is proposed. Moreover, a complete procedure to obtain likelihood-based consistent estimates of the dynamic common factors is given.

A Monte Carlo simulation study is carried out to be able to provide a comparison of the parameter estimates, both for the joint estimation and for the subsequent estimation procedure used in previous works. Also a real data application to electricity prices in the EU is provided.

6. REFERENCES

- [1] Alonso, A.M., García-Martos, C., Rodríguez, J. and Sánchez, M.J. (2011). "Seasonal Dynamic Factor Analysis and Bootstrap Inference: Application to Electricity Market Forecasting" *TECHNOMETRICS*, 53 (2), 137-151.
- [2] Bai, J. (2003), "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 1, 135-171.
- [3] Caines, P.E. (1988), "Linear Stochastic Systems," New York: John Wiley & Sons.
- [4] Dempster, NM and Rubin, DB (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39 (1), 1-22.
- [5] García-Martos, C., Rodríguez, J. and Sánchez, M.J. (2011). "Forecasting electricity prices and their volatilities using Unobserved Components", *Energy Economics*, 33 (6), 1227-1239.
- [6] Peña, D. and Poncela, P. (2004), "Forecasting with Nonstationary Dynamic Factor Models," *Journal of Econometrics*, 119, 291-321.

- [7] Peña, D. and Poncela, P. (2006), "Nonstationary Dynamic Factor Analysis," *Journal of Statistical Planning and Inference*, 136, 1237-1257.
- [8] Peña, D. (2021) "Some recent methods for analyzing high dimensional time series," *Spanish Journal of Statistics*, 3, 1, 7-35.
- [9] Poncela, P., Ruiz, E. and Miranda, K. (2021). "Factor extraction using Kalman filter and smoothing: This is not just another survey," *International Journal of Forecasting*, 37, 1399-1425.
- [10] Stock, J. H. and Watson, M. (2002), "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167-79.

Jordan National Interactive Data Centre Data as a Service

Faisal AlAlwan (Department of Statistics)

1. INTRODUCTION

The Hashemite Kingdom of Jordan is pursuing a comprehensive economic transformation plan, driven by digital transformation and the knowledge economy. Data as a Service (DaaS) plays a crucial role in providing accurate, real-time data to both government and private institutions, enabling informed decision-making. As global economies increasingly rely on data-driven models, Jordan's investment in DaaS aligns with international trends.

DaaS enhances key sectors such as education, health, energy, and finance by offering flexible, user-friendly data access. It promotes transparency, fosters innovation, and supports an attractive investment climate. The establishment of the National Interactive Data Center is a key initiative, advancing Jordan's digital infrastructure and boosting economic resilience. This project positions Jordan as a regional leader in leveraging data for sustainable development, innovation, and modernization.

1.1. Global Projects

To advance the establishment of an interactive statistical data center in Jordan, it is beneficial to draw connections to successful global projects that have demonstrated tangible results:

- A. **Eurostat (European Centre for Statistics):** Enhances data-driven decision-making, aiding governments and businesses in developing economic and social policies.
- B. **World Bank's Data for Development Initiative:** Serves as a model for Jordan's data center, following a similar approach of providing DaaS to support local and regional development.

1.2. The Reality of Data Management in Jordan

While data management in Jordan has seen remarkable development in recent years, several challenges remain that hinder its full potential:

- A. **Technical Infrastructure:** Jordan's infrastructure has improved significantly due to government and private sector investments in ICT.
 - Cloud Infrastructure: There is a need for further enhancement of cloud adoption to support the growing data demands in Jordan.
 - Internet Coverage and Speed: Jordan boasts a good level of internet coverage, although continuous improvements in speed and accessibility are necessary to support data initiatives effectively.
- B. **Government Policies and National Initiatives:** The Jordanian government has launched several initiatives aimed at enhancing data utilization across various sectors:

- Digital Transformation Program: This program includes the development of digital and statistical platforms designed to improve data collection, management, and analysis at a national level.
- Open Data Initiatives: These initiatives encourage the publication of public data to promote transparency and enhance informed decision-making in both the public and private sectors.

1.3. Private Sector and Data Use

The private sector in Jordan has started adopting modern data management technologies, particularly in sectors like telecommunications, financial services, and retail. However, challenges persist, especially in adopting big data technologies due to high costs and a shortage of skilled technical personnel.

2. CHALLENGES FACING DATA MANAGEMENT IN JORDAN

Despite tangible progress, several challenges continue to impede the development of data management in Jordan:

- Lack of Technical Skills: Although educational and training programs exist, the demand for experts in data management and analysis exceeds the current supply.
- Data Protection and Privacy: With the growing reliance on digital systems, there is a pressing need to develop comprehensive laws and regulations to ensure data protection in accordance with international standards.
- Lack of System Integration: The lack of integration between various data management systems hampers the ability to effectively analyze and utilize data, limiting its full potential.

3. FUTURE OPPORTUNITIES

Despite these challenges, data management in Jordan presents significant opportunities for future growth and development:

- Accelerated Digital Transformation: Jordanian organizations are increasingly willing to invest in advanced data management technologies, accelerating the pace of digital transformation.
- Data as a Service (DaaS): Jordan has the potential to position itself as a regional hub for providing DaaS, offering data and related services to neighboring countries to support their economic development.
- Collaboration with the Educational Sector: Jordanian universities and educational institutions have begun introducing specializations in data analytics and artificial intelligence, which will help bridge the gap in technical skills and foster a more robust data management ecosystem.

4. METHODOLOGY

4.1. **Establishment of the Center:** The center was created using theoretical and applied analyses to provide Data as a Service (DaaS). The methodology includes:

- Preparatory Stage: Analysis of the current state, preparation of work plans, technical and administrative requirements, and governance structures.
- Execution Stage: Tendering, procurement, platform development, and a trial launch on existing infrastructure.

4.2. Data Collection Phase

- Primary Sources: Data was collected from the Department of Statistics, with a trained team compiling data tables and setting up electronic links.
- Secondary Sources: Interviews with IT and statistics experts and officials to understand technical and structural requirements.

4.3. Technical Framework

- Infrastructure: Analyzed cloud infrastructure, security, privacy, and data requirements.
- Data Management System: Interactive data storage and analysis.
- APIs: Open interfaces for user access and integration.

4.4. Challenges and Opportunities

- Challenges: Infrastructure limitations, data protection, and privacy concerns.
- Opportunities: Modernizing the economic sector and building a knowledge-based economy.

4.5. **Implementation Plan:** A roadmap covering infrastructure development, staff training, and the rollout of DaaS services.

5. TANGIBLE AND EXPECTED RESULTS

5.1. Tangible Results

- Improved Data Access: Easier access to big data, enabling flexible data search, analysis, and manipulation.
- Increased Decision-Making Efficiency: Advanced tools will enhance the accuracy and efficiency of decision-making.
- Development of a Knowledge-Based Economy: Data collection and distribution will support startups, innovation, and more efficient government policies.
- Strengthened Cross-Sector Cooperation: DaaS will foster better collaboration between public and private sectors, based on reliable data.

5.2. Expected Results (Medium and Long-Term)

- Enhanced Data Quality: Improved accuracy and real-time updates will lead to better forecasting and analysis.
- Innovative Solutions to Challenges: The project will aid in addressing social and economic challenges with targeted solutions.

- Improved Operational Efficiency: Advanced data analytics will help organizations reduce waste and improve performance.
- Support for Innovation: Data will be leveraged to develop AI-based applications, fostering technological advancements in Jordan.

5.3. Expected Challenges

- Infrastructure Constraints
- Privacy and Security Concerns
- Sustainability Costs

6. CONCLUSION

The establishment of an interactive statistical data center in Jordan is a vital and pivotal step toward enhancing the country's ability to make data-driven decisions and achieve sustainable economic development. The tangible and expected results of this project will contribute significantly to creating a knowledge-based environment, boosting Jordan's ability to compete in the digital economy both regionally and globally.

7. REFERENCES

- [1] Department of Statistics, Jordan - www.dos.gov.jo
- [2] Ministry of Digital Economy and Entrepreneurship, Jordan - www.modee.gov.jo
- [3] Jordan's Economic Modernization Vision - <https://www.jordanvision.jo>
- [4] Eurostat - <https://ec.europa.eu/eurostat>
- [5] National Center for Health Statistics (NCHS), U.S. - <https://www.cdc.gov/nchs>
- [6] Data for Development Initiative, The World Bank - <https://datatopics.worldbank.org>
- [7] White, T. *Hadoop: The Definitive Guide*.
- [8] Thereska, E., & Bhatti, S. *Data as a Service: A Cloud Computing Perspective*.
- [9] The World Bank. *Big Data and Innovation: Assessing the Potential for Data-Driven Innovation in the Middle East*.

Large language models for COICOP classification in Norway's household budget survey

Boriska Toth (Statistics Norway)

1. INTRODUCTION

There is much excitement about the potential of large language models (LLM's) to revolutionize statistics production. In the past few years, NSI's have reported progress towards achieving far greater levels of automation with new LLM-based approaches like zero-shot/few-shot learning [1], [2], [3]. Norway is running its next household budget survey in 2026 and is exploring new LLM-based systems for the classification of purchased items to 5-digit COICOP codes (i.e. "apple juice 1L" → 01.2.1.0). As a smaller country, we are especially interested in contributing experiences regarding:

- adapting approaches to small languages
- the challenges of working with receipts data (i.e. abbreviated text and OCR noise)

This work is partly covered by two Eurostat grants: the *Innovative tools and sources for HBS 2026* (SMP-ESS-2023-HBS-IBA)¹, and the wide ranging *AIML4OS* project across the ESSnet system ("one-stop-shop for AI"). This is ongoing work, and below we describe the context of our work, our implementation approaches, and initial findings. Results of the full implementations will be shared at the conference presentation.

1.1. The household budget survey in Norway

The **most recent HBS 2022** featured several major innovations: 1) a modernized data collection wherein respondents could choose to scan receipts for processing with OCR, or enter purchases in a web app (about 90% of the time they chose to scan), and 2) machine learning paired with a custom-built web application for human-in-the-loop COICOP coding. Overall, classifying with machine learning was a success, and Statistics Norway experienced a significant savings in resources [4]. We found that 23,000 out of 60,000 unique items passing through the human-in-the-loop system could be classified automatically with quality matching human coding. Random forest using n-grams was found to work the best. Nevertheless, manual coding and editing remained a major effort.

HBS 2026. We expect to have a similar size survey and plan to continue data collection with both scanned receipts and a manual option in a web app. We are investigating a more sophisticated software for scanning receipts, with layout and text learned simultaneously, and will possibly use an interactive app where the respondent can see and mark suggested COICOP classifications and edits right away.

1.2. Large language models for classification

We focus on transformers (ChatGPT, BERT) that include an attention mechanism to dynamically encode words based on their context in very nuanced ways. We use the prevailing

¹ Due to a delay in when Norway can implement the next HBS, we applied for and were granted a delay in grant funds, with the grant period now starting in Jan 2025. As a first step to implementing an innovative new HBS in 2026, we are studying the performance of various LLM's on HBS 2022 data. Due to the delay, we haven't had the resources to complete these studies at the time of writing.

retrieval augmented generation (RAG) approach, which ensures that the chatbot is pointed to updated and reliable information through inserting into the prompt the most relevant snippets of information retrieved from a trusted database. This paradigm has proven very powerful in enabling new classification systems to sidestep the traditional need for manually labelled training sets. We believe that a RAG-based system in which a classification system can be updated seamlessly by simply replacing it in the knowledge base will be very advantageous for the upcoming transition to a new COICOP scheme.

Embedding models and language models are trained on overwhelmingly more data in English than in small languages, and there is often a need to adapt an existing LLM to work well with a particular language, domain, and task. However, in experimenting with prompting ChatGPT-4 for COICOP classification, we found outstanding modelling of Norwegian text, and furthermore, knowledge of items and brand names in the Norwegian market from its training on Norwegian websites. Thus we will insert prompts into ChatGPT in Norwegian to make use of the pretraining specific to the Norwegian context.

We implement a less massive, but nevertheless powerful, large language BERT model that also uses a transformer architecture. It generates representations of words that capture the context before and after each word in the input sequence. Adding a classifier layer has shown very good performance [2], [3].

Finally, we also plan to implement as a comparison the not-large language model of FastText to generate word embeddings in Norwegian for Random Forest classification. This approach has proven successful at other NSI's for working with the often abbreviated format of receipts data [5], [6].

2. METHODOLOGY AND IMPLEMENTATION

2.1. Generative predictive transformers (GPTs)

We chose OpenAI's ChatGPT-4 with RAG as our main classification tool, due to the massive pretrained knowledge and familiarity. We are implementing COICOP classification with the publicly available ClassifAI system developed by UK's Office of National Statistics (sic-soc-llm-datasciencecampus.github.io). ClassifAI is a pipeline offering a complete implementation of: making an embedding from a classification index using a HuggingFace model, generating a prompt to an LLM that injects the most likely candidate classification codes from RAG, interfacing with a ChatGPT or VertexAI LLM through an API, and finally parsing the output.

We eventually want to experiment with various ways of prompting. For example, we could include information about purchased items beyond the item name, such as the store. We can also use a richer classification index that includes examples or descriptions corresponding to COICOP codes and not just definitions. Additionally, for AIML4OS work package 12, we will be studying how to get information about ChatGPT's confidence in various predictions and calibrate a confidence measure. The analysis of confidence and quality will inform how much need there is for human-in-the-loop coding.

We try two basic approaches to prompting: RAG, and hierarchical prompting without RAG. In the latter approach, the chatbot is first shown all COICOP codes at a higher level, i.e. 2-digit and asked to choose. In each successive iteration, items are coded at a more refined level. Below we describe experiments where we used 3 hierarchical steps to classify the ~300 COICOP classes. The main advantage of hierarchical prompting is it avoids the computational demands of an embedding model.

An in-house GPT. Privacy and general data protection concerns have prohibited the use of external commercial LLM's like ChatGPT for statistics production at many NSI's. COICOP

classification involves a non-sensitive data source consisting of isolated purchased items. Nevertheless, in rare cases we have found scanned receipts to have sensitive personal information. Given the continuing debate surrounding the use of commercial LLM's, we feel it is important to study the performance of a downloadable GPT as soon as time and computing resources permit. We picked **NorskGPT-Llama-7b** from HuggingFace. This is a 7 billion parameter Llama2 model with supervised instruction fine-tuning in Norwegian. Computation is a challenge with the 14GB model size.

IT needs. Statistics Norway is in the process of strengthening IT infrastructure support for LLM work regarding GPU's, disk space, and network access to GPT API's. We will have the resources to test and report on results from various models before NTTS 2025. To facilitate collaborations with other countries in AIML4OS, we are exploring the use of Onyxia (<https://github.com/InseeFrLab/onyxia>) as an open access cloud environment for modern data science.

2.2. BERT

We choose NorBERT, a large-scale BERT model trained in Norwegian in 2023 <https://github.com/ltgoslo/NorBERT?tab=readme-ov-file>. Given the extensive training on web data, the model should capture knowledge about Norwegian brand names and products. Nevertheless, we wish to fine tune to HBS data, so we can use the 40,000 manually labelled items from HBS2022 both for language model fine tuning and for training the classification layer. Since NorBERT is available as a Hugging Face model, we use off-the-shelf functionality for these tasks.

2.3. FastText

We also plan to implement a much more lightweight language model, FastText, to make word embeddings that we can then feed into a Random Forest classifier. FastText has been found to be very effective for COICOP classification on receipts data [4], [5], which contains many abbreviated words and different forms for writing the same word, due to FastText's ability to work with subword ngram units. We use a Norwegian FastText model, Norsk aviskorpus/NoWAC, which was trained on newspaper and web data.

3. INITIAL RESULTS

We performed initial experiments on COICOP classification with ChatGPT-4 using the hierarchical approach. We validate performance using items from HBS2022 that were manually coded in our human-in-the-loop process. We used a Norwegian version of the prompt shown below and a 3-step hierarchal process. In the example in the figure, the chatbot is given a list of items with 3-digit code 01.1 and asked to classify at the 4-digit level. It is also given the definition of each possible COICOP category to choose from.

✎ You are a classification assistant for respondent data for use by Statistisk sentralbyrå. Below you are given a list of item descriptions, and a list of possible COICOP classifications. Your task is to output a COICOP code corresponding to each item. Here is an example of the output you should generate for each item description in the correct format:

===Output format===
Item: Coca Cola 1l
COICOP: 01.2

===List of item descriptions===
'barilla spagetti 500g', 'avocado', 'Starbucks coffee', ...

===List of possible COICOP classifications===
cereals; 01.1.1, live animals; 01.1.2, fish and seafood; 01.1.3,

Figure 1. Chat-GPT prompt with hierarchical classification.

We selected 200 each of scanned/OCR items and manually entered items. We found very strong performance above 90%. Next, we selected 200 items whose text had been automatically or manually edited in our processing pipeline. We ran both the pre-editing and post-editing version of the text through ChatGPT-4 and found mediocre performance before editing but very strong performance of 94% post editing. These results are extremely encouraging especially in light of the fact that we used a quick-and-dirty implementation with very basic prompting. Only the category definitions (i.e. “fish and seafood”) were shown next to the category labels. Adding example items, or using RAG to suggest candidate categories, can significantly improve these results. Further experiments are called for due to the limited number of items checked.

Table 1. Percent of items assigned a correct 5-digit code by ChatGPT-4 in the 3-step hierarchical procedure. 200 items were checked in each case.

OCR	90%		Pre-edit	59%
Manual	94%		Post-edit	94%

4. SUMMARY

We described the work at Statistics Norway in transitioning from a classic (supervised) machine learning, human-in-the-loop system for COICOP classification to one based on large language models. In our presentation we will go over results from trying 3 implementation approaches (chat-GPT, BERT, and FastText). These different approaches represent tradeoffs between classification performance vs ease of implementation (especially regarding computation). We have begun to implement these approaches and evolving infrastructure support will facilitate further work. In preliminary experiments using a hierarchical approach to classification, we found chat-GPT to have excellent support for the Norwegian language and the domain of purchased items. Overall, across NSI’s early applications of LLM’s show great promise in vastly reducing the burden of manual labelling. Further work on obtaining quality or confidence measures in addition to generated codes will support the use of LLM’s in statistics production.

5. REFERENCES

- [1] R. Davies and F. McEvoy, Faster, bigger, cheaper: how AI can improve UK price data. Economics Observatory Working Paper, 16th May 2024.
- [2] G. Hess, Use of a large language model to derive the economic sector of businesses from unstructured text on economic activities. Conference on Foundations and Advances of Machine Learning in Official Statistics, Wiesbaden, 3rd-5th April 2024.
- [3] L. Fiedler, B. Hofmann, K. Loogman, and T. Scherl, Domain adaptation of a BERT Model for analyzing job advertisements at the German Federal Employment Agency. Destatis Conference on Foundations and Advances of Machine Learning in Official Statistics, 3rd-5th April 2024.
- [4] B. Toth, R. Mustad, and S. Jentoft, Machine learning methods to support a modernized household budget survey. New techniques and technologies for statistics conference, Brussels, 7-10 March 2003.
- [5] H. Martindale, E. Rowland, T. Flower, and G. Clews, Semi-supervised machine learning with word embedding for classification in price statistics, *Data & Policy* (2020), 2:e12. doi:10.1017/dap.2020.13.
- [6] A. Montbroussous and M. Monziols, Classification of scanner data into COICOP: a machine learning approach. Ongoing work. Paper for meeting hosted by Statistics Canada and Bank of Canada, Ottawa, 13-15 May 2024.

Measurement effects in the French Time Use test survey

Loreline Court & Simon Quantin (Insee, Direction de la Méthodologie et de la Coordination Statistique et Internationale)

1. INTRODUCTION

1.1. Time use survey and Time use test survey

The harmonised European time use survey (HETUS) is a household survey collecting the data needed for quantified analyses of daily life. The aim is to collect precise information on how individuals use their time on each activity in a day. In previous surveys, respondents filled in their timetable in a paper diary. As part of INSEE's program to develop multimode household surveys, the next edition of the time use survey in 2026 will adopt a competitive multimode protocol. Respondents will have to choose between filling in their diary online or on paper. This change of protocol meets a number of challenges: to limit the drop in the non-response rate, particularly among the youngest respondents; to control the response burden on respondents (by allowing the diary to be filled in *via* smartphones); to control survey costs, by mobilizing interviewers to the respondents for whom their added value is greatest.

A mode test, purely methodological, was carried out in 2023: its aim was to verify the quality of the data in the digital diary, and to measure the comparability of its data with that of the paper-based collection method. In order to detect any possible fashion effect on logbook completion, the 2023 test protocol adopts a crossover design approach: the same individuals fill in both a day on the paper diary and a day on the digital diary one week apart. Households are randomly assigned to a sub-sample defining the order in which they complete the test (digital logbook followed by paper, or vice versa). The test sample is made up of 2,100 dwellings, selected by a systematic sampling stratified by commune group.

This protocol is a AB/BA cross-over trial. The main advantage is that each individual is its own control (within-subjects comparison). The estimators are more efficient, for the same sample size. However, this protocol is suitable under strong hypothesis. In particular, we must consider that the responses given in the second survey reflect neither a persistent impact of the first collection mode (carry-over effect), nor an exogenous effect of the survey linked to the period (period effect). Usually, such effects are neutralized before the cross-test, by allowing a long enough "washout" period between the two surveys (7 days in our test). In addition, individuals may end up not taking part in the test, or stopping after only one diary: taking non-participation into account greatly complicates the analysis of AB/BA cross-tests.

In this article, we will focus on the effect of Internet mode on certain durations of interest and ask:

- How can it be estimated from this protocol?

- If the effect is significant, can it be explained and corrected?

2. METHOD

The model implemented is one often used in the analysis of crossover protocols and by doing so, we can estimate the effect of responding online on several durations of interest. For each individual i , period j , and mode k , it can be written as follows:

$$Y_{ij} = \mu + \pi_j + \tau_k + (\tau \pi)_{kj} + \xi_i + \epsilon_{ij}$$

However, this model is based on strong assumption of no correlation between non-participation and our variables of interest, no period effect and no carry-over effect. The relevance of the results needs to be discussed by conducting a between subjects analysis.

To do so, we will consider only the responses to each respondent's first diary, after matching: this is equivalent to comparing the responses given by different respondents, surveyed with two modes (paper and internet). This eliminates some possible effects of the crossover trial protocol (period effect, carry-over effect and period-by-mode interaction effect). By comparing the responses of individuals perceived to be similar on certain observed characteristics, we might have possible bias due to an unobserved characteristic. Indeed, despite random assignation of surveyed individuals to sequence/mode, respondents may still differ due to an endogenous selection bias, which implies a possible composition bias after matching.

In both analysis, endogenous selection (or missing not at random (MNAR) issue) remains. The sensitivity analysis model proposed by Rosenbaum ([1], [2], [3]) and implemented in this study provides a partial response to this possibility. It involves estimating the magnitude of the unmeasured characteristic required to modify the conclusions about the existence of a measurement effect, assuming the absence of hidden bias.

3. RESULTS

Regardless of the method implemented, responding *via* the Internet results in under-reporting of time spent on leisure, eating, sleeping, working and travelling. On the other hand, the effect is not significant for declarations of time devoted to domestic chores or "physiological and personal" time. The effect of the internet mode is particularly strong for leisure and sleep, since on average, respondents declare almost one hour less on this mode than when filling in a paper diary.

Table 1. Estimates of the effect of internet mode on different durations

	Cross-over trial analysis	Between subject analysis
Leisure	-51 [-63;-38]	-45 [-67;-24]
Household chores	-6 [-16;-3]	6 [-11;23]
Eating	-15 [-22;-9]	-16 [-28;-4]
Sleeping	-56 [-72;-40]	-53 [-79;-26]
Working and/or studying	-8 [-35;-5]	-22 [-53;0]
Travelling	-20 [-27;-14]	-32 [-43;-20]
Personal and physiological time	-1 [-21;18]	-5 [-36;25]

For each variable where the mode has a significant effect, we try to provide a credible explanation.

- Under-reporting of leisure time online would be due to the questionnaire's design. Leisure activities are at the bottom of the screen, and some of the subcategories are hard to find leading to respondents not always using the right categories to enter certain leisure activities (such as watching TV, which is often declared in "personal time" rather than "leisure"). By re-imputing them, the negative effect is rectified.
- Under-reporting of sleeping time online could be explained by respondents online behaviour: the diaries are to start at 4am and finish at 4am the following morning, but some online respondents no longer declare any activities between midnight and 4am, leading to a drop in time period covered and probably estimated sleeping time.
- The lower time spent on meals when declaring activities online is explained by the lower number of meals declared. It may be due to either a difference in behaviour regarding the declaration of short periods (few online respondents report short eating periods such as coffee breaks) and/or problems to code "eating" activities on paper that artificially increase the number of meals.
- Under-reporting of travel time online is the hardest effect to explain. It might be a codification problem from the paper form so this should be seen as an over-reporting of travel time on paper. Activities such as "going to the restaurant" or "go shopping" were more likely to be wrongly coded as travel time.

4. CONCLUSIONS

We benefit from the crossover design of the test to implement two analyses to question the existence of a measurement effect on declared time spent on different activities. Both have different advantages and drawbacks and give coherent results. However, none addresses the issue of endogenous selection, leading us to consider a sensitivity analysis. This approach considers the impact of the possibility of endogenous selection on the conclusions of the between analysis. Our main conclusions:

- It is difficult to reject the hypothesis of a measurement effect on the duration of certain activities, even assuming an endogenous selection bias.

- The measurement effect often only impacts the durations of a small number of respondents.
- Depending on the activity, the reasons for over- or under-reporting may be: an effect of the website's ergonomics, a problem completing the diary, a change in behaviour or a problem with the coding of paper activities.

Regarding to the Time use survey, changes are planned to the design of the site and work is underway on the various activity codes. The coding of paper activities will also be improved for the 2026 survey. Finally, before the final survey, a new full-scale test will be carried out, taking these modifications into account.

5. REFERENCES

- [1] P. R. Rosenbaum, *Observational Studies*, Springer series in Statistics (2002).
- [2] P. R. Rosenbaum, *Design of observational studies*, Springer series in Statistics (2010)
- [3] P. R. Rosenbaum and J. H. Silber, Amplification of Sensitivity Analysis in Matched Observational Studies, *Journal of the American Statistical Association* 104 (488):1398–1405 (2009)

Measuring data science skill levels in an NSI using text mining of production code

Susie Jentoft (Statistics Norway)

1. INTRODUCTION

Like many NSIs in Europe, Statistics Norway is in the process of transitioning towards using more open-source software and IT solutions. A new data platform (Dapla) within the organization has seen a change to separating data and code to separate tools and environments; these have traditionally been stored within the same storage area. Production code for producing official statistics is now versioned with git and stored on GitHub while data is store in a cloud-based storage system. This transition is a significant change for employees and requires new data science skills and changes in workflow.

Data science competencies including programming and IT skills, have previously been measured through simple surveys, targeted towards division leaders. While this provides a reasonable insight, there are some disadvantages to this approach including mismatches between leaders understanding of skill-levels, and time/resource requirements for implementing questionnaires in n NSI. Using new technologies like GitHub for code storage allows alternate ways to measure competency levels.

In this study, we investigate an alternate approach to measuring competency using GitHub's API to retrieve production code and basic text mining tools to assess the repositories. We looked at all production repositories within the organization and establish indicators for competency and compliance with established standards. These fast, progress reports may be a useful tool measuring progress towards our ambitious transition goals.

2. METHODOLOGY

Measurable indicators were first described in words based on the four areas of competency from Statistic Norway's learning plan for transitioning to a new data-platform. The areas include: 1) understanding the Dapla Lab framework and documentation, 2) programming in Python/R, 3) interacting with data, and 4) using Git and GitHub. Measurable indicators were all metrics that could be extracted from information on GitHub. This allows quick reports that can be regularly extracted to easily show developments and trends over time and identify specific areas for additional support.

The metrics were divided into three categories: general, code and activity metrics. General metrics include organization level information such as the number of production repositories on the organizations GitHub. Code-based metrics included those based on what has been written and included in the production programs on GitHub. For example, if code used specifically recommended packages or if the repository contained a README file. Activity metrics include those that describe activity levels of the repositories and code, for example how many users have made changes/commits to a repository. Table 1 provides some examples of learning goals, and metrics used to measure them.

Programs to collect the data using GitHub's API were written in R. R-packages *httr* and *jsonlite* were used and reports were generated using *Quarto*. Reports were published internally within the organization using a website set up on GitHub pages.

Table 1. Examples of a selection of learning outcomes and their metrics

Learning outcome	Learning area	Metric type	Metric
Understand Markdown syntax and are able to write code documentation in this format.	Dapla Lab and documentation	Code	Percent of production repositories with at least one README (not in root folder).
		Activity	Number of unique GitHub users that have made at least one commit to a README file in the last 2 years.
Able to use git commands: <i>add</i> , <i>commit</i> , <i>push</i> and <i>pull</i>	Git and GitHub	Activity	Number of unique GitHub users that have made a commit to at least one production repository in the last 2 years.
Able to fetch classifications using Statistics Norway API	Programming	General	Number of production repositories that use recommended packages for fetching classifications: <i>klass</i> (python) or <i>klassR</i> (R).
Read and write data to storage buckets (Google cloud storage)	Data interaction	Activity	Number of unique users with commits to code that reads and/or writes data to buckets.

3. RESULTS AND PRACTICAL APPLICATION

Programs for data retrieval and reports were established and run in 2024. General findings from the first report indicate that of the 119 production repositories, only 60 percent were of a size indicating that they had been worked on beyond the initial repository setup (71 repositories). Around 36 percent were also setup to run on an older IT-platform, indicated only initial work has started. Statistics Norway has a goal to move the production of 120 statistics to the new data platform by the end of 2025. We measured 71 active repositories, indicating good progress towards this goal, however, in some cases, one repository may represent several statistics, while in others, several repositories may contribute to only one statistic. This is therefore not a direct measure of how many statistics are in the process of establishing their code in R/python on the new platform.

The first report indicated 120 unique users had contributed to production repositories in the last 2 years and 75 in the last 2 months. This may indicate that there is a high turnover of staff with technical git and programming skills, however, it can also simply indicate the nature of working on statistics that are not produced and working on continuously (for example statistics that are produced yearly). Further work on fine-tuning the indicators and metrics to better reflect the learning outcomes is needed.

A recommendation within Statistics Norway is to use virtual environments for running production code using poetry (for python) and renv (for R). Our initial report showed 71 percent of repositories had establish poetry environments, whereby only 7 percent had established renv environments. This could indicate that more training is required for supporting

statisticians in setting up (particularly R) environments. Further work on narrowing the metrics to determine if python, R or both are used within the repository, would provide a more accurate measure for determining if this is an area needing attention.

The average number of unique users making commits to a repository (within the last 2 months) was found to be 0.96. This may indicate risks of knowledge bottlenecks, and that only a few people are individually managing several repositories alone. If these individuals with programming and git skills become unavailable due to illness, resignation, or other reasons, the statistical production process may be vulnerable. However, using a mean metric in this case is not easy to interpret and further metrics should be established to understand vulnerability further.

4. MAIN FINDINGS

The move to a new data platform where code is versioned with git and stored on GitHub allows new possibilities for monitoring. This study has shown that API-based retrieval of repositories and code with simple text mining is reasonable easy to set up and provides fast indicators, using few resources to gain a picture of progress. Repeating the process allows quick updates and can be a valuable tool to gauge progress towards an organisation's transition goal. General information on GitHub activity levels may be useful to measure technical skill levels through time, and also competency vulnerabilities. Metrics can be used to target further areas for training courses and workshops. Assessing how well learning activities translate into improvements in production code, may also be possible through these indicators. However, indicators need to be carefully considered and adjusted to best reflect the learning outcomes to be measured.

Methods for integrating survey data and big non-survey data

Elena Viti (EMOS Programme, University of Pisa & University in Trier)

1. INTRODUCTION

This thesis addresses the critical challenge of integrating data from diverse sources, focusing on the combination of probability samples and big data. Data integration offers significant advantages, including cost reduction through the supplementation of existing surveys and enhanced accuracy of estimates by mitigating the limitations of individual data sources. Crucially, combining information from multiple surveys can be beneficial for addressing both sampling and non-sampling errors. By strategically using one survey to compensate for information lacking in another, we can improve the precision and reliability of estimates. This is particularly relevant when dealing with finite populations where single surveys may not provide sufficient data to accurately measure certain phenomena. The choice of appropriate integration methods, however, depends heavily on the type of sample considered. Probability samples, with known selection probabilities, allow for rigorous design-based inference, while big data, frequently a non-probability sample, poses distinct challenges concerning representativeness and data quality. This research systematically explores these scenarios: the integration of multiple probability samples, and the integration of probability samples with big data, examining how methodological choices must adapt to these varying data characteristics.

2. OBJECTIVE

The research aims to: 1) Compare existing integration techniques for probability samples using macro and micro approaches; 2) Develop a framework for integrating probability samples with big data, adapting existing calibration methods to handle challenges like measurement error and duplicate data; 3) Evaluate the performance of these approaches through simulation studies using synthetic data, investigating the impact of variable selection and data characteristics on estimation accuracy.

3. METHODS

The study systematically compares two established approaches for probability sample integration: a macro approach, which aggregates summary statistics from multiple surveys, and a micro approach, which creates a single synthetic dataset by combining individual-level data. For the macro approach, the generalized regression estimator (GREG) is employed, exploring both proportional and optimal weighting strategies September 2024based on sample sizes and the inclusion of control variables. The micro approach utilizes multiple imputation techniques to address missing data within the probability samples. The integration of probability samples with big data is addressed using a framework that adapts existing calibration weighting techniques. This framework does not introduce a new calibration method but rather presents a novel application of existing methods to the specific challenges posed by combining probability samples with big data. The framework addresses non-probability

sampling by treating the big data as a separate population. The calibration process is modified to accommodate the frequent presence of measurement errors and duplicate entries in big datasets. A regression-based calibration method was implemented, adjusting weights to align sample totals of auxiliary variables with known population totals where possible. The methodology uses the information in big data to improve estimates from the probability sample by creating a new variable (δ_i) which indicates whether a unit i from sample A is also present in big data B. This allows calibration to improve the estimation from probability sample A by incorporating data from big data B, which may contain information not present in sample A. The AMELIA synthetic dataset is used to generate a series of simulations evaluating the performance of these integration approaches across varied conditions, such as differing sample sizes, data quality (presence of duplicates and measurement errors), and the number and type of auxiliary variables employed in the calibration. The simulations also explore scenarios involving monotone and non-monotone missing data patterns in the probability samples.

4. RESULTS

The simulation results demonstrate the relative strengths and weaknesses of the different integration methods. For the integration of multiple probability samples, the macro approach, employing the GREG estimator, proved most efficient when common variables exhibited strong correlations with the target variables. The micro approach, while effective in creating a unified dataset, showed sensitivity to inconsistencies in data quality across the different probability samples. Regarding the integration of probability samples with big data, the results highlight the significant influence of data quality (presence of duplicates and measurement errors) and the choice of auxiliary variables on estimation accuracy. The framework employing calibration weighting generally improved estimates compared to relying solely on the probability sample, particularly when known population totals for auxiliary variables were available. However, scenarios with a high proportion of duplicates or significant measurement error within the big data yielded less precise estimates.

5. CONTRIBUTION

The use of big data sources in estimation is nowadays a relevant topic also in the content of official statistics. However, from a methodological point of view, it is important to evaluate the impact of the use of big data in the estimation process. In the thesis the focus was on evaluating the impact of the use of big data in context of integrating data from diverse sources. The thesis results highlight that the effectiveness of the integration framework was shown to be highly dependent on the extent to which the auxiliary variables in the big data captured variation in the target variable not already explained by the probability sample. The simulations reveal a September 2024 trade-off: including more auxiliary variables improves accuracy when known population totals are available but can lead to instability if correlations become too strong or if totals are unknown. The optimal choice of auxiliary variables and the robustness of the integration method is heavily influenced by the quality and characteristics of the big data itself. Therefore, the thesis results suggest that the including big data in the estimation process should be carefully evaluated.

6. REFERENCES

- [1] Yang, Shu and Jae Kwang Kim (2020). "Statistical data integration in survey sampling: A review". In: Japanese Journal of Statistics and Data Science 3.2, pp. 625–650.

- [2] Tam, Siu-Ming and Frederic Clarke (2015). "Big data, official statistics and some initiatives by the Australian Bureau of Statistics". In: *International Statistical Review* 83.3, pp. 436–448.
- [3] Yang, Shu and Peng Ding (2019). "Combining multiple observational data sources to estimate causal effects". In: *Journal of the American Statistical Association*.
- [4] Kim, Jae-Kwang and Siu-Ming Tam (2021). "Data integration by combining big data and survey sample data for finite population inference". In: *International Statistical Review* 89.2, pp. 382–401

Micro-level bank card data for the use of official statistics

Natasha Mifsud & Brandon Sacco (Structural Business Statistics Unit, National Statistics Office, Malta)

1. INTRODUCTION

The rapid growth of digital technologies, especially in electronic payments, presents new opportunities for measuring digitalisation. Various studies suggest that electronic payments are increasing in popularity and dominance driven by the growth of Fintech (Khando, Islam, & Gao, 2023). Access to bank card data offers another innovative perception of economic activity with minimal time lag. This non-traditional data source can significantly enhance economic measurement by providing real-time, detailed insights and reduce administrative burden on businesses and consumers. The full benefits of this new source of data depend greatly on its level of granularity. However, countries across Europe exhibit varying degrees of detail, which limits its adoption (Eurostat, 2024).

Traditional sources, such as surveys and administrative records, are still heavily relied upon within statistical offices (Yung, 2021). By obtaining this alternative data source, the limitations of traditional data sources can be reduced. This paper outlines how the National Statistics Office of Malta (NSO) is currently exploring the utilisation of granular bank card data, and details how the office is using this alternative data in experimental statistics with a vision of implementing it into regular statistical production. This data is provided with a time lag of less than twenty days after end of month from all licensed card acquirers and it is provided by the Central Bank of Malta.

2. METHODOLOGY

Since gaining access to micro-level bank card data in 2023, with coverage backdated to 2022, standardisation and quality assurance have been prioritised to ensure accuracy, consistency, and timely outputs. To streamline this process, the Structural Business Statistics Unit within the NSO developed an array of Python scripts. This approach ensures that the data is robust and consistently available for experimental statistics.

2.1. The impact of bank card data in identifying local units

By having access to bank card data at the point-of-sale (POS) terminal level of detail, the statistical office is able to estimate the number of branches or locations from which merchants operate, allowing for the mapping of towns and villages where merchants predominantly maintain brick-and-mortar establishments. This provides valuable insights into the geographic distribution of commercial activities.

For countries in which no administrative register of local units¹ for active enterprises exist, supplementary surveys are used to gather information on business branches. These surveys may lack full coverage of the statistical business population, face high non-response rates, and rely on businesses to provide accurate information. Micro-level bank card data may help

¹ Local unit is defined as an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place (Eurostat Glossary).

bridge the gap in resources and reaffirm information obtained from traditional sources. Using such proxy, several previously unidentified local units may be discovered, increasing the overall population of identified local units and enhancing regional accounts. Turnover values can be identified and allocated by locality, serving as a proxy for distributing other key variables such as Gross Value Added.

2.2. Integrating bank card data in short-term business statistics

Granular bank card data offers potential beyond identifying local units, particularly for Short-Term Business Statistics (STBS). The trade industry (NACE G), the largest industry domain in the EU (Eurostat, 2024), provides an ideal scenario for exploring this innovative source, particularly within the retail industry (NACE 47). Being client-facing by nature, retailers adopt digital technologies to allow customers the flexibility of purchasing by card or cash. For such companies, records of all online and POS monthly turnover transactions are available within this data source. This allows for a more comprehensive tracking of business activity across both physical and digital channels.

Before integrating bank card data into a selected number of current statistical practices, it is deemed necessary to juxtapose the 2022/2023 bank card with STBS indices. The sample data for the short-term retail turnover indicator, which is collected through questionnaires, was reproduced using the bank card data for this analysis. Merchants missing from the bank card data or with over 20% missing records were excluded for fair comparison. The absence of bank card records for some merchants may be attributed to several factors, including a lack of terminals, no online presence or terminals administered by acquirers not recognised by the Central Bank of Malta. Missing monthly records were populated using the Donor Imputation Method with temporal adjustments based on the monthly growth rates.

Following the data cleaning process, indices were derived for the STBS sample, the corresponding bank card sample, and the complete bank card population. These indices were plotted from January 2022 to July 2023 to visually analyse the proximity of their directional trends. Hypothesis testing was conducted using the paired t-test with the null hypothesis stated as follows: there is no statistically significant difference between the true means of the official re-referenced STBS index and the corresponding bank card data index over the period from January 2022 to December 2022.

To effectively integrate micro-level bank card data into traditional data collection techniques in STBS, it is crucial to also ensure that bank card terminal technology exhibits a substantial penetration rate in the market. This is essential to avoid bias and ensure the reliability of the data collected.

2.3. Penetration rate

Following the publication of the 2022 Structural Business Statistics (SBS) in July 2024, the 2022 bank card data was used to derive an essential metric in assessing market activity, known as the Penetration Rate. The penetration rate may be defined as the proportion of legal units that are linked with the bank card data relative to the total number of merchants within the SBS population. The SBS population serves as a benchmark for the measurement of such metrics due to its comprehensive and official record of all market activity, reflecting the tangible economic footprint through active businesses within the Maltese economy. This metric will allow for a conclusion to be made whether such data is suited as a proxy for market activity. If such technology exists and has a high presence in the market, it is necessary to perform adequacy testing by assessing the weighted average bank card usage rate of all merchants in a particular industry. If it is seen that other methods of payment such as cash prevail, the utility of this data may diminish.

2.4. Weighted average bank card usage rate

Following the identification of market adoption of bank card technology among merchants, the bank card usage rate by consumers was examined. The metric known as the Weighted Average Bank Card Usage Rate is derived to better understand actual sales proxies from bank card transactions. This metric represents an adjusted average usage rate of bank card payments across different merchants, where each merchant's impact on the average is proportional to its size (weighted by its turnover). For equitable comparison, only merchants present in each month of the 2022 bank card data were included in the statistical analysis.

3. MAIN FINDINGS

One of the key derivations of bank card data is its ability to serve as a source for identifying local units for regional accounts. In 2023, the population of local units increased by approximately 11% as a result of utilising this data source.

From the identified local units, a bar chart (figure 2) was drawn to illustrate the additional local units by locality. This chart highlights the local units identified through bank card data, which were not captured by traditional data sources.

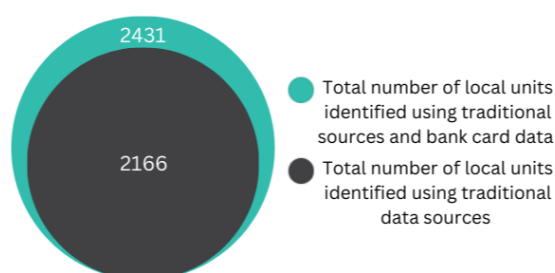


Figure 1. Local units population



Figure 2. Local units identified via bank card data

Following this, the proximity of the directional trends between the STBS and bank card data

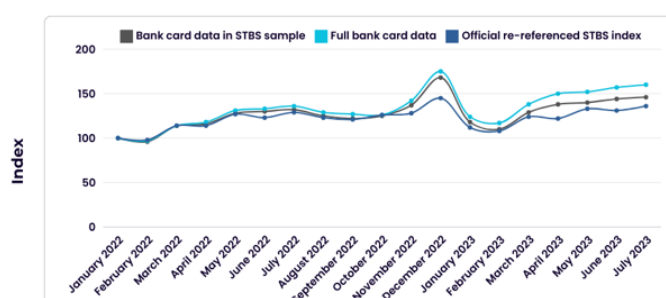


Figure 3. Line chart of indices

was examined. The strong alignment between the bank card data in the STBS sample and the full bank card dataset is a testament to the sample's representativeness of the entire market. The assumptions of the paired t-test were evaluated to assess the validity of results. The normality of the differences between the official re-referenced STBS index and the corresponding bank card

index was assessed using the Shapiro-Wilk test, which indicated that the differences were normally distributed (p-value 0.1107). Observations were paired as the indices were derived using the same merchants. The indices can be assumed independent based on the nature of the data sources. Therefore, the paired t-test was determined appropriate. It resulted in a t-statistic of 1.97 with p-value 0.077. This suggests that there is no statistically significant difference between the true means of the indices. Thus, observed differences do not appear to be systematic or attributable to any underlying cause, concluding that the market for NACE 47 is well represented by the bank card data.

However, it remains necessary to verify whether the market has truly been effectively penetrated by electronic payment solutions. By concentrating on client-facing markets, the penetration rate was

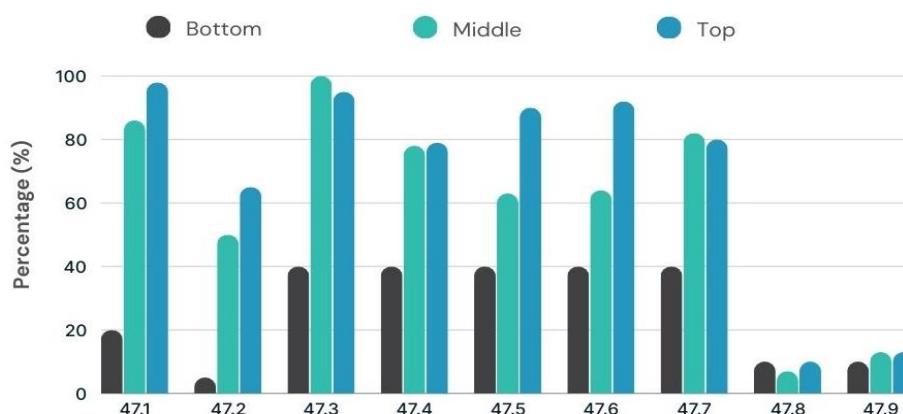


Figure 4. Penetration rate by merchant classification

calculated for each NACE group within NACE 47 to study the number of merchants providing locals the flexibility of payments. Merchants were classified by utilising turnover data from the SBS population. The cumulative percentile method was applied to segment the merchants into bottom, middle, and top-tier classes (top-tier being from the 20th to the 100th percentile). Figure 4 demonstrates a clear trend; merchants with a higher cumulative percentile based on turnover are significantly more likely to offer bank card payment options to their customers.

Nevertheless, accurately identifying the specific usage patterns within each retail market is crucial to determine the true extent of bank card utilisation. Top-tier merchants were prioritised due to their performance stability and reduced likelihood of underreporting economic activities, thus projecting a more accurate bank card usage rate (Heinemann & Stiller, 2024). In the top-tier class, large supermarket chains have a clear tendency for higher rates of bank card usage. For instance, the average bank card usage rate across the largest five supermarket chains in Malta reached 69.6%. Some markets may exhibit lower bank card usage rates due to the nature of the goods sold. Industries specialising in high-value or bulky items, such as motorboats, often show reduced rates of card payments. This is likely influenced by the cultural norm of completing such transactions via cheques or bank transfers.



Figure 5. Bank card usage rate

Figure 6 depicts the bank card usage rate distribution. It aims to illustrate the variation in bank card usage across different top-tier merchants. Each point in the boxplot represents a top-tier merchant classified under the relevant NACE group. The boxplot for NACE 47.3 (Retail sale of automotive fuel in specialised stores) reveals a notably low bank card usage rate compared to other markets. This is largely due to the cultural preference in Malta for using cash when fuelling vehicles.

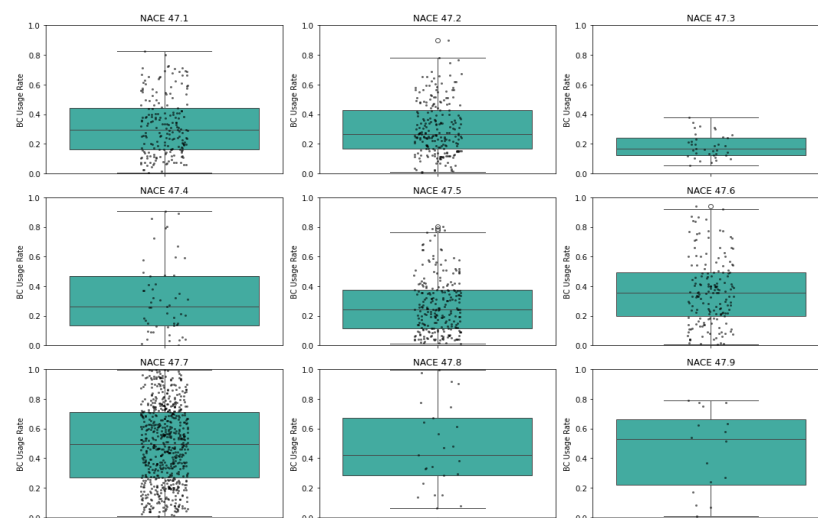


Figure 6. Boxplot illustrating the distribution of top-tier merchant bank card

4. CONCLUSION

The integration of granular bank card data through micro-data linking offers statistical offices timely insights; reducing reliance on traditional, costly surveys. This approach modernises statistical practices by potentially enhancing precision and introducing improved quality control mechanisms, especially in high-volume sectors like retail, while supporting the evolving monitoring needs of the digital economy.

ACKNOWLEDGEMENTS

The authors are thankful to the Central Bank of Malta, particularly Victoria Briffa, for their support and provision of timely data. They would like to extend their thanks to Sam Sacco and Kimberlin Lia from the Short-Term Business Statistics Unit, and Christianne Micallef and Oriana Parisi from the Business Register, Research and Innovation Unit, for their valuable contributions.

5. REFERENCES

- [1] Eurostat. (2024). *Task force of the production of statistics based on bank card data*. European Commission.
- [2] Heinemann, M., & Stiller, W. (2024). Digitalization and cross-border tax fraud: evidence from e-invoicing in Italy. *International Tax and Public Finance*, 1-43.
- [3] Khando, K., Islam, M. S., & Gao, S. (2023). The emerging technologies of digital payments and associated challenges: A systematic literature review. *Future Internet*, 15(1), 21. <https://doi.org/10.3390/fi15010021>
- [4] Yung, W. (2021). The evolution of official statistics in a changing world. *Harvard Data Science Review*, 3(4). <https://doi.org/10.1162/99608f92.48466abe>

MILC vs. Tree-MILC: a simulation study comparing correction methods for multi-source statistics

Danielle Remmerswaal (Utrecht University & Statistics Netherlands)

Laura Boeschoten (Utrecht University) & Ton de Waal, CBS)

1. INTRODUCTION

National Statistical Institutes (NSIs) use surveys and administrative registers to produce statistics. Both types of data sources are not error-free. Numerous methods exist to correct statistics for errors in single-source data. Linking variables from multiple sources, which is possible on the unit-level by NSIs, together in a combined dataset creates many opportunities such as the application of latent variable modelling (LVM) to correct for measurement error. The Multiple Imputation of Latent classes (MILC) method is a specific use of LVM in combination with multiple imputation (MI) to correct measurement errors and produce consistent estimates. Estimates often suffer from multiple error sources simultaneously. However, it has not been investigated if and how MILC could simultaneously handle multiple error sources. In this paper we propose an alternative approach where both selection and measurement error are estimated and corrected by a separate LC and MI step in a tree-like structure: tree-MILC. We compare tree-MILC with an approach where both selection and measurement error are estimated and corrected for with a single LC and MI step (regular MILC). The performance of MILC and tree-MILC are compared in a simulation study, where the level of selection and measurement errors are varied. Overall, tree-MILC outperforms MILC at estimating statistics for combined datasets in terms of accuracy measures and confidence interval measures. The differences in performance are bigger for datasets with higher levels of errors.

2. METHODOLOGY

2.1. MILC

The MILC method, developed by [1] is used to identify and correct classification errors in combined datasets. MILC is a combination of LC analysis (see e.g. [2]) and multiple imputation (MI) (see [3]). An important requirement is that the variables in a unit-linked combined dataset measure the same concept.

2.2. Tree-MILC

We developed an adaptation of the MILC method, tree-MILC, as a potential solution for situations with multiple sources of errors. Tree-MILC is inspired by latent class tree (LCT) analysis as developed by [4] and uses the concept of applying LC models sequentially on subsets. Tree-MILC has for the most part the same procedure as MILC, but has an extra LC model and MI step for a subset of the data, based on the outcome of the previous LC model, before the final estimates are calculated.

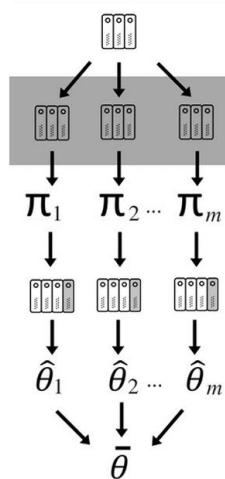


Figure 1. Graphical overview of the MILC method:

At step 1 M bootstrap samples are taken from a dataset containing multiple indicator variables. At step 2 a LC model is applied to each bootstrap sample and probabilities (denoted with π) are obtained. At step 3 M imputations (denoted with the extra vertical bar) for the latent variable are created and placed next to the dataset. At step 4 the estimates of interest (denoted with $\hat{\theta}$) are calculated from the imputations. At step 5 the estimates are pooled (denoted with $\bar{\theta}$).

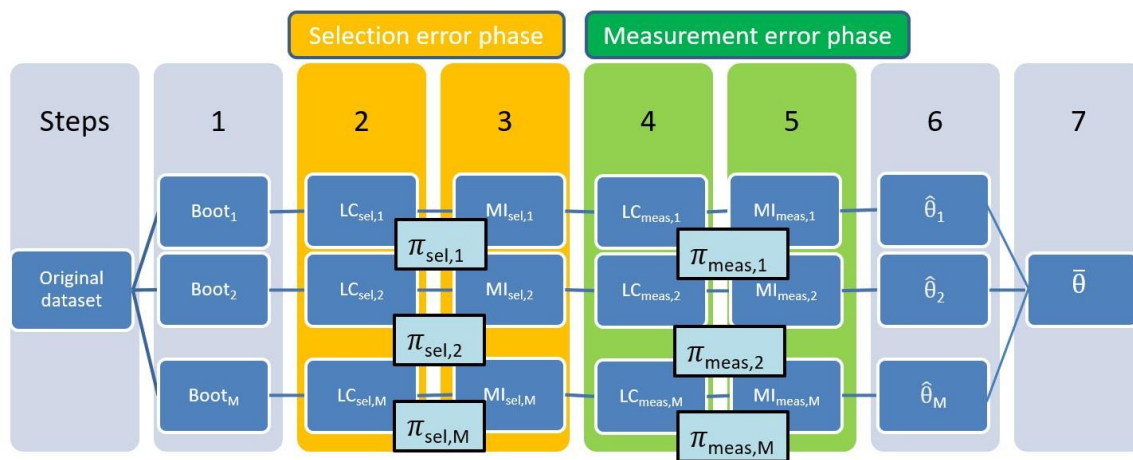


Figure 2. Schematic overview of steps of tree-MILC:

At step 1, M bootstrap samples are taken from a dataset containing multiple indicator variables. At step 2, a LC model is applied to each bootstrap sample and posterior probabilities (denoted by π_{sel}) are obtained. With the imputations calculated in step 3, subsets are created to which the LC model in step 4 is applied. At step 5 imputations are calculated for the subset using the obtained posterior probabilities (denoted by π_{meas}). The imputations of the subset replace the imputations of the entire dataset from step 3 where appropriate. At step 6 estimates of interest (denoted with $\hat{\theta}$) are calculated for the dataset with the combined imputations of both phases. At step 7 the estimates are pooled (denoted with $\bar{\theta}$).

2.3. Simulation study

We generated 1000 populations of 5000 units and applied both methods (MILC and tree-MILC) to each of the generated populations. All populations contained four variables: three polytomous indicator variables (Y), each with four categories, measuring the latent variable (X) with measurement error, and one polytomous covariate (Z), again with four categories. We used the R package poLCA [5] to generate the data.

Table 1. Population values latent class sizes

Class	$P(Y_i = 1)$	$P(Y_i = 2)$	$P(Y_i = 3)$	$P(Y_i = 4)$
proportion	0.15	0.34	0.30	0.21

When evaluating the performance of MILC and tree-MILC, we are interested in whether the imputed latent variable has similar properties as the simulated true latent variable. Therefore, we inspect the size of the LCs; the pooled proportions of each class averaged over all the simulations. We compare these pooled proportions to the true latent variable size.

To see how MILC and tree-MILC perform on datasets with different structures, we generated data with eight different variations on the described theoretical population. The simulation conditions differ by the level of selection error (5% or 20%) and the level of measurement error (5% or 20%) for the indicator variables and the source of error strongly associated with the covariate (measurement or selection error).

We use accuracy-based and confidence interval performance measures to compare the two method. Accuracy based: root mean squared error (RMSE), mean absolute relative error (MARE). In the latter the estimates are divided by their population values as to reduce the effect of the estimate's size when comparing estimates of several sizes. The two confidence interval performance measures: the coverage of the 95% confidence interval (CI) of the population estimate, and the average width of the 95% CIs of each simulation iteration.

Table 2. data simulation conditions

				Measurement error			
				5%		20%	
				Covariate			
				Weak	Strong	Weak	Strong
Selection error	5%	Covariate	Weak	-	2	-	4
			Strong	1	-	3	-
	20%	Covariate	Weak	-	6	-	8
			Strong	5	-	7	-

Note: the simulation conditions are numbered to ease the display of the results in figure 3.

3. RESULTS AND PRACTICAL APPLICATION

In the boxplots in figure 3 you can see an overview of the deviation and the spread of the results for the class sizes per simulation condition for both methods. The first impressions from the boxplot are the following: for low selection errors both methods perform equally well, for high selection errors tree-MILC performs better. Class 1 seems to be the class most affected by higher errors; both methods perform worse on class 1 than on the other classes in most

conditions. The spread of the results for class 1 is larger than that of the other classes, there is more uncertainty, especially for more challenging conditions.

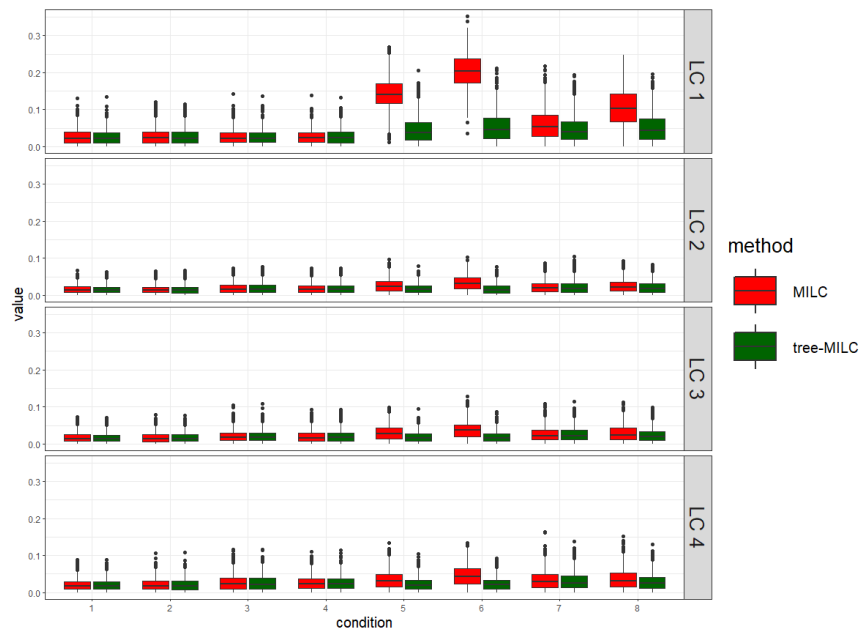


Figure 3. mean absolute relative error (MARE) of class sizes per method (MILC and tree-MILC) for each data simulation condition.

4. MAIN FINDINGS

Overall, both methods perform comparably well when the errors in the combined dataset are small. As expected, MILC performs worse in terms of accuracy and the coverage of the 95% CI on classes subjected to higher levels of errors. The performance of MILC is indeed related to the 'difficulty' of the simulation condition. Different than expected is that conditions five and six were more challenging conditions for MILC than conditions seven and eight. Although the variation of its estimates is higher for the conditions with high selection error, tree-MILC does perform well in terms of accuracy and coverage of the 95% CI. In these cases, tree-MILC is a good alternative to MILC.

5. REFERENCES

- [1] L. Boeschoten, D. Oberski, and T. De Waal, "Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC)," *J. Off. Stat.*, vol. 33, no. 4, pp. 921–962, Dec. 2017, doi: 10.1515/jos-2017-0044.
- [2] J. K. Vermunt, J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma, "9. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis," *Sociol. Methodol.*, vol. 38, no. 1, pp. 369–397, Aug. 2008, doi: 10.1111/j.1467-9531.2008.00202.x.
- [3] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. in Wiley Series in Probability and Statistics. New York: John Wiley and Sons., 1987.
- [4] M. Van den Bergh, "Latent class trees," PhD Thesis, Tilburg University, 2018. [Online]. Available: https://pure.uvt.nl/ws/portalfiles/portal/20026000/Van_Den_Bergh_Latent_05_01_2018.pdf

- [5] J. Lewis and D. Linzer, “poLCA: An R Package for Polytomous Variable Latent Class Analysis,” *J. Stat. Softw.*, vol. 42, no. 10, pp. 1–29, 2011, doi: 10.18637/jss.v042.i10.

Mixture models and graphical tools to analyse the perception of work-related stress across EU countries

Stefania Capecchi & Francesca Di Iorio (Department of Political Sciences, University of Naples Federico II)

1. INTRODUCTION

Occupational stress is currently considered a primary social issue as well as an extensive problem of public health. Latest research on risk factors at work displays that about 25% of workers claim to experience work-related stress at least “most of the time”. A similar proportion also claim that their job may negatively affect their physical and psychological health and overall well-being [1]. Self-employed workers have strongly suffered since the early stages of the pandemic from a significant decrease in working hours and consequently in their earnings.

This study draws on data from the 2022 European OSH Pulse survey, commissioned by the European Union information agency for occupational safety and health (EU-OSHA) as a part of the European Commission’s Strategic Framework on Health and Safety at Work. Aim of the survey was to explore the different impacts on both the self-employed and employees regarding the reported levels of work-related stress, access to psychosocial risk prevention measures and their disposition to discuss mental health and stress-related issues at work. Such evidence is usually collected by means of self-reported answers to several questions on working conditions which are often gathered on Likert scales. Implementing the CUB models, a class of mixture models suitable for the analysis of ordinal data, the paper analyses the perception of occupational stress across European Union countries. Our preliminary results would indicate that, after two years from the outburst of the COVID-19 pandemic, some consequences on workers’ mental health remain, and that the impact on the self-employed was even higher. Those findings can be conveniently displayed in the parametric space (the unit square, in this case) and compared by respondent groups, characteristics, countries of origin, and so on.

2. METHODOLOGY

Data stem from the Flash Eurobarometer - OSH Pulse survey, commissioned by EU-OSHA), which was implemented in April-May 2022 [1], aimed to detect how the COVID-19 has impacted on workers’ health and well-being. A probability sample of employed workers of 27,250 individuals aged 16 and over in the EU member states (EU-27, total sample size: 25,683) and two associated countries (Iceland and Norway, total sample size: 1,567) was interviewed over the telephone including both employees and self-employed individuals.

The survey encompasses all the economic sectors investigating psychosocial risk factors, stress, and mental health, OSH preventive measures, opinions and experience of OSH in the workplace, and digitalisation. We focus on question E2_4, which reads: “My work stress has increased as a result of the COVID-19 pandemic”). Answers are collected on a 4-point Likert scale (1=strongly agree; 4=strongly disagree. For a better readability of results, we needed to reverse the original wording scale).

Consolidated statistical analysis of preferences, opinions, judgments, etc. expressed by means of ordinal variables is afforded by Generalized Linear Models [2]. We choose to

implement a relatively novel approach based on the hypothesis that self-assessments may be interpreted as the result of a mixture of the agreement towards the item and the uncertainty/heterogeneity within the response pattern. The selected class is that of CUB models, the acronym standing for a convex Combination of a discrete Uniform and a shifted Binomial random variable. Proposed in 2003 by [3] and then developed in many directions and with various applications (for an extensive review, see [4]), CUB model assumes that the probability of a response R_i for the i -th individual is given by:

$$Pr(R_i = r | \pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i) \frac{1}{m}, \quad r = 1, \dots, m,$$

where parameters $\pi_i \in (0, 1]$ and $\xi_i \in (0, 1)$. The larger $1 - \xi_i$ the more the distribution is concentrated along the highest scores. Thus, high level of $1 - \xi_i$ indicates a positive tendency in the data with respect to the analysed topic; in this sense $1 - \xi_i$ is a *feeling* parameter. Mixing weight $1 - \pi_i$ is the uncertainty parameter. Richer CUB model are obtained including explanatory variables in *feeling* and/or *uncertainty* components directly depending on respondents' profiles. In the selected mixture, the indecision is explicitly considered and may be related to the available respondents' covariates, when significant.

Covariates and parameter are connected by the usual logit link function such that: $\text{logit}(1 - \pi_i) = -\beta_0 - y_i \beta_1$ and : $\text{logit}(1 - \xi_i) = -\gamma_0 - w_i \gamma_1$ where y_i is the row vectors of p covariate values for the i -th subject for his/her uncertainty, and w_i is the row vectors of q covariate values driving his/her feeling. Likelihood methods are the preferred estimation procedures for CUB models and currently estimation packages are available in R [5], Gretl [6] and STATA [7]. For this paper CUB Gretl package has been modified to consider sample weights and in the present study the standard error estimates are obtained from Sandwich matrix and by Bootstrap Standard errors [8]. More importantly, the visualization of the estimated models and the effects of covariates on the ordinal responses can be depicted using several graphical devices, these representations conveying a straightforward interpretation of the results.

3. RESULTS AND PRACTICAL APPLICATION

CUB models without covariates (CUB 0,0) were estimated separately by country, distinguishing between employees and self-employed, the latter representing about 16% of the overall sample, yielding for each country the coefficients that enable the representation in the unit square (Panel A, Figure 1). It can be observed from Panel A that nearly all countries display some intermediate levels of feeling, while a strong disparity emerges between the countries in terms of uncertainty. In particular, self-employed interviewees from France show a high level of feeling, reaching 0.74, with a high level of uncertainty. Through multidimensional scaling and grouping by means of a complete linkage cluster, four groups of countries clearly emerge (Panel B, Figure 1). German self-employed respondents, French non-self-employed respondents and Austrians interviewees display a distinct response pattern, which is characterised by high uncertainty. Estonian respondents and self-employed Swedes, in addition to their high uncertainty, are the ones featuring the lowest feeling.

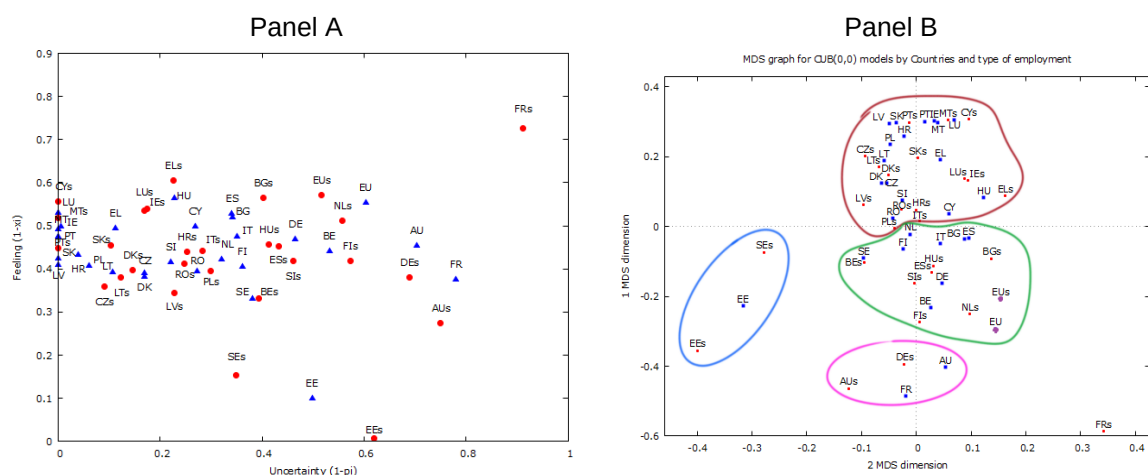


Figure 1. CUB (0,0) by country and type of employment

Moreover, to examine the effect of gender and job typology on the increase of occupational stress due to COVID-19, we performed some additional analyses. A CUB model with covariates was estimated for the whole sample with gender (reference category: female) for the feeling component, and both gender and type of job (reference category: employee) for the uncertainty component (Figure 2, Panel A). The average estimated probabilities are graphically represented distinguishing by gender and job type (Figure 2, Panel B).

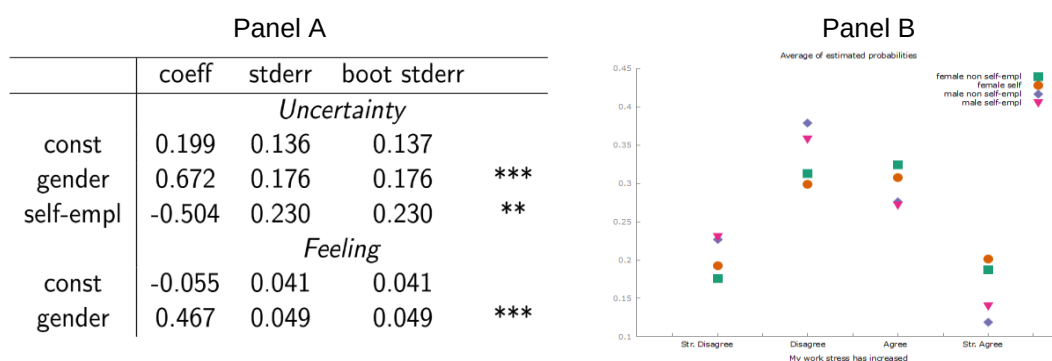


Figure 2. Estimated CUB models by gender and type of employment

The plot highlights that women, regardless of their job type, are more likely than men to strongly agree with the item, i.e., to have experienced increased stress as a result of covid, whereas non-self-employed males reach a probability of disagreeing with the item of about 37%. Our findings indicate that, across the 27 EU member states, the COVID-19 epidemic has had a more significant impact on the mental health of the self-employed as compared to that of the employees. This is even more true in the health and education sector as well as in countries where the measures to contain the spread of the pandemic were more stringent. Moreover, we found that self-employed women have been more affected by the pandemic, in line with extant research.

4. REFERENCES

- [1] EU-OSHA. Flash Eurobarometer – OSH Pulse survey. (2022)
<https://osha.europa.eu/en/facts-and-figures/osh-pulse-occupational-safety-and-health-post-pandemic-workplaces>

- [2] McCullagh, P., and Nelder, J. (1989). Generalized linear models (2nd ed.). London: Chapman and Hall/CRC
- [3] D. Piccolo, On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, (2003) 85-104
- [4] D. Piccolo, R. Simone, The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Stat. Methods Appl.* 28, (2019) 389–435.
- [5] R. Simone, FastCUB: Fast EM and Best-Subset Selection for CUB Models for Rating Data. R Package Version 0.0.2 (2020). <https://CRAN.R-project.org/package=FastCUB>
- [6] R. Simone, F. Di Iorio, R. Lucchetti, CUB for Gretl. In: Di Iorio, F., Lucchetti, R. (eds.) *Gretl 2019: Proceedings of the International Conference on the GNU Regression, Econometrics and Time Series Library*, (2019) 147–166. feDOA University Press, Naples
- [7] G Cerulli, R. Simone, F. Di Iorio, F., D. Piccolo, C.F. Baum, The CUB STATA module: mixture models for feeling and uncertainty of rating data, *The Stata Journal*, 22(1), (2022), 195–223.
- [8] S. Koelikov, Resampling variance estimation for complex survey data, *The Stata Journal*, 10(2) (2010), 165–199

Mobile Network Operator Data as a New Source for Transport Planning and Mobility Analyses

Markus Zwick & Yannik Buhl (Destatis)

1. INTRODUCTION

Mobile network operator data (MND) are an increasingly important data source for official statistics (Saidani et al. 2022; Hadam 2022), e.g. for population statistics, mobility analyses or policy decisions in general. To this end, Germany's Federal Statistical Office (Destatis) continuously cooperates with universities and mobile network operators (MNO) towards developing new solutions to address the qualitative challenges of MND, preparing it for a use in official statistics. At NTTS 2025, Destatis focuses on the application of MND in the field of traffic and mobility planning as well as on new methods to resolve existing quality issues with MND. In doing so, it is building on the work of the past years and current research commissioned by the German Federal Ministry of Digital and Transport.

2. CURRENT RESEARCH INTO MND

Past research has shown that current methods of extrapolating MND to the total population suffer from a lack of additional information to compensate existing structural differences between MNO. Thus, Destatis – currently leading a national research consortium of mobile network operators, data providers and universities – aims to improve efficiency of extrapolation methods for MND, e.g. in order to map traffic networks more precisely. To be able to improve this situation, it has been shown that there is a need for better information on the socio-demographic background of mobile phone users. This missing data is collected by a large-scale survey conducted within the research project, and it is linked with MND from two network operators: A major challenge in the previous work had been the limitation to one MNO (Telefónica), which tends to have an urban, young customer base. Now, with Deutsche Telekom as a second MNO involved in the project, significant new insights from its customer structure are expected, since Telekom tends to have a more rural and slightly older clientele. By analysing their different customer structures, it will be possible to gain important insights into the selectivity of the data by region and by socio-demographic characteristics. Through combining survey data with traditional (official) data, MND and small area estimation methods, the research consortium aims to compensate the known quality limitations of mobility data as best as possible.

A second strand of research involves using statistical simulation in the context of MND to investigate how the consistency of complex traffic forecasts can be improved. Microsimulations (e.g., Li and O'Donoghue, 2013; Münnich et al., 2021; Münnich et al., 2013; Burgard et al., 2020) are used to investigate how sample information can contribute to a significant improvement in the estimation quality of the resident population and mobility behaviour at regional level using MND. To this end, the aforementioned dedicated survey is used to obtain essential information for optimising the simulations, which in turn can be used to design future sampling concepts. Furthermore, methodological and practical questions regarding the optimisation of existing large-scale mobility surveys in Germany are also integrated into the research project.

Lastly, the findings from the large-scale survey, the resulting analyses of distortions and the possibilities for compensation become the basis for the development of a quality concept for MND-supported traffic data. On the one hand, this concept analyses the process of MND evaluation of the data providers. On the other hand, the limits within which MND-supported traffic data can be used for different transport planning issues are described in detail. As a result, the consortium will provide regionally detailed origin-destination matrices that are adjusted for distortions and selectivities as far as possible, having leveraged the results of the large-scale survey and the comparative calculations for both mobile network operators. These quality-optimised and quality-assured origin-destination matrices – segregated by mode of transport – are finally made available as part of a pilot application for mobility simulations and forecasts in selected regions.

3. SUMMARY

For the past years, Destatis has made extensive efforts researching mobile network data and working towards an implementation in official statistics (Hadam 2022). Recent research with a focus on using MND for mobility and transport planning has revealed missing information that prevents the creation of reliably extrapolated statistics. At NTTS 2025, Destatis presents its current research that aims to tackle these shortcomings by deploying a large-scale survey with focus on the population's use of mobile phones, combining the results with traditional and non-traditional data and leveraging methods of small area estimation and microsimulation – all with the ultimate goal of enabling the future usage of MND by European national statistical offices.

4. REFERENCES

- [1] Saidani, Y., Bohnensteffen S., Hadam S. 2022. „Qualität von Mobilfunkdaten – Projekterfahrungen und Anwendungsfälle aus der amtlichen Statistik.“ WISTA – Wirtschaft und Statistik, 74 (5), 55–67.
- [2] Münnich, R., Burgard, P., J., Vogt, M. 2013. „Small Area-Statistik: Methoden und Anwendungen.“ AStA Wirtschafts- und Sozialstatistisches Archiv, 6 (3/4), 149–191.
- [3] Burgard, J. P., Dieckmann, H., Krause, J., Merkle, H., Münnich, R., Neufang, K. M., Schmaus, S. 2020. „A generic business process model for conducting micro-simulation studies.“ Statistics in Transition New Series, 21 (4), 191–211.
- [4] Li, J., O'Donoghue, C. 2013. “A survey of dynamic microsimulation models: uses, model structure and methodology.” International Journal of Microsimulation, 6 (2), 3-55.
- [5] Münnich, R., Schnell, R., Brenzel, H., Dieckmann, H., Dräger, S., Emmenegger, J., Höcker, P., Kopp, J., Merkle, H., Neufang, K., Obersneider, M., Reinhold, J., Schaller, J., Schmaus, S., Stein, P. 2021. „A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model.“ Methods, data, analyses, 15 (2), 241-263.
- [6] Hadam, S. 2022. Einsatzmöglichkeiten von Mobilfunkdaten in der amtlichen Statistik. Dissertation, Supervisors: Prof. Dr. Markus Zwick & Prof. Dr. Timo Schmid, <http://dx.doi.org/10.17169/refubium-38281>.

Mobile Network Operator data for Tourism Statistics: facing the challenge of cross-roaming

Tiziana Tuoto (Istat), Egle Rüütli (Positium), Juan Arranz (Nommon),
Florabela Carausu (GOPA Worldwide Consultants GmbH), Loredana Di
Consiglio (Istat), Kadri Arrak (Positium), Miguel Picornell (Nommon),
Margus Tiru (Positium) & Ricardo Herranz (Nommon)

Instructions: This research initiative investigates potential challenges in producing inbound tourism official statistics based on mobile network operator (MNO) data. The supporting analyses aim at contrasting the results derived from a single MNO with those produced from multiple MNOs, exploring at the same time different data fusion methodologies. To the best of our knowledge, the proposed investigation is the first of its kind, and it aims at enhancing the accuracy and reliability of MNO data as a source for official statistics, particularly in the context of tourism statistics. Ultimately, it contributes to the broader discussion on how emerging data sources can complement and enrich existing traditional data and methods in order to improve the quality of official statistics in Europe and worldwide.

1. INTRODUCTION

The data collected by mobile network operators (henceforth MNOs) has emerged as an appealing source for producing socio-economic indicators. Its ability to provide continuous, large-scale, and geographically detailed information makes it an interesting complement to traditional sample surveys and administrative data to produce reliable timely statistics. In recent years, several National Statistical Institutes (NSIs) within the European Statistical System (ESS) have started to conduct exploratory activities aimed at using MNO data for the development of innovative statistical products or for complementing existing statistics in areas like urban planning, mobility studies, and tourism analysis.

Parallel to the national initiatives, Eurostat has financed a number of projects to firstly explore, and then facilitate the re-use of MNO data for statistical purposes by NSIs, in compliance with the European Code of Practice (CoP) principles (i.e. ensuring the soundness and transparency of the methodology, comparability of the results among countries and domains, independency and impartiality of the NSIs towards stakeholders, including private companies acting as data providers, privacy and confidentiality of personal and business sensitive data, etc.). One Eurostat-funded project currently running is the [Multi-MNO](#) project, aiming at developing, implementing and demonstrating a proposal for a reference standard processing pipeline for the future production of official statistics in Europe based on MNO data from multiple operators. The Multi-MNO project explicitly takes into account the combination of information from multiple MNOs, both within and across countries. This multi-MNO orientation brings a series of benefits by ensuring a better representativeness of the total population (i.e. reducing the risk of population coverage bias in the final statistics), improving the robustness of the final statistics to anomalies, glitches and interruptions of data availability caused by technical flaws of individual MNOs, and ensuring equal treatment of competing MNOs. The combination of data from multiple MNOs also prevents the introduction of asymmetries and differences in treatment, reducing the risk of interfering with competition dynamics between MNOs and offering an extra layer of protection for the business sensitiveness of the information from a single MNO. In addition, this approach also ensures that no single MNO holds undue influence, thereby enhancing the NSI's independence from any MNO or brand.

However, despite the benefit of the multi-MNO approach and of the MNO data potential in general, MNO data presents its own challenges, especially when one is interested in producing official statistics on a regular basis. In this work, we concentrate on one specific challenge that may arise when using data from multiple MNOs to measure inbound tourism. Inbound tourism indicates the stays at a specific destination within a country by visitors who are not residents of that country, e.g. Italians usually living in Rome are inbound tourists of Belgium when attending in person the NTTS conference in Brussels.

Beyond tourism, cross-roaming can also distort other critical indicators such as population mobility, commuter patterns, and the spatial distribution of populations within border regions. For instance, in regions where cross-border commuting is frequent, the loss of data when users switch networks can lead to inaccurate representations of cross-border worker flows, impacting infrastructure planning and public service allocation.

2. SCOPE AND METHODOLOGY

The purpose of our contribution is to investigate to which extent inbound tourism indicators can be affected by the cross-roaming phenomenon. In the visited country (Belgium in the example above), foreign SIMs (the Italian SIMs to continue with the example) are served by the local operators (e.g. Base, Orange, Proximus) through agreements of international roaming. We use cross-roaming to indicate when, during the international roaming, a single SIM is served by multiple local MNOs, e.g. during the NTTS week the Italian SIM is taken over by both Orange and Proximus.

The cross-roaming is particularly problematic because the data observed by the single local MNO becomes fragmented when the device connects to different operators, while traveling within the visited countries. The situation is similar to what we can observe in border areas between countries, when the devices can be taken over alternatively by the local MNO or the MNOs of the border country, fragmenting the complete observation of any single MNO. In cross-roaming, this results in gaps in the visited single MNO dataset, while the complete observations of the user are available in principle only to the original mobile operator. The cross-roaming phenomenon might affect the inbound tourism statistics, when calculating indicators such as the total arrivals and number of nights a tourist stays in a country. When a tourist's mobile device initially taken over by a given MNO of the visited country switches to a second network, due to cross-roaming, this can be incorrectly registered in a two-fold perspective: (1) as their departure from the country by the first operator, even if they have not actually left, and (2) as a new arrival in the country by the second operator, even if it is continuation of the stay. This can lead to an underestimation of the total number of nights spent by tourists in a destination country and, conversely, an overestimation of the tourist arrivals in a country.

Despite the intuition behind the situation potentially generated by the cross-roaming phenomenon, a clear assessment of its size and implications on inbound tourism statistics has not been investigated so far, to the best of our knowledge.

This due to the fact that it requires, among others, a good cooperation between different operators within the same country and in different countries, as well as the application of the same algorithm among multiple operators for the extraction of information on the touristic indicators (e.g. the same operationalisation of the concept of arrivals and nights spent) from the MNO data. The multi-MNO project has a privileged set up in terms of the two factors mentioned above, since it actually focuses on the definition of a common standardised reference pipeline for processing MNO data and actively involves 5 operators from 4 different European countries for the test of the pipeline. In addition, inbound tourism statistics have been selected among the use cases for which the project will take care of defining standard algorithms. Hence, in this work, we propose to investigate the following questions:

- (I) Is the cross-roaming phenomenon a serious issue affecting the estimation of inbound tourism indicators, such as arrivals and nights spent?
- (II) If yes, can the use of single MNO data mitigate the effects of cross-roaming when it comes to the evaluation of inbound tourism indicators?
- (III) Alternatively, can we handle the cross-roaming with statistical methods applied to multiple MNO aggregated data or this requires a privacy preserving micro-linkage approach among multi-MNO data?

To provide answers to the previous questions, we analyse data from several operators and from several countries in Europe. The same algorithm to assess the inbound tourism indicators is applied across the MNOs and countries. The data are processed from the subscribers' MNO perspective, that is the origin MNO, and the visited multi-MNO observations are re-created from this, allowing us to compare the complete set of observations, with the single visited MNO, and with different potential data fusion of multiple visited MNO. The selected reference period is long enough to largely cover more than twice the average stay of European inbound tourists in Europe, in the different countries.

In terms of output indicators, we consider mainly the number of arrivals and nights spent. The nights spent are considered both as overall value and as its distribution in number of nights per tourist, in order to evaluate the distribution preservation of the single MNO data compared to the complete observation.

Regarding the statistical methods to recover the complete observation from the single MNO and multi-MNO data, we will consider two approaches, namely: (1) when the MNO data are intended to be used as auxiliary information in a super- population modelling setting, and (2) when the MNO data are intended to be used as primary data sources in a quasi-randomisation approach. In both cases, we explore the possibility to obtain results derived by the complete data observation, applying a data fusion methodology as mixture models for time series and for count data. The validity of some model assumptions is also tested.

3. MAIN FINDINGS AND CONCLUDING REMARKS

The results of our investigation will be presented during the conference, duly anonymised per MNO and per countries. All the required business secrecy safeguards will be applied. Nevertheless, meaningful findings will be highlighted and discussed, stressing conclusions that seem to be general across operators and countries. Due cautions and limits of the analysis will be remarked as well, hoping that further investigations from different research groups will follow.

For instance, a first relevant limitation of the analysis we carried out is related to the lack of information to detect the potential heterogeneity given by some spatial effect. Indeed, it seems intuitive that the cross-roaming can be related to the places visited by the roamers, e.g. it could be unusual in big cities and places where all the operators have a high coverage of the areas, while it could be concentrated in remote areas covered only by a given operator. Unfortunately, the subscribers' MNO perspective adopted in this analysis does not allow us to know in which places the roamer has moved across the visited country, preventing us the possibility to investigate further the connection between international cross- roaming and territory. A similar reasoning applies to time heterogeneity, i.e. the possibility that the international cross-roaming is dependent on the time, e.g. the night of a special event which overloads the antennas in a given area, causing the re-distribution of the roamers in close antennas potentially taken by different operators.

By addressing the cross-roaming problem, we aim to enhance the accuracy and reliability of MNO data as a source for official statistics, particularly in the context of tourism. Ultimately, our findings contribute to the broader discussion on how emerging data sources can complement and enrich traditional data and methods to improve the quality of official statistics in Europe.

ACKNOWLEDGEMENTS

The work was funded by the European Commission Project Multi-MNO - Service Contract Number – 2021.0400. The authors express their gratitude, for the inspiring discussions on cross-roaming and MNO data in general, and on methodologies for data integration, to the Eurostat experts and to the experts of Work Package 3 of the European Commission co-funded Project “MNO-MINDS” - 101132744 — 2022-IT-TSS-METH-TOO, respectively.

Mobility Research with a Smartphone App

Hanja Maksim & Marcus Roller (Federal Statistical Office)

1. INTRODUCTION

In Switzerland, the Federal Statistical Office (FSO) and the Federal Office for Spatial Development (ARE) have been conducting the Mobility and transport microcensus (MTMC) every five years since 1974 to survey the mobility behaviour of the Swiss resident population.

Until now, the survey has been carried out by means of computer assisted telephone interviews (CATI method). There are currently a number of challenges, such as a declining number of landlines in Swiss households and a decreasing willingness to participate in telephone surveys of the Swiss population. Thanks to digitalisation, however, there are alternatives to this survey method.

The FSO and ARE therefore carried out a pilot project with the research company intervista AG in 2022, in which mobility was surveyed using a smartphone app. This pilot study focused on the following questions:

- Can such a complex and comprehensive survey be carried out entirely via a smartphone app?
- What hurdles need to be overcome?
- How high is the willingness to participate in such a study with a smartphone app?
- How many people would prefer a survey via an app to a CATI survey?

2. METHODOLOGY

The regular MTMC CATI survey provides answers to questions such as: How far do the Swiss residents travel every day? What means of transport do they use? What are the purposes of their trips? The results of this survey are used, for example, to make targeted, fact-based decisions on transport policy or spatial planning, as well as for the research. To obtain the data, telephone interviews lasting around 25 minutes were used to record all the trips of the selected target persons on an allocated reference day. Using a routing tool, the departure and arrival points are geocoded directly during the interview, i.e. associated with geographical coordinates. The precise route is also recorded.

The pilot survey with the app had to take into account all these complexities of the CATI survey system. The survey with smartphone therefore includes a questionnaire section and a mobility tracking section, using the device's movement sensors. A validation module ensures that the data is as correct as possible. Once developed, around 7,000 people were invited to participate. Alternatively, they could also take part in a telephone survey. The data collected presented the challenge of finding a way of comparison with the CATI data, given that the people who chose the app or the CATI did not have the same socio-demographic characteristics.

3. RESULTS AND PRACTICAL APPLICATION

Overall, it was shown that recruitment for a survey using a smartphone app is possible. However, there are differences between the two samples (see e.g. Figure 2). This applies, for example, to the measurement of transport behaviour and the composition of the sample. Over 20% of the people invited installed the app on their smartphone. A survey of the participants also revealed that a relevant proportion favoured participation via app over a telephone interview.

This illustrates the high level of acceptance of digital survey methods in modern social research.

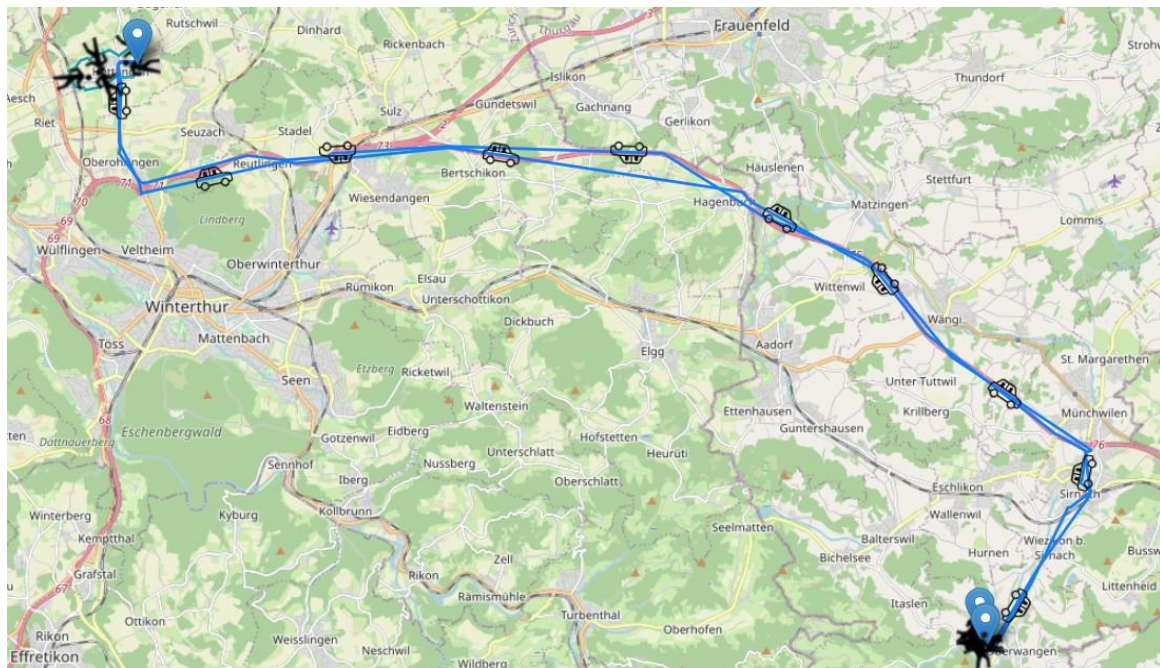


Figure 1. Illustration of measured mobility behaviour

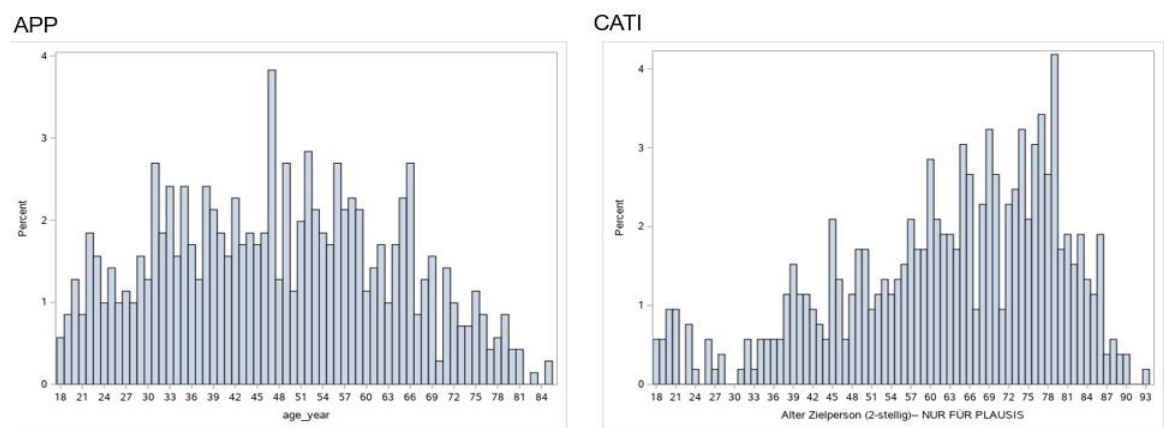


Figure 2. Age structure of participants

4. MAIN FINDINGS

A smartphone app can be used to successfully collect data on the mobility behaviour of individuals.

While the technology used by intervista has already been used in various private research projects since 2018, this project has shown that this methodology is also fundamentally suitable for official statistics surveys. This is why a second pilot project will be carried out next year, with the aim of implementing this new channel as part of official statistics in Switzerland and as an alternative to CATI.

5. REFERENCES

- [1] Footprints Research Panel, intervista
- [2] Population's mobility behaviour, Federal Statistical Office

Model explainability: a use case on FastText for automatic coding

Meilame Tayebjee (Innovation Lab, Insee)

1. INTRODUCTION

At INSEE, the FastText model is used across various divisions for different use cases, with the primary application being automatic coding — a recurring topic of innovation for National Statistical Institutes (NSIs). The model has been deployed in production across multiple registers, achieving a high level of automation. However, when the model suggests an unexpected but plausible label, managers have expressed the need to understand why the model predicted that label — this is a typical example of the growing demand for model explainability. Beyond mere automation, providing users with clear insights into the model's decision-making process has become a key objective. This paper presents a use case of an explainability pipeline integrated into FastText's automatic coding, leveraging the integrated gradients [1] method.

2. METHODOLOGY

2.1. Explainability: definition

Model explainability is strategically important from both the user's perspective (increasing confidence in the predictions, enhancing understanding, and promoting better usage) and the data scientist's perspective (facilitating improved model training and performance). We can distinguish between two types of model explainability:

- *Global explainability*: This approach seeks to explain the model's behaviour across the entire input space. Techniques include sensitivity analysis, embedding space visualizations, and assessing feature importance across all predictions.
- *Local explainability*: This focuses on understanding why the model made a specific prediction and which input influenced that outcome.

In this paper, we will concentrate on local explainability.

2.2. Integrated gradients and Captum

Given an input $x = (X_1, \dots, X_n)$ and the corresponding output from the model, Sundararajan et al. [1] propose a method to evaluate the attribution of a given coordinate i , that is how much X_i did influence the prediction (compared to the other X_j). The method is based on the gradient of the output with regard to X_i , for several input z lying on the straight path between a baseline x' and x : all these gradients are then averaged. We refer to the article for further details. The Captum library in Python [2] provides an easy way to use the integrated gradients method with any trained PyTorch model.

Using our custom PyTorch implementation of FastText, we integrated this feature into our pipeline. In our case, the inputs are a list of tokens. At inference time, each token in a given input sentence is assigned an influence score **on a given prediction** (the first predicted code in terms of confidence might be driven by certain tokens, and the second one by other ones –

as we will see in the Results section). We then apply a softmax function to normalize these scores between 0 and 1, ensuring that they sum to 1 at the sentence level.

2.3. Final visualization

To achieve effective visualization, the first step is to map each token to its corresponding processed word. A processed word is essentially the original word transformed by removing stop words, converting to lowercase, and applying stemming. This mapping utilizes the "tokenized" sentence, where tokens derived from a specific processed word appear together in sequence, ordered from the shortest to the longest token. Additionally, the symbols ">" and "<" indicate the end and beginning of a word, respectively. The process starts by associating the first token with the first word. Iteration continues until encountering a ">" in a long token, signalling the end of the current word's token group. If the next token is short and begins with "<", the transition to the next processed word occurs.

We then provide two visualization modes:

- *Letter Level:* After establishing the mapping between processed words and their corresponding token groups, we distribute each token's attribution across the letters it contains. A detailed example will be provided in the Results section.
- *Word Level:* To aggregate at the original word level, each processed word is matched to its original counterpart. This involves iterating through the original words, converting them to lowercase, and applying stemming to closely align them with the processed words. Using Python's difflib library, the Ratcliff-Obershelp [3] algorithm identifies matches based on a similarity score exceeding a specified threshold. Once a match is found, the original word receives the sum of the attributions of all tokens linked to the matched processed word. If no match is found, the original word is identified as a stopword and is assigned an attribution score of 0.

Example:

Input: Rénovation bâtiments dont: électricité, plomberie, serrurerie, menuiserie

Processed words: renov bat dont: electricite, plomberie, serrurerie, menuiser

Beginning of the tokenized sentence:

'<re', 'ren', 'eno', 'nov', 'ov>', '<ren', 'reno', 'enov', 'nov>', '<reno', 'renov', 'enov>', '<renov', 'renov>', '<ba', 'bat', 'at>', '<bat', 'bat>', '<bat>' etc.

It is very easy to identify the end of word and to identify the block of tokens associated to renov which is itself matched to Rénovation.

3. RESULTS AND PRACTICAL APPLICATION

We provide here the output of our explainability pipeline, for the example input given above that is taken from the French business activity register (SIREN). Based on a sentence describing the activity of the company, a label is given by the model (from a list of approximately 700 labels).

The following visualizations are taken from the deployed API [Explicabilité | Codification APE RIAS/SSP Lab](#), where one can test the explainability pipeline with any custom input.

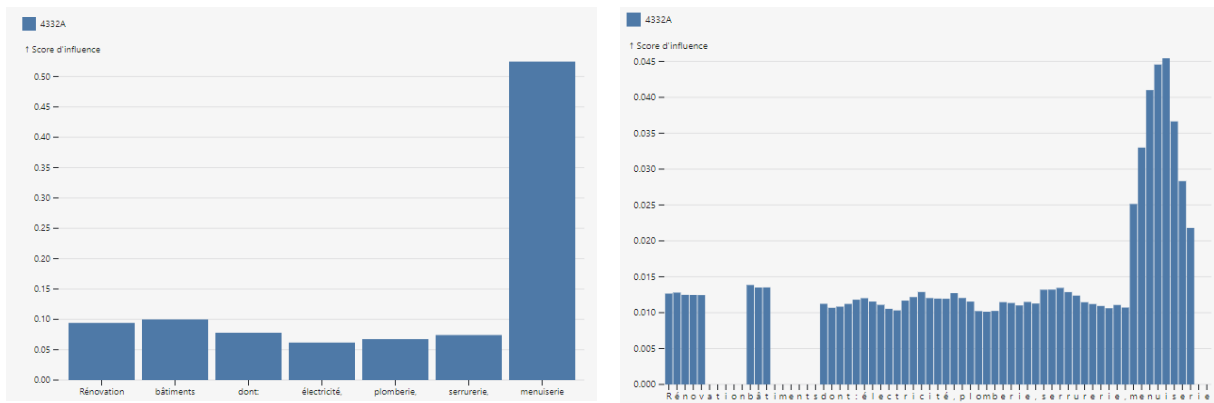


Figure 1. Influence scores for prediction 4332A NAF code at word (left) and letter (right) levels

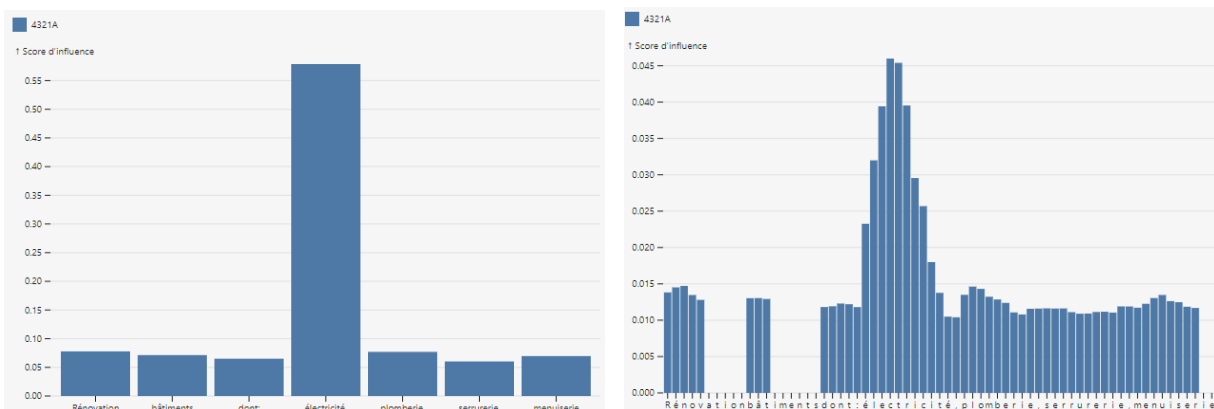


Figure 2. Influence scores for prediction 4321A NAF code at word (left) and letter (right) levels

It is here interesting to see that:

- for prediction 4332A (*Wood and PVC joinery*, most confident prediction), it is the word *Menuiserie* (*joinery/carpentry*) that drove the prediction,
- for prediction 4321A (*Electrical installation work on all premises*, third most confident prediction), it is the word *Electricity*.

4. MAIN FINDINGS

This streamlined pipeline requires no retraining and can be seamlessly integrated into a monitoring API for registry managers. The API provides real-time predictions and explanations for custom inputs, enabling managers to better understand the model's predictions and what influenced each one.

Overall, this work represents an initial step toward the broader integration of machine learning model explainability within NSI, particularly in the domain of automatic coding, addressing both crucial ethical concerns, such as accountability, and practical challenges faced by register managers.

5. REFERENCES

- [1] Mukund Sundararajan, Ankur Taly and Qiqi Yan, *Axiomatic Attribution for Deep Networks*, International Conference on Machine Learning, 2017

- [2] Narine Kokhlikyan and Vivek Miglani and Miguel Martin and Edward Wang and Bilal Alsallakh and Jonathan Reynolds and Alexander Melnikov and Natalia Kliushkina and Carlos Araya and Siqi Yan and Orion Reblitz-Richardson, *Captum: A unified and generic model interpretability library for PyTorch*, 2020
- [3] John W. Ratcliff and David E. Metzener, *Pattern Matching: The Gestalt Approach*, Dr. Dobb's Journal, page 46, July 1988.

Multi-Mode Protocol for the French Household Budget Survey: More Choices, More Respondents?

Claire-Lise Dubost & Simon Moreau (Insee)

1. INTRODUCTION

The Household Budget Survey (HBS) aims to gather a comprehensive overview of households' expenditure and resources. Traditionally, the HBS collected data from a representative sample through two face-to-face interviews and paper diaries, where participants manually recorded their expenses for a week. This protocol places a significant burden on respondents, as both interviews are very long to capture all expenses. To make the survey more acceptable to respondents, reduce costs and interviewer burden, and improve data quality and timeliness, several changes will be introduced in the next survey protocol.

In this context, an experiment with around 2,500 sampled households was conducted in France to test two new tools: a web-based questionnaire to shorten the second interview when completed, and a smartphone application for recording diary entries.

By comparing results between two independent samples which used different protocols, the 2024 experiment aimed to assess the capacity of these new tools to support data collection, and the soundness of the new protocol. Did respondents accept these new tools? Is the data comparable to that collected under the previous protocol? Has the overall response rate been affected by the new protocol?

2. METHODOLOGY

2.1. Protocol of the 2017 HBS

Historically, the HBS gathered data from a representative sample through two face-to-face interviews, supplemented by paper diaries. The two interviews collected information on household composition, characteristics, and resources, along with a comprehensive review of all expenditure, except for the most frequent ones. These frequent expenses were documented separately in an expense diary over the course of seven days.

Two visits were necessary because the questionnaire would be too lengthy for a single visit, and the second visit allowed for verification and collection of the expense diaries.

2.2. Protocol of the 2024 experimental HBS

Two innovative tools are introduced for the experimental survey: a web-based questionnaire and a mobile app.

The web questionnaire aims to reduce the interview time during the second visit, thereby improving the survey's acceptability for the respondent. It covers a small portion of the second interview, focusing on housing expenses and high-value expenditure. If the web section is not completed between the two visits, the questionnaire is administered during the second face-to-face interview.

The second innovative tool is a smartphone application, replacing the paper diary for collecting expenses over the week. Based on the open-source @HBS developed by Statistics Netherlands (CBS), it allows participants to manually record expenses by selecting from a product list, and to upload receipts using their phone's camera.

2.3. Methodology of the 2024 experimental HBS

The experimental HBS, conducted between May and July 2024, was designed to answer two main questions:

- Are the new tools suitable for collecting HBS data?
- Is the new protocol overall more efficient than the previous one?

The test was conducted on two distinct samples. The first sample (831 households) was offered the traditional collection method only, without web-based questionnaire or diary application, to serve as a reference sample. The second sample (1669 households) was given the option to use the web-based questionnaire and/or the application, with these two choices being independent. The experimental survey was made compulsory to ensure a sufficient response rate, enabling statistical exploration across all modes.

At the end of the first visit, in the reference sample, respondents were shown how to complete the weekly paper diary. In the second sample, respondents were shown not only this paper tool, but also how to connect to the web questionnaire and use the app, allowing them to choose.

A methodological questionnaire was administered during the second visit to get feedback on the usability of new modes and their features.

3. RESULTS AND PRACTICAL APPLICATION

3.1. New collection tools allow to collect data in an efficient way

The first measure of the new tools' validity is the proportion of respondents who used them, their satisfaction and the quality of the data collected:

- 42 % opted to use the app instead of the paper diary;
- 40 % opted to use the web-based questionnaire.

Other multi-mode experiments have shown that use of the Internet questionnaire is lower when the link is given by paper than by e-mail [1]. This may therefore explain why the use of the web was not so widespread. The use of new tools was also age-dependent, particularly for the diary app.

During the second interview, respondents who used the new tools were asked about their ease of use. The app was rated "very easy" or "easy" by 95 % of users. Additionally, 97 % of those who used the web-based questionnaire found it easy or very easy to complete. Further information was collected on the respondents who did not use the tools to improve the acceptability in the future survey.

3.2. Drop-off rates are higher in the multimode protocol than in the reference protocol

One counter-intuitive finding from the 2024 test is that offering a wider choice of data collection tools to the respondents had a negative impact on the overall response rate.

While 51 % of sampled households completed the first interview (consistent across both samples), the proportion completing the second interview and diary dropped to 41% under the multimode protocol, compared to 46% under the reference protocol (Table 1).

Table 1. Response rate at each step of the test survey by protocol

	Global collection rate (N=2500)	Reference protocol (N=831)	Multimode protocol (N=1669)
Interview 1 completed	51 %	51 %	51 %
Interview 1 and 2 completed	45 %	47 %	44 %
Interview 1 + 2 + diary completed	43 %	46 %	41 %

Source: Insee, experimental HBS survey 2024

One explanation could be that the protocol including the new collection methods is more complex. Thus, its presentation at the end of the first visit might have confused respondents, increasing their perceived burden [2], leading them to abandon the survey.

Irrespective of the protocol, attrition decreases with age, confirming a classic result for survey participation rates (Figure 1). However, attrition is higher in the multi-mode protocol for all age groups: 5 points higher for the under-40s, 40-50s and 60s and over. The difference is very slightly higher for the 50-60 age group (6 points), but age does not seem to be a determining factor in discouragement for the second visit associated with the multi-mode protocol.

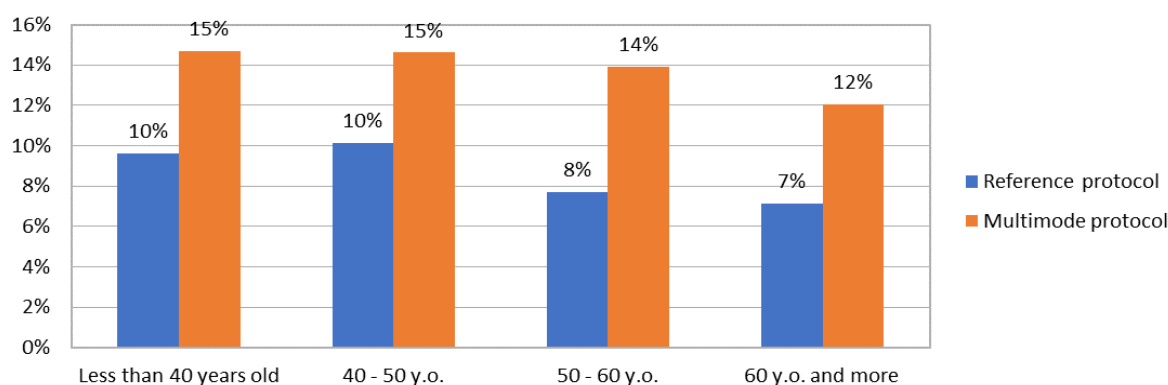


Figure 1. Attrition rate by age of the respondent and protocol

Note: Among people aged under 40 who responded to visit 1, 10% did not make the second visit under the reference protocol, and 15% under the multi-mode protocol.

Source: Insee, experimental HBS survey 2024

More generally, a logistic regression on the determinants of attrition within each sub-sample does not reveal many significant characteristics, apart from the fact of being unemployed. This characteristic is correlated to a higher probability of attrition in the multimode protocol.

Another global explanation may be that this multimode protocol also confuses the interviewers because of the complexity.

3.3. Data processing is more efficient with the app

In its paper version, the expense diary allows participants to manually record products or submit receipts. The app also enables the use of receipts by allowing participants to take a photo. In 2017, 80% of the expenses were reported via receipts in the diaries. It is then important to ensure the quality of the photos collected with the application

During the experimental survey, approximately 1000 photos were collected via the app, of which 83 % were complete and readable receipts with item details, 14 % were credit card receipts, 1 % were incomplete, and 2 % were blurry or unreadable. Since these photos are immediately available in the database, their processing is significantly easier compared to physical receipts collected in the paper diaries, which have to be collected at the household and sent to an operator in order to be scanned.

Manually entered expenses in the app are recorded via a list selection, allowing for direct coding of expenses within the COICOP classification. When the household could not find the product on the list, an addition could be made, which required recoding; however, this accounted for only 16% of the products reported manually in the test.

Data from the app was available in real-time, enabling validation and processing before fieldwork was completed.

4. MAIN FINDINGS

The web-based questionnaire and diary application proved to be a valid collection tool for the French HBS: more than 40 % of the respondent chose to use them, and the data collected was usable for the survey. Data-collection is faster and more efficient with the new tools, reducing interviewer burden and survey costs.

However, the higher drop-off rate among respondents offered the new collection tools indicates that not all respondents or interviewers were comfortable with increased choice, which could lead to screen which respondent are offered which tool. There is thus a trade-off between presenting a simplified protocol to the respondent and burdening interviewers with too many different scenarios.

5. REFERENCES

- [1] S. Chikhi, Du papier à internet: les apprentissages du multimode dans l'enquête mensuelle auprès des sortants de l'IAE. 12Ème Colloque francophone sur les sondages, Mar 2023, Aubervilliers (Campus Condorcet), France.
- [2] T. Yan and D. Williams, Response Burden – Review and Conceptual Framework *Journal of Official Statistics*, Vol. 38, No. 4, 2022, pp. 939–961,
- [3] Mixed-Mode: Past, Present, and Future. In : *Survey Research Methods*. [en ligne]. 13 août 2018. Vol. 12, n° 2, pp. 75-89.
- [4] Le multimode dans les enquêtes ménages: une collecte modernisée, un processus complexifié, F.Beck, L. Castell, S. Legleye, A. Schreiber, *Courrier des Statistiques* n° 7, 2022, pp. 7-28
- [5] Position paper on mixed-mode surveys, F. Beck, G. Brihault et al., statistical working paper, Eurostat, 2022

Municipal-Level Estimation of Tourism Perception: A Machine Learning-Based Approach to Small Area Estimation

Nina Niederhametner & Rebecca Daul (Statistics Austria)

1. INTRODUCTION

Understanding the perception and acceptance of tourism among local residents is crucial for sustainable tourism management. In this talk, we want to present Austria's new "Tourism Acceptance" survey and its methodological framework, which includes an innovative model estimating the level of tourism acceptance in Austria's municipalities by measuring how residents perceive the number of tourists in their home town. Drawing on a sample of approximately 12,000 respondents per year, the Tourism Acceptance survey asks participants about their opinions on tourism's impact on and significance for the economy, labor market and leisure activities in their place of residence, as well as whether they think the number of tourists is too low, excessive or acceptable. The sample is drawn from the Central Residence Register using a stratified random selection (1st level: federal province of residence, 2nd level: age, 3rd level: sex) and later extrapolated accordingly. Thus, the smallest regional breakdown of the Tourism Acceptance data, based on the sample alone, is on NUTS-2 level (federal provinces). However, tourism intensity and its impact on the local population varies in different regions of Austria. Consequently, there is a need to monitor tourism acceptance and perception in local communities on a smaller regional level than the federal province, but interviewing a representative sample in each region or municipality in Austria is neither affordable nor feasible. Therefore, we have developed a Small Area Estimation Model, which builds on possible tourism indicators that can influence tourism acceptance in the municipalities.

2. METHODOLOGY

To estimate the perception of the entire population, we propose to employ novel machine learning models, leveraging auxiliary data to impute the response variable for non-surveyed citizens on a unit level. We do this by linking each respondent's answer to auxiliary administrative data, including demographic information (age, sex, place of residence, income), employment sector (NACE classification), and municipal-level data (tourism-related profits, number of overnight stays per capita). These auxiliary variables are available for the entire country's population. An XGBoost [1] model is then trained on these auxiliary variables to predict a person's answer for the question "How do you personally perceive the number of tourists in your place of residence?". The model predicts a value of 0 for too few or acceptable number of tourists, and 1 for too many tourists. We then predict the answers of residents not included in the survey. This allows us to aggregate both the predicted estimates and, if applicable, the actual survey responses of all residents within the same municipality, providing a single acceptance estimate per municipality. Additionally, we use a bootstrap approach to estimate the errors, ensuring reliable confidence intervals for our estimates.

This approach parallels methods used in Small Area Estimation (SAE), extending the analysis beyond the sample to provide comprehensive estimates.

3. RESULTS AND PRACTICAL APPLICATION

The survey results offer valuable insights into tourism acceptance, highlighting spatial variations and underlying socio-economic factors influencing residents' perspectives on tourism. The findings are particularly relevant for policymakers seeking to balance tourism development with residents' quality of life. This model-based approach not only enhances our understanding of local tourism acceptance but also demonstrates the potential of integrating survey data with auxiliary information to provide detailed, population-wide insights.

4. MAIN FINDINGS

Since this is a new survey still under development, we are currently able to share only preliminary findings derived from the survey's pre-test, which was done in the third quarter of 2023. The final results for 2024 will be published in May 2025.

Our initial results align with previous studies published by the Austrian Federal Ministry for Labour and Economy [2]. We observe that residents in municipalities with higher tourist numbers tend to perceive the influx as excessive. Figure 1 illustrates the estimated percentage of residents in each municipality who feel that the number of tourists in their home region is too high. Notably, it is mostly residents in larger cities that attract tourists for their urban attractions, such as Vienna and Salzburg, as well as in rural areas known for their natural scenery and winter skiing, like Tyrol, who commonly perceive that the number of tourists is too high.

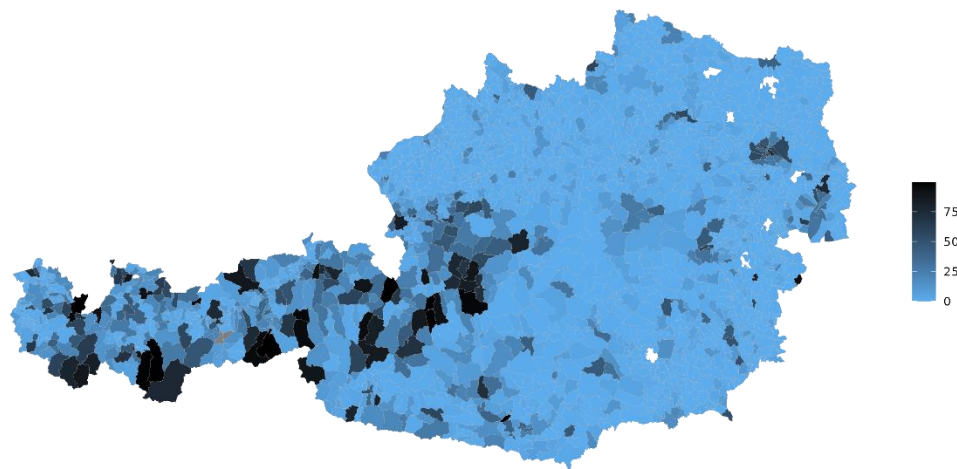


Figure 1. Estimated percent of residents perceiving the number of tourists as excessive.

5. REFERENCES

- [1] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
- [2] Austrian Federal Ministry for Labour and Economy. (2023). Tourism acceptance in Austria: Accessible report 2023. <https://www.bmaw.gv.at/Themen/Tourismus/tourismuspolitische-themen/nachhaltige-entwicklung/tourismusakzeptanz.html>

New challenges demand new solutions

Liina Osila (Statistics Estonia)

1. INTRODUCTION

In recent years, National Statistical Institutes (NSIs) have been operating in an environment characterized by dynamic data ecosystems. More and more, statistical offices are being asked to provide new services and take on tasks beyond their traditional role of creating and sharing official statistics.

As the growth of statistical institutions and the production of statistics increasingly depend on how stakeholders and users perceive them, a more user-centric approach should be adopted to meet their needs and expectations. This means focusing on decreasing the administrative burden and providing innovative data services. For that Statistics Estonia has put its emphasis on the increasing importance of data governance, constant development of data services and improvement of data literacy.

Statistics Estonia has declared this year the year we focus on saving data providers time by decreasing data providers administrative burden i.e. by making data submitting more convenient and by analyzing where it is possible to take into use administrative sources and big data instead of questionnaires. In addition, we seek opportunities how to provide added value to already collected data for reuse.

New diverse data sources put pressure to adopt innovative technical solutions for automatization, but they also illustrate the importance of data governance to ensure data interoperability, data quality and machine to machine data transfer. In addition, the opportunities of AI, should be further researched to seek ways to innovate and increase efficiency in statistics production.

Statistics Estonia has acknowledged the new challenges ahead and the importance of data governance as the foundation to compile fast and high-quality data services. Statistics Estonia is the state level coordinator for data governance.

To have a holistic and user-centric approach to data services Statistics Estonia is implementing service-based management that should enable better and faster service quality, but also demand more complex roles and skills from employees.

New Challenges in Official Statistics: Big Data Analytics and Multi-level Product Classification of Web-Scraped Data

Juliana de Freitas Ulisses Machado (EMOS Programme, University of Porto)

1. INTRODUCTION

The rapid growth of data generated on the web offers unprecedented opportunities to modernize data collection methods across various domains within official statistics. As more detailed and timely information becomes available through these non-traditional data sources, there is significant potential to enhance traditional statistical indicators by integrating high-frequency and granular data. However, much of this web-generated data is unstructured, presenting unique challenges for official statistics. Without additional processing, it is not easy to utilize such data for precise statistical analysis and reporting. Banco de Portugal had been collecting retail data daily through web scraping from several retail brands, but this data was initially not usable for economic research because the products lacked official classifications. This exponential growth of collected data underscored the need to understand the state-of-the-art in Big Data technologies and, mainly, to automatize the classification of the collected products according to the European Classification of Individual Consumption According to Purpose (ECOICOP), thereby enabling these databases to be used for economic research. Therefore, the primary objectives of this study are to conduct a comprehensive literature review on Big Data technologies, explore the use of web scraping in official statistics, and examine techniques for multi-level product classification. Additionally, the study aims to develop a machine learning-based classification pipeline to automatically categorize web-scraped retail data into 71 categories of ECOICOP related to food and beverage. In addition to developing a practical classification pipeline, this study addresses technical questions essential for refining model performance and applicability. Specifically, it investigates which machine learning models are most effective for short-text classification at this specific case, and whether language-specific large language models outperform cross-language models in this context.

2. OBJECTIVES

The main objectives of this study are as follows:

1. **Conduct a Comprehensive Literature Review:** Explore state-of-the-art technologies related to Big Data, investigate the use of web scraping in official statistics, and examine the latest techniques for short text classification to be used in this research.
2. **Utilize Iterative Labeling and Model Training Processes:** Given the absence of a labeled dataset, develop an iterative approach that begins with manually labeling a small dataset to train initial models. These models are then used to label larger datasets iteratively, improving the classification model's accuracy and efficiency over time.
3. **Develop a Machine Learning-Based Classification Pipeline:** Build a complete pipeline that applies Machine Learning techniques to automatically classify webscraped retail food and beverage data according to the European Classification

of Individual Consumption According to Purpose (ECOICOP). This allows the web scraped data to be used for economic research and CPI nowcasting.

4. **Determine the Most Effective Model for Short Text Classification:** Evaluate various machine learning models to identify which is best suited for the classification of food and beverage products using only the name and brand.
5. **Compare Language-Specific and Cross-Language Models:** Investigate whether large language models specifically trained for Portuguese, such as BERTimbau, demonstrate superior performance compared to cross-language models like XLM-RoBERTa for this classification task

3. METHODS

This research employs a two-phase methodology: a literature review focused on Big Data technologies and techniques for multi-level product classification, followed by the implementation of machine learning models for product classification

Phase 1: Literature Review

The initial phase involved conducting a comprehensive literature review to understand the current state of Big Data technologies. Subsequently, research was undertaken to examine the use of web scraping in official statistics, providing a deeper insight into this topic. Finally, an exploration of the most suitable techniques for multi-level product classification was conducted:

1. **Big Data Technologies:** The literature review explored frameworks and methodologies relevant to managing large-scale data, including data storage and processing techniques. The aim was to assess how these technologies could be utilized to handle the exponential growth of data volumes at Banco de Portugal.
2. **Web Scraping in Official Statistics:** This section of the review focused on the role of web scraping as a non-traditional data collection method within official statistics.
3. **Multi-level Product Classification Techniques:** The literature review examined various techniques for classifying products at multiple levels. This classification presents distinct challenges in natural language processing, particularly due to the characteristics of short-text classification, which often suffers from a lack of context.

Phase 2: Machine Learning Implementation

The second phase focused on developing and evaluating machine learning models to classify web-scraped food and beverage retail data according to 71 categories of European Classification of Individual Consumption According to Purpose (ECOICOP). A significant challenge encountered in this phase was the absence of a pre-labeled dataset, which necessitated a careful approach to building an effective classification system. The following key steps were implemented:

1. **Addressing the Lack of Labeled Data:** To overcome the challenge of unlabeled data, an iterative labeling process was established. A small dataset was obtained from a single day's worth of web-scraped data from one supermarket, which, although limited in size, provided a diverse sample of food and beverage products. This dataset included various product categories and brands, creating a representative foundation

for manual labeling. The labeled subset served as the starting point for training initial machine learning models.

2. **Data Preprocessing:** The dataset, provided by Banco de Portugal, consisted of food and beverage product titles, brands, and other relevant details scraped from various Portuguese supermarkets. Extensive preprocessing was conducted to clean and standardize the data, preparing it for classification. This involved removing noise, normalizing text, and tokenizing product descriptions.
3. **Model Selection and Training:** Various machine learning models were evaluated for their effectiveness in short-text classification. Traditional algorithms such as Support Vector Machines (SVM) and XGBoost were tested alongside advanced deep learning models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and large language models, including BERT, XLM-RoBERTa, and BERTimbau. The latter, specifically designed for the Portuguese language, leverages transfer learning to enhance performance in classification tasks involving Portuguese text. Additionally, pre-trained word embeddings specific to Portuguese were incorporated to ascertain whether this strategy could further improve the effectiveness of the deep learning models. The goal was to determine the most effective approach for accurately classifying the short product titles in the food and beverage category.
4. **Model Evaluation:** To determine the most effective model, the dataset was strategically partitioned, with 80% allocated for model training and the remaining 20% designated for testing. The evaluation metrics for model performance included accuracy and the F1 Macro score. Accuracy measures the proportion of correctly classified 5 instances among the total instances, while the F1 Macro score evaluates the balance between precision and recall across all categories. This distinction is especially important in this study due to the presence of 71 categories, some of which had fewer instances. This balance is crucial for ensuring that all categories are represented fairly in the classification process, as high accuracy alone can be misleading when dealing with imbalanced datasets. The time processing of each model was also considered.
5. **Iterative Labeling and Model Refinement:** After training the initial models on the manually labeled dataset, they were applied to datasets obtained from other supermarkets for a single day. An iterative process was employed to classify products from these supermarkets. The best-performing model from the initial evaluation was first used to classify products from the second supermarket, with its predictions reviewed and any inconsistencies manually corrected.

The corrected dataset from the second supermarket was then combined with the first to retrain the model, thereby enhancing its performance. This process was repeated for all six supermarkets, ensuring accurate ECOICOP category labels through expert review and correction. Ultimately, nearly 100,000 unique labeled products from all supermarkets were accumulated for this single day, allowing the final model to be retrained and used to label all products collected throughout 2022

4. RESULTS

The results revealed that the CNN and BERTimbau models were the top performers for this classification task. The CNN model presented a robust accuracy of 96.8% and an impressive F1 Macro score of 87.3%, achieved in only 2.1 minutes. In comparison, the Portuguese-specific BERTimbau model achieved a slightly higher accuracy of 97.3%; however, its F1 Macro score of 72.2% lagged behind that of the CNN model. Furthermore, the processing time for BERTimbau was significantly longer, clocking in at 192.5 minutes. These findings suggest that while BERTimbau excels in accuracy, the CNN model provides a more balanced

performance across categories, making it the preferred choice for short-text classification in this context. With regard to large language models, BERTimbau significantly outperformed the others in terms of accuracy, achieving a rate of 97.3%, and for the F1 Macro score, obtaining a rate of 72.5%. Cross-language models such as BERT and XLM-RoBERTa, on the other hand, exhibited lower levels of accuracy, scoring 94.9% and 92.2% respectively, along with diminished F1 Macro scores (63.9% for BERT and 58.9% for RoBERTa). The results clearly demonstrate the significance of language-specific large language models. BERTimbau achieved a remarkable F1 Macro score that was 12% higher than that of BERT, despite both models having the same architecture. This stark difference highlights the importance of language specific training in constructing highly efficient and effective language models

5. REFERENCES

- [1] Aluko, V., & Sakr, S. (2019, December). Big SQL systems: An experimental evaluation. *Cluster Computing*, 22, 1347–1377. <https://doi.org/10.1007/s10586-019-02914-4>
- [2] Aparicio, D., & Bertolotto, M. I. (2020). Forecasting inflation with online prices. *International Journal of Forecasting*, 36(2), 232–247. <https://doi.org/10.1016/j.ijforecast.2019.04.018>
- [3] Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., & Trunfio, P. (2022, December). Programming big data analysis: Principles and solutions. *Journal of Big Data*, 9. <https://doi.org/10.1186/s40537-021-00555-2>
- [4] Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation - Volume 6* (p. 10). USA: USENIX Association.
- [5] Faridoon, A., & Imran, M. (2021, November). Big data storage tools using NoSQL databases and their applications in various domains: A systematic review. *Computing and Informatics*, 40(3), 489–521. https://doi.org/10.31577/cai_2021_3_489
- [6] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- [7] Harchaoui, T. M., & Janssen, R. V. (2018). How can big data enhance the timeliness of official statistics? The case of the U.S. Consumer Price Index. *International Journal of Forecasting*, 34(2), 225–234. <https://doi.org/10.1016/j.ijforecast.2017.12.002>
- [8] Jahanshahi, H., Ozyegen, O., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2021). Text classification for predicting multi-level product categories. *arXiv*. <https://doi.org/10.48550/ARXIV.2109.01084>
- [9] Lehmann, E., Simonyi, A., Henkel, L., & Franke, J. (2020, December). Bilingual transfer learning for online product classification. In *Proceedings of Workshop on Natural Language Processing in E-commerce* (pp. 21–31). Barcelona, Spain: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.ecomnlp-1.3>
- [10] Raasveldt, M., & Mühleisen, H. (2020). Data management for data science - towards embedded analytics. In *Conference on Innovative Data Systems Research*.
- [11] Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In R. Cerri & R. C. Prati (Eds.), *Intelligent Systems* (pp. 403–417). Cham: Springer International Publishing.
- [12] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... Stoica, I. (2016,

October). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65. <https://doi.org/10.1145/293466>

New index for evaluation livestock farms digitalization

Colomba Sermoneta & Khadija Sossey Alaoui (Eurostat)

1. INTRODUCTION

Digitalisation is increasingly becoming one of the key elements to ensure the transition of production systems to higher levels of competitiveness [1]. In Europe, the European Commission has identified "fully connecting farmers and the countryside to the digital economy" as a key objective, aiming to foster a smarter, more modern, and sustainable future for food production and farming (European Commission, 2017, p. 7) [2].

To assess the impact of digitization on enterprises, specific indices have been defined to measure results and identify opportunities for improvement [3,4]. However, in the agricultural sector, particularly within the livestock industry, existing approaches for monitoring the effects of digital transformation on sector dynamics remain limited [5].

This paper points out that it is possible to create a specific, purposeful index that will give a realistic view on the degree of digitalization in each individual livestock farm. The indicator will be flexible enough over time to be adjusted by introducing or modifying new dimensions and easily calculable even by the farms themselves in order to proceed with a self-assessment.

2. METHODOLOGY

2.1. The survey

To calculate the initial indicators a specific module has been introduced in the survey on livestock. This survey is carried out every six months and detects the number of cattle, buffaloes, pigs and sheep and goats in livestock farms on 1 December and 1 June of each calendar year. The livestock farms were selected through a simple stratified random sampling method, with the farms being stratified by NUTS2 and livestock classes. This stratification was based on the total number of animals owned by each farm.

The assignment of sample units between layers is obtained by means of the Bethel algorithm which represents a generalization of Neyman's Optimal Allocation. This method ensures cost minimization and ensures that the Coefficient of Variation (CV) of the estimates of the variables of interest is less than or equal to specific pre-set thresholds which have been set at 0.7%. The variables used for the sample allocation (variables of interest) are: total number of cattle, total number of pigs, total number of goats and sheep. The sample consists of about 10,000 livestock farms.

2.2. Digitalization Index

Composite indicators are commonly used to evaluate effects in a relative context, as they can observe and explain complex multidimensional phenomena. A composite indicator of digitalization was calculated based on the "base" indicators collected in the survey.

The main indicators have been classified into sub-dimensions or "derived" indicators. Each of these sub-dimensions has been weighted according to its influence on the digitization process. The proposed digitalisation index essentially follows the structure of the IDSME is composed

of three (main) dimensions, each divided into sub-dimensions which in turn include one or more indicators as showed in Table1.

Dimensions, by their nature, must be related to have positive effects. For instance, it's difficult to imagine using the Internet without being connected to it. The dimensions are selected to represent logical units that can be compared with results from other indices, such as broadband subscription rates, online sales, etc. Due to the different dimensions of individual indicators, it is necessary to normalize their values. The data obtained from the study were analyzed using JMP v16 software for statistical analysis.

Table 1. Structure of digitalization index

DIMENSION	SUB-DIMENSION	INDICATOR	MIN	MAX
1. CONNECTION	1.a INTERNET CONNECTIONS (W=35%)	1.a1 BROADBAND CONNECTION	0	1
		1.b WEBSITE (W=20%)	0	1
	1.c SOCIAL NETWORK (W=10%)	1.c1 POSSESSION OF THEIR SOCIAL PAGES (FACEBOOK, INSTAGRAM, TIK TOK, ETC.)	0	1
		1.d DYNAMIC WEBSITE (W=25%)	0	1
	1.e INFORMATIVE WEBSITE (W=10%)	1.d1 ONLINE SELLING	0	1
		1.d2 WEB_PERSONALIZATION	0	1
		1.e1 WEB_INFO_PRODUCTS	0	1
		1.e2 WEB_TRACEABILITY_PRODUCTS	0	1
		1.e3 WEB_QUALITY SUSTAINABILITY SAFETY	0	1
		1.e4 WEB_OTHER_SERVICES	0	1
2. USE OF DIGITAL TECHNOLOGY	2.a BASIC CLOUD (W=35%)	2.a1 USE OF CLOUD_COMPUTING SERVICES	0	1
		2.a2 CLOUD_EMAIL	0	1
		2.a3 CLOUD_OFFICE SOFTWARE	0	1
		2.b ADVANCED CLOUD(W=65%)	0	1
		2.b1 CLOUD-HOSTING	0	1
		2.b2 CLOUD_STORAGE	0	1
		2.b3 CLOUD_FINANCE SOFTWARE	0	1
		2.b4 CLOUD_CRM SERVICES	0	1
		2.b5 CLOUD_OTHER_SERVICES.	0	1
3. USE OF PRECISION ANIMAL FARMING SYSTEMS	3.a PLF BASIC (W=25%)	3.a1 USE OF PLF	0	1
		3.a2 MILKING ROBOTS	0	1
	3.b PLF ADVANCED(W=75%)	3.b1 INFORMATION_SYSTEMS	0	1
		3.b2 FOOD QUALITY TOOLS	0	1
		3.b3 MILK QUALITY MEASUREMENT	0	1
		3.b4 REMOTE SENSORS	0	1
		3.b5 DEVICES TO ANIMAL'S BEHAVIOURS	0	1
		3.b6 DEVICES FOR THE HERD MANAGEMENT	0	1
		3.b7 OTHER_SIST_PRECISION	0	1

Not all indicators have the same importance, so weighting is required. One of the main challenges lies in determining the appropriate weight values. Initial weights are typically based on literature, personal experience, and assumptions. The hypothesis is that, once processed and properly weighted, these dimensions can generate a digital readiness index for livestock farms.

3. RESULTS AND PRACTICAL APPLICATION

Factor Analysis has been used to define the weights that the level I indicators have in the determination of the readiness index. The weights to be attributed to the first level for the calculation of the synthetic indicator are reported in the Table

Table 2. Weight assigned to each dimension.

DIMENSION	WEIGHT
USE OF DIGITAL TECHNOLOGY	51%
USE OF PRECISION LIVESTOCK FARMING	46%
CONNECTION	33%

Therefore, the following formula was used for the i-th livestock farm:

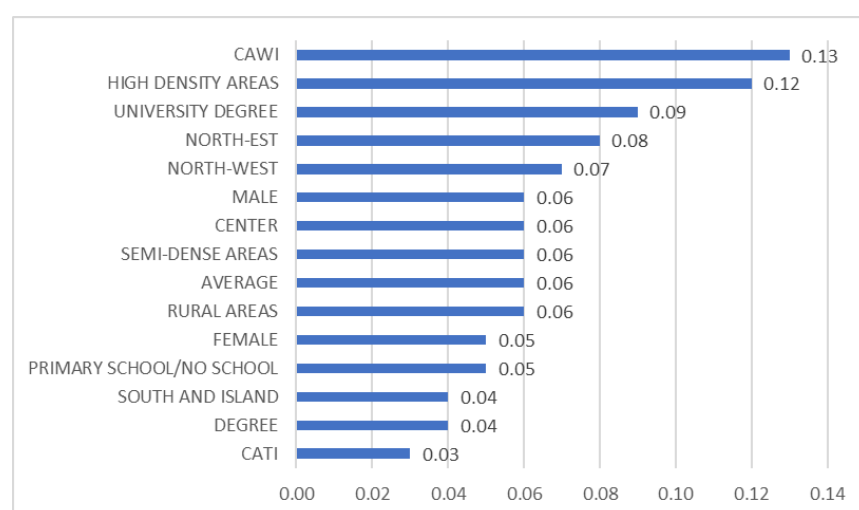
$$Index = (Connection \times 0.33) + (Use\_Digital\_Technologies \times 0.51) + (Precision\_Livestock\_Farming \times 0.46)$$

Finally, it was necessary to normalize the units before aggregation. Normalization is performed using the min-max method:

$$XN = \frac{(X - xmin)}{(xmax - xmin)}$$

In this way, the normalized value of the indicator varies between 0 and 1. The index for 2020 assumes an average value of 0.06. The assessment of the situation at national level is represented (Figure 1).

Figure 1. DIGITALIZATION INDEX BY BUSINESS CHARACTERISTICS OF AGRICULTURAL HOLDING



Agricultural holdings that sent the questionnaire using CAWI methods are more digitized than those using CATI.

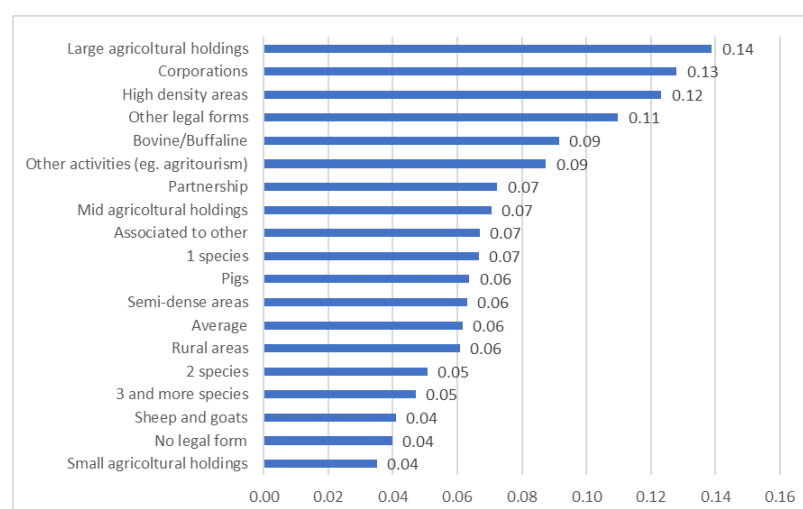
The territory is representative of the phenomenon,

confirming the digital divide between the northern, central and southern areas of the country. Using the

degree of urbanization as an indicator, the index assumes a value above or like the average in densely populated or intermediate-density areas, adjacent to large cities or small urban centres where there is no connection problem. Conversely, this is not the case in rural areas. To confirm the hypothesis of a link between territory and digitization, a test was carried out.

The chi-square test is statistically significant ($p\text{-value} < 0.05$), so we reject the null hypothesis of independence between the degree of digitization and the area, concluding that there is a relationship between the two. Finally, it should be noted that the digitization of livestock farms is strongly linked to the profile of the farm manager. In particular, if the company is run by a man and the educational qualification is high, the propensity to run digitized companies increases. Variables related to structural characteristics also strongly influence digitization (Error! Reference source not found.):

Figure 2. DIGITALIZATION INDEX BY AGRICULTURAL HOLDING BY FARM STRUCTURE



Large agricultural holdings are digitized almost four times more than small family-run companies which, on the contrary, have a very low propensity for computerization. sole proprietorships are the least likely to digitize compared to corporations or partnerships. This agrees with the size of the company, as corporations are larger than other types with an average size of around 5 employees. In addition, farms that keep

cattle and buffaloes are the most digitally advanced, followed by the farms that keep pigs and at the tail end the farms with sheep and goats. Furthermore, those who breed only one species of livestock are clearly more specialized and tend to adopt more technology for the management of the farm, which consequently leads to greater digitization. Finally, livestock farms associated with other companies or organizations and carrying out other remunerative activities tend to be much more digitized than others.

4. MAIN FINDINGS

The main results show that the broadband connection is still limited in rural areas. The most digitalized livestock farms are big and located in the Northern territory of Italy, in high density areas, are corporation, and the owner is male with a university degree.

The results were obtained using an “ad hoc” module within an existing survey, demonstrating how the burden on respondents and the cost of conducting a specific survey can be significantly reduced.

“..... New indicators that allow users to determine to what extent births of new firms are driving innovation would also be useful dynamics indicators” (Organisation for Economic Co-operation and Development-Eurostat, 2005) [6].

The indexes based on the ranks have led in a simple and intuitive way to the same results obtained with more complex statistical techniques (logistic regression, analysis of the main components, etc.) and must be largely used to simplify the work of operators or in any case non-experts in the livestock sector. These indices can also be used to detect other phenomena that vary very quickly over time such as those related to the Sustainable Development Goals or the agri food chain.

5. REFERENCES

- [1] O. Korzhyk, J. V. Gomes, and G. Jo~ao. A comparative study of different digitalization indexes. In *Effective AI, Blockchain, and E-Governance Ap-plications for Knowledge Discovery and Management*, pages 238–267. IGI Global, 2023.
- [2] J. Peterson. Juncker’s political european commission and an eu in crisis. *JCMS: Journal of Common Market Studies*, 55(2):349–367, 2017.

- [3] M. Bogavac and Z. Ćekerevac. Idsme index–new method for evaluation of smes digitalization. MEST Journal, 7(2):9–20, 2019.
- [4] A. De Carolis, M. Macchi, E. Negri, and S. Terzi. Guiding manufacturing companies towards digitalization a methodology for supporting manufactur-ing companies in defining their digitalization roadmap. In 2017 international conference on engineering, technology and innovation (ICE/ITMC), pages 487–495. IEEE, 2017.
- [5] C. Sermoneta. La diffusione delle tecnologie nelle aziende zootecniche – Statistica Report Istat (2020)
- [6] V. Gineviciene and M. Tvaronaviciene. Trends and level of development: view to new eu members. Journal of Business Economics and Management, 6(2):113–121, 2005.

New results for time-dependent models: a perspective for seasonal adjustment

Guy M  lard (Universit   libre de Bruxelles)

1. INTRODUCTION

In a few recent papers, ARIMA models with time-dependent (td) coefficients varying in a deterministic way have been investigated [1, 2, 3, 4]. Let us call them tdARIMA. Except for that time dependency, the modeling is very similar to what is done in JDemetra+ or X-13ARIMA-SEATS with the possibility of regressors, detection of outliers, etc. In [5], the authors presented a study of U.S. industrial production monthly time series where it was shown that these time-dependent models fit better the data, at least for one-half of the series. There is another approach in the literature based on locally stationary processes, due to Dahlhaus [6], where the asymptotics are inline, see an application to (presumably seasonally adjusted) official data in [7]. The problem with that approach is that it is incompatible with seasonality whereas the tdARIMA approach is compatible with seasonality. It had rather complex conditions but, following the path started in [8] for models with constant coefficients, simpler sufficient conditions based on the roots of the tdAR and tdMA polynomials can be used instead. This opens the door to new advancement. In this presentation, we will propose a generalization of SEATS [9, 10, 11, 12, 13] based on the tdARIMA models to obtain a seasonal adjustment method. That generalization was never considered by the authors of tdARIMA models. At this stage, there will not be a full implementation, only a list of steps to be accomplished and difficulties to solve.

2. METHODOLOGY

This proposition is based on a relation between tdARIMA models, on one side, and seasonal adjustment based on signal extraction.

2.1. tdARIMA models

Let $\{y_t, t = 1, \dots, n\}$ be a time series of length n . A tdARIMA(p, d, q) model is defined by the equation

$$\nabla^d y_t = \sum_{k=1}^p \phi_{tk}(\beta) \nabla^d y_{t-k} + e_t - \sum_{k=1}^q \theta_{tk}(\beta) e_{t-k}, \quad (1)$$

Contrary to the standard case, the coefficients $\phi_{tk}(\beta)$ and $\theta_{tk}(\beta)$ are here time-dependent, meaning that they are a function of time t and a (small number) parameters β . Let $\phi_t(\beta) = 1 - \sum_{k=1}^p \phi_{tk}(\beta) L^k$ and $\theta_t(\beta) = 1 - \sum_{k=1}^q \theta_{tk}(\beta) L^k$ be, respectively, the (time-dependent) autoregressive (AR) and moving average (MA) polynomials. Usually, there is also a seasonal difference operator ∇_s , where s is the seasonal period. The AR and MA polynomials can each be factorized into regular and seasonal polynomials.

Let us consider a simple example taken from the dataset used in [7] of U.S. industrial production time series, more precisely series cX5421T of business supplies except energy, motor vehicles & parts, etc. The final model obtained is defined by

$$(1 - \phi_1 L - \phi_2 L^2) \nabla \nabla_s y_t = (1 - \theta_{t,12} L^{12}) e_t, \quad (2)$$

where $\theta_{t,12} = \theta'_{12} + \theta''_{12}(t - 193)$, with the following values for the parameters estimates (with the Student t values): $\phi_1 = 0.192$ (3.7), $\phi_2 = 0.223$ (4.3), $\theta'_{12} = 0.545$ (12.3), and $\theta''_{12} = 0.00172$ (4.2). The main difference with respect to the standard constant ARIMA models is that the roots of the polynomials can now be time-dependent, like the MA polynomial here.

2.2. Seasonal adjustment based on signal extraction

Gómez and Maravall [9] have proposed a method of seasonal adjustment based on ARIMA models and signal extraction. That method called SEATS for Signal Extraction in ARIMA Time Series is based on signal extraction. The series is considered (after a suitable transformation) as a sum of up to four components, a permanent (trend) component, a seasonal component, a transitory component, and noise. The models for the first three components are obtained from the global ARIMA model and the last one corresponds to a white noise.

3. RESULTS AND PRACTICAL APPLICATION

The idea is to extend the signal extraction procedure to tdARIMA models. For that purpose, the differences and the AR polynomial are combined in a full AR polynomial. The extension is not straightforward at all for several reasons: first, the starting point is the factorization of the full AR polynomial; second, the AR part of the components is based on that factorization, especially the roots; third, the method is heavily based on (generalized) spectral arguments, as well for the principles than interpretation. Of course, we can consider time-dependent regular and seasonal polynomials but the roots will be time-dependent, like in the case of the example in Section 2, making it difficult to apply strict threshold limits for the attribution of the roots to the components.

4. MAIN FINDINGS

We describe a tentative extension of SEATS for tdARIMA models both conceptually and on a few examples, like the case of a simple tdARIMA(0,1,1) non-seasonal model, a time-dependent extension of the airline model, and the case of the model presented in (2).

The idea is to adapt, as much as possible, the steps of the SEATS method to tdARIMA models. Consequently, the component models will be time-dependent in general and the estimation of the components will also be obtained using a tdARIMA model. There will be some difficulties like the attribution of the roots to the components and the justification of the Wiener-Kolmogorov filter.

It is not possible to show a complete implementation at this time. The extension will take the form of Excel files where all the computations will be done. There is also nothing about the diagnostics, sliding spans and revisions, and the final estimation error.

5. REFERENCES

- [1] R. Azrak and G. Mélard (2006) Asymptotic properties of quasi-likelihood estimators for ARMA models with time-dependent coefficients. *Statistical Inference for Stochastic Processes* 9: 279--330.
- [2] A. Alj, K. Jónasson, and G. Mélard (2016) The exact Gaussian likelihood estimation of time-dependent VARMA models. *Computational Statistics and Data Analysis* 100: 633--644.

- [3] A. Alj, R. Azrak, and G. Mélard (2024) General estimation results for tdVARMA array models. *Journal of Time Series Analysis*, forthcoming.
- [4] R. Azrak and G. Mélard (2022) Autoregressive models with time-dependent coefficients – A comparison between several approaches. *Stats* 5: 784–804.
- [5] G. Mélard (2023) ARMA models with time-dependent coefficients: official statistics examples. In *Time Series Analysis – New insights*, Abdalla R, El-Diasty M, Kostogryzov A, Makhutov N (eds.). IntechOpen, London: 18–35.
- [6] R. Dahlhaus (1997) Fitting time series models to nonstationary processes. *Annals of Statistics* 25: 1–37.
- [7] R. Killick, M. I. Knight, G. P. Nason, M. A. Nunes, and I. A. Eckley (2024) Automatic locally stationary time series forecasting with application to predicting UK gross value added time series. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 2024, qlae043: 1–16.
- [8] G. Mélard (2022) An indirect proof for the asymptotic properties of VARMA model estimators. *Econometrics and Statistics* 21: 96–111.
- [9] V. Gómez and A. Maravall (2001) Seasonal adjustment and signal extraction in economic time series. In: D. Peña, G. C. Tiao, and R. S. Tsay (eds) *A course in time series analysis*. Wiley, New York: 202–246.
- [10] G. L. Mazzi and D. Ladiray (2018) *Handbook on Seasonal Adjustment*, Luxembourg: Publications Office of the European Union.
- [11] European Union (2024) *ESS Guidelines on Seasonal Adjustment*.
- [12] B. C. Monsell (2009) Update on the Development of X-13ARIMA-SEATS, <https://www.census.gov/library/working-papers/2009/adrm/monsell-01.html>.
- [13] K. Webel (2013) Time series-dependent selection of an appropriate seasonal adjustment approach, *Proceedings 59th ISI World Statistics Congress*, 25-30 August 2013, Hong Kong.

Nowcasting gas demand based on timely available data sources

Maximilian Ofner, Daniel Strenger (Statistics Awards for Nowcasting, Graz University of Technology)

1. INTRODUCTION

The amount of natural gas consumed is a critical metric for economists and policymakers as it serves as both an indicator of industrial activity and a proxy for greenhouse gas emissions. High levels of gas consumption often signal increased industrial production, energy demand, and economic growth, offering valuable insights into the economy's performance. At the same time, natural gas, is a significant source of carbon dioxide and methane emissions. Tracking its consumption helps policymakers assess progress toward climate goals, design effective energy policies, and balance economic growth with environmental sustainability.

Given its significance, the timely availability of official natural gas consumption data is essential. Eurostat provides such data through the indicator *Inland gas consumption – monthly data* (NRG_CB_GASM) [1]. Currently, these figures are gathered by reporting countries using separate, dedicated questionnaires, which are then aggregated and submitted to Eurostat. However, this process introduces a delay of approximately two months before the indicator is published, limiting its usefulness for real-time decision-making and rapid policy responses.

The European Statistics Awards Energy Nowcasting Challenge – GAS [2] sought to identify innovative methodologies and external data sources to enhance the timely availability of official gas consumption figures. Teams from around the globe were invited to submit their estimates of monthly gas consumption for selected countries within the European Statistical System. Participants had the freedom to choose their data sources and methods. The special challenge of nowcasting is to provide the estimates without delay – the submission for each month had to be made at the last day of that month.

We describe our method of estimation that provided the most accurate estimates among all submissions and in addition was awarded the Reproducibility Award by a panel.

2. METHODOLOGY

We use data from the European Network of Transmission System Operators for Gas (ENTSOG) [3]. More specifically, ENTSOG publishes the amount of natural gas that passes through certain stations in the transmission system. This data is available in high resolution and with a delay of usually less than a day and therefore well suited for nowcasting purposes.

Since natural gas passes multiple stations – possibly in various countries – before being actually consumed, to identify the specific stations within each country that provide meaningful information on actual consumption. In most countries, several *distribution* and *final consumer* points exist. These points represent stations, where gas is distributed among smaller units (e.g., households) and single consumers of large amounts of gas (e.g., large industrial sites), respectively.

On the last day of each month, we downloaded and aggregated the flow data from all relevant stations in each country. To account for the final hours or days of the month that were unavailable at the time of calculation, we scaled the aggregated values to ensure accurate estimates.

In principle, estimating total gas flow to distribution and final consumer points by using the average flow from the available days should yield reasonably accurate estimates of total consumption. However, this approach may not fully account for systemic factors such as gas losses during transport or unidentified consumption sites. To address these potential gaps, we fitted a linear regression model for each country, using the aggregated flows described above as predictors and the official gas consumption indicator, NRG_CB_GASM, as the response variable.

3. RESULTS

The linear model shows an excellent fit with an R^2 -value in excess of 0.99 for many countries. The mean squared relative error over several countries, that was used as an evaluation metric for the nowcasting challenge was well below 0.001 in all but one month.

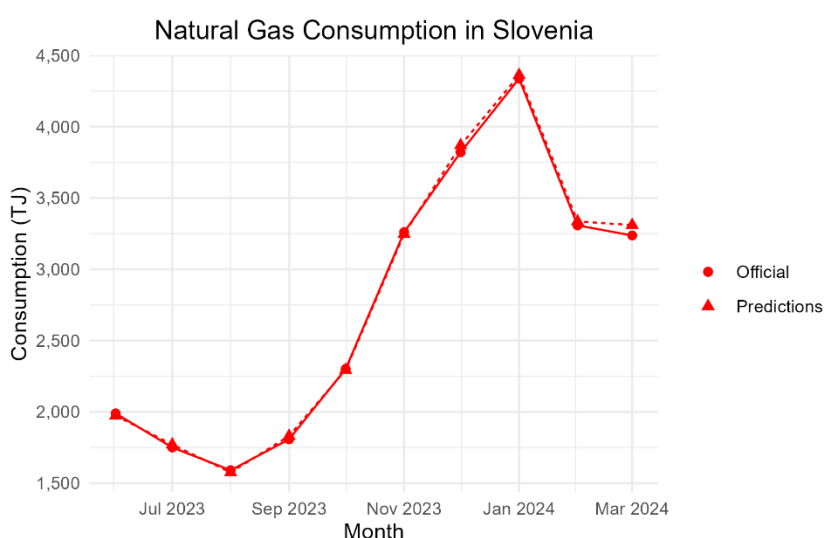


Figure 4. Nowcasts compared to official values of NRG_CB_GASM for Slovenia

4. MAIN FINDINGS

The simplicity and strong performance of our method highlight two key insights:

1. The importance of data sources: Identifying and selecting appropriate data sources is often as critical as choosing the right model. This is particularly true for nowcasting tasks, where the timely availability of data is essential for accuracy and relevance.
2. The effectiveness of simple models: Simple models, such as linear regression, frequently deliver reliable predictions. Their straightforward structure ensures robustness and minimizes the risk of overfitting, making them well-suited for many real-world applications.

We conclude that similar approaches hold significant promise for improving the timeliness of official statistics in various domains. By leveraging data that is strongly correlated with the variable of interest — especially when such data is already collected for legal or physical reasons — these methods can provide practical and scalable solutions for real-time estimation.

5. REFERENCES

- [1] Eurostat, Supply, transformation and consumption of gas - monthly data, https://ec.europa.eu/eurostat/databrowser/view/nrg_cb_gasm/default/table?lang=en
- [2] Eurostat, European Statistics Awards Energy Nowcasting Challenge – GAS, <https://statistics-awards.eu/competitions/6>
- [3] ENTSOG, Transparency platform, <https://transparency.entsog.eu/#/map>

One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics: Reflections on the First Year

Eimear Crowley (Central Statistics Office, Ireland)

1. INTRODUCTION

1.1. Background

April 2024 saw the launch of a four year ESSnet project named 'One-Stop-Shop for Artificial Intelligence and Machine Learning for Official Statistics' (AIML4OS). The project, which involves 16 countries, has an overall aim of developing a single entry point for ESS staff to access a suite of resources, methodologies, guidelines and use cases for implementing AI/ML based solutions in the production of official statistics. The project also aims to build communities around open-sourced solutions which will share knowledge and develop innovative solutions. [1]

1.2. Key Objectives

The key objectives of this project are to:

- Develop, maintain and evolve a coherent set of relevant capabilities – including methodologies, guidelines, sandboxes, labelled data, processes, methodological, implementation and quality frameworks for implementing AI/ML based solutions in official statistics across the ESS,
- Set up a platform/hub providing a single entry point for ESS staff to access relevant capabilities,
- Provide support and guidance for the integration and maintenance of relevant AI/ML based solutions in ESS organisations through training and active and efficient support,
- Build communities around open-source solutions developed and maintained by ESS members,
- Share ideas, experiences, success stories and lessons learned to stimulate innovation based on the use of AI/ML and
- Enable and facilitate the transition from development and experimentation of AI/ML based solutions to actual production.

2. METHODOLOGY

2.1. Project Structure

The AIML4OS project is broken down into 13 work packages (WPs). There are six support WPs (WP1-6) and seven use cases (WP7-13) which are listed below.

WP	Name
1	Project Management and Coordination
2	Communication and community engagement
3	ESS AI/ML lab: Technical infrastructure and organisational setup
4	AI/ML state-of-play and ecosystem monitoring
5	Standards, methodological and implementation frameworks
6	Knowledge repository and training material
7	AI/ML on earth observation data, satellite imagery
8	Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on editing
9	Statistically valid and efficient editing and imputation in official statistics by AI/ML – with a special focus on imputation
10	From Text to Code – Experiences and Potentials of the Use of AI/ML for Classifying and Coding
11	Use Case Applying ML for estimating firm-level supply chain networks
12	Use of generative large language models in statistics
13	Synthetic data - SDC issues introduced by using AI/ML

3. RESULTS AND PRACTICAL APPLICATION

Deliverables planned for submission during Year 1 of the project are described in Table 1 below. These will be discussed along with any other significant work completed as part of the project.

Table 1. Deliverables planned for submission during Year 1.

WP	Deliverable	Description	Submission date
3	D3.1 - ESS AI/ML Lab	An AI/ML platform available on internet and suitable to train AI/ML models on nonsensitive data. The main application will be available at least in Dutch, English, Finnish, French, German, Italian and Norwegian.	30 June 2024
3	D3.2 - Deployment of open resources with the ESS AI/ML Lab	This website (in English) will offer a technical documentation on how to deploy open resources using the ESS AI/ML Lab.	31 December 2024

4. MAIN FINDINGS

This presentation will reflect on the first year of the AIML4OS project. Deliverables achieved and tasks completed to date will be discussed. The key objectives of this project, as outlined in Section 1.2 of this abstract, will be also considered when reflecting upon the first year of the project.

5. REFERENCES

- [1] One-stop-shop for Artificial Intelligence and Machine Learning for Official Statistics (AIML4OS). Available at: <https://cros.ec.europa.eu/dashboard/aiml4os> (Accessed 24 September 2024).

Open Infrastructure and Open Source Datalab to Strengthen Collaboration and Innovation

Romain Lesur & Frédéric Comte (Insee)

1. INTRODUCTION

National Statistical Institutes (NSIs) have a long history of cooperation and joint work on statistical methods and standards. However, in an environment characterised by a number of technological breakthroughs and the development of numerous public and private data sources, a fundamental shift in how we approach collaboration and technology is needed in order to develop common solutions and products. This presentation explores a new paradigm that uses open source tools, collaborative infrastructures, and cross-institutional cooperation to advance data science in official statistics.

The main message of this presentation is the importance of collaboration as a driver of innovation. Within statistical institutes, the relationship between statisticians, data scientists and developers is crucial. Equally important is the promotion of collaboration between NSIs and public sector bodies. These relationships allow institutions to pool resources, share knowledge, and develop solutions that can be applied to a wide range of contexts. Collaboration not only accelerates innovation but also ensures that developments are widely accessible and benefit a wider community.

2. OPEN INFRASTRUCTURE TO EASE COLLABORATION

To facilitate this collaborative environment, we have built an open infrastructure based on the Onyxia open-source platform (<https://datalab.sspcloud.fr>). This infrastructure allows statistical institutes to collaborate on non-sensitive data in a flexible and scalable way. One major characteristic of this platform is that it is DevOps oriented, allowing statisticians, data scientists, ML engineers and developers to collaborate and prototype end-to-end pipelines and data-centric products.

Onyxia can be installed on-premises, ensuring that institutions retain full control over their data while leveraging the benefits of open-source innovation. This dual capability allows statistical institutes and public sector organizations to work together seamlessly while respecting data privacy and security concerns.

3. PRACTICAL EXAMPLES

A key element of this open infrastructure is the ability to easily share results and solutions across platforms and organisations. This is exemplified by the development of Lomas, a differential privacy platform developed by the Swiss NSI, *Bundesamt für Statistik* (BFS). The development of Lomas has adopted the DevOps and agile principles, resulting in a production-ready product. Lomas demonstrates the flexibility of our platform, as it can be deployed on any similar infrastructure, and provides a blueprint for how open infrastructure solutions can be developed collaboratively and scaled effectively across the public sector.

4. TAKE-AWAY MESSAGES

This presentation also looks at the broader context of open source software in the official statistic. As the demand for transparency and efficiency increases, open source solutions offer a crucial way for NSIs to innovate without being restricted by proprietary software. By embracing an open ecosystem, NSIs can share methodologies, tools, and best practices more widely, leading to faster adoption of innovative solutions across the statistical community.

Finally, the presentation will argue that open infrastructure and open source tools are key enablers of collaboration and innovation in the public sector. This paradigm not only addresses the challenges of limited resources and siloed operations but also enhances the ability of statistical institutes and other public organisations to respond to the evolving needs of society. By working together on shared platforms such as Onyxia, we can develop solutions that are both scalable and sustainable, driving forward the next generation of official statistics and public sector data science.

Open source software sharing through the awesome list of official statistics software

Olav ten Bosch & Mark van der Loo (Statistics Netherlands)

1. INTRODUCTION

The use of mature open source software for the production of official statistics has many advantages. It saves costs, adds to quality, creates flexibility and may have a positive effect on standardisation. However, it may be difficult to know what software is already available, what their maturity is, and what can be used for which tasks. Hence, it is crucial to share knowledge on existing open source software solutions among statistical organisations.

In this presentation we explain the “*awesome list of official statistics software*” created in 2017 to share knowledge on open source software in official statistics. We describe its origin, its growth, its current status and some of the spin-offs of this list, such as the experiences flowing into the ESS principles on open source software. Moreover, we touch upon possible extensions and improvements.

2. THE AWESOME LIST OF OFFICIAL STATISTICS SOFTWARE EXPLAINED

Software sharing is not new. The exchange and reuse of statistical software among statistical institutes has proven to be valuable for long in domains such as statistical disclosure control, data editing, data collection, and statistical dissemination. In these domains a few well-known solutions have been used by many offices. However, these days the software landscape for official statistics is much more complex and dynamic. Numerous small software packages with specialized statistical functionality are being developed and continuously improved. This makes it much harder to maintain a good overview of the statistical software landscape and increases the risk of redeveloping solutions that already exist. Hence, there is a need to spread knowledge about the existence and use of generic official statistics software.

A popular way of maintaining such knowledge is the so-called awesome list concept [1]. In 2017, during the UNECE Statistical Data Editing Conference, this concept was adopted to create the so-called “[awesome list of official statistics software](#)” [2]. The initial goal was to remember the software presented at that specific conference, but over time the list was extended with many other software presented at conferences or suggested by members of the statistical community. This made the list essentially a community approach to facilitate open source software sharing.

The list reached 100 contributions by 2019, and currently lists 135 open source packages that are easy to download and install, have at least one stable release, and are used in statistical production in at least one Statistical Office. Packages that facilitate automated access to Official Statistics output are included as well. The list itself has a Creative Commons license, is developed in the spirit of open source and receives many contributions from collaborators internationally. Each item has a link to the software download and a short description. The items may come from different communities, may have been developed in different programming languages and may be distributed via different package systems. Despite their different origin, every item on the list is accompanied with up to three badges showing the

latest version, the last commit and the license. These are automatically derived from the packaging system metadata. Figure 1 shows some examples of entries on the list.

- GitHub v.2.1.1 last commit june 2021 license GPL-3.0
 Python [Social-Media-Presence](#). A script for detecting social media presence on enterprises websites. By Statistics Poland.
- CRAN 1.1.5 – 7 months ago license GPL-3
 R package [validate](#). Data validation checks such as on length, format, range, missingness, availability, uniqueness, multivariate checks, statistical checks and checks on SDMX codelist. See [Cookbook](#). By Statistics Netherlands.
- GitHub v2.2.5 last commit last saturday license EUPL-1.2
 Java application [JDemetra+](#). The seasonal adjustment software officially recommended for the European Statistical System.
- GitLab v24.1.0 last commit today license MIT License
 Node.js and other [Stat Suite](#). An SDMX-based platform to build tailored data portals, topical or regional data explorers, or lightweight reporting platforms. [Documentation](#). By [SIS-CC](#).
- CRAN 2.1.3 – 8 months ago license GPL (>= 2)
 R package [simPop](#). Simulation of synthetic populations from census/survey data considering auxiliary information.

Figure 1. Examples of items on the list

To give the user an indication of the use of each element on the list, it is organized according to the Generic Statistical Business Process Model (GSBPM). Figure 2 shows how the 135 items are distributed across GSBPM processes.

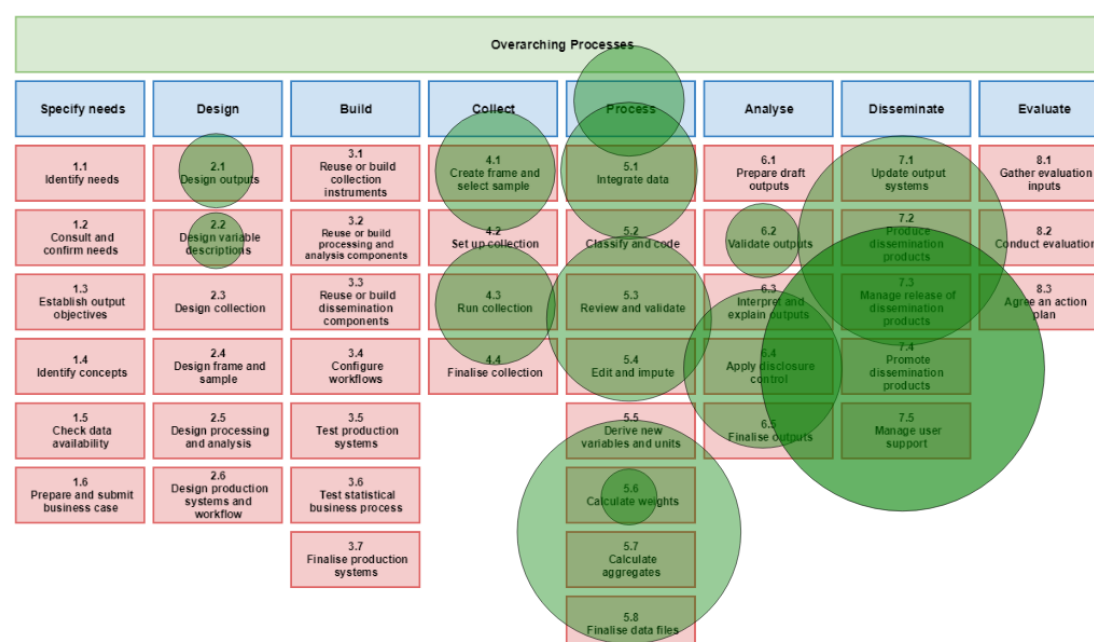


Figure 2. Software on the list by GSBPM process phase

The presentation of the list is derived from a machine readable YAML file. This information, including pointers to the package repos, in combination with package system APIs, allows for automatically deriving statistics on the official statistics software landscape. Figure 3 shows the distribution of programming languages of items on the list. The vast majority of items is written in R, which shows the excellent software sharing methods in this community. Figure 4 shows the licenses used on the list. GPL is the most popular license followed by MIT and EUPL.

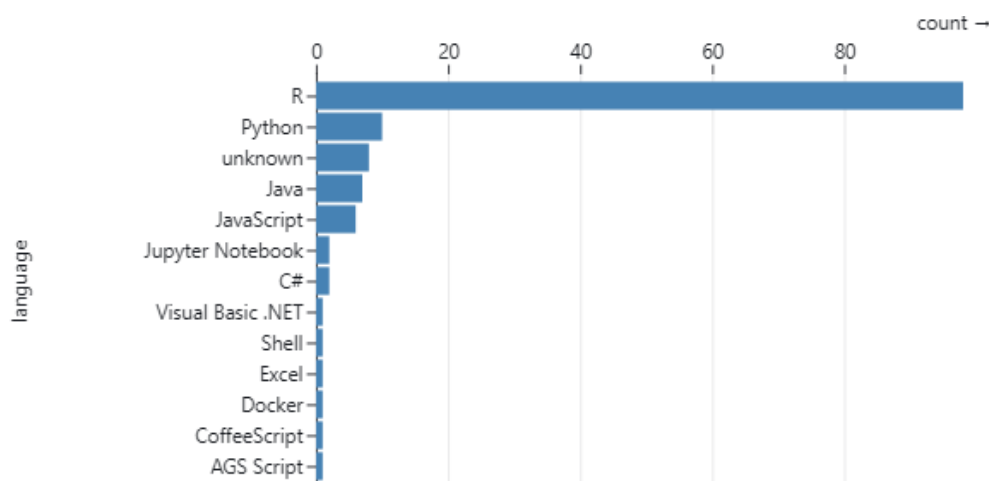


Figure 3. Programming languages used on the list

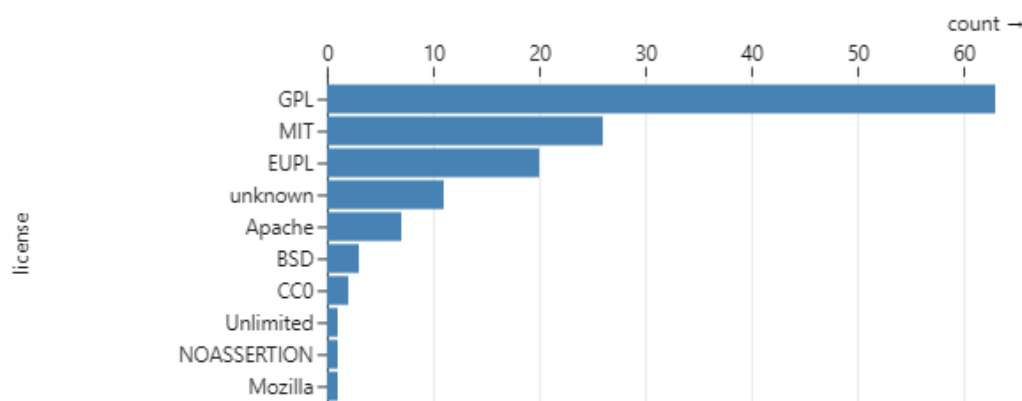


Figure 4. Licenses used on the list

3. THE LIST USED IN THE ESS PRINCIPLES ON OPEN SOURCE SOFTWARE

There have been some spin-offs of the awesome list initiative. First of all the experiences in the official statistics open source community have been written down into a list of best practices, which found their way into the *ESS principles on open source software* [3], put together by the 'group on Open Source for Official Statistics (OS4OS)', a number of statistics offices together with Eurostat and OECD. They put together seven principles that reflect the way of working of successful software projects found on the awesome list. For completeness we repeat the principles here:

1. OSS by default: in the production of official statistics we prefer the use of open source software solutions over closed software solutions. Moreover, we share our software solutions as open source.
2. Work in the open: We start our projects in the open from the beginning and clearly mark maturity status.
3. Improve and give back: We rather improve existing open source solutions than decide to create new solutions and we give our improvements back to the respective open source community.
4. Think generic statistical building blocks: In our open source work we strive for re-usable generic functional building blocks that support well-defined methodology in statistical processes.
5. Test, package and document: We test, package and document our open source software for easy-re-use.
6. Choose permissive: We choose the most permissive OS license possible for sharing our software.
7. Promote: We invest in promoting new developments or improvements on our open source software within the ESS community and where applicable in a wider context.

One notable other spin-off is the adoption of concept by the NTTS 2021 organisers to publish the [links to all source code](#) accompanying NTTS 2021 presentations.

Having a good overview of essential open source software also creates other opportunities. One of the largest subsections of the list is “access to official statistics”. The software in this category supports to access official statistics databases and has been used to identify FAIRness (Findable Accessible, Interpretable, Re-usable) of the official statistics landscape [4].

4. SUMMARY AND OUTLOOK

Ideas for extending the functionality are documented on GitHub itself containing ideas such as adding compatibility, maturity and popularity of the items. Download figures were previously added, but later removed as they are not representative across platforms and languages. A more fair indicator of popularity takes multiple aspects of use, maturity and functionality into consideration. All in all, we argue that the awesome list of official statistics software has proven to be a useful tool for sharing knowledge on existing open source software in official statistics and we hope it can play this role also in future innovation projects.

5. REFERENCES

- [1] <https://github.com/sindresorhus/awesome>
- [2] <http://awesomeofficialstatistics.org>
- [3] <https://os4os.pages.code.europa.eu/pbbp/principles.html>
- [4] O. ten Bosch, E. de Jonge, H. Laloli, To be FAIR, what is missing in Official Statistics?, COSMOS conference, Paris, march 2024

Optimizing Text Classification with Deep Learning: The Impact of Preprocessing on ATECO Code Assignment in Italian Business Descriptions

Giulio Massacci, Mauro Bruno, Francesco Ortame & Giorgia Sacco (Italian National Institute of Statistics, Istat)

1. INTRODUCTION

Text classification is a crucial task in natural language processing (NLP) with numerous applications, including the automated assignment of industry codes to businesses based on their descriptions. In this study, we explore the use of deep learning models to classify company descriptions and map them to the corresponding new re-classification of Italian ATECO codes (2025 version) based on the European NACE Rev 2.1 classification of economic activities¹, which categorize economic activities in Italy. We implement and compare several state-of-the-art models, based on Transformer architectures like BERT. Models are evaluated on a dataset of Italian business descriptions, using various text preprocessing techniques to assess how the results change based on different model inputs. The results demonstrate the effectiveness of different deep learning techniques, highlighting the strengths and limitations of each model in capturing the semantic nuances required for accurate ATECO code classification. Furthermore, the challenge will be to evaluate the ATECO classification results by comparing them with internal and external data sources. This includes internal company data like financial reports and external references such as online job portals. The aim will be to assess alignment between the model's predictions and these benchmarks, identifying discrepancies. This comparative analysis will be crucial for validating the model's performance and ensuring the classification results are reliable and applicable in real-world contexts.

2. METHODOLOGY

This study focuses on a text classification task, precisely mapping Italian company descriptions to the corresponding on the Italian ATECO codes, a system used to categorize economic activities in Italy.

Due to the unavailability of the specific classification under the new ATECO 25 system a priori, resulting from the absence of ground truth, this study aims to investigate the behaviour of deep learning models as a function of varying input configurations. The analysis will consider both long and short job descriptions, as well as the presence or absence of data preprocessing techniques. Model performance will be assessed through similarity metrics, such as vector dot product, to evaluate the correspondence between job descriptions and the most appropriate ATECO 25 classification codes. The study will further examine how these similarity measures fluctuate in relation to the different input variables. The methodology involves using state-of-the-art deep learning models, primarily based on Transformer architectures such as BERT (Bidirectional Encoder Representations from Transformers).

¹ https://eur-lex.europa.eu/eli/reg_del/2023/137/oj

Outlined below are the steps followed to produce the results analysed in this study:

- **Data Collection and Preprocessing:** The dataset consists of Italian business descriptions, including long textual data from companies' financial statements and short descriptions gathered through a statistical survey completed by the companies themselves. Preprocessing steps include tokenization, stop-word removal, and normalization of the text (e.g., lowercasing, removing special characters) to prepare the input for the models. Given the domain-specific nature of the ATECO codes, additional preprocessing steps include handling abbreviations, domain-specific terms, and possibly applying custom embeddings to capture relevant features of economic activities.
- **Models:** The study explores available pre-trained Transformer-based architectures, with BERT as the core model. Other potential Transformer variants (e.g., RoBERTa, bge-m3) may also be considered to evaluate their performance on the task. In only one case, a fine-tuned model on the specific task of ATECO classification is built, to compare it with the previous ones, if there are improvements with the inclusion of context information. This involves adding a classification layer on top of BERT to output the corresponding ATECO category for each company description. Hyperparameter tuning is performed to optimize model performance, adjusting learning rate, batch size, and the number of epochs.
- **Evaluation and Analysis metrics:** The evaluation of the models is conducted by closely examining similarity metrics to assess the quality of the outputs. The focus is on understanding how these metrics fluctuate in response to various text preprocessing techniques, such as tokenization, stop-word removal, and normalization, while keeping the model parameters constant. This approach allows for an analysis of how different preprocessing methods influence the model's performance, without introducing variability from changes in the model's internal configuration.

3. RESULTS AND PRACTICAL APPLICATION

The results of this study demonstrate the impact of text preprocessing and description length on the classification accuracy of job descriptions into ATECO codes. The presented analysis is based on the bge-m3 transformer model [8]. The normalized dot product, used as the similarity metric, revealed that preprocessing has a more significant effect on short texts, where the removal of noise and the extraction of key information enhances classification performance.

Below the key findings, based on [Table 1] and [Table 2] results:

- **Short Texts Benefit Most from Preprocessing:** For short job descriptions (e.g., 1-15 characters), preprocessing techniques such as stop-word removal, stemming, and synonym inclusion led to higher similarity scores with the correct ATECO codes. Specifically, the percentage of short texts achieving a similarity score of 0.6 or higher increased by 20% after preprocessing, indicating a substantial improvement.
- **Longer Texts Show Stability:** In the case of longer texts (26-71 characters or more), preprocessing had a less pronounced effect. The similarity scores remained relatively stable, with minor improvements seen in the distribution of similarity values. This suggests that while longer texts contain more inherent information for the model to leverage, preprocessing still contributes to a slight refinement of the results.

The practical application of these findings is evident in scenarios where brief and inconsistent job descriptions are provided, such as short entries or survey responses. By employing appropriate preprocessing techniques can significantly improve the ATECO code assignments, leading to better alignment between job activities and the assigned economic sectors.

This analysis requires further investigation through a more detailed examination of the obtained results, assessing whether the activities identified by the algorithm accurately correspond to real-world conditions by cross-referencing them with the real outcome. Nevertheless, the primary objective here is to emphasize the internal behavior of these algorithms as a function of variations in the provided input.

Table 1. Distribution of the length of business activity description texts WITHOUT pre-processing and similarity index with the closest ATECO code using bge-m3 model.

String length	bge-m3 Model similarity							Total
	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
[1-15]	4%	33%	42%	19%	1%	0%	0%	100%
[16-20]	1%	14%	39%	40%	5%	0%	0%	100%
[20-25]	0%	8%	42%	41%	9%	0%	0%	100%
[26-32]	0%	4%	38%	46%	11%	0%	0%	100%
[33-36]	0%	2%	33%	49%	15%	0%	0%	100%
[37-42]	0%	2%	26%	51%	19%	2%	0%	100%
[43-48]	0%	1%	25%	55%	18%	1%	0%	100%
[49-57]	0%	1%	20%	54%	24%	2%	0%	100%
[58-71]	0%	0%	16%	51%	30%	3%	0%	100%
[72+]	0%	0%	10%	47%	36%	7%	0%	100%
Total	0%	6%	29%	45%	17%	2%	0%	100%

Table 2. Distribution of the length of business activity description texts WITH pre-processing and similarity index with the closest ATECO code using bge-m3 model.

String length	bge-m3 Model similarity							Total
	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
[1-15]	0%	4%	49%	39%	8%	0%	0%	100%
[16-20]	0%	0%	30%	56%	13%	1%	0%	100%
[20-25]	0%	0%	25%	60%	13%	2%	1%	100%
[26-32]	0%	0%	22%	61%	16%	1%	0%	100%
[27-36]	0%	0%	16%	62%	19%	2%	0%	100%
[37-42]	0%	0%	14%	61%	21%	4%	0%	100%
[43-48]	0%	0%	12%	57%	25%	6%	0%	100%
[49-57]	0%	0%	9%	56%	30%	6%	0%	100%
[58-71]	0%	0%	5%	54%	32%	7%	1%	100%
[72+]	0%	0%	2%	45%	43%	10%	1%	100%
Total		0%	18%	55%	22%	4%	0%	100%

4. MAIN FINDINGS

Preprocessing plays a pivotal role in improving the accuracy of ATECO code classification, and the study demonstrates that the application of those techniques significantly enhances the efficiency and performance of models, particularly when dealing with short textual inputs. By standardizing the text through processes such as tokenization, stop-word removal, and normalization, the models are better equipped to capture the underlying patterns within the data, leading to more accurate and robust outputs. This improvement is especially evident in

scenarios where short texts, such as brief job activity descriptions, are analysed. These texts, often containing sparse information, benefit from preprocessing steps that reduce noise and highlight the most relevant features, allowing the models to focus on key linguistic and semantic elements. Our findings suggest that adopting tailored preprocessing strategies is crucial for optimizing model performance, especially when dealing with limited or condensed text data, making it an essential step in natural language processing tasks. Future research may explore more advanced techniques, such as custom embeddings or domain-specific normalization, to further refine model efficiency across different text lengths and domains.

5. REFERENCES

- [1] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). "Deep Learning Based Text Classification: A Comprehensive Review." *ACM Computing Surveys*, 54(3), 1-40.
- [2] Howard, J., & Ruder, S. (2018). "Universal Language Model Fine-tuning for Text Classification." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328-339.
- [3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*.
- [4] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2017). "Bag of Tricks for Efficient Text Classification." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427-431.
- [5] Ferragina, P., Scaiella, U. (2012). "Fast and Accurate Annotation of Short Texts with Wikipedia Pages." *IEEE Software*, 29(3), 70-75.
- [6] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, (2024). M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. University of Science and Technology of China

Perceived health and life expectancy in EU countries: a machine learning approach for small samples

Simona Cafieri (Istat) & Francesco Petrucci (Sapienza University)

1. INTRODUCTION

The aim of this study is to analyse health in all 27 EU countries and how different types of causes may affect it. The following variables have been considered initially:

- Percentage of individuals who perceive their health as 'Bad or very bad';
- Percentage of individuals who suffer from obesity (according to BMI Index);
- Percentage of smokers;
- Percentage of individuals with depression symptoms;
- Percentage of individuals with more than weekly heavy drinking episodes;
- Percentage of individuals who practice aerobic sports;
- Percentage of individuals who consume zero portion of fruits and vegetables a day;
- Tonnes of nitrogen oxides in the air by square kilometre of national surface area;
- Tonnes of sulphur oxides in the air by square kilometre of national surface area;
- Tonnes of PM2.5 in the air by square kilometre of national surface area;
- Percentage of individuals at poverty risk;
- Percentage of students between the individuals in [15-24] age interval;
- Percentage of individuals who live in a poor condition housing;
- Rate of occupation.

The values considered for each variable, the single EU country being the statistical observation, refer to the years 2014 and 2019, for which EHIS data are available at the time of the study, and they were collected from EUROSTAT database. We want to reach our aim by building a model with a machine learning algorithm in which the first variable of the list is explained by the others using the 2014 data. Then, forecasting with the model the values of 2019, we evaluate the model obtained, comparing predicted and observed values of the response variable in 2019 and measuring the difference between them. We'll show how different ways of treating data can empower AI algorithms of machine learning and lead to solve problems caused by a little sample size. Secondly, to compare perceived health, represented by the first variable of the list, to a more objective indicator of health, we will show the same analysis, with same independent variables and same methods, conducted on a different response variable: life expectancy at age 0.

2. METHODOLOGY

To seek the best model for the forecast, we use three machine learning algorithms, which are regression tree, random forest and neural networks, pre-processing the data in different ways and obtaining different datasets:

- Original data,
- 'Augmented' data,
- 'Augmented' data with PCA,

where the augmented data consists of a new dataset obtained from the application of a data augmentation function. This function creates $n-1$ copies of the original sample and it sums pseudo-random values from a Normal variable with 0 mean and desired standard deviation.

Each estimated model's forecast performances are evaluated using the root mean squared error (RMSE) and the explained variance. We consider both measures relatively to train set, test set and finally to observed 2019 data.

In each one of the trained models, the number of principal components, where the PCA is used, is determined considering Cross Validation combined with explained variance. The transformation matrix obtained is then applied to the new data of 2019 to calculate the predicted values, using the model trained with 2014 principal components' scores.

3. RESULTS AND PRACTICAL APPLICATION

For the first part of the study, we concentrate on perceived health.

Training a model on the original data leads to very poor results in terms of forecast performance both for regression tree and random forest algorithms, even trying with different sets of parameters. The RMSE is very high and for the regression tree model we have a negative explained variance (it happens if the error is too high).

27 observations are evidently too few to let such algorithms estimate a flexible model. We show that just by employing data augmentation and widening the sample, the results improve, but further steps can be made by using PCA.

On a 10 times augmented sample, we estimate two models considering eight principal components. We obtain, with the random forest algorithm, the model which supplies the best predictions for the response variable in 2019.

Table 1. Regression tree and random forest with data augmentation and PCA - Forecast performances

	RMSE		Explained variance	
	Cross Validation	2019 data	Cross Validation	2019 data
RT	3.07	2.75	50%	45%
RF	2.44	2.41	70%	58%

The main methodological result is that the best model obtained with random forest can be outperformed by far by using neural networks.

The same path is followed about life expectancy. We start with simple models with regression tree and random forest, and we finally reach good predictions complicating them, but eventually the best predictions are obtained with neural networks.

4. MAIN FINDINGS

Calculating the predictions with the best model obtained from the point of view of predictive capacity and comparing them to the real ones, we are able to make some considerations. The predicted values for some countries are very close to those observed, but for others the forecasts are less precise. In this analysis, in fact, the focus was placed not only on the absolute result, but also on the possibility of improving a machine learning model through the use of techniques to manipulate the data, especially in the case, such as the present one, in which the sample size is small.

Anyway, the predicted values for the two variables of interest are near enough to the real ones to find a clear negative correlation between them, as we could expect. Moreover, we are able to represent an evolution trend from 2014 to 2019, which is well caught by our predictions.

5. REFERENCES

- [1] <https://ec.europa.eu/eurostat/data/database>
- [2] <https://www.istat.it/informazioni-sulla-rilevazione/indagine-europea-sulla-salute-ehis/>
- [3] Breiman, L., Random forests. Machine learning 45(1): 5–32. (2001).
- [4] Breiman, L., Friedman, J. et al., Classification and regression trees. (1984).
- [5] Deprez, P., P. Shevchenko and M. Wüthrich. 2017. Machine learning techniques for mortality modeling. European Actuarial Journal 7(2):337-352.
- [6] Hainaut, D. (2018). A neural-network analyzer for mortality forecast. ASTIN Bulletin, 48, 2, 481-508.
- [7] F. Chollet, J.J. Allaire. Deep Learning with R (2018).
- [8] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning - data mining, inference and prediction. Springer, 2008.
- [9] G. James, D. Witten, T. Hastie, R. Tibshirani (2017) An introduction to statistical learning: with applications in R. Springer texts in statistics. Springer, Berlin.

Predicting Travel Purpose in a Smartphone-Based Travel Survey

Solichatus Zahroh (EMOS Programme, Utrecht University)

1. INTRODUCTION

A travel survey records people's movement patterns in a specific area and was valuable in many research fields. By analysing some supplementary variables, one could uncover distinctive patterns demonstrated by travellers, regardless of the distance or the duration of their journey. However, the general population travel survey was burdensome for the respondents as each respondent should document a full-day trip, including the precise start times, end times, and locations. Predicting travel purposes automatically would be beneficial since it is difficult to accurately recall, and report stops and tracks from memory in traditional travel surveys.

In travel behaviour analysis, trip purpose was a fundamental yet complex research issue. Understanding the meanings of activities within the context of a trip was often essential. Most existing methods, however, depend on obtaining sensitive information from passengers, such as their home addresses or daily travel logs from surveys, to generate precise conclusions. Consequently, it was seldom applicable in real-world circumstances due to the reluctance of certain respondents to offer the data (Liao et al., 2022).

The growing popularity of smartphones was enabling the emergence of smart city applications (Soares et al., 2019). Smartphones used various sensors, including the Global Positional System (GPS), Global System for Mobile Communications (GSM), and accelerometer, to gather real-time data for location tracking and detection (Yu et al., 2012). The collection of GPS sensor data was advantageous for discovering movement patterns and portraying movement behaviour (Calabrese et al., 2013).

Statistics Netherlands (*Centraal Bureau voor de Statistiek* or CBS) implemented an innovation in travel survey data collection in 2020 (McCool et al., 2021). Automatically collecting GPS sensor data could alleviate the burden on respondents by eliminating the necessity for them to remember and report all of their daily trips. Using smart devices to track visit times and stop durations also enabled the analysis of passengers' behaviours, which is critical for predicting the purpose of their trip (Kakar, 2020).

An automated travel diary was created for the survey respondents based on the GPS sensor data. Each day was divided into distinct segments representing stationary periods (stops) and periods of travel (tracks). Through the personal movement investigation data gathered by multi-day use of the CBS app, movement patterns of travellers could be uncovered. Such an app-based approach could reduce recall bias, collected a lot of data with minimal effort and was less time-consuming (Zhou et al., 2022). The travel survey community, therefore,

considered that GPS data would emerge as a key method for future data collection, offering a solution to current challenges (Bricka et al., 2012).

The current approach was to simply ask someone why they travelled. Detailed activity-related data, such as nearby places and past choices made by other travellers, was absent from the current techniques (Cui et al., 2018). Examining the historical places people have visited could replicate the personalised preferences of travel destinations. Furthermore, publicly accessible information about nearby places, such as Open Street Map (OSM) data, could provide a more comprehensive insight into the functionalities of a certain site.

Online resources like OSM and Google Places API, as well as offline land use data collected by CBS, were the two primary sources of information regarding POIs. This dataset contained information regarding specific geographical places within a given area. Google Places API requests deliver nearby search, text search, radar search, and place details queries. In contrast, OSM provides free access with limited coverage and a more lenient license compared to Google Places. OSM relies on crowd-sourced data and has strong community support. Additionally, OSM may be downloaded and used offline. However, there was no standardized format of using OSM therefore making it impossible to verify the accuracy of OSM.

2. OBJECTIVE

The goal of this study is, therefore, to use GPS data, external spatial and temporal patterns data, and socio-demographic characteristics to automate trip purpose prediction. It is expected that it is no longer necessary to directly ask respondents about the purpose of their trip. One could gain a more profound understanding of travel patterns within a particular geographic area and time frame.

This study aims to answer the general research question “How well can we predict the travel purpose using sensor data from a smartphone-based travel diary study?” which can be translated into two sub-questions as follows:

1. To what extent are external spatial and temporal patterns data helpful in predicting travel purposes?
2. To what extent do individual behaviours and characteristics influence the accuracy of travel purpose prediction?

3. METHODS

To answer the aim of this study, several machine learning models were used. Before the training process, several data cleaning was being done. The collected data consisted of 21,397,699 location observations. Multiple observations were obtained at one location within a time frame of approximately 5 seconds, and the most accurate set of location measurements was selected. After eliminating redundant and low-quality data, there were 12677 locations from 456 users. About half of the observations were classified as stops (stationary period) and the other half were tracks (moving period).

The information about the distance of each track and the duration of the stops and tracks were added as features, along with the time of day (in 24 hours) and day of the week of each stop and track. Since the total number of visits to regularly visited places, such as one's house, was rather low, numerous stop locations in close proximity to one another were merged, resulting in an accuracy of 55 meters.

Multiple bounding boxes with varying radiuses were determined from OSM for different tags associated with trip purposes. Four distinct radiuses—25, 35, 50, and 200 metres—were given to evaluate discrepancies that may arise when a stop was located at a position that did not align with OSM data. A large number of features could not be solely dropped into the model since the information of four distinct radiuses of the bounding box contains overlapping information. In order to maximise training time, reduce noisy attributes, and prevent overfitting the data, the model was constrained to only one radius per tag.

The analysis excluded data with insufficient quality, which included less than one hour and less than 2000 observations. Furthermore, noisy data that might adversely affect the algorithm's stop-track categorization was excluded using median smoothing by manual and automatic detection. Data acquired beyond the specified reference periods, which might be either one or seven days, and tracks more than 24 hours were also omitted. The locations outside the Netherlands were also excluded. The final dataset for training process consisted of 4961 locations (Figure 1).

This dataset was divided into training and testing sets with a ratio of 80:20. Due to the complexity of the data, neural network model as one of the machine learning models was used to analyse the data. However, given the diverse sources of error and uncertainty involved in the study, it may not be the most effective technique to choose a single model for both development and application (Cheng et al., 2019).

For comparison, various machine learning models, including Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB), were evaluated. The test sets of various models were assessed using two metrics, specifically balanced accuracy and F1-score.

4. RESULTS

The majority of stops occurred between the hours of 8 a.m. and 4 p.m. coinciding with typical working hours. Home accounted for most stops (30.3%), while education (0.7%), sport (1.2%), and transit (1.4%) had the fewest stops. The car was the most prevalent form of transportation, accounting for 33.7% of the tracks, while the tram was the least preferred option, representing only 0.6%.

The initial analysis was performed without OSM data, and only included GPS data and socio-demographic variables (gender, age, income group, household numbers, household types, socio-economic groups, car ownership, and working hours). The training and testing data had approximately 65% accuracy. The balanced accuracy of the assessment results ranged from 53.9% for the education category to 90.9% for the home category. The test's accuracy decreased by 12% when it incorporated data from OSM within four distinct radiuses. In the

subsequent trial, the tags were categorized and examined as count data and percentages. A percentage represents the ratio of the total number of POIs in a specific category to the total number of all POIs. This was done since tag selection was previously based on the labels of trip purposes. Various combinations of category-only, percentage-only, category and individual tags, and percentage and individual tags were utilized. Only the model accuracy of the percentage-only model decreased compared to the prior model.

The optimal model was an ANN model that utilized OSM data, using both percentage and individual tags. In addition, weather information was incorporated to enhance the model. Regrettably, the inclusion of weather data did not enhance the performance of the model. The RF, XGB, SVM, and NB models were trained using the same set of training control parameters and data sets for comparison. The RF model's accuracy was perfect for the training model, but not for the testing model. The XGB model also achieved near-perfect training accuracy, given that both models are tree-based. XGB's accuracy was greater than 90% for the training set, but not for the test set. The SVM model produced similar balanced accuracy compared to the ANN. Considering that SVM was initially developed to enhance the training of ANN. Unfortunately, the NB model performed poorly. The classifier exhibited a lack of ability to differentiate between classes and produced almost random predictions, leading to an accuracy rate of approximately 50% (Table 1). Table 2 shows the confusion matrix of ANN and XGB models as the best model (similar balanced accuracy for the test data). From confusion matrix, we can see that "other" category mostly misclassified as visit or shop and "pick-up" had lower accuracy despite of low number of misclassification due to small sample size.

In conclusion, a smartphone-based travel diary study could predict the travel purpose pretty well. The model's accuracy was a bit lower than in past studies due to its ability to classify a greater number of trip-purpose labels than in previous research. The model achieved satisfactory accuracy on the initial attempt when trained without OSM data, and some sociodemographic factors were essential. However, only the respondents' ages, which serve as indicators of individual qualities, became significant factors in the integration of OSM data. The overall visit frequency to the same area was crucial, representing 45% in the ANN model and 25% in the XGB model. This demonstrates that a data collection period of seven days was better than one day. It demonstrated individuals' ability to adjust and thrive in various situations. The study concludes that individual behaviors (visiting the same location with the same purpose) were more accurate predictors of travel purpose than individual characteristics (administrative data).

Stop duration was the sole factor responsible for all variability in all models. The same location might serve multiple purposes, depending on the length of the visit. This suggested that we can use temporal patterns to identify the purpose of the trip. In OSM, the number of recreational facilities were important data. The presence of a higher number of shops and sports facilities within a 25-metre radius significantly enhanced the probability of making a halt. Spatial data aided in predicting the purpose of trips.

Recording spatial and temporal patterns diminished the importance of certain characteristics. This was a promising indication for CBS to accurately predict a trip's purpose in real-time. The absence of a request for the users' sociodemographic characteristics prevented immediate access to this information in real-time. Nevertheless, people might visit the same location several times on different occasions. In order to accurately determine the purpose of the trip, it was important to include additional information, such as the respondents' occupation or their participation in hobbies-related memberships. Utilizing spatial and temporal patterns was

valuable for predicting trip purposes, and individual behaviors had minimal influence on the accuracy of trip-purpose prediction. Some types of stops would be more important for making accurate predictions than others.

Based on the challenges we encountered, this study offered several recommendations for future research. Firstly, the data collection time should not be limited to one season, data collection for the whole year can be an option. Secondly, due to timing errors, inaccuracies in both the satellite and receiver clocks, as well as relativity effects, can result in position errors of up to two meters, more than one observation should be selected per location. Thirdly, indoor and outdoor activities should be divided and more sample sizes per class are needed to reach model convergence. Additionally, certain information, such as respondents' profession and hobbies-related subscriptions, should be included to predict specific purposes. Lastly, exploration of individual characteristics, such as the tendency for teachers to visit educational places more frequently than others, must be done. Moreover, we might improve the model by opting for better tags in OSM and leveraging the most up-to-date OSM data. The predicted increase in the quantity of data points was expected to facilitate the training of more complex models, hence enabling the detection of variations in unique behaviors across many seasons and the identification of weather dependencies.

TABLES AND FIGURES

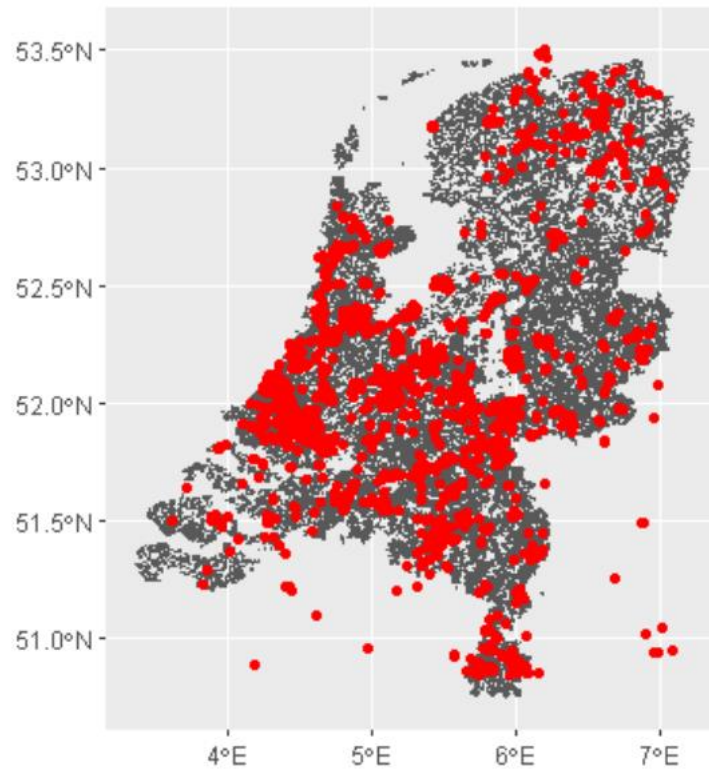


Figure 1. Mapping of the Location Data

Table 1. Balanced Accuracy of Test Model (in %)

Model	ANN	RF	XGB	SVM	NB
Overall	72.3	77.5	77.7	70.8	42.7
Pick-up	75	78.2	75.3	70.7	50
Edu	83.3	80.3	83.1	80.4	50
Others	62.9	69	74.1	64.4	50
Transit	80.1	79.1	83.3	71.7	50
Sport	71.1	73.1	81.2	68.1	50
Home	91.3	92.8	93.2	90.7	52.6
Visit	67.5	62.3	74.6	67.3	50
Work	84.1	90.8	87.9	83.9	49.9
Shop	81.5	86	83.8	80.1	55.5

ANN model is the ANN_6 model (ANN model with OSM data from 1 radius as count data and percentage without weather data). The other models represent the best models with the optimum parameters tuning.

Table 2. Confusion Matrix of the Best Model
ANN Model

Obs Pred	Pick-up	Edu	Other	Transit	Sport	Home	Visit	Work	Shop
Pick-up	41	0	4	0	1	5	4	3	7
Edu	0	12	0	0	0	0	0	1	0
Others	9	1	22	3	3	3	7	4	10
Transit	2	0	1	21	0	0	1	8	3
Sport	3	0	3	0	13	0	1	4	0
Home	4	2	8	2	2	377	15	14	12
Visit	3	1	10	0	3	11	24	2	4
Work	10	1	6	5	4	7	3	116	4
Shop	6	1	19	3	4	3	7	7	88

XGB Model

Obs Pred	Pick-up	Edu	Other	Transit	Sport	Home	Visit	Work	Shop
Pick-up	42	0	4	1	1	5	4	6	8
Edu	0	12	0	0	2	1	0	2	0
Others	8	1	38	1	3	4	5	3	11
Transit	1	0	2	23	0	0	0	1	5
Sport	6	0	2	0	19	0	0	0	0
Home	3	1	3	2	2	384	17	13	6
Visit	3	1	7	0	1	5	32	2	4
Work	6	2	7	5	1	5	4	127	3
Shop	9	1	10	2	1	2	0	5	91

5. MAIN REFERENCES

- [1] Atwal, K. S., Anderson, T., Pfoser, D., & Züfle, A. (2022). Predicting building types using OpenStreetMap. *Scientific Reports*, 12(1), 19976. <https://doi.org/10.1038/s41598-022-24263-w>
- [2] Bricka, S. G., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, 21(1), 67–88. <https://doi.org/10.1016/j.trc.2011.09.005>
- [3] Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26, 301–313. <https://doi.org/10.1016/j.trc.2012.09.009>
- [4] Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*, 14, 1–10. <https://doi.org/10.1016/j.tbs.2018.09.002>
- [5] Cui, Y., Meng, C., He, Q., & Gao, J. (2018). Forecasting current and next trip purpose with social media data and Google Places. *Transportation Research Part C: Emerging Technologies*, 97, 159–174. <https://doi.org/10.1016/j.trc.2018.10.017>

- [6] Kakar, A. (2020). Trip Purpose and Prediction. *International Journal of Engineering Research*, 9(10), 278–285.
- [7] Liao, C., Chen, C., Guo, S., Wang, Z., Liu, Y., Xu, K., & Zhang, D. (2022). Wheels Know Why You Travel: Predicting Trip Purpose via a Dual-Attention Graph Embedding Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1), 1–22. <https://doi.org/10.1145/3517239>
- [8] McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. *Journal of Official Statistics*, 37(1), 149–170. <https://doi.org/10.2478/jos-2021-0007>
- [9] Soares, E., Revoredo, K., Baiao, F., A. De M. S. Quintella, C., & V. Campos, C. A. (2019). A Combined Solution for Real-Time Travel Mode Detection and Trip Purpose Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 20(12), 4655–4664. <https://doi.org/10.1109/TITS.2019.2905601>
- [10] Zhou, Y., Zhang, Y., Yuan, Q., Yang, C., Guo, T., & Wang, Y. (2022). The Smartphone-Based Person Travel Survey System: Data Collection, Trip Extraction, and Travel Mode Detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(12), 23399–23407. <https://doi.org/10.1109/TITS.2022.3207198>

Prediction of crop yields at field scale from earth observation data

Damien Babet (The French ministry of Agriculture and Food)

1. INTRODUCTION

We feed an XGBoost model with satellite imagery (Sentinel 2), meteorological and geographical data, and information on land plots and crops, and train it on yields from farm survey data over 5 years (2017-2021).

Test set yield prediction at the field scale is good, particularly for soft winter wheat, the most abundant crop in the training data. We test same-year prediction and year-to-year forecast. Out-of-year prediction performance is not good enough to improve on existing expert and survey forecasting, likely due to the lack of historical depth in the data. Same-year yield prediction, however, shows much higher performance. We detail two valuable use of same-year, field-level yield prediction: diffusion of a field-level map of yields for researchers, and as auxiliary variable for random sampling optimization.

2. METHODOLOGY

2.1. Data

We mobilize Sentinel 2 satellite open data imagery from the Copernicus programme: multispectral imaging with 10m to 60m resolution and 5 days revisit frequency, with data calibration and correction (filtering out clouds, etc.). Monthly averages, min and max of radiometric indices (ndvi, ndwi, etc.) are computed.

Additional features include geographical data (coordinates, slope, orientation, altitude) and meteorological data (monthly indicators such as number of Growing Degree Days, cumulative rain, etc.).

Field plot geometry and crop come from detailed farmers administrative declarations in the European Common Agricultural Policy grants scheme. Inputs are aggregated at the field scale before training.

Training data come from yearly statistical surveys on agricultural yields for major crops, with 17000 farms surveyed each year [1]. Yields are known only at the farm scale, not the field scale, and survey respondents exhibit a rounding behaviour in their responses. However, we can leverage a bigger training sample than related literature [2] thanks to the direct access to yield survey individual data.

2.2. Training and evaluation

The predictor model is xgboost [3], known to deliver good results on average-size training sets.

We base features selection on agronomic expertise rather than data-driven selection, because of the average size of the ground-truth training data.

The process, however, has high computing needs, mostly concentrated in the satellite data preparation step.

We measure yield prediction performance on a 20% out-of-sample test set for same year-prediction, and out-of-year test set for forecasting. We report performance at the field scale, and for regional and national averages.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Better performance in same-year prediction

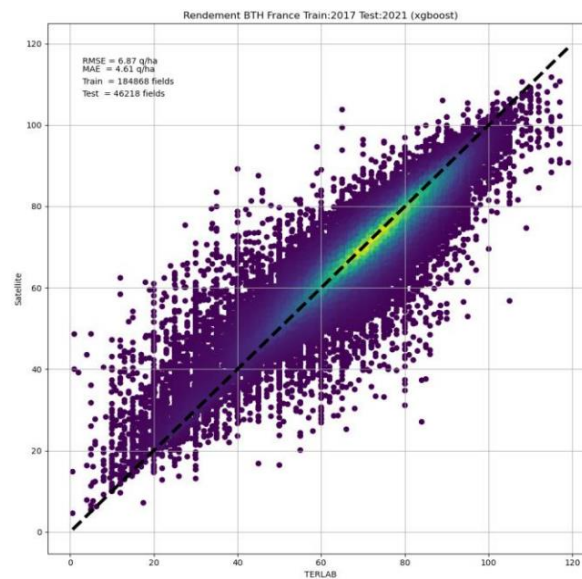


Figure 1. Plot yield for soft winter wheat, real (x-axis) and predicted (y-axis), same-year out-of-sample test set.

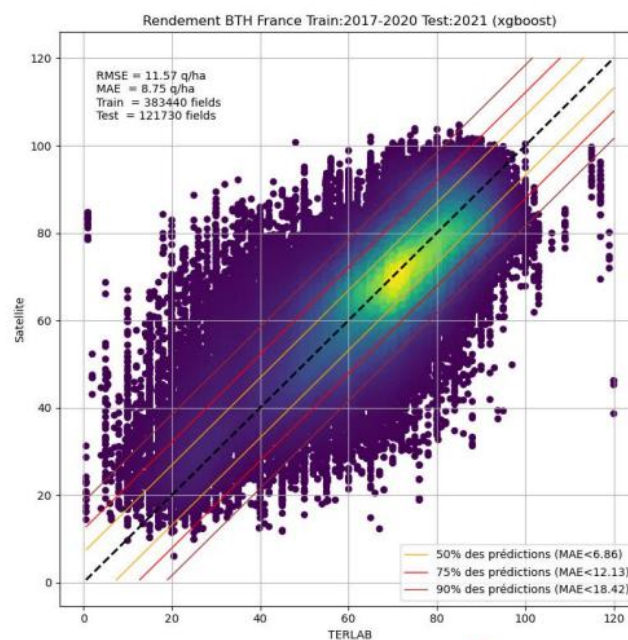


Figure 2. Plot yield for soft winter wheat, real (x-axis) and predicted (y-axis), out-of-year test set.

RMSE and MAE are 50 to 100% higher in out-of-year predictions, an increase out of proportion with the reduced training sample size (four years instead of five, figures 1 and 2). The model likely lacks historical depth to learn year-to-year fluctuations. Within-year yields are correlated and likely do not provide enough implicit information on yearly variations such as large or lasting meteorological events or pests. It is possible that the characteristics of the model that help it attain good performances on same-year prediction are responsible for overfitting and lack of generalization power in an out-of-year setting [4].

3.2. No bias in aggregated prediction

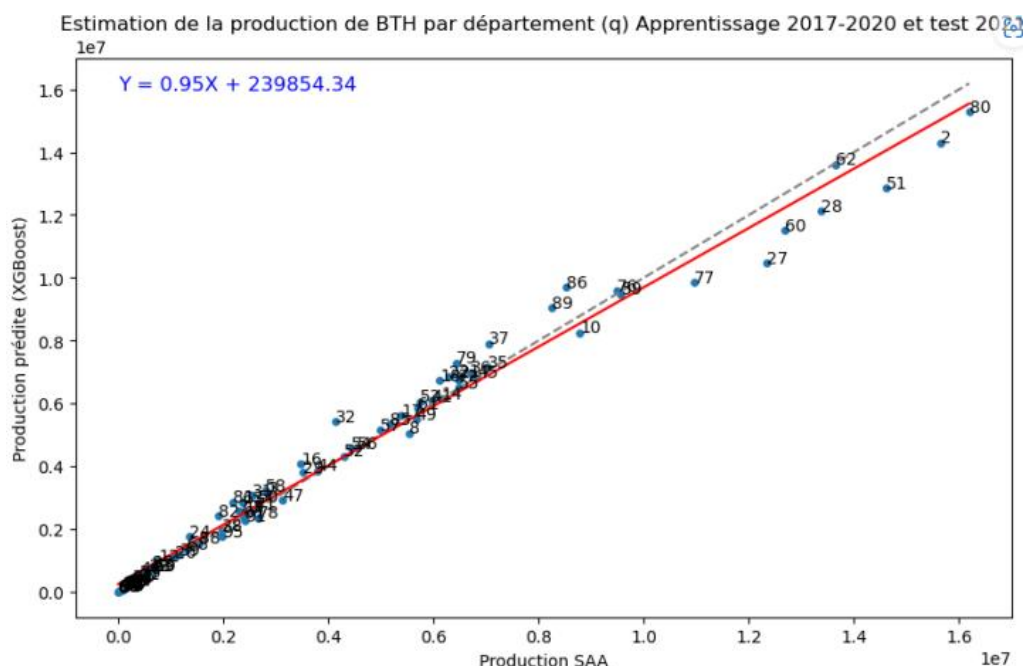


Figure 3. Average regional production for soft winter wheat, real (x-axis) and predicted (y-axis), out-of-year test set.

We find no obvious bias in aggregated predictions, a result indicative both of an efficient training of the model and of the good statistical properties of the survey that makes the training sample (figure 3).

3.3. Planned applications

We develop two main applications. First, we will produce for diffusion a map of predicted yields per field. Such a map has two advantages over the existing survey individual data: it can preserve privacy (by omitting surveyed farms or replacing known values with cross-validated predictions). It can provide a much larger, quasi-exhaustive dataset (in the order of millions of fields rather than tens of thousands), that can be matched to other statistical sources, analysed for fine-grained phenomenon, or be used in the training of other models.

The second application is to inform random sampling for future surveys, including the survey used here for training. Predicted yields are precious auxiliary data, correlated to other variables of interest, most obviously real yields.

4. REFERENCES

- [1] INGLADA, Jordi. Exploitation de l'enquête TERLAB pour l'estimation du rendement des cultures à la parcelle à partir de séries temporelles Sentinel-2. 2020. Thèse de doctorat. CESBIO.

- [2] ZHAO, Yan, POTGIETER, Andries B., ZHANG, Miao, et al. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. *Remote Sensing*, 2020, vol. 12, no 6, p. 1024.
- [3] CHEN, Tianqi et GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In : *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. p. 785-794.
- [4] DESLOIRES, Johann, IENCO, Dino, et BOTREL, Antoine. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture*, 2023, vol. 209, p. 107807.

Privacy-Enhancing Technologies for Synthetic Data Creation with Deep Generative Models

Alessio Crisafulli Carpani (EMOS Programme, Università di Bologna)

1. INTRODUCTION

In light of the recent technological advancements, our society has evolved into a prolific source of data, which is then gathered, processed, and subjected to analysis, effectively converting our society, economy, and physical environment into expansive reservoirs of data, resembling what could be termed as “data fountains” (Ricciato 2019).

The utilization of data, particularly datasets containing micro-level, individual-specific information, have drawn significant attention in the realm of data mining research. In today’s world, numerous real-world systems heavily depend on machine learning (ML) models to carry out a diverse range of tasks, including uncovering novel data patterns and facilitating recommendation systems. However, a significant challenge arises, as many of these ML algorithms have an insatiable demand for data, often necessitating the inclusion of personal sensitive information, in spite of the fact that these systems are vulnerable to privacy breaches (Shokri 2017). Thus, the organizations responsible for these technologies must strike a delicate balance between complying with GDPR and EUDPR and minimizing the risks associated with data loss, theft, or misuse, and cater to the needs of the “modeler”, whose aim is to optimize such systems.

On the other side, National Statistical Institutes' (NSIs), alongside other relevant institutions, have the critical responsibility of providing reliable, pertinent, timely, and high-quality data to support evidence-based decision-making. Nevertheless, to respond effectively to emerging issues, NSIs often require supplementary data from secondary sources, including administrative or private sector data. This scenario calls for a coordinated international response, necessitating timely access to new data sources and potentially sensitive data shared among multiple partners, some of whom may be in different countries. However, due to legitimate privacy concerns, unrestricted access to all data cannot be granted to these partners.

Entities that lack access to extensive data-gathering resources, including researchers, small businesses, and ordinary individuals, face difficulties in accumulating sufficient data for training specific types of models. In such cases, generating synthetic data offers a more accessible alternative to acquiring original data, combining two aspects: usefulness for the statistical analysis (data augmentation) and the preservation of confidentiality. There are numerous scenarios September 2024 in which companies employ synthetic data to make information available for processing, especially in a post-GDPR world when regulations or privacy concerns impose restrictions on accessing the original data.

Within this context, generative models have emerged to create synthetic samples across various domains. Ideally, these models should prevent the exposure of individual-specific information from the training data. Unfortunately, recent literature has shown that this assumption is not consistently met, particularly with Generative Adversarial Networks (GANs), which lacks robust privacy guarantees. Nevertheless, there is a critical need to strike a balance between our responsibilities as data stewards and the importance to advance data mining research. In this regard, Privacy-Enhancing Technologies (PETs) can help mitigate these

challenges by imposing privacy constraints on models or more generally in algorithms, enabling their use and sharing without compromising the confidentiality of the training data.

2. OBJECTIVE

One promising approach in this domain involves modelling the data-generating distribution by training a generative model on the sensitive data, introducing the mechanism of Differential Privacy, a mathematical foundation for quantifying and achieving privacy in data analysis. This privacy-preserving model is then shared along with its private parameters, allowing anyone to generate a synthetic dataset that closely mirrors the original training data without compromising the robust protection of privacy.

This research is dedicated to exploring the latest techniques in the field of Privacy-Enhancing Technologies, in particular of Differential Privacy, by injecting these constraints into generative neural networks to create differentially private synthetic datasets and investigating the trade-off between data utility and privacy preservation through state-of-the-art programming libraries.

Beyond this, applications of the differential privacy mechanism were also studied with other supervised learning algorithms, to demonstrate that the contribution of the individuals' data is also masked out within these models, thus preventing data-leakage, and leading to comparable performance with the non-privatised models.

3. METHODS

In this work, the two primary approaches focus on incorporating differential privacy directly into the training process of Generative Adversarial Networks, which offer a distinct advantage by introducing noise within the latent space, rather than directly altering the data as with other output privacy techniques. These approaches allow us to ensure privacy while minimizing the overall loss of information.

At the core of the GAN framework lies the concept of adversarial training, wherein the generator aims to produce samples that are indistinguishable from real data and the discriminator strives to differentiate between genuine and generated samples, leading to data samples that accurately capture the characteristics of a desired target distribution. Differential privacy can be seamlessly integrated into the discriminator, introducing gaussian noise into the stochastic gradient descent (SGD) algorithm (Martin Abadi 2016). Similarly, this can be applied to Conditional Tabular GANs (CTGANs), a more robust version of the neural network that models the conditional probability distribution among the rows of tabular data, hence more able to mitigate the effect of heterogeneity, imbalance, or more generally highly sparse vectors. Another method to guarantee the privacy of the training data is to transfer the knowledge from an ensemble of "teacher" models to a "student" model during the learning process. This is achieved through the Private Aggregation of Teacher Ensembles (PATE) mechanism, which replaces the standard architecture of the GANs' discriminator (Jordon 2022).

Therefore, various generative neural networks, including DP-GANs, DP-CTGANs, PATE-GANs, and PATE-CTGANs, were implemented with different privacy budget parameters to study the trade-off between data privacy and utility. The experiments were conducted on two open-source datasets, reflecting this real-world scenario in which this sensitive type of data is made publicly available and could also be used as input for different privacy attacks. The quality of the datasets generated by these differentially private generative models is assessed by training a set of binary classification models, including Logistic Regression, Gaussian Naive Bayes, Random Forest, AdaBoost, Bagging of Decision Trees and Gradient Boosting, which

have been evaluated using the accuracy, the area under the receiver operating characteristics curve (AUC), the recall and the F1 score. The same evaluation was also considered using the repeated k-fold cross validation, on three different training/testing settings:

1. Setting TRTR: the models are trained on the real training set and assessed on the real testing set, the standard setting for benchmarking purposes.
2. Setting TSTR: the models are trained on the synthetic training set and assessed on the real testing set, to determine how well the synthetic data is able to capture the relationship between the variables.
3. Setting TSTS: the models are trained on the synthetic training set and assessed on the synthetic testing set, to evaluate the consistency of relative performance with TRTR.

To evaluate the utility of the synthetic data, also the synthetic ranking agreement (SRA) and the propensity score mean-squared error (pMSE) were computed. Whereas for the privacy risk assessment, multiple membership inference attacks (MIA) were conducted at each privacy budget level.

Moreover, to demonstrate the efficacy of differential privacy (DP) and to evaluate the deviation in performance in supervised learning algorithms, when privacy constraints are applied, two models - Logistic Regression and Gaussian Naive Bayes - were trained, both involving the inclusion and exclusion of differential privacy

4. RESULTS

From the experimental results, it is possible to notice that the classification scores for synthetically generated datasets (TSTS) are close to those of non-private datasets (TRTR). In particular, the DP-CTGANs or PATEGANs were found to be more robust to overfitting and therefore more able to capture the whole variability of source data, compared to the DP-GANs which often led to a generalization of the training data (**Errore. L'origine riferimento non è stata trovata.**).

Setting	ϵ	PATEGAN							
		Baseline Scores				Cross Validation Scores			
		Accuracy	AUC	F1	Recall	Accuracy	AUC	F1	Recall
TRTR		82.18%	79.76%	71.44%	75.28%	77.79%	88.50%	61.48%	64.68%
TSTR		83.15%	65.90%	34.36%	32.56%	66.23%	50.23%	10.20%	18.91%
TSTS	0.1	84.52%	71.83%	53.48%	49.73%	76.91%	70.19%	27.29%	22.05%
	10	85.62%	75.06%	91.13%	95.44%	76.99%	72.72%	85.70%	91.15%
	100	77.53%	73.56%	61.97%	65.96%	74.70%	76.39%	43.25%	39.73%

Table 1. Evaluation Results for PATE-GAN

When we consider the privacy budget of a model, we can observe that greater values of this parameter (ϵ) imply fewer privacy constraints being imposed, which in turn, suggests a more transparent generation of data. Consequently, it was observed that increasing this parameter, led to an improvement in the model's performance on synthetic data utility metrics (**Errore. L'origine riferimento non è stata trovata.**).

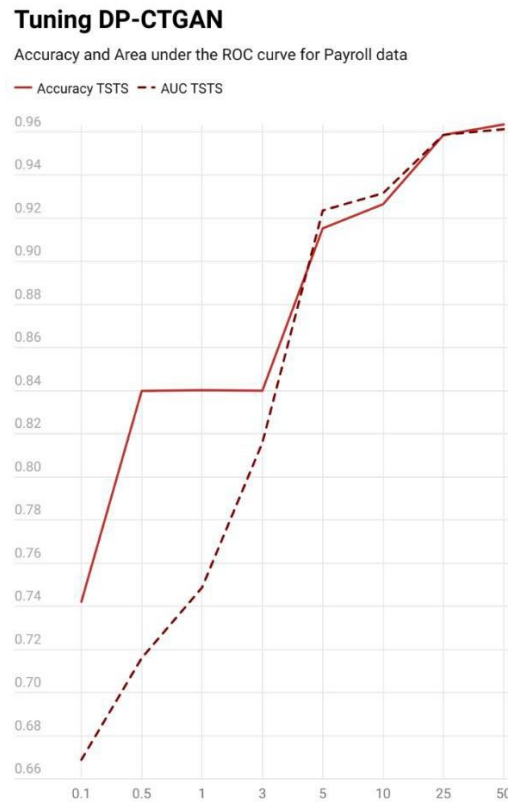


Figure 1. INCREASING THE PRIVACY BUDGET, THE MODEL PERFORMANCE INCREASES

However, from a privacy risk perspective, this implicates a higher vulnerability to privacy attacks. This was evidenced by the results of membership inference attacks performed on the synthetic datasets, which reported significantly lower success rates on those generated with a low privacy budget ([Figure 2](#)).

Membership Inference Attacks

Share of synthetic records closer to the training than the holdout dataset

0.1 10 100 No Privacy

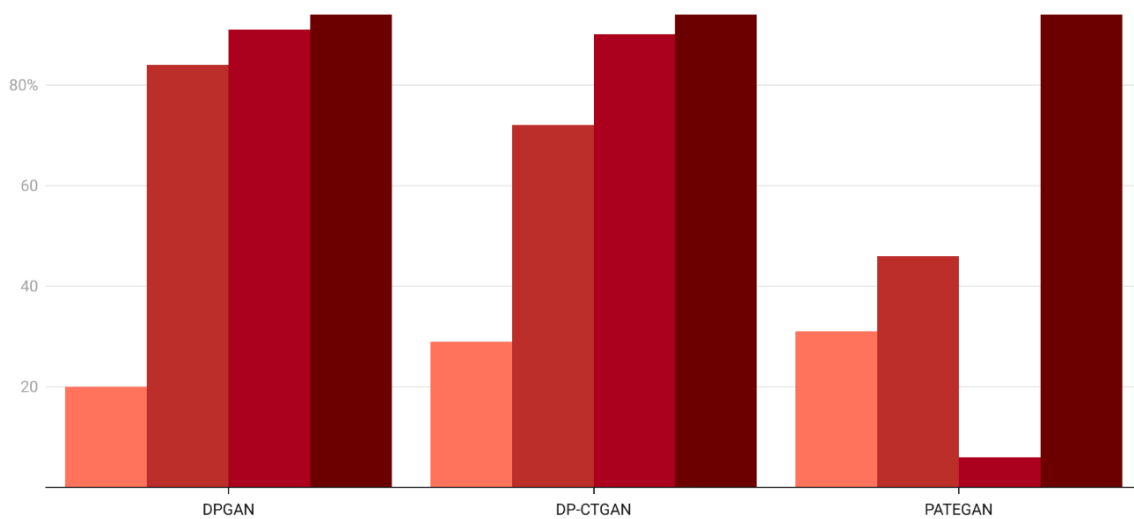


Figure 2. Privacy Attacks Results (MIA)

Furthermore, the experiments conducted with differentially private classification models reported good performances. As expected, the DP scores were found slightly lower but still very competitive with the non-private counterparts, as their averages ranged between 60%-70%. Particularly, the score of the DP-version of the logistic regression decreased in accuracy and AUC of just 6 and 4 basis points respectively.

However, training generative adversarial networks, especially when integrating the PATE framework, comes with several challenges. First, this includes the necessity of working with very powerful systems to meet the intense computational costs required to correctly train these networks. Second, GANs are unstable during training and the large bias produced by the critic in the gradient of the generator, when mixed with the imposed gradient noise by the differential privacy, can increase training instabilities. Future work could focus on implementing DP within other generative models and compare the results, by trying over different datasets with higher complexity and assessing them through more types of privacy attacks.

5. CONTRIBUTION

This thesis, which is the outcome of a research project in the field of Privacy- Enhancing Technologies for Official Statistics (PET4OS) conducted with the Italian National Institute of Statistics (ISTAT), revealed the efficiency and feasibility of applying differential privacy for multiple purposes. This study has the potential to:

- Enhance the understanding of the latest techniques for generating synthetic data while respecting the principles of differential privacy and furnish guidance to researchers, organisations, and policymakers on the practical application of differential privacy, also in supervised learning tasks.
- Provide insights about the trade-off between data utility and privacy preservation, specifically in the context of generative models and how to investigate it.
- *Contribute to the development of best practices for leveraging synthetic data in data-driven tasks while adhering to stringent privacy regulations.*

The synthesizer approach's main advantage is that the resulting dataset can be shared and used for analytical purposes any number of times without increasing the risks associated with privacy loss. Another advantage is that the synthesizer allows producing any arbitrary amount of data derived from the original dataset's distribution, a promising approach for data augmentation to improve a model performance. The python modules developed for this study can be adopted by NSIs to generate new privatised data, investigate the parameters and metrics, share the models and allowing external organisations to make inference on the synthetic dataset. For instance, numerous potential use cases could be identified between the European Statistical System (ESS) and the European System of Central Banks (ESCB). Furthermore, this can serve as starting point for further investigations or can be extended with more generative models or functionalities.

Another example, in a post-GDPR world, the processing of customer data involves stringent compliance and governance requirements for companies. In this scenario, the data curator initially encodes private data into a generative model, subsequently this model is shared with an analyst, who can use it to create data that resembles the original dataset. This provides organisations or companies with greater flexibility and freedom to process data in a secure manner.

6. REFERENCES

- [1] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. 2023. "I know what you trained last summer: A survey on stealing machine learning models and defences." doi:10.1145/3595292.
- [2] European Commission. 2020. "White Paper on Artificial Intelligence: a European approach to excellence and trust." https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en#related-links.
- [3] Jordon, James and Yoon, Jinsung and Schaar, Mihaela van der. 2022. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees." Edited by International Conference on Learning Representations. <https://openreview.net/forum?id=S1zk9iRqF7>.
- [4] Joshua Snok, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and specific utility measures for synthetic data." *Journal of the Royal Statistical Society* 181.
- [5] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar,. 2016. "Deep
- [6] learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM)*. doi:10.1145/2976749.2978318.

- [7] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar. 2018. "Scalable private learning with pate." <https://arxiv.org/abs/1802>.
- [8] OECD. 2023. "Emerging privacy-enhancing technologies." doi:<https://doi.org/10.1787/bf121be4-en>.
- [9] Ricciato, Fabio, et al. 2019. "Trusted smart statistics: Motivations and principles." *Statistical Journal of the IAOS*
- [10] (IOS Press) 35 (4): 589-603. <https://cros.ec.europa.eu/system/files/2023-12/sji190584.pdf>.
- [11] Shokri, Shmatikov, Marco Stronati and Congzheng Song and Vitaly. 2017. "Membership Inference Attacks against Machine Learning Models." (arXiv). <https://arxiv.org/abs/1610.05820>.
- [12] United Nations Committee of Experts on Big Data and Data Science for Official Statistics. 2023. "United Nations Guide on Privacy-Enhancing Technologies for Official Statistics." <https://unstats.un.org/bigdata/task-teams/privacy/guide/index.cshtml>.

Processing of receipts in Household budget survey

Urška Pirnat & Črt Grahonja (Statistical office of the Republic of Slovenia)

1. INTRODUCTION

Household budget surveys (HBS) are national surveys conducted across all European countries, focusing on household expenditure on goods and services (Eurostat, 2024). These surveys provide vital data that informs economic and social policies, enabling governments to understand consumption patterns and make evidence-based decisions. Data collection in the HBS typically involves gathering information through various methods, including interviews, diaries, and the collection of receipts. However, traditional data collection methods can be time-consuming and prone to human error.

The ESSnet HBS app project aims to modernize and streamline the collection and processing of household budget survey data by integrating advanced technologies. By utilizing Optical Character Recognition (OCR) and machine learning techniques, the project automates the extraction and categorization of information from receipts, transforming how data is collected and analysed. This innovative approach not only enhances efficiency and accuracy but also reduces the burden on respondents, allowing for more reliable and comprehensive data on household expenditures. The use of automation in data processing ultimately supports the goal of producing high-quality statistics that can drive informed policy-making across Europe.

1.1. A smart household budget survey – Dutch models

We adopted the Dutch models for receipt processing, which utilize Python 3+ as the core programming environment. These models were trained on a dataset of 300 images, allowing the process to effectively read the images and identify the contours of receipts. Key transformations in the workflow include converting the image to grayscale, shrinking its size, detecting borders, and performing necessary rotations or skew corrections to ensure optimal alignment. Furthermore, the contrast and brightness of the images are enhanced, followed by line detection to improve accuracy in the subsequent OCR extraction.

For ECOICOP classification, we combined FastText with rule-based methods and natural language processing (NLP) techniques to clean and pre-process product descriptions. This involved removing unwanted symbols, numbers, and supermarket names, as well as replacing units with general tags such as UNIT_SOLID or UNIT_LIQUID. The model is capable of classifying items up to the 4th ECOICOP level, encompassing 112 codes, and operates within a Human-in-the-Loop system that has been trained on 200,000 examples of supermarket receipts. This comprehensive approach ensures both accuracy and reliability in the classification of household expenditure data.

2. METHODOLOGY

Beyond the ESSnet HBS app, our team explored various OCR solutions to enhance the Household Budget Survey. We solicited users to send receipts or images, which were subsequently stored in SQL and Oracle databases for detailed analysis. Initially, we employed a custom Python OCR script; however, we faced significant challenges, including the necessity

for an extensive training set and performance-related issues. As a result, we opted to adopt Azure OCR, capitalizing on its performance and scalability advantages.

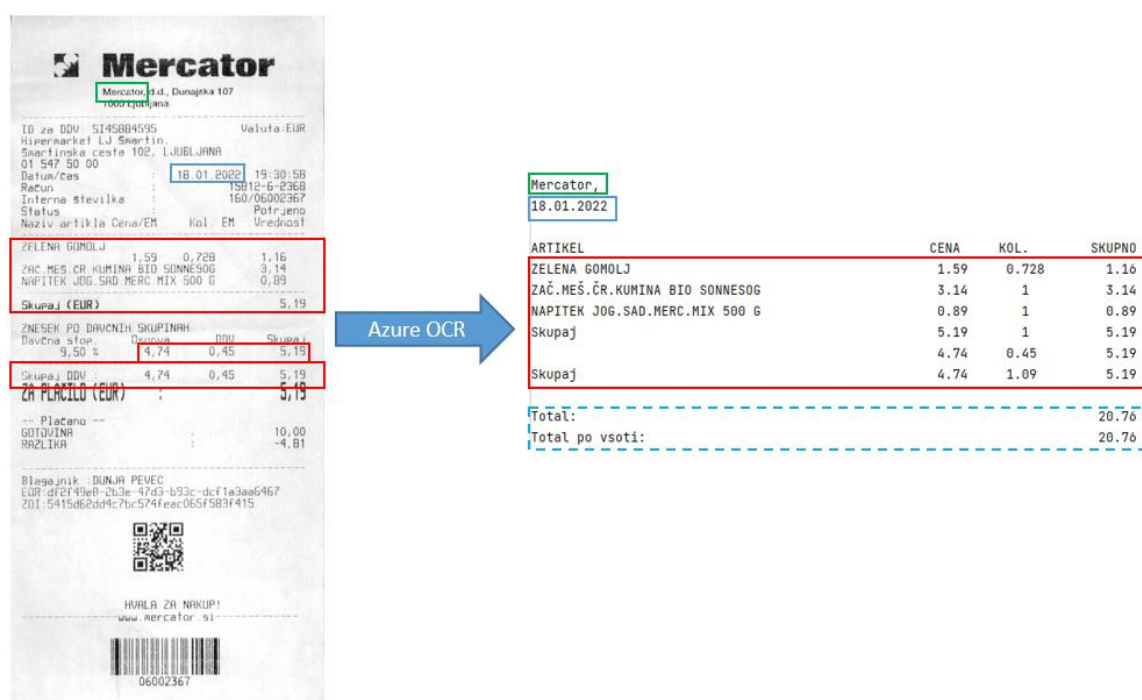


Figure 1. The process of scanned receipts

Our OCR implementation for the HBS began with a daunting 2,576 target codes at the 5th and 6th levels of ECOICOP. To streamline the process, we filtered these codes by eliminating non-essential items such as services and parking tickets. We particularly focused on excluding food and drink entries from restaurants, bars, and pubs, which enabled us to merge classifications and ultimately arrive at a final count of 191 codes that primarily represent food, drink, and tobacco articles. This classification utilized a hierarchical combination of various models to ensure accurate categorization.

During training, we tested several machine learning models, including decision trees, logistic regression, multi-layer perceptrons, support vector machines, and multinomial Naïve Bayes. In the end, only three models proved useful for the classification process: multilayer perceptron, logistic regression and decision tree. Our training set consisted of nearly 8 million articles, supplemented by an additional 1 million prepared for each classification step, which included numerous duplicates and "perturbed" duplicates to simulate potential OCR errors. Data cleaning was performed meticulously, eliminating empty articles and receipts from known establishments, as well as removing erroneous items such as sums and taxes. The workflow for classification commenced with a rule-based selection among three models, where any predicted class scores below 70% necessitated a manual check via a Human-in-the-Loop (HitL) process for untrusted classifications, ensuring the accuracy and reliability of the final results.

3. RESULTS AND PRACTICAL APPLICATION

By refining the classification of 160,000 articles from 20,000 receipts into 191 relevant codes, the project enhances the accuracy and efficiency of categorizing household expenditures. This streamlined approach allows for better insights into consumer behaviour and spending patterns. Based on our findings, the future goal should focus on improving the preparation of training sets, including manual annotations, upsampling small classes, and downsampling

larger ones, leading to a more balanced and representative dataset. This ensures that machine learning models can be trained more effectively, resulting in higher classification accuracy. Implementing the HitL process for 20,000 manual re-checks ensures that classifications meet a quality threshold. Providing suggestions for validation allows for continuous improvement of the models, enhancing reliability in the classification of household expenditures. The use of a hierarchical combination of machine learning models facilitates a structured approach to classification, allowing for nuanced categorization of articles while managing the complexity of a large number of target classes.

While the overall findings offered numerous lessons learned, we were not entirely satisfied with the training sets, recognizing the considerable effort required to enhance them. Additionally, we explored applying our models to scanned data from major retailers, encompassing 15 million articles, which also benefited from having GTINs for improved classification accuracy. Overall, the results of this project not only improve the operational aspects of the HBS but also contribute to a better understanding of consumer spending, ultimately benefiting both researchers and policymakers.

4. MAIN FINDINGS

The main findings of the project reveal that the initial classification framework of 2,576 ECOICOP codes was effectively streamlined to 191 relevant codes, enhancing the focus on critical expenditure categories like food, drink, and tobacco. The importance of improved data quality was highlighted through better training set preparation, resulting in more accurate manual annotations and balanced class distributions via targeted upsampling and downsampling strategies. Implementing a Human-in-the-Loop process for 20,000 manual re-checks demonstrated that suggested corrections significantly enhance model performance and address uncertainties in predictions. Additionally, a hierarchical combination of various machine learning models proved effective in managing the complexity of classification tasks.

Our future goals focus on enhancing the preparation of training sets through improved manual annotations, which will involve upsampling small classes and downsampling larger ones to achieve a balanced dataset. We aim to refine our approach to hierarchical classifications, assessing whether there are too many or not enough steps in the process, and clearly defining the thresholds for classification accuracy. Additionally, we will determine whether the output should consist solely of suggestions or if it should provide actual values for a more precise classification system.

5. REFERENCES

- [1] Eurostat (2024). Household budget survey. Retrieved from: <https://ec.europa.eu/eurostat/web/microdata/household-budget-survey>

Protecting high-resolution grid data with additive noise while retaining fitness for use

Martin Möhler (Destatis)

1. INTRODUCTION

In the 2021/22 European census round, countries are producing demographic aggregates for geographic grid cells. These data products pose challenges with respect to confidentiality, since they are unprecedented in terms of spatial granularity and can, furthermore, be combined with aggregates for administrative regions to derive values for even smaller areas by a process called ‘geographic differencing’ [1]. Germany and several other countries protect grid data with the Cell Key Method (CKM) [2], a disclosure control method based on additive noise [3] that is suited to protect against geographic differencing risks [4]. We assess the protective effect of CKM w.r.t. geographic differencing, as well as its impact on analytical validity of the resulting ‘noisy’ grid data product, focusing on the highest resolution (100m by 100m).

2. METHODS

2.1. Additive noise

For a given grid cell $j = 1, \dots, M$ we do not publish the true subpopulation count X_j (e.g. number of people aged 60+), but rather its perturbed counterpart $X'_j = X_j + \Delta X_j$, where $\Delta X_j \in [-D..D]$ is an additive noise component, drawn from a noise distribution with constant variance $Var(\Delta X) = V$ and constant maximum absolute value D .

2.2. Measuring loss of fitness for use

Adding noise may inhibit the utility of grid data for further analysis and the question arises, how to assess the extent of inhibition. Following [5], a high-resolution grid is intended for “supporting analysis closely related to specific policy objectives”. This implies that a potential inhibition of utility should not only be assessed by some average-case metric, like mean squared error (MSE), for instance. One might also look at common use cases and compare their results before and after noise addition. Here, we choose to compare accessibility of subpopulations to common points of interest (POI), as measured by routed travel distance. Accessibility for a cell is measured as travel distance from the cell’s centroid to the POI that is reached fastest [6]. We classify the continuous travel distance estimate into accessibility categories and aggregate cell-level (sub-)population counts by category. This yields relevant statistics like, for instance, the number of people aged 60+ that live at distances of 5 to 10min by car to the next pharmacy. The deviation between these statistics before and after noise addition quantifies the loss of fitness for use.

Here, we drew 100 random areas of 100km² extent (100 by 100 grid cells) and aggregated for each the number of units per accessibility zone. We compared the aggregate from original cell

values ($\sum X_j$) to that from cell values with added noise ($\sum X'_j$). Routed distance was measured in meters for pedestrian accessibility and in minutes for accessibility by car.

3. RESULTS

3.1. Protection afforded by additive noise against geographic differencing

We considered an encompassing region A and an encompassed region B , as in [1], where one is an administrative area (e.g. LAU), the other a union of m grid cells. The value of the differenced region $X_{A \setminus B} = X_A - X_B$ is the target. With additive noise we get $X'_{A \setminus B} = (X_A + \Delta X_A) - (X_B + \Delta X_B) = X_{A \setminus B} + \Delta X_A - \Delta X_B$, where either ΔX_A or ΔX_B is the noise component of the value in the administrative area and the other is the sum of m noise components from as many grid cell values. As a result, the differenced value is perturbed by $m + 1$ noise components and is therefore a random variable with $X'_{A \setminus B} \in [\max(0, X_{A \setminus B} - (m + 1) \cdot D) .. X_{A \setminus B} + (m + 1) \cdot D]$, $\text{Var}(X'_{A \setminus B}) = (m + 1) \cdot V$ and $E(X'_{A \setminus B}) = X_{A \setminus B}$. Since the published data allows only a single realization, the last property is not problematic for disclosure, but reflects preservation of utility. The higher m , the higher the uncertainty around the true differenced value.

3.2. Fitness for use results

Exemplarily, we show below results for the accessibility of pharmacies. For example, the first graph in the top left of Fig.1 shows for the 100 selected areas the number of people that live up to 200m (pedestrian walk) from their closest pharmacy; the horizontal axis shows the count from original (unperturbed) grid cells, the vertical axis that from perturbed grid cells. The dashed line in each graph marks perfect equivalence. Fig.2 shows the breakdown by travel time with a car (excerpt).

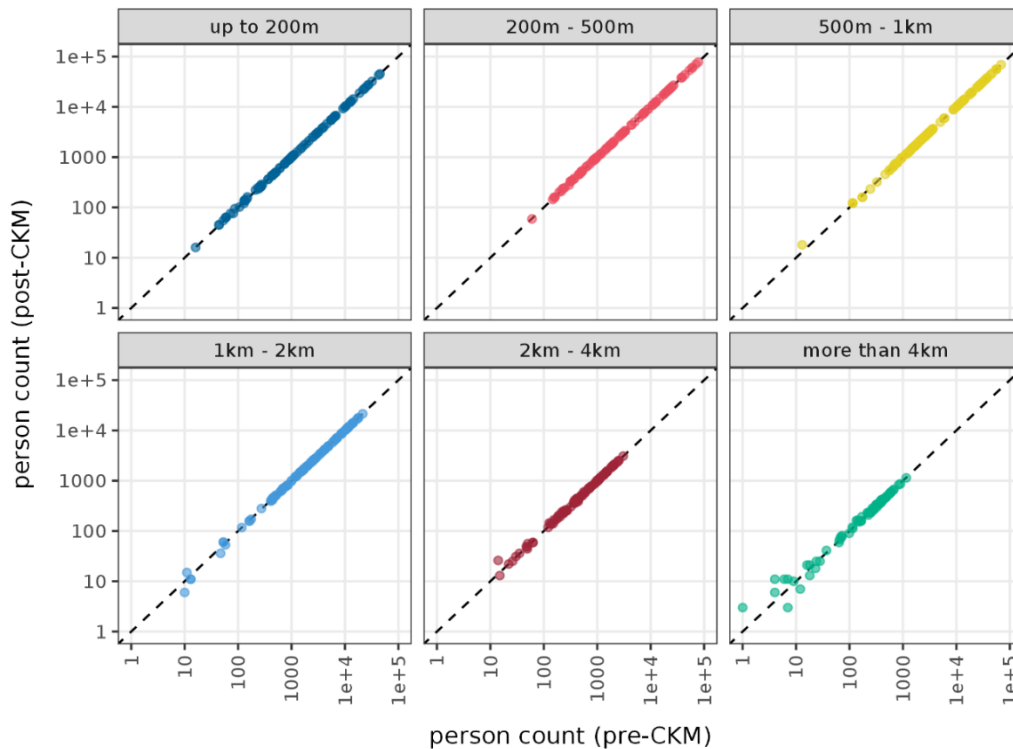


Figure 1. Population count by routed travel distance (pedestrian) to nearest pharmacy before and after applying CKM for 100 random focus areas

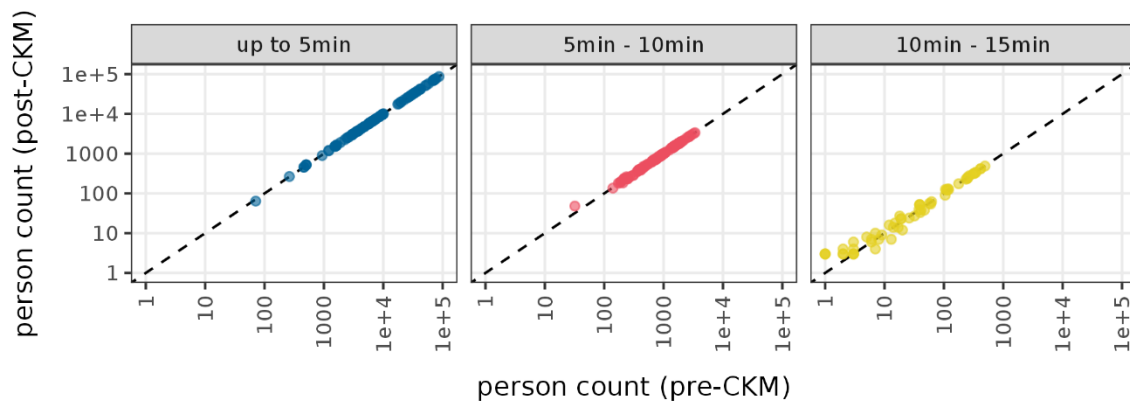


Figure 2. Population count by routed travel time (car) to nearest pharmacy before and after applying CKM for 100 random focus areas (excerpt)

We find average deviations after aggregation of between ca. $1.5 D$ and $7.1 D$. It can be seen that relative error is almost always small (mind the log-scale), except for the case where only very few inhabited cells with a small combined population are aggregated.

4. CONCLUSIONS

CKM is well suited for protecting data against geographic differencing. The fitness for use of high-resolution demographic grid data is overall well preserved after its application. Notable relative error can result, if the number of cells aggregated is very small, or if the average population size per cell is low. For robust analysis, we suggest building accessibility zones such that each category includes a medium to high number of grid cells.

Further risk and utility aspects of CKM for high-resolution grid data are discussed in [4].

5. REFERENCES

- [1] V. Costemalle, Detecting geographical differencing problems in the context of spatial data dissemination, *Statistical Journal of the IAOS*, 35(4), 2019, pp. 559-568.
- [2] M. de Vries, M. Golmajer, R. Tent, S. Giessing and P.-P. de Wolf, An overview of used methods to protect the European Census 2021 tables, *UNECE Expert meeting on statistical data confidentiality*, Sep. 26-28, Wiesbaden, Germany, 2023.
- [3] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. Schulte Nordholt, G. Seri, P.-P. de Wolf, R. Tent, A. Młodak, J. Gussenbauer and K. Wilak, *Handbook on Statistical Disclosure Control*, 2nd Edition, 2024, <https://sdctools.github.io/HandbookSDC/>
- [4] M. Möhler, J. Jamme, E. de Jonge, A. Młodak, J. Gussenbauer and P.-P. de Wolf, *Guidelines for Statistical Disclosure Control methods applied on geo-referenced data*, 2024, <https://github.com/sdcTools/GeoSpatialGuidelinesSources/releases>
- [5] H. Poelman, Grid-based analysis in regional and urban policy: a state of play, *EFGS Conference presentation*, Sep. 7-8, 2021.
- [6] E. Miller, Measuring accessibility: Methods and issues, *International Transport Forum Discussion Papers No.2020/25*, OECD Publishing, Paris, 2020.

R package for creating multi-resolution grids of census and sample data

Jon Skoien (ARHS Developments), Nicolas Lampach, Helena Ramos,
Rudolf Seljak (Eurostat), Renate Koeble (ARHS Developments), Linda See (IIASA)
& Marijn van der Velde (European Commission - JRC)

1. INTRODUCTION

Data from agricultural censuses and surveys are used to monitor the structure of agriculture over time. Based on a collection of individual information at the farm holding level, these statistical data are often released as aggregates of coarse resolution administrative zones to comply with statistical confidentiality requirements. One example is the European farm survey (also known as the Farm Structure Survey (FSS)), which consists of a census of millions of farm holdings across Member States (MS) every ten years and sample surveys that take place every 3 to 4 years. Publicly available data are only released at NUTS2 or coarser resolution (NUTS1 or national).

The Implementing Regulation (EU) 2018/1874 allows for dissemination of gridded data, but requires for confidentiality reasons that each grid cell has a minimum of 10 holdings, in addition to “the normal disclosure control mechanisms for tabular data”. An additional requirement is then that an aggregated value can only be disseminated if the values from the two largest producers in the grid cell constitute less than 85% of the overall value. Lastly, for sample data, there is also a requirement that the expected estimation error of the aggregated value must be less than 35% of the value itself.

Regular grids will either imply that the grid cells will have to be large everywhere, or that a large number of the grid cells will have to be suppressed. The Implementing Regulation suggests the use of nested grids with 1 km, 5 km, 10 km or larger, for dissemination of the aggregated data. However, it does not specify how to construct these grids. The 1 km grid reflects the resolution of the records in the FSS. The reported geo-location should be the coordinates of the lower left corner of the 1 km grid cell where they are actually located.

Grids with hierarchical resolutions, often referred to as multi-resolution grids or quadrees, are not new [1], [2]. However, creation of such grids have traditionally included some problematic issues, such as some very large grid cells on coast lines and borders, different grids for different variables and limited options for applying different restrictions for different applications.

Here we will present an R package with the name MRG that overcomes these issues. It can:

- apply a contextual suppression which will suppress some values if (and only if) this gives increased resolution in the neighborhood
- create a joint grid for a random number of variables, where all of them will respect the confidentiality rules
- apply different restrictions, either through an adaption of the existing ones, through modification of variables, or by adding user defined functions.

2. METHODOLOGY

Figure 1 presents a simplified example of the considerations while creating a multi-resolution grid with the methodology implemented in the R package. There is only one restriction in this case, each grid cell needs at least 10 holdings. The grid to the left shows the number of holdings in the original 1 km grid. There are some empty cells, the rest of the grid cells are all confidential, as they have less than 10 holdings. The values have been hierarchically aggregated to larger grid cell sizes in the right grid, to make sure that each green grid cell respect the confidentiality rule. However, there is one suppressed value in the right grid. There is only one holding in this grid cell, and we could have created a larger grid cell with 12 holdings in the lower left corner. But for many applications it is better to suppress this value, and rather keep the remaining three smaller grid cells, with 11, 0 and 0 holdings, respectively. This feature is controlled by a parameter that specifies the minimum share of the potential new cell for a grid cell to be aggregated. For example, if the limit is 0.1, a grid cell would only cause an aggregation to lower resolution if the value in the grid cell is more than 10% of the value in the lower resolution grid cell. Instead, it would be left as it is, and will be suppressed in the post-processing step. This is the case in the example below, where the suppressed grid cell only represents 1/12 of the holdings of the potential larger grid cell. Normally, a smaller value would be recommended, to avoid that too many grid cells are suppressed.

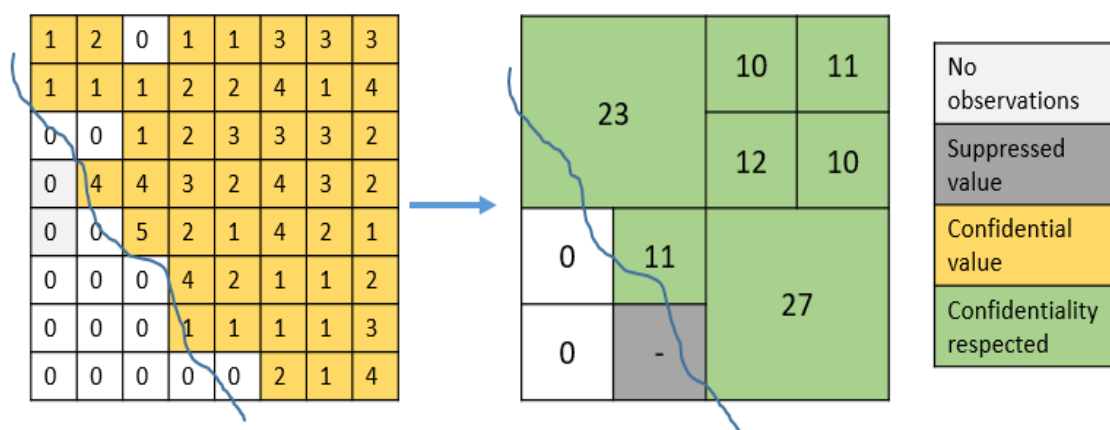


Figure 1. Example of creation of a multi-resolution grid, where each grid cell must have a minimum of 10 holdings

The package includes a synthetic example data set, where typical FSS variables have been generated for Denmark. These will help in understanding the methodology in the package.

3. RESULTS AND PRACTICAL APPLICATION

Figure 2 shows two example grids. In the left panel, only the frequency rule has been applied, whereas also the dominance rule has been applied to produce the right grid. The red circles show some changes before the grid, where grid cells have been aggregated because the largest farms were above the 85% limit.

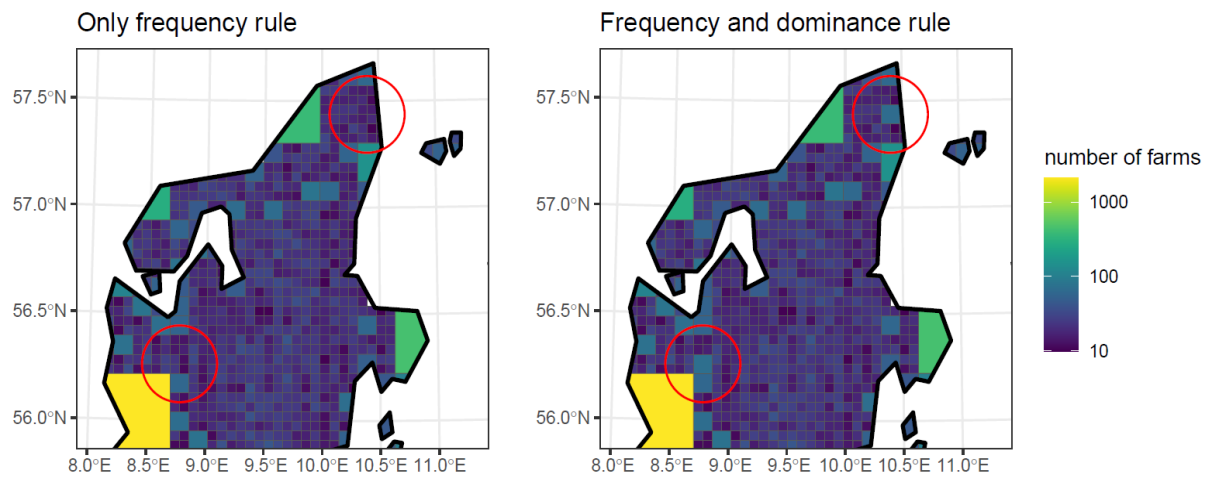


Figure 2. Example grids of farm numbers, indicating the number of farms per grid cell. The panel to the left only applies the frequency rule (at least 10 holdings per grid cell), whereas also the dominance rule for utilized agricultural area (UAA) has been applied for the right grid (the UAA of the two largest farms should not include more than 85% of the UAA of the grid cell).

We can notice some large grid cells on the coast in Figure 2, which typically occur when the hierarchal aggregation first aggregates a grid cell with empty grid cells in the sea, before it is finally aggregated to a large grid cell including many smaller and non-confidential grid cells inland. This is what we would intend to avoid using the before mentioned contextual suppression. Figure 3 has the same figure in the upper left panel, where no suppression has been applied. We can notice some larger grid cells within the circles. With increasing suppression level in the following panels, we can see how these are replaced with smaller grid cells. However, we can also notice that there are more suppressed grid cells (shown with red color) with increasing suppression level. The optimal choice will depend on application.

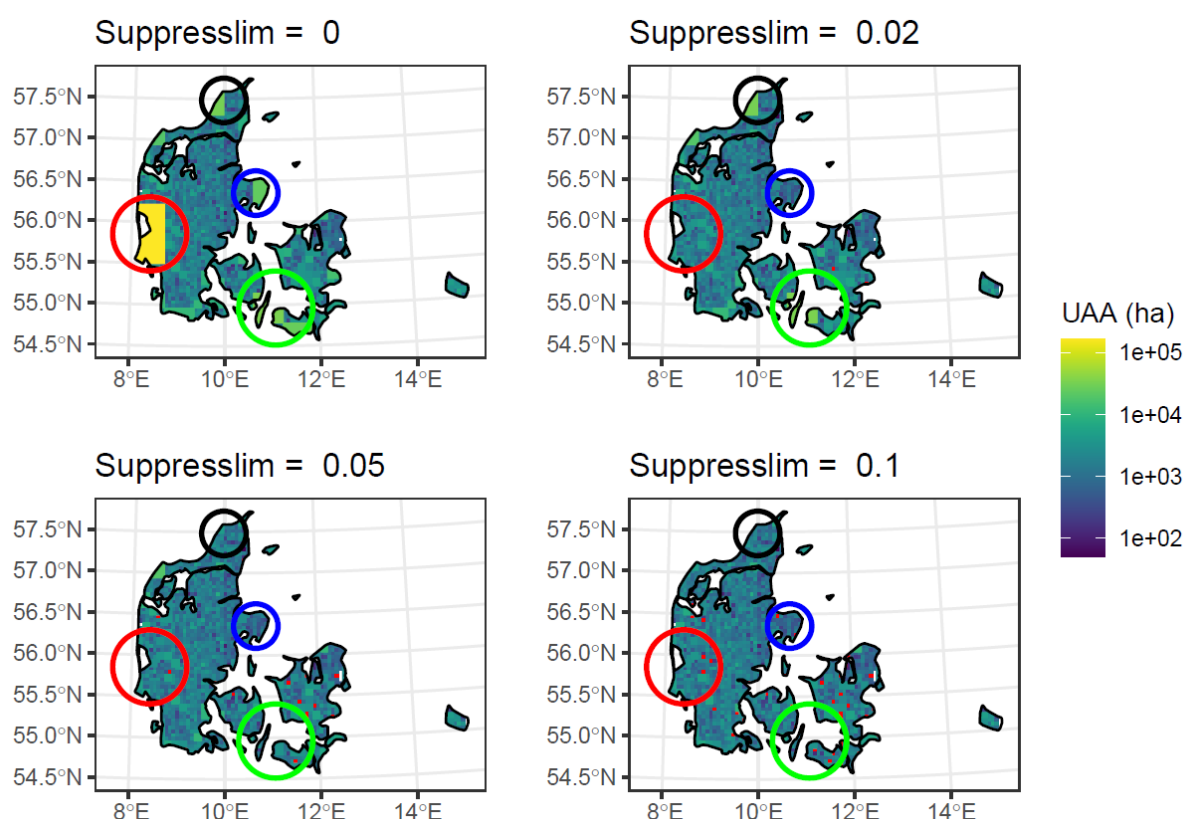


Figure 3. The distribution of grid cells after applying different degrees of contextual suppression

The package also has some other features that are useful for gridding

- Gridding of variables together
- Reliability estimation – estimate of expected error
- Parallel processing for some of the computation
- Post-processing, including rounding to the requested level
- Possibilities for intermediate plots and diagnostic output

4. MAIN FINDINGS

This package will make it possible to disseminate confidential data with a resolution that has not previously been possible. First of all, the method has already been used to create a first set of maps based on the European wide Farm Structure Survey, which will then be accessible for further analyses. Second, as the package is freely available from the Comprehensive R Network (CRAN), it can be applied by all services that would like to share aggregated confidential micro-data.

5. REFERENCES

- [1] M. Behnisch, G. Meinel, S. Tramsen, and M. Diesselmann, Using quadtree representations in building stock visualization and analysis, *Erdkunde*, 67-2 (2013), 151–166.
- [2] R. Lagonigro, R. Oller, and J. C. Martori, AQuadtree: An R Package for Quadtree Anonymization of Point Data, *R Journal*, 12 (2020), 1–17.

Recent Guidelines for Statistical Disclosure Control

Peter-Paul de Wolf (Statistics Netherlands)¹

Abstract: From 01-09-2020 until 31-08-2024 a project took place that was partly financed by a Eurostat Grant (B4449-4655-2019-Methodology): STACE, Statistical methods and tools for time series, seasonal adjustment and statistical disclosure control. One of the packages was the continuation of a Centre of Excellence on Statistical Disclosure Control. This paper presents the main outcomes of that project and discusses some of the lessons learned.

1. INTRODUCTION

On September 1st, 2020 the STACE project started for a duration of four years. This project – Statistical methods and tools for time series, seasonal adjustment and statistical disclosure control– was co-funded by Eurostat via grant B4449-4655-2019-Methodology. The project consisted of two main work packages: a center of excellence on statistical disclosure control (CoE on SDC) and a center of excellence on time series analysis and seasonal adjustment (CoE on TSA). This paper will focus on the achievements of the CoE on SDC.

The CoE on SDC was in some sense a continuation of the work accomplished in a previous European project, the Framework Partnership Agreement no 11112.2014.005-2014.533. In the consortium that formed the CoE on SDC in the STACE project, the national statistical institutes (NSIs) of the following countries were represented: The Netherlands (work package leader), Austria, France, Germany, Iceland, Poland and Slovenia. Additionally, Bulgaria was part of the CoE on SDC in its role as overall project coordinator.

The main goals of the CoE on SDC could be formulated as:

- Provide user support for and maintenance of SDC software tools
 - Provide support to producers of European statistics with regard to the implementation of these tools
 - Provide guidance (guidelines) on SDC methodology
- To achieve those goals, the project group intended to

- Maintain the current User Group on SDC Tools and provide that User Group with means for communication among each other as well as with the developers of software to apply SDC
- Maintain and extend the GitHub sources repositories
- Set up a facility to exchange knowledge between users
- Exchange knowledge between developers

In the next section we will discuss the main tangible results. In section 3 we will mention some lessons learned and contemplate a bit on possible future continuation of the work of the CoE after the project has ended on August 31st, 2024.

¹ The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands, neither does the European Commission accept any responsibility for use that may be made of the information this paper contains.

2. RESULTS

The main tangible results of the project are three guidelines: a guideline for SDC methods applied to geo-referenced data, a guideline for SDC methods applied for census and demographics data and an upgrade of the general handbook on SDC. We will discuss each guideline in the next subsections.

Additionally, during the project several new versions of software tools were released. The most recent releases can be found via <https://sdctools.github.io/UserSupport/Releases.html>

A less tangible result of the project is the way the User Support was continued. The GitHub repository <https://github.com/sdcTools/UserSupport> is used to file issues and ask questions about SDC software. During the project 60 issues were created and 40 issues were closed. Most of the still open issues concern requests for new features or improvements. The User Group on SDC Tools has grown to 102 members (status 3/12/2024).

2.1. Guidelines for SDC Methods Applied on Geo-Referenced Data

Users of statistical data are often interested in spatial distribution patterns. However, when a location is too detailed (e.g., an address) or when its neighbourhood has few inhabitants (e.g., one isolated household), the displayed information at that location is very disclosive. Traditional SDC methods are generally applied but often do not take into account the geo-spatial character of the data and the output. For that reason the guidelines for SDC methods applied to geo-referenced data were written.

The guidelines start with some preliminary concepts and definitions targeted at geo-spatial disclosure control. As with all SDC methods, the methods proposed in the guidelines will influence the risk of disclosure as well as the conservation of information. The guidelines thus assign two chapters to these two aspects, discussing the particularities of disclosure risk and information loss in case of geo-spatial data/output. After these introductory chapters, several SDC methods specifically targeted at publishing data with geo-spatial character are described. Finally, a chapter on some practical user cases is provided to show how some of the proposed methods can be applied. The methods described in the guidelines can be used to protect publishable output, i.e., tabular data or data plotted on a (cartographic) map. The methods are not intended to protect microdata to be released as microdata.

A detailed table of content is:

1 Introduction

2 General Considerations

2.1 How identifying are geo-references?

2.2 Outputs based on geo-referenced data

3 Measuring Disclosure Risk for Geo-Referenced Data

3.1 Sensitivity measures for aggregates

3.2 Differencing issues

4 Information Loss Measures for Geo-Referenced Data

4.1 Measures of distributional distance

4.2 Measures of spatial association

4.3 Local Information Loss

4.4 Information Loss Measures and the Modifiable Area Unit Problem

5 Disclosure Control Methods for Geo-Referenced Data

5.1 Quadtree-based methods

5.2 Spatial Smoothing Methods

5.3 Cell Key Method

5.4 Targeted Record Swapping

5.5 Auxiliary Methods

5.6 Methods Overview

6 Case Study: Population grids with the Cell Key Method

- 6.1 Motivation
- 6.2 Assessing risk
- 6.3 Applying CKM
- 6.4 Assessing protection results
- 6.5 Concluding thoughts

Bibliography

List of software tools

Glossary

The latest version of the guidelines can be found here:

<https://github.com/sdcTools/GeoSpatialGuidelinesSources/releases>

2.2. Handbook on SDC – second edition

In a European project launched in 2005 (Grant agreement no. 25200.2005.001-2005.619) a handbook on Statistical Disclosure Control was written. The current project upgraded that handbook. To be able to do that, we contacted the original authors (as far as possible) and asked for their permission, which we got.

The original handbook from 2006 was a Word document. Firstly, we changed the format to quarto markdown. That way it is relatively easy to produce several output formats, including a quarto book (i.e., a html-based format) and pdf-format. Secondly, we updated some information and added some new information on new methods as well as on some new insights about disclosure risk measures and information loss measures.

The advantage of a quarto book available online, is that it is relatively easy to continuously update the content. That way it can be a “live” document. Obviously, the versioning should then be clear. Technically that is taken care of by using the GitHub versioning system. However, it should also be reflected in the “released” versions of the handbook. This is something that needs to be defined clearly by the maintainers of the GitHub repository.

The quarto book of the latest release of the handbook can be found here: <https://sdctools.github.com/HandbookSDC>.

2.3. Guidelines for SDC Methods Applied for Census and Demographics Data

At least once every ten years all countries in the world are recommended by the UN to conduct a Population and Housing Census. In the European Union all member states have to conduct such a census in years ending on a ‘1’. Historically, censuses were the only source of information about the number of people residing in a country. A number of countries used these censuses to set up population registers. Over time both the censuses and population registers contain more and more variables. In a European project on harmonizing the SDC of census tables, two methods were proposed that add noise to the data, as opposed to the more traditional suppression of information or coarsening of variable categories.

Since census tables and general population demographics data are related, the guidelines contain an overview and comparison of different available SDC methods. A detailed table of contents is:

1 Introduction

2 Overview of SDC methods for Census and Demographic Table

- 2.1 Current methodology
- 2.2 Software available for current methodology
- 2.3 Comparison of the SDC methods

3 Consistency and Disclosure Risk Issues

- 3.1 Introduction

- 3.2 Differences in used SDC methods
- 3.3 Differences in reference dates
- 3.4 Concluding remarks
- 4 Defining parameters**
 - 4.1 Parameter definition by trial and error
 - 4.2 A deeper look at the parameters of the Cell Key Method
- 5 Communication of SDC methods to Data Users**
 - 5.1 Introduction
 - 5.2 Examples from Statistics Slovenia
 - 5.3 Examples from Statistics Netherlands
 - 5.4 Examples from Official Statistics in Germany
 - 5.5 Examples from Statistics Austria

The latest version of the guidelines can be found here:

<https://github.com/sdcTools/CensusDemographyGuidelinesSource/releases>

3. LESSONS LEARNED AND OUTLOOK

The outbreak of Covid influenced the progress of the project. Not only the communication between the partners was more difficult, it also influenced the timing of the intended output of the project. E.g., the census was postponed due to Covid and hence the work on guidelines related to the census could take place later than expected. One lesson learned is thus that this kind of “unexpected circumstances” could or should be included in the project proposals in some way, e.g. by allowing some flexibility in the agreed output.

A more practical lesson to be learned is that getting in contact with original authors (like in our case when upgrading the 2006 version of the Handbook on SDC), can be quite difficult. To make this a bit easier in the future, we have now published the three guidelines under the Creative Commons License BY-SA (<https://creativecommons.org/licenses/by-sa/4.0/>). That means that people are free to Share (copy and redistribute the material in any medium or format for any purpose, even commercially) and Adapt (remix, transform, and build upon the material for any purpose, even commercial) under the following terms: Attribution (you must give appropriate credit, provide a link to the license and indicate if changes were made; you may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use) and ShareAlike (if you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original).

For a possible continuation of the CoE on SDC we would like to mention that financial support for a longer period is preferred, especially concerning the development and maintenance of SDC tools. However, since a CoE on SDC should be able to act on topical issues as well, a possible future project should facilitate a flexible way of determining topics and their priorities.

ACKNOWLEDGEMENTS

We would like to thank the main authors of the mentioned guidelines for their project participation, resulting in the release of the guidelines (alphabetical order):

Fabian Bach (Eurostat), Sarah Giessing (Destatis), Manca Golmajer (SURs), Johannes Gussenbauer (Statistics Austria), Julien Jamme (INSEE), Edwin de Jonge (Statistics Netherlands), Nynke Krol (Statistics Netherlands), Rob van de Laar (Statistics Netherlands), Andrzej Młodak (Statistical Office Poznań), Martin Möhler (Destatis), Eric Schulte Nordholt (Statistics Netherlands), Reinhard Tent (Destatis), Marieke de Vries (Statistics Netherlands),

Kamil Wilak (Poznań University of Economics) and Peter-Paul de Wolf (Statistics Netherlands).



**Co-funded by
the European Union**

Redesigning the online French Census questionnaire: a user-centered methodology and its limitations

Dirand Julia & Sébastien Hallépée (Insee)

1. INTRODUCTION

In the context of online questionnaires used for data collection, respondents often use a variety of devices—such as computers, smartphones, and tablets—to submit their answers. User Experience (UX) design plays a critical role in ensuring that respondents can complete surveys easily and effectively across all these devices. The need for adaptive and responsive survey designs is therefore essential.

Prior to 2024, the online French Census questionnaire was optimized only for desktop use, creating barriers for many respondents, particularly in regions where mobile access is more prevalent. The 2024 redesign of the online French Census questionnaire focused on applying UX design principles and methodology to ensure the questionnaire was fully responsive and accessible across all devices, particularly mobile. Care was taken to implement these improvements without significantly changing the questions, as the French Census operates on a five-year rolling system. This method surveys a portion of the population annually, and the data from each year is combined to provide an overall demographic picture. Maintaining consistent questions is crucial to avoid introducing statistical bias and ensuring comparability of data across different years. The key objective was to reduce partial non-response by improving the user experience, to make it easier for respondents to complete the questionnaire in its entirety while preserving the integrity of long-term data comparisons.

We will explore the methodology used, the impact of the redesign and the challenges that emerged post-implementation. Key findings demonstrate the successes of making the questionnaire responsive, but also reveal certain limitations, emphasizing the importance of ongoing testing and refinement.

2. METHODOLOGY

Before embarking on the redesign, the 2023 version of the online French Census questionnaire, which was optimized only for desktop use, was thoroughly analysed. Usability challenges were identified through quantitative analytics and an initial round of usability testing. This was followed by a prototyping phase, after which a second round of usability testing was conducted.

2.1. First round of usability testing

The first round of testing aimed to evaluate the actual 2023 questionnaire on both desktop and mobile devices. Usability tests were conducted to observe how respondents navigated and interacted with the existing system. During this phase, several key usability challenges were uncovered:

- Navigation difficulties: respondents had trouble efficiently navigating the questionnaire due to complex layouts and unclear progression through the form. On mobile, the lack of optimized

mobile-first design resulted in even greater frustration. Navigation was cumbersome, with some elements being hard to locate.

- **Unresponsive elements:** on mobile devices, users encountered several unresponsive elements, particularly buttons and input fields that were difficult to engage with, causing delays and frustrations. Additionally, the interface required horizontal scrolling, which is particularly problematic on mobile devices. Horizontal scrolling disrupts the natural flow of reading and navigation on smaller screens, making it harder for users to interact with content efficiently. It also increases the likelihood of missing important information, as users may not realize there is additional content off-screen.
- **Touch input difficulties:** on mobile, respondents encountered difficulties when using touch input for interacting with form elements. Buttons and fields were too small or poorly positioned. This resulted in increased completion time and the likelihood of abandoning the questionnaire altogether.

Based on this feedback, refinements were planned to improve the overall user experience and ensure adaptability across various devices and screen sizes. These changes would aim to create a more intuitive and accessible experience across all platforms, especially for mobile users, who represented a growing portion of the respondent base.

2.2. Prototyping

Based on the initial analysis and feedback, the UX designer launched a prototyping phase to address usability issues and ensure the questionnaire was adaptable to mobile devices. Adopting a "mobile-first" approach, we prioritized optimizing the mobile experience before adapting to larger screens like desktops and tablets, ensuring respondents on smartphones could navigate intuitively. The prototypes focused on a clean interface with minimal text input, vertical scrolling, and clear call-to-action buttons. To build trust among respondents unfamiliar with INSEE, the redesign integrated the French Government's design system, offering a familiar interface similar to other public services. After finalizing the prototypes, another round of testing was conducted to evaluate the changes.

2.3. Second round of usability testing

This phase tested the new prototypes to confirm that the design changes were effective across all devices. The tests demonstrated significant improvements in mobile navigation and performance, leading to further refinements and validation of the design for both platforms.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Indicators from the French Census between 2023 and 2024

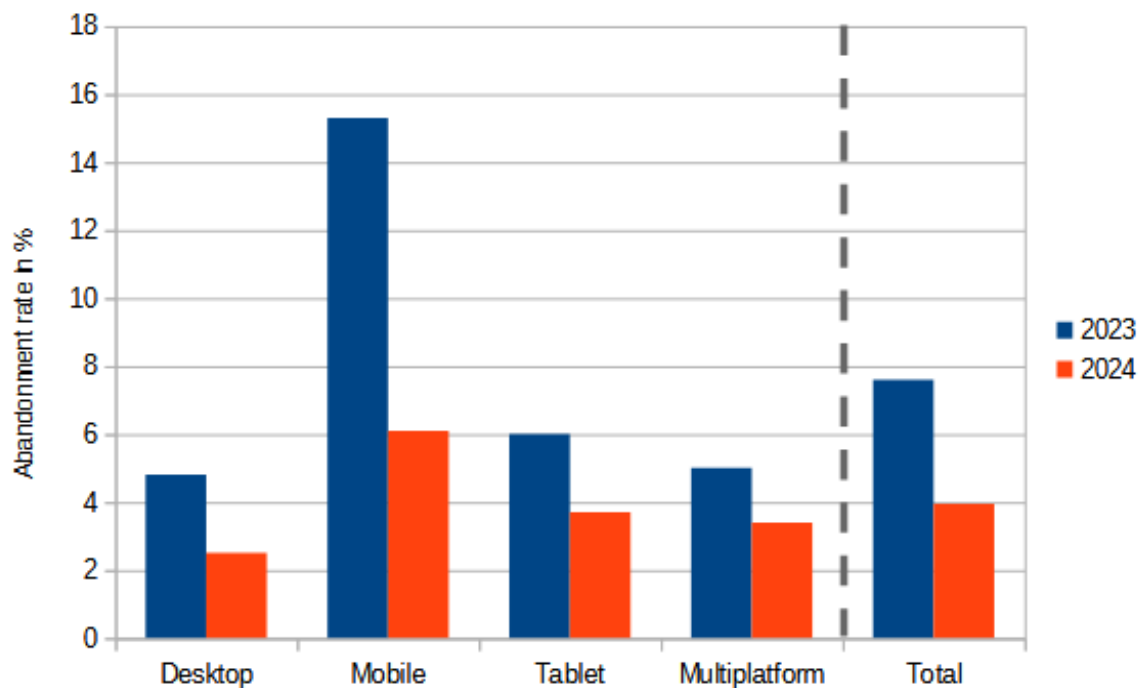


Figure 1. Online French Census survey abandonment rate by devices between 2023 and 2024.

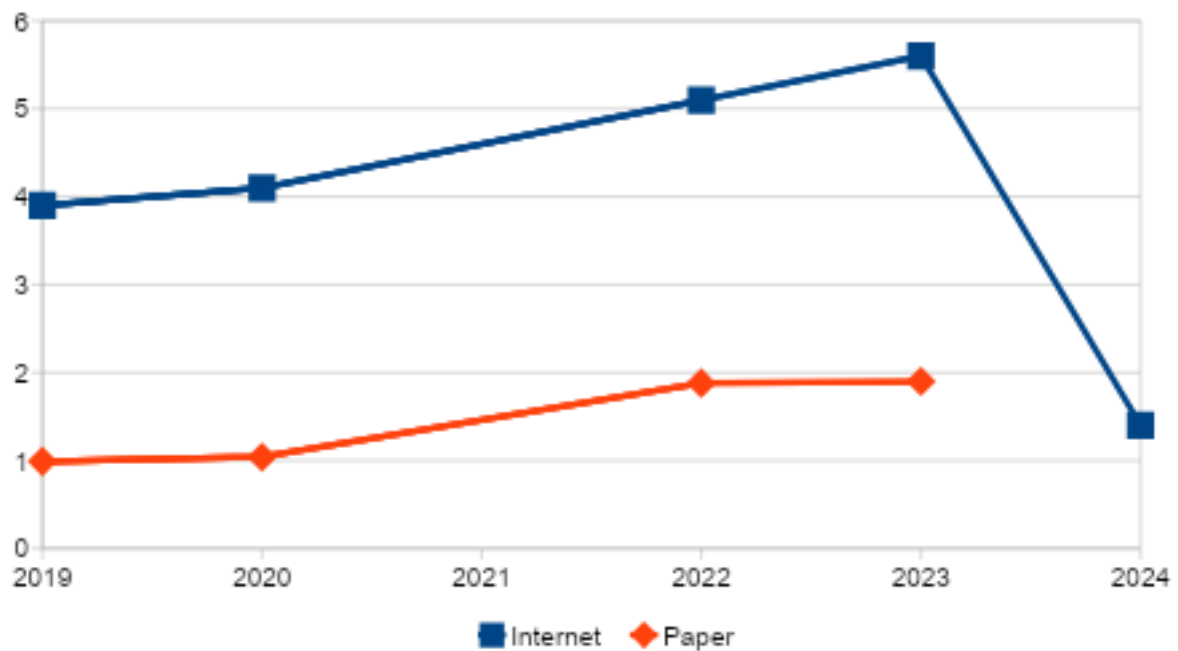


Figure 2. French Census non-response rate on paper and internet over the years

Note: The non-response rate for paper in 2024 is not yet available as the results from the paper questionnaires have not been processed.

4. MAIN FINDINGS

The redesign process was guided by key quantitative metrics to assess its impact on user behaviour and data quality. Several indicators were tracked to compare the performance of the new design against the 2023 version:

- completion rates: monitored to evaluate how the redesign improved users' ability to successfully complete the questionnaire across different devices, with a particular focus on mobile users.
- abandonment rates: tracked the number of users who started but did not finish the questionnaire, providing insight into usability and engagement.
- non-response rates by device: Measured non-responses to assess the effectiveness of the questionnaire across various platforms.
- device-switching behaviour: tracked respondents who switched devices mid-survey, which could signal frustration or usability issues on one platform.

These metrics revealed the positive impact of the redesign, such as increased completion rates, lower abandonment rates, and a reduction in non-responses.

However, the limitations of usability testing methodology were also exposed. Some changes, such as the inversion of first and last name fields, were not flagged during the testing phase but led to a significant increase in input errors during data collection. In 2023, there were 66,000 instances of respondents reversing their names out of 7 million responses, compared to 410,000 inversions in 2024 with the same number of respondents. This issue, though relatively minor, highlights the importance of complementing usability testing with quantitative performance evaluations. Certain problems can only be detected through large-scale data analysis, underscoring the need for a combination of qualitative and quantitative testing to capture a more comprehensive range of issues.

Regularizing Probability Sample Estimates Through an Angle-Based Similarity Approach

Jacob Westlund (EMOS Programme, Leiden University)

1. INTRODUCTION

Today, estimation using probability samples (PSs) is ubiquitous within National Statistical Institutes (NSIs) and using PSs is considered the gold standard due to their strong theoretical properties and relative ease of application. However, collecting PSs is a relatively costly and time-consuming method of acquiring data (Bakker et al., 2014; Van den Brakel, 2019). This has been further exacerbated in recent decades by survey response rates steadily declining, also affecting NSIs who now find it harder to contact and convince individuals to participate in their surveys (De Leeuw & de Heer, 2002; Luiten et al., 2020). For NSIs, such a decline has a twofold impact. First, non-response threatens the estimates' validity by increasing the risk of bias and potentially altering the selection mechanism outside the researcher's control (Bethlehem, 2009). Second, and more practically, non-response also increases the work needed to reach the same sample sizes, which means that either the observed sample sizes will have to be reduced, or more resources need to be allocated to each probability survey sample (Luiten et al., 2020).

Because of this, there has been a growing interest within NSIs in replacing or supplementing the PSs with non-probability samples (NPSs), which are samples obtained outside the probability sampling framework, through for example register data from the tax office or other administrative authorities. This type of register-based NPS is very appealing since it is a very cheap method for acquiring large amounts of data, with no additional burden or requirements for respondents (Van den Brakel, 2019). The hope is that by incorporating these new types of data sources, NSIs can both improve the quality of their estimates and reduce their costs. However, like traditional PS estimates, these new NPS estimates are no silver bullet as they come with their own flaws. Most problematic is the fact that NPSs come without a known sampling designs or inclusion probabilities, and that they tend to suffer from a combination of selectivity and coverage issues. These two factors combined mean that NPS estimates tend to be biased and, lacking a sampling frame, there is no way for NSIs to correct this bias. This is highly problematic for NSIs who place a large value on the unbiasedness of estimates, often preventing NPS estimates from being used directly in official statistics (Bakker et al., 2014).

Therefore, NSIs are currently facing a problem. Both statistically and financially, relying purely on traditional PSs is becoming a prohibitively expensive and troublesome approach. However, transitioning to the alternative NPSs risks producing biased estimates, which is highly problematic for NSIs whose main purpose is to produce accurate descriptions of a country. Nevertheless, given the high potential of NPSs, the question is raised if there is not some sort of method by which the bias of NPS estimates could be mitigated, directly or indirectly, allowing NPSs to be utilized in official statistics.

Valliant (2020) and Valliant et al. (2018) outline three traditional correction methods, quasi-randomization (also known as propensity score adjustment), superpopulation modelling, and doubly robust estimation. Quasi-randomization is a design-based method where lacking real inclusion probabilities, a PS is used as a surrogate to in some way estimate pseudo-inclusion probabilities for the NPS. Using a common set of auxiliary variables, the PS can be used to estimate the assumed existing but unknown inclusion probability in the NPS. These new

pseudo-inclusion probabilities can then be used to re-weight the NPS estimates to reduce or even remove the bias (Elliott & Valliant, 2017; Valliant et al., 2018). Superpopulation modeling breaks with the design-based approach and rather treats the NPS as just a sample from a theoretically infinite “superpopulation”, where the outcome of interest follows some unknown probability distribution. The goal of superpopulation modeling is then, given a set of auxiliary variables that explain the selectivity, to use the NPS to model the relationship between the auxiliary variables and estimates of interest. Coefficients can then be extracted and applied to a wider population for population-level statistics that should account for the sample selectivity (Elliott & Valliant, 2017; Valliant et al., 2018). Finally, there is the doubly robust method, which is a combination of quasi-randomization and superpopulation modeling. (see for example Chen et al. (2020)).

These three correction approaches can work, however, they rely on similar and often practically problematic assumptions.

Rather than utilizing the NPS estimates alone, several authors have attempted to harness the information from an NPS to improve the estimates of a related PS instead.

Although the methods differ, the general idea is to maintain approximate unbiasedness of the combined estimates through the PS, whilst leveraging the potentially larger sample size of the NPS to reduce their variances. Elliott and Haviland (2007) and Villalobos-Alíste (2022) looked at integrating an NPS with a PS through a composite estimator. Disogra et al. (2011) proposed an estimation method called “blended calibration” where a calibrated PS is combined with an uncalibrated NPS. The combined sample is then calibrated again using differentiator variables from the PS alone, resulting in a final estimate from the combined sample. Finally, Wiśniowski et al. (2020) applied a Bayesian approach to the incorporation problem, using an NPS to construct priors which are used to estimate a posterior distribution in combination with the PS.

2. OBJECTIVE

The above methods do have their use cases, yet none is designed for producing robust results in a scenario with an unbalanced sample distribution (large NPS and small PS), significant NPS selectivity, and a limited number of auxiliary variables. Correction methods are limited by the lack of auxiliary variables to correct for the NPS selectivity, whilst integration methods on the other hand are more widely applicable, but their results are mixed given very biased NPS estimates and a small PS.

Seeing the larger potential in using integration methods, this thesis seeks to contribute to the literature by proposing an alternative type of integration estimator to the composite, blended calibration, and Bayesian approach. The estimator draws from the wider literature on penalized regression, which has seen earlier success in integrating estimates from heterogeneous data sources (see Li et al. (2014), Liang et al. (2020), and Tian and Feng (2023)). The general idea of penalized regression for data integration is to use regression with additional penalties. These penalties constrain a set of target estimates of interest towards a set of auxiliary estimates. While we do not care about the auxiliary estimates, they share some similarities with the underlying target estimands. The hope is that this will increase the accuracy of the target estimates by leveraging information about the magnitude and direction of the target estimands of interest from the auxiliary estimates, reducing their variance for only a marginal increase in bias.

Therefore, given its earlier success in general data integration but unknown utility in official statistics, the overarching goal will be to apply a specific type of penalized regression in a context more relevant to NSIs and to answer the question of: How can penalized regression be used to incorporate non-probability samples into official statistics?

3. METHODS

To answer the above-mentioned question, this thesis applies the Angle-Based Transfer Learning Estimator (ABTLE) which is an extension of the traditional ridge regression estimator proposed by Gu et al. (2022). The method is similar to superpopulation modelling and the Bayesian approach of Wiśniowski et al. (2020) in that it is also model-based. However, rather than directly correcting the selectivity, or using the NPS to construct a prior, the ABTLE uses the estimates of the NPS as part of a penalty, seeking to constrain the PS estimates by rewarding them for aligning angle-wise to the estimates of the NPS. This allows for a correction of the PS estimates should the two differ, borrowing the stability of the NPS estimates whilst still through the PS estimates ensuring some protection against the potential bias of the NPS estimates.

The ABTLE assumes a regression model for the vector of target variables y

:

$$y = X\beta + \varepsilon$$

where X is matrix with auxiliary data, β is a vector of regression coefficients, and ε is a vector with normally distributed residuals with mean zero. The vector β can be estimated based on data from PS and on data from NPS. ABTLE and the later mentioned RRE and DBTLE basically start with estimating β using data from PS and adjust this estimate using data from NPS.

The ABTLE has three main advantages over the aforementioned: 1. It is data-cheap, meaning that it only requires estimates directly relevant to the target estimates from the NPS. No actual microdata is required nor are any additional covariates to explain the selectivity of the NPS needed. 2. Given sufficient angle similarity, the degree of bias in the NPS is less impactful on the estimator's quality, making it robust against a very biased NPS, even with a small PS. 3. There is no theoretical risk of negative transfer, meaning the estimate should never be worse than just using the PS estimates in terms of MSE, as long as sufficiently correct penalty parameters are applied.

This thesis applies and evaluates the ABTLE through a simulation study on a wide range of practical scenarios with differing sample sizes, degrees of bias, multicollinearity, sample correlation, and residual variance. This allows for the evaluation of the estimation method in general but also highlights in what context the method is most or least useful. Although the primary metric of interest is the average root mean squared error (ARMSE), the mean of average bias of the estimator in various scenarios is also of interest given the importance of unbiased estimates in official statistics. The second supplementary goal is to compare this estimation method to other benchmarking estimators such as the Bayesian approach proposed by Wiśniowski et al. (2020) and the ridge regression estimator, to highlight if and when the proposed ABTLE is the appropriate choice.

The ABTLE is compared to standard ridge regression (RRE) and to Distance-Based Transfer Learning Estimator (DBTLE; see Tian and Feng, 2023). Both RRE and DBTLE are special cases of ABTLE, so in principle the results of ABTLE should be at least as good as those of RRE and DBTLE. In practice, this is not always the case due to computational problems when estimating model parameters.

4. RESULTS

The results suggest that the ABTLE can aid in improving the accuracy of ARMSE relative to the PS estimates. The actual improvement however depends a lot on population and sample parameters. Generally, the ABTLE tends to perform better (relatively) in scenarios where the baseline PS estimates were less accurate (higher multicollinearity and larger residual variance).

Regarding the bias, the ABTLE never really outperformed the other estimators, consistently showcasing a higher or merely similar bias as the benchmarking methods. Nevertheless, this is not necessarily problematic since this was most prevalent at smaller levels of bias. As the bias in the NPS estimates increases, the relative bias of all other estimators tends to converge at a really low relative bias to the NPS estimates, showcasing that the ABTLE is also robust when incorporating very biased NPS. The ABTLE only really outperformed the other estimators in scenarios of high correlation but even then the bias was comparable with other methods.

As anticipated, and consistent with the findings of Gu et al. (2022), the results indicate that significant performance improvements in relation to the RRE can be achieved by expanding the penalization scheme to not only penalize magnitudes but also reward aligning it anglewise with auxiliary estimates. Unless the accuracy of the Maximum Likelihood Estimate of the regression coefficient based on the PS only (PS MLE) was already accurate, then even with a flawed penalty parameter estimation this was shown to be true. The results also reveal that a distance-based penalization scheme lacks flexibility as is evident from the fact that the DBTLE tended to underperform compared to the ABTLE. It should be noted that, unlike the results of Gu et al. (2022), it remains unclear from this research whether this issue stems from over or under-penalization of the DBTLE. Finally, relating the results to the research of Wiśniowski et al. (2020), they are in line with their findings. We can see from the relative performance that although the Bayesian estimator works relatively well (sometimes even better than the ABTLE), it is vulnerable to bias in the NPS estimates. As the bias increases to the maximum, the Bayesian approach is never better than the ABTLE, offering further evidence of the limitations of their approach in specific scenarios.

Contrary to the results of Gu et al. (2022), this research also found that the ABTLE can result in negative transfer. Specifically, incorporating the NPS sometimes led to a deterioration in ARMSE relative to the RRE and PS MLE when the accuracy of the PS MLE was already high. This is caused by over-penalization which was not reported as a potential problem in the research of Gu et al. (2022). The likely cause behind both these differences lies in the different simulation designs employed here and by Gu et al. (2022). By chance, it is possible (and was observed) that a sample (or a fold during cross-validation) exhibited a very high correlation or residual variance, inflating the model's perception of the optimal value of the penalty parameters. This in turn can lead to over-penalization and the occurrences of negative transfer, which does not occur if the theoretically optimal values are used.

Tying the discussion back to the research question of How can penalized regression be used to incorporate non-probability samples into official statistics? the results do highlight a potential type of scenario where the ABTLE does fit. Given its relative success, the ABTLE seems to be a good alternative to the correction methods outlined in Section 2 in scenarios where there are few to no additional covariates to explain the selectivity in the NPS and where the bias of the NPS estimates are large. There is no such thing as the best estimator for every case (Hastie et al., 2009). However, given this type of scenario, correction methods cannot be used, the Bayesian approach and the composite estimator have been shown to be vulnerable to bias. Limited data, high bias scenarios can therefore be seen as a type of scenario where the ABLTE is most likely to be the best choice.

Another more practical scenario where the ABTLE could be a good alternative is in situations where the auxiliary data is sensitive, and sharing micro-level data is not feasible. In such cases, sharing estimates might be more viable which benefits the ABTLE that unlike all other comparable methods only requires auxiliary estimates from the NPS. However, in its current form, there are still clearly some scenarios where ABTLE is to be avoided. With small samples, the estimation of penalty parameters becomes increasingly uncertain, and it is not guaranteed to result in a performance increase. In such scenarios, it might then be preferable to simply use ridge regression since it is more stable, and unlike all other estimators was the only method that never led to deteriorating estimates.

5. CONTRIBUTION

This type of estimation method tends to stem from biostatistics and has to the author's knowledge never been applied in a setting such as the one described above. It is true that one type of estimator proposed by Wiśniowski et al. (2020) is the Bayesian interpretation of the ridge regression estimator (a type of penalized regression), however, it was not a major part of their discussion or research.

6. REFERENCES

- [1] Bakker, B. F., Van Rooijen, J. & Van Toor, L. (2014). The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-Based Social Statistics. *Statistical Journal of the IAOS*, 30 (4), pp. 411–424. <https://doi.org/10.3233/SJI-140803>.
- [2] Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. John Wiley & Sons.
- [3] Chen, Y., Li, P. & Wu, C. (2020). Doubly Robust Inference with Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115 (532), pp. 2011–2021. <https://doi.org/10.1080/01621459.2019.1677241>.
- [4] De Leeuw, E. & De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, J. Eltinge & R. Little (Eds.), *Survey Nonresponse*, pp. 41–54. Wiley.
- [5] Disogra, C., Cobb, C., Chan, E. & Dennis, J. M. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. *Section on Survey Research Methods – JSM Proceedings*, pp. 4501–4515.
- [6] Elliott, M. & Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology*, 33 (2), pp. 211–215.
- [7] Elliott, M. & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, 32 (2), p. 249–264. <https://doi.org/10.1214/16-STS598>.
- [8] Gu, T., Han, Y. & Duan, R. (2022). Robust Angle-Based Transfer Learning in High Dimensions. <https://arxiv.org/abs/2210.12759>.
- [9] Li, C., Yang, C., Gelernter, J. & Zhao, H. (2014). Improving Genetic Risk Prediction by Leveraging Pleiotropy. *Human Genetics*, 133 (5), pp. 639–650. <https://doi.org/10.1007/s00439-013-1401-5>.
- [10] Liang, M., Park, J., Lu, Q. & Zhong, X. (2020). Robust and Flexible Learning of a High-Dimensional Classification Rule Using Auxiliary Outcomes. <http://arxiv.org/abs/2011.05493>.
- [11] Tian, Y. & Feng, Y. (2023). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*, 118 (544), pp. 2684–2697. <https://doi.org/10.1080/01621459.2022.2071278>.

- [12] Valliant, R. (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, 8 (2), pp. 231–263. <https://doi.org/10.1093/jssam/smz003>.
- [13] Valliant, R., Dever, J. A. & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples* (2nd ed.). Springer International Publishing. <https://doi.org/10.1007/978-3-319-93632-1>.
- [14] Van den Brakel, J. (2019). *New Data Sources and Inference Methods for Statistics*, Statistics Netherlands. <https://www.cbs.nl/en-gb/background/2019/27/new-data-sourcesand-inference-methods-for-statistics>.
- [15] Villalobos-Alíste, S. (2022). *Combining Probability and Nonprobability Samples on an Aggregated Level*, Master thesis, Utrecht University.
- [16] Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A. & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology*, 8 (1), pp. 120–147. <https://doi.org/10.1093/jssam/smz051>.

Research Network: Anonymization for Secure Data Use

Jannek Mühlhan (Destatis) & Markus Zwick (Statistisches Bundesamt)

1. INTRODUCTION

Digital systems are ubiquitous today: smartphones, smart homes, medical technology, and increasingly digitally connected vehicles are just a few examples. Digital systems of all kinds have become an integral part of modern life. The operation of these systems generates extensive data that can be used in a variety of ways. By automatically analyzing customer, business, and process data, supermarkets can replenish their inventories, trucking companies can optimize their routes, energy grids can be used more efficiently, and industrial plants can be maintained more effectively. Data analytics improves traffic planning and control, enables autonomous driving, and supports personalized healthcare. In short, data is the foundation of the digital society; it is generated everywhere and enables more and more digital services and data-driven business models. However, digital data cannot be used at will. Data protection ensures that everyone's right to privacy is protected when data is used. This tension between data use and data protection means that the potential of digitization and the applications, services and business models based on it have not yet been fully exploited in Germany.

The "Research Network Anonymization for Secure Data Use" of the German Federal Ministry of Education and Research (BMBF) within the framework of the Federal Government's IT security research program "Digital. Secure. Sovereign" is financed by the European Union – NextGenerationEU and aims to secure data protection while exploiting new informational potential through digitization. The core task of the researchers in the is therefore to develop technologies, processes and methods to anonymize personal data and thus make it usable for data-based services and innovations in a legally compliant manner. The anonymized data will also be used to develop and test services and business models that can be used in areas such as mobility, medicine, or administration and that will drive digitalization forward.

The research network consists of 5 different competence clusters and 17 associated research projects that develop innovative and efficient new processes, methods and technologies to promote the anonymization and use of data in compliance with data protection regulations. The Federal Statistical Office of Germany (Destatis) leads and coordinates the competence cluster "Anonymization of integrated and georeferenced data" (AnigeD).

The talk will introduce the challenges in the anonymization process and its possible impact on the quality of statistical products and publications. It will provide an overview about the different research strands within in the network and summarize the research results of the AnigeD cluster with respect to the main research strands within the AnigeD project: 1) evaluation of anonymized data according to formal criteria, 2) anonymization through synthetic data, 3) anonymization of georeferenced data, and 4) open software tools for anonymization. A detailed presentation of the progress made in the four areas will be given in separate presentations by cluster participants.

The AnigeD competence cluster is part of the "Research Network Anonymization for Secure Data Use" of the German Federal Ministry of Education and Research (BMBF) within the framework of the Federal Government's IT security research program "Digital. Secure. Sovereign".

2. RESEARCH NETWORK ANONYMIZATION FOR SECURE DATA USE

The research network Anonymization for Secure Data Use consists of 5 different competence clusters and 17 associated research projects, each dealing with different focal points regarding anonymization of data in order to provide them for scientific or development of products and services.

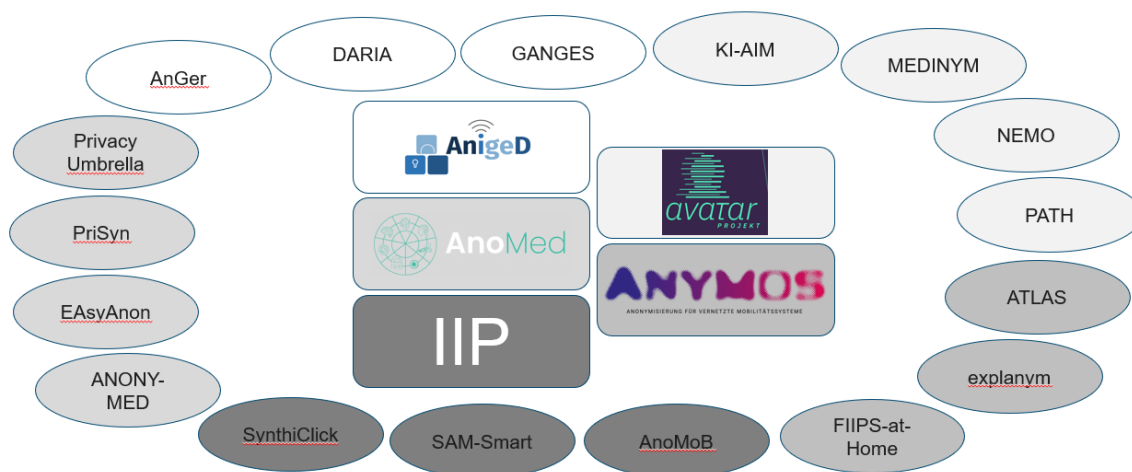


Figure 1. Research Network Anonymization for Secure Data Use

2.1. Anonymization of Integrated and Georeferenced Data (AnigeD)

Data-driven information plays a central role in politics, business, science and public life. With digitisation and the exponential growth of stored data, as well as new analytical methods such as machine learning, the possibilities for evidence-based decision making have expanded and evolved significantly. A key challenge in integrating disparate data sets from different data custodians is the protection of personal privacy and trade secrets within organisations. This currently hinders both the wider use of data as a product and the use of integrated data in policy advice and scientific research. Methods of anonymisation and statistical confidentiality face the challenge of finding a compromise. On the one hand, they need to protect the information of the data subjects, while on the other hand, the chosen methods should still offer sufficient analysis and information potential for the anonymised data. In the past, it has been shown that common anonymisation strategies for individual data in economic statistics led to de facto or absolute anonymised data sets, which were severely limited for scientific analysis due to the reduced or even distorted information potential.

The cluster aims to develop strategies for the protection of personal and company-related data when using complex integrated data sets. Not only the integration of different data via direct identifiers or probabilities is relevant, but also the integration and linking of data via regional information in the form of geo-referencing. AnigeD is conceived as an open competence cluster for research on anonymisation techniques for geo-referenced and linked data. The aim is to network with interested institutions. AnigeD with the research focus on anonymisation of integrated and geo-referenced data is divided into the following research areas 1) evaluation of anonymised data according to formal criteria, 2) anonymisation through synthetic data, 3) anonymisation of geo-referenced data, and 4) open software tools for anonymisation.

2.2. Anonymisation of personal health data through the generation of digital avatars (AVATAR)

A large amount of personal health data is currently available, but cannot be utilised due to existing data protection requirements. Our motivation in the project is to resolve this

contradiction through a new anonymisation approach and to be able to use available data for better patient care, innovative product developments and efficient and short development times.

The use of health data promises great potential for medicine and research, however the protection of personal data is extremely important in this area. The aim of the BMBF-funded project AVATAR is to resolve such contradictions by anonymizing patient-related data records and merging them into so-called avatars. The research institutions, clinics and healthcare companies cooperating in the project aim to make an important contribution to improving patient health in the long term. The cluster aims to develop of new procedures for anonymizing patient data for research, development and application, simplify data requests and data donation for legally compliant data use and educate on the benefits of donating health data, creating incentives for data donation

2.3. Anonymization of medically relevant data (AnoMed)

The amount of sensitive, medically-relevant data collected in healthcare settings and by smart devices is growing rapidly and must be handled with care. At the same time, AI and big data methods make it possible to process this data efficiently, enabling new individualised precision medicine and long-term observations of the effects of drugs, therapies and behaviours. But how can the anonymity of the patients behind the data be preserved? As a catalyst for anonymisation research for medical applications, the AnoMed cluster will both present the challenges of medical applications to anonymisation researchers and show medical users the dangers of deanonymisation and the possibilities of the latest anonymisation techniques. The AnoMed cluster will explore new, provably secure and useful anonymisation solutions for medical applications, and will establish a benchmarking platform for the international research community, where researchers can upload their anonymisation solutions and put them through their paces.

The platform will provide the basis for evaluating different anonymisation metrics. It will provide interfaces for medical researchers and companies to enter gold standards, define metrics and store data securely. A set of medical use cases will be built where stakeholders can interact with medical data and the latest anonymisation research results, both protection mechanisms and attacks. Technology transfer will be accelerated through concrete attacks on medical challenges and a comprehensive science communication campaign. This will make the dangers tangible for companies, medical professionals and citizens, and demonstrate that modern anonymisation techniques provide effective protection against these dangers.

2.4. Anonymization for connected mobility systems (ANYMOS)

Future mobility solutions will be data-driven and highly user-centric. Personal data is often indispensable, but subject to data protection regulations. The ANYMOS competence cluster uses specific use cases to investigate the use of anonymisation techniques while maintaining the utility value of the data. The advantage of anonymisation in reducing uncertainties about the data protection compliant handling of personal data must not be at the expense of the usability of the data for specific applications.

The ANYMOS competence cluster pursues the long-term goal of establishing anonymisation as an enabling technology in order to reduce uncertainties about the need to apply data protection regulations when exchanging and using data and to maintain Germany's strong position in the automotive and public transport sectors, including in data-driven innovations. As an instrument to reduce uncertainties, ANYMOS will develop and establish a process model that helps companies to identify anonymisation needs and possibilities for use cases in the field of mobility, to select suitable anonymisation methods according to the state of the art, to apply them correctly and to systematically identify and evaluate re-identification risks.

2.5. Intelligent intermodal commuter traffic (IIP)

Private transport has increased significantly in recent years, especially in urban areas. The use of alternative means of transport is still not attractive enough for many people. Intelligent, intermodal mobility networks can counteract this problem by optimising the combined use of different modes of transport.

The Intelligent Intermodal Commuter Traffic (IIP) competence cluster aims to research technologies, processes and methods for anonymising or depersonalising personal mobility data. To this end, data will be collected in the sovereign context of a municipality and dynamically anonymised, i.e. for specific applications and external solution providers. Existing official and publicly available data will be brought together on a platform and supplemented by new data sources from crowdsourcing and citizen participation. The researchers will develop and test a toolkit for different levels of anonymisation that can be used in practice for different applications. This will enable the mapping of objects such as cars, buses, people, cyclists, e-scooters, etc. across data sources without violating the privacy of road users. The developed solutions will be made available to the public as open source.

3. RESULTS AND PRACTICAL APPLICATION

The talk will provide an overview over the network and its various research strands regarding the anonymization for secure data use. The talk aims to inform about the most recent methods for anonymization of data and its advantages, possible drawbacks and important pitfalls. One goal is to make the confidentiality experts within the European Statistical System aware of the German Research Network Anonymization for Secure Data Use and connect the respective stakeholders.

RUMS – how to compare structures of enterprise groups?

Adrian Urban & Simon Rommelspacher (Federal Statistical Office Germany)

1. INTRODUCTION

The statistical business register is the backbone of official business statistics and covers all the different units used in business statistics [1]. One of these unit types is the enterprise group, which is becoming increasingly important in the economy. Enterprise groups are often multinational, with structures spread over many countries.

An enterprise group is defined as a set of more than one legal unit (LEU) linked by control relationships. A LEU is controlled by another LEU if more than 50 % of the voting rights are held by that other LEU. Within an enterprise group, every LEU is controlled by another LEU (parent). There is one LEU at the top of the control chain. This is called the global group head (GGH). The GGH and the controlled LEUs together form the enterprise group [2].

An enterprise group in a statistical business register usually has specific properties, e.g. turnover, employees, main economic activity (e.g. NACE-Code), identification number and so on. Within an enterprise group some LEUs act in specific roles like GGH or global or regional decision centre (head office).

In both data production and assessment processes, e.g. for statistical publications, it is essential to compare enterprise group data either between different registers in order to evaluate the registers' different data sources, or at different points in time within the same register in order to evaluate how enterprise group structures have changed over time.

2. METHODOLOGY

We developed RUMS, a similarity metric for comparing enterprise groups, that is designed to fulfil the following expectations.

- a. If two enterprise groups are completely different, i.e. have no overlap in the characteristics considered, then RUMS should be 0.
- b. If two enterprise groups partially overlap in the properties considered, i.e. are similar to each other, RUMS should be between 0 and 1.
- c. If two enterprise groups are completely identical in terms of the properties considered, then RUMS should be 1.
- d. The higher the similarity between enterprise group A and enterprise group B, the higher RUMS should be. This means that if enterprise groups A and B are more similar than A and C, $RUMS(A,B) > RUMS(A,C)$ must always be true.
- e. RUMS should be flexible enough to expand and reduce the properties considered.
- f. It should be possible to flexibly weight the considered properties in RUMS, depending on the expected quality and importance of the properties in the populations being compared.

Requirements e and f are intended to allow the widest possible implementation for a variety of use cases in official statistics.

RUMS uses quantitative characteristics available in the German NSBR. To calculate RUMS between enterprise groups A and B, the number of overlapping LEUs $L(A \cap B)$, the number of

overlapping employees $E(A \cap B)$ and the turnover of the overlapping LEUs $T(A \cap B)$ are related to the total number of LEUs, employees and turnover in enterprise groups A and B respectively. The general formula is:

$$RUMS(A, B) = \begin{cases} a * \frac{L(A \cap B)}{L(A)} + b * \frac{L(A \cap B)}{L(B)} + c * \frac{E(A \cap B)}{E(A)} + d * \frac{E(A \cap B)}{E(B)} + e * \frac{T(A \cap B)}{T(A)} + f * \frac{T(A \cap B)}{T(B)}, & L(A \cap B) \neq 0 \\ 0, & L(A \cap B) = 0 \end{cases}$$

Under the constraint: $a + b + c + d + e + f = 1$ and if one denominator equals 0 then the whole fraction equals 1. E.g. if $E(A)=0$ then $\frac{E(A \cap B)}{E(A)}=1$.

If one denominator equals 0 then there is one of the enterprise groups with no employees or turnover. In this case we let the sub-indicator take the value 1 so that such constellations remain calculable and do not result in this property not being considered or other properties having to be weighted higher as a result. This would distort the comparability of RUMS values.

The three properties considered (LEUs, employees and turnover) and their intersections in RUMS formula are each set in relation to the total of both enterprise groups and are thus considered in two sub-indicators, so that additional LEUs on either side reduce the similarity. This allows one of the two populations to be weighted more strongly if it is considered to be of higher quality.

The properties that are considered in RUMS can be reduced or extended as required. For example, it would be possible to compare the roles of the LEUs and add a partial indicator that indicates whether the two enterprise groups have the same GGH. However, after some empirical testing with the data from the use cases in chapter three, we found that it is more useful to analyse such a same GGH/different GGH similarity after calculating RUMS. It would also be possible to look more closely at the control relationships in the groups and see how many direct relationships are identical. This procedure would be much more complicated and in our point of view the information on the group membership is more important than the individual relationship. The properties of employees and turnover are generally precise and complete in the registers considered, so that the intersections of the three properties LEUs, employees and turnover were considered in RUMS for the current use cases. This may be different in other registers so that these sub-indicators can either be deleted or replaced or the weighting can be changed by adjusting the parameters a-f.

Determining the weights is a very difficult question. We carried out simulation studies to analyse the effects of different weights. The populations were analysed to see how often several enterprise groups are similar and in how many of these cases there is no dominance of a comparison, so that the weighting is actually decisive for deciding which enterprise groups are more similar to each other. For these constellations, a grid was created for each parameter so that each parameter could assume 100 different values [3]. Each possible constellation was then simulated and a calculation was made to determine which comparison wins more often. In addition, a geometric procedure was used to determine the value of the parameters at which the decision changes. With both methods it could be shown that the equal weighting of the sub-indicators produces a result that is always a very probable result even with non-equal weightings. However, this weighting could be investigated further applying these or other methods and also with more different use cases and data. We currently recommend an equal weighting of the parameters as long as there are no indications that one of the considered populations or properties is of poorer quality.

3. RESULTS AND PRACTICAL APPLICATION

The use cases of RUMS can be separated in the necessity to compare groups from different points in time or to compare groups from different data sources.

Use Case 1: Comparison of groups at different points in time

At the beginning of every new reference year, the German NSBR receives data on enterprise groups. To process this data, a decision must be made for each existing enterprise group in the business register from the old reference year as to whether this enterprise group is to be continued and with which new group information of the new reference year. For this purpose, the enterprise group data of the reference years 2021 and 2022 were prepared and the RUMS-formula with equal weighting for the parameters a-f was applied. The enterprise groups at the two different points in time can now be divided into three parts.

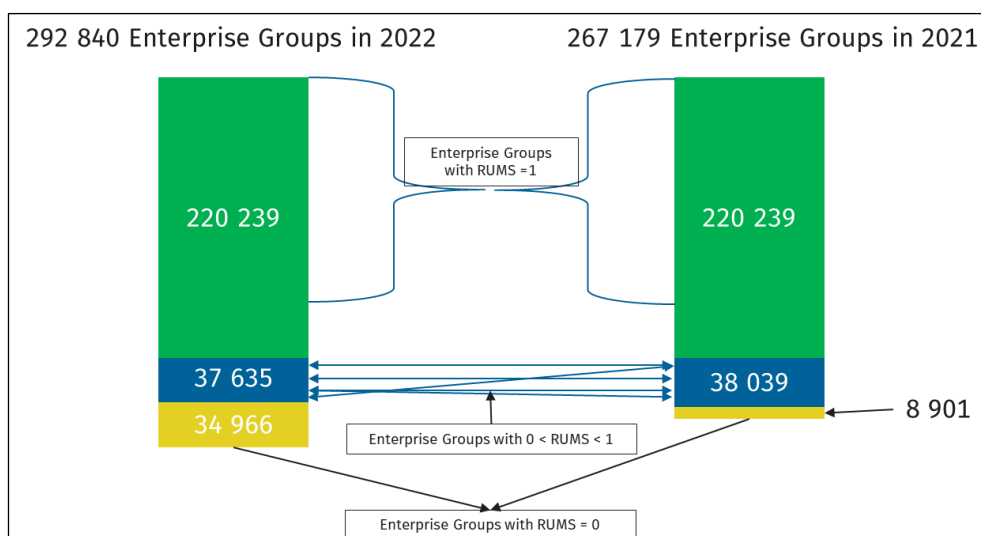


Figure 1. Comparison of enterprise groups between reference years 2022 and 2021

Firstly, there are the identical enterprise groups in figure 1 (the upper bars) with a RUMS=1. For these groups, continuity is out of question since no changes concerning their group structure occurred in the two reference years.

Secondly, the lower bars of figure 1 shows enterprise groups with no similarity to any of the enterprise groups in the other reference year, i.e. RUMS is 0. These are groups that are either completely new or no longer exist in 2022.

Lastly, the most interesting enterprise groups are in the middle bars in figure 1. They show a varying degree of similarity with one or more enterprise groups from the other reference year. Their RUMS-values vary between 0 and 1. For these groups, RUMS can help to decide which new group information of reference year 2022 belongs to which group from the old reference year 2021.

With the help of RUMS, the annual processing of new information could be automated to a large extent and many enterprise groups could be continued in a reasonable way, meaning that they keep the same ID.

The second use case will be shown during the presentation at the NTTS conference.

4. MAIN FINDINGS

As enterprise groups are often multinational and may have very large structures, it is increasingly important that their data are of high quality in the statistical business registers in order to be a good backbone for all business statistics. With the help of RUMS, enterprise groups as complex data objects are to be made comparable within statistical business registers, as long as they meet the described requirements. This allows changes over time or deviations in different registers to be quantified and thus better analysed in order to increase quality.

To measure and compare the similarities of entire populations or registers, we developed an extension of RUMS called RUMSA. We will show this new measure during the presentation at the NTTS conference.

ACKNOWLEDGEMENTS

The authors would like to thank Roland Sturm and Martin Beck for their valuable reviews and continuous motivation.

5. REFERENCES

- [1] Snijkers G, Bavdaž M, Bender S, Jones J, MacFeely S, Sakshaug JW, et al. Advances in Business Statistics, Methods and Data Collection. 2023. p. 1-22.
- [2] Nielsen PB, Plovsing J. Concepts Used in Statistical Business Registers in View of Globalisation and the Information Society. *International Statistical Review* 1997;65(3):351-63.
- [3] Permanyer I. Assessing the Robustness of Composite Indices Ranking. *Review of Income Wealth*. 2011;57(2):306-26.

SAS to R Training: A Collaborative Approach to Promote a Learning Culture in Statistical Programming

Kathryn Foskin (Central Statistics Office, Ireland)

1. INTRODUCTION

The pace of technological advancement has had a significant impact on NSI operations in recent years. Up-to-date programming skills across various packages are now critical to the traditional statistical role. The decision by the Central Statistics Office (CSO) Ireland to move from SAS to R meant that R programming became a core competency for statisticians and upskilling in R a significant focus for statistical staff. This transition is more than a language change; it represents a significant change in the CSO's tools, processes, and workflows, and it requires a shift in mindset and culture for those involved. R is an open-source product with a vast array of packages available, and its customisability can be daunting for those who are used to the uniformity and consistency of SAS.

Beyond the specific demands of R, the SAS to R project team recognised the need to take a broader view, to build and support a workforce that could adjust to an evolving landscape and had the openness and capacity to learn and adapt to technologies and techniques as they emerge in the future. This approach included building awareness of open-source solutions, making them part of the culture and fabric of the organisation. For the SAS to R project, this meant considering ways to develop as a learning organisation and support the learning culture within the office. Training in this context required a holistic approach to consider both the technical demands of the transition and the psychological and pedagogical needs to build a robust learning culture for statistical programming.

This paper will describe the CSO approach to training in R to enable the SAS to R transition, with broader consideration of future adaptability. This transition is about developing adaptable programming and problem-solving skills and an overall learning culture that will allow us to adjust to future technological evolution.

2. METHODOLOGY

2.1. Learning Culture

Learning organisations are those where "people continually expand their capacity to create the results they truly desire, where new and expansive patterns of thinking are nurtured, where collective aspiration is set free and where people are continually learning how to learn together" (Senge, 1990, p. 3). Organisational learning is about more than the sum of the parts of individual learning and the organisational environment or "culture" needs to be one of continuous learning and growth. A learning culture can be defined in many ways, but most tend to refer to an environment "in which there are as set of shared beliefs, values and attitudes favourable to learning" (OECD, 2010, p. 25). But what is an environment "favourable to learning"? Gephart et al. (1996) define a learning culture in terms of values and beliefs that "promote enquiry, risk-taking and experimentation" as well as allowing "mistakes to be shared and viewed as opportunities for learning", while Dimovski et al. (2007) refer to a "a set of norms, values and underlying assumptions about the functioning of an organisation that

support more systematic, in-depth approaches aimed to achieve deuterio, strategic or generative learning..." (cited in OECD, 2010, p. 25). When developing an R training framework, we considered a learning culture to be one that encourages and empowers people to be curious and have space to experiment and explore, make mistakes, and learn from them, from both a team and an individual perspective. The training was designed to support continuous learning, promote psychological safety, and support the learning journey.

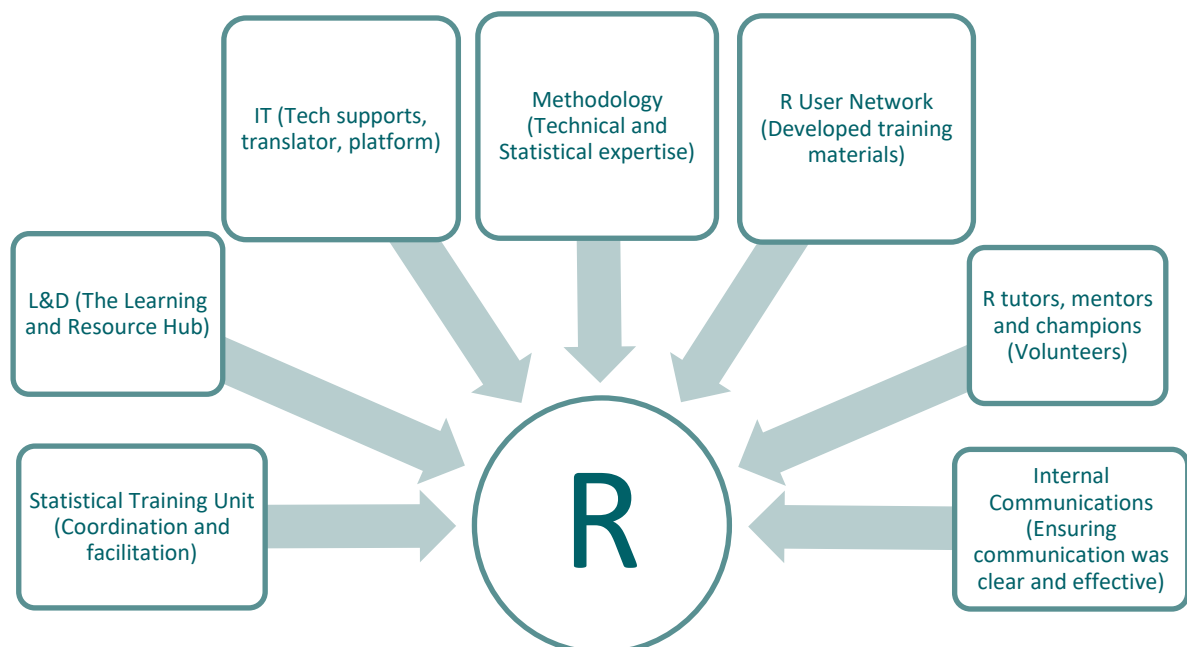
2.2. SAS to R Training

The SAS to R training in CSO is underpinned by the core principles of the 70:20:10 model, which recognises the importance of knowledge acquired through job-related experiences and social interaction as well as formal educational events in learning and development (Training Industry, 2015). Using this theoretical model as a general guide, we are flexible in the distribution of learning depending on the learners in question. For example, formal training has a higher weight for beginners, while eLearning and self-learning materials become more suitable as learners move into intermediate and advanced training. In the provision of training and knowledge-sharing in the SAS to R context, the following were considered:

- The skills of the learners and processes they needed to migrate;
- The need for just-in-time training where possible;
- Information needs to be available and accessible when needed;
- The support that could be made available to help with the transition and
- Consideration of the different types of learners and how attitudes and behaviours will affect the success of the project.

2.3. Collaboration

The SAS to R project team recognised the extensive knowledge of R that was already building up within CSO and the enthusiasm from those experts and early adopters to share knowledge. The SAS to R team leveraged this capacity to develop learning opportunities through collaboration with:



The team collaborated to make available various learning and migration supports, including self-paced learning material, in-house Beginner R training with mentors and social support, SAS to R translator tool (Trans-compiler), drop-in clinics, and divisional and one-to-one support and guidance for on-the-job transition.

3. RESULTS AND PRACTICAL APPLICATION

The suite of learning materials and support available to staff has resulted in R skills development in CSO, and responses towards the upskilling approach have been largely positive to date. The benefits of this approach are:

- Volunteer R champions, tutors and mentors have the opportunity to develop and demonstrate communication, presentation, leadership, and collaboration skills. At the same time, the learners can engage with them in informal and on-the-job learning contexts and staff have follow-up support after the training course if needed.
- Technical supports (such as the Trans-compiler), learning and development supports (e.g. testing guidance, code reviews, methodological support) and social learning supports (champions and mentors) give people a psychological safety net to experiment and make the changes needed.
- Problem-solving challenges, group work and the use of relevant, work-related data in R beginner training enable learners to adapt the theory into practice.
- Just-in-time training is preferable. Some people completed training but did not use it immediately and needed to return for further training some months later. Making eLearning and self-learning options available can be very useful here as staff can manage their learning at a time and pace that works for them.
- Learners can become mentors, further embedding continuous learning within the organisation.
- Knowledge-sharing and collaboration opportunities are essential to reinforce the sense of continuous learning and development.

4. MAIN FINDINGS

Some key findings and recommendations from the process thus far are:

- **The importance of finding your champions.** Leverage internal skills, harness enthusiasm, and give people the opportunity to contribute. People value the opportunity to demonstrate and share their technical knowledge and develop skills such as leadership, collaboration and communication. Also, when others see people participating in the project they are encouraged to join in, building confidence towards becoming mentors and trainers. Platform successes and skills through existing networks and knowledge-sharing arenas, such as Communities of Practice¹ and CSO's "GMF Bitesize"².
- **Remove barriers to learning, large or small.** This enables people to focus on learning and knowledge-sharing. Barriers can be technological, financial, administrative, psychological or structural. Some are easier to solve than others, but all can impact the ultimate goal. For example, having an administrative team responsible for the smooth running of training courses and sharing of learning materials allows the learners to focus on learning. In addition, ensuring eLearning options are available and accessible means that staff can learn at a time that works for them.
- **The process is iterative and evolves.** The training needs of staff will change, as will technology. It is important to review the content and the process regularly and be informed by feedback from stakeholders. In CSO, feedback is sought after each training session, and adjustments are made accordingly. For example, we received feedback that our intensive

¹ A Community of Practice (CoP) is a network of enthusiasts who share and create knowledge around a common interest area. It is a group of practitioners with a common sense of purpose who agree to work together to solve problems, share knowledge, cultivate best practice and foster innovation.

² CSO GMF Bitesize is an online knowledge-sharing session for CSO staff run on a regular basis. Any staff member can submit topics to the bitesize and attend.

beginner R training programme, while ideal for some, is too time-intensive for others. This has led to the consideration of a more modular approach in 2025. CSO also conducted annual R surveys in 2023 and 2024 to identify training needs.

- **The importance of psychological safety.** Programming competency can vary widely amongst statistically competent staff members, which can be exposed during the R transition process. Psychological safety is vital to ensuring a high-quality transition and a sustained learning culture in general. In creating learning supports, the organisation must empathise with the individual challenges of learning R, particularly for someone unfamiliar with multiple programming languages who may have only ever used SAS. For trainers and mentors, this may require some repetition, basic explanations, and additional support for some staff members. At an organisational level, it may be in allowing staff the time needed to upskill, having realistic expectations of what can be achieved and ensuring appropriate supports are available.
- **Leverage eLearning options if possible.** Making eLearning and self-learning materials available and accessible is extremely useful in ensuring staff have access to training when needed, thus promoting continuous learning. It is essential that people know what is there and how to access it, and it should be reviewed regularly. Sometimes there is a danger of information overload, so ensure that the mode of provision is clear and offers guidance on how staff find what they need. In CSO, the Learning and Resource Hub developed by the Learning and Development team was used to enable access to resources.
- **People need time to learn.** Ensuring managers and leaders recognise this and allow this time is critical. Learners need time to immerse themselves in R, learn on the job, make mistakes, and learn from them. Ensure that when they hit a blockage or a challenge, they know where to go and trust the support they get so learning doesn't stall. To allow this time, the organisation must recognise the long-term added value of investing in training.
- **Make training relevant to the role by including project work based on data familiar data.** CSO beginner R training included project work on students' own data, which they presented back to the group. This was a challenging but beneficial element of the training.
- **Consider resource allocation.** New recruits often approach programming for statistics in innovative ways. There can be benefits in pairing up with a more experienced statistician who needs to develop programming skills but has the subject matter knowledge. In CSO's SAS to R project, the availability of R champions, trainers and mentors has been crucial in embedding R skills. Ideally, that mentor is a colleague working in their area.
- **Balance in-person and online training.** In the CSO experience, in-person attendance in the early days of the intensive beginner R course allowed participants to immerse themselves, while online allowed some flexibility and sharing screens for project work. Feedback from some participants in the beginner R training in CSO was that in-person training was beneficial.
- **Not everything works, and allowing for trial and error is essential.** For example, drop-in clinics were poorly attended in CSO, so it was decided to cease them after a trial period. Also, different types of training work better for different people, and people may need to try various formats to see what works for them. Having alternative modes and formats allows learners to find what works for them.
- **Cater for the "Just get it done" people as well as the explorers.** Some learners want to be guided to the one tool that suits their needs rather than be given a suite to choose from or a single package for a specific activity. Work done by the SAS to R project on packaging up the packages and aligning them with the GSBPM was helpful for them.

5. REFERENCES

- [1] Senge, P.M. (1990), *The Fifth Discipline: The Art and Practice of the Learning Organization*, London, Random House.
- [2] OECD (2010), *Innovative Workplaces: Making Better Use of Skills within Organisations*, OECD Publishing. https://www.oecd.org/en/publications/2010/11/innovative-workplaces_g1g11bef.html
- [3] Gephart, M., V. Marsick, M. Van Buren and M. Spiro (1996), "Learning Organizations Come Alive", *Training & Development*, December, pp. 35-45.
- [4] Dimovski, V. M. Skerlavaj, R. Skrinjar, J. Jaklie and M. Stemberger (2007), "Organizational Learning Culture as the Link between Business Process Organization and Organizational Performance", working paper, Faculty of Economics, University of Ljubljana.
- [5] Training Industry. 2015. The 70:20:10 Model for Learning and Development. Retrieved from: <https://trainingindustry.com/wiki/content-development/the-702010-model-for-learning-and-development/>

Small area estimation using incomplete auxiliary information

Andrius Čiginas, Donatas Šlevinskas & Ieva Burakauskaitė (Vilnius University; State Data Agency, Statistics Lithuania)

1. INTRODUCTION

The additional information available in probability sample surveys is essential in obtaining as accurate parameter estimates in the finite population and its domains as possible. Having auxiliary data related to the study variables at the unit or domain (area) level provides a range of models to choose from that can improve the direct design-based estimates. Estimation approaches supported by models in different ways are well developed for estimating parameters in the population or its larger domains [1] and in small population areas [2].

The classical literature on survey statistics usually deals with idealized additional information: the values of auxiliary variables are assumed to cover the survey population at a detailed or at least some aggregated level. However, even the administrative data sources commonly used in official statistics often cover only a part of the population. Therefore, standard methods such as calibration estimators [3] cannot be applied immediately. In addition to administrative data, many other potentially valuable sources do not meet the ideal (completeness or coverage) conditions, such as sample data from other probability surveys, non-probability samples, and big data samples. Previously, efforts to effectively integrate these additional data types for population parameter estimation were sporadic. However, research in this area has advanced rapidly in recent years, as reviewed by [4].

In the context of estimation in small population domains, when using aggregated auxiliary data in area-level models such as the famous Fay–Herriot (FH) model [5], there are also specific challenges to employing incomplete auxiliary information. The authors of [6] use unbiased estimators of auxiliary domain means obtained from the same or other preferably larger probability samples as covariates in area-level models. They show that a naive application of the FH model can lead to worse results than direct estimation. They propose to use a modified version of the FH model, allowing for measurement error in the auxiliary aggregates. A drawback of this approach is that the estimated covariates may be unavailable for some domains if the auxiliary variables are taken from samples that are not large enough. Although non-probability samples can be much larger than probability samples, it can be challenging to ensure that the estimators of auxiliary domain characteristics based on them would be approximately unbiased or to assess the biases of such estimators. This difficulty is due to the typically unknown selection mechanism of the non-probability samples [7]. These biases cannot be ignored [8] and make the method of [6] not immediately applicable because it requires the estimated mean squared errors for the estimated covariates. Among exceptions is the application of [9], where a big data covariate is based on a non-probability sample treated as a simple random sample. Some other recent applications that use area-level big data as additional predictors in various models are reviewed in [10]. However, it should be noted that some applications ignore population coverage errors in certain big data variables, such as those compiled from Google Trends, mobile operators, or social network data.

The authors of [11] assume the linear regression relationship between the variable from a big non-probability sample and the study variable. This is a unit-level measurement error model. To estimate the population total, their idea is to stratify the population into a big data stratum

and a missing data stratum. Then, they applied the calibration estimation method with different conditions imposed on these artificial strata to exploit the big data sample as complete auxiliary information. Such [3] type calibration could be applied separately to each population domain or small area. However, this calibration approach is unsuitable if the underlying unit-level dependence between the variables is not linear.

2. METHODOLOGY

We use the idea of [11] but through the model-calibration (MC) approach of [12], which allows more general unit-level models. The MC method to improve the direct probability sample-based estimates in small areas is based on the predictions of the study variable in the big data stratum, which are further used in calibration constraints. Under certain conditions, the MC estimators of the population totals are asymptotically design-unbiased [12]. Due to this property, our estimation approach has the second step modeling the MC estimates using exactly known area-level covariates.

3. RESULTS AND PRACTICAL APPLICATION

Our study is motivated first by the problem of combining the probability sample data on job vacancies with online job advertisements (OJA) information and administrative data to improve the estimates of job vacancy totals in small population domains like municipalities. In this application, the OJA data is a non-probability sample covering only a limited part of the survey population, and their relation to the study variable is not linear. We also consider other practical problems of official statistics with auxiliary non-probability samples from administrative sources.

4. MAIN FINDINGS

In sample surveys, an additional variable closely related to the study variable may not be fully known across the population. We present a reasonably general methodology for how a variable observed in a non-probability sample can be used to refine the estimation of totals (or means) in small population domains. This is the approach to solve the problems of data incompleteness and bias.

In our applications, we find that the overall improvement in accuracy over direct design-based estimates depends on how many areas are sufficiently covered by the non-probability sample. The applications also show how vital administrative information commonly used in official statistics can be when utilized in small area estimation models.

5. REFERENCES

- [1] C.-E. Särndal and B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- [2] J. N. K. Rao and I. Molina (2015). *Small Area Estimation* (2 ed.). New Jersey: John Wiley & Sons.
- [3] J.-C. Deville and C.-E. Särndal (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* 87 (418), 376–382.

- [4] S. Yang and J.-K. Kim (2020). Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.* 3, 625–650.
- [5] R. E. Fay and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* 74 (366), 269–277.
- [6] L. M. R. Ybarra and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* 95 (4), 919–931.
- [7] C. Wu (2022). Statistical inference with non-probability survey samples. *Surv. Methodol.* 48 (2), 283–311.
- [8] X.-L. Meng (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* 12 (2), 685–726.
- [9] S. Marchetti and C. Giusti and M. Pratesi and N. Salvati and F. Giannotti and D. Pedreschi and S. Rinzivillo and L. Pappalardo and L. Gabrielli (2015). Small area model-based estimators using big data sources. *J. Off. Stat.* 31 (2), 263–281.
- [10] J. N. K. Rao (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* 83 (1), 242–272.
- [11] J.-K. Kim and S.-M. Tam (2021). Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.* 89 (2), 382–401.
- [12] C. Wu and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.* 96 (453), 185–193.

Smart Features in Surveys: Respondent Navigation and Its Impact on Outcomes - Insights from the French Time Use Survey Web App Test

Fanny Mikol (French National Statistical Institute, Insee)

1. INTRODUCTION

In 2023, France conducted an experiment to test a multimode approach for its future Time Use Survey (TUS). Traditionally, the TUS collected paper diaries from a representative sample, with participants manually recording their daily activities. These records were then scanned and encoded to match an established nomenclature.

For the 2025 TUS, Insee – the French National Statistical Institute (NSI) – decided to prioritize digital diaries for activity recording, offering paper diaries as a second option. This shift aligns with the broader trend of survey digitalization in NSIs, primarily due to reduced collection costs and the introduction of new tools designed to lower respondent burden.

However, such a multimode protocol requires thorough testing before implementation, particularly to ensure that web diaries maintain – if not improve upon – the data quality achieved with paper diaries. In the 2023 TUS pilot data collection, respondents were asked to complete both paper and digital diaries, reporting on the same day for both modes, but with a one-week interval between entries.

This study focuses on data collection from the digital diaries, addressing several key questions: what do the navigation data reveal about respondents' behaviour within the app? How do these navigation patterns relate to main TUS outcomes? And more precisely, are there systematic patterns between app features and outcomes? Based on these assessments, how can the application be improved to better align with respondents' actual activities? These questions are important for improving smart features in surveys, and the insights gained could be beneficial for other innovative survey approaches.

2. METHODOLOGY

In the digital diary, respondents had three main options for recording their chosen activities:

- Button navigation: users could navigate through a series of buttons to reach the final activity. For instance, to record *"sleep"*, they would click on *"personal care time"*, then *"being in bed"*, and finally *"sleep"*.
- Search field ("Suggester"): respondents could enter keywords into a search field, prompting the application to suggest relevant activities. For example, entering *"dodo"* (in French) would generate *"sleep"* as a suggestion.
- Open field entry: users could directly input their activity, allowing them to describe their activity in their own words, similar to a paper diary.

2.1. Navigation Variables

These three options were designed to maximize accurate activity recording. The combination of digital diary data and associated paradata provided valuable insights into respondents' navigation patterns within the app. Key variables analysed included:

- Registration mode: indicating whether the activity was recorded via buttons, suggerter, or open field.
- Navigation history: a string tracking user's path through the app (button clicks, selector use...) before final activity selection.
- Search field entries: all words entered into the search field during navigation, if used.
- Final open field entry: the last entry in the open field, if this was the final registration mode for the chosen activity.

2.2. Encoded variables

To analyse the relationship between navigation paths and main TUS outcomes, we needed to encode the activity variables into an established TUS nomenclature (personal care time, professional time, leisure time, etc.). This encoding process process differed based on the data entry method:

- Buttons and suggerter: the encoding was straightforward, as each button or suggestion were built to match one and only one category in the nomenclature.
- Open field entries: a language learning model was employed to encode the “written” entries into the nomenclature, like with paper diary activities.

2.3. Strategies to identify flaws in the app's tools

First, we conducted an exploratory analysis of the buttons' sequence, search field suggestions and open field entries using navigation variables:

- We examined cases where respondents clicked buttons but ultimately chose another input method. This helped identify potentially misspecified, inaccurate, or incomplete buttons.
- We focused on instances where users attempted to use the search field before resorting to the open field. Analysing the last searched string provided insights to enhance the suggestion database.
- We examined frequent open field terms, which helped us identify missing occurrences in both the button interface and search field suggestions.

Secondly, to complement our exploratory analysis, we developed a systematic inferential approach using logistic regression to analyse how activity selection in the web app was influenced by various factors. The model predicted whether an individual i selected activity k (from the 130 activities in the TUS nomenclature), based on:

- Activity frequency observed in paper diaries (as reference)
- Button accessibility (depth in interface hierarchy and screen position)
- Presence of “other” in button label (testing potential “catch-all” effects)
- Respondent characteristics, particularly button usage intensity

The analysis covered over 10,000 observations (670 individuals who completed a digital diary * 130 activities).

3. RESULTS AND PRACTICAL APPLICATION

3.1. Interaction between recording mode and type of activities selected

Table 1 illustrates the distribution of recording modes by activity type. Overall, buttons were the predominant choice, accounting for 76% of activity selections. However, this aggregate figure masks significant variability across different activity types.

For activities such as personal care, work/study, and caring for others, buttons were used about 80% of the time, whereas for leisure activities, they were used only 62% of the time. Indeed, the placement of the *"leisure"* button at the bottom of the eight categorical buttons likely influenced its lower usage. Users had to scroll to access it, which proved challenging depending on screen type and computer proficiency.

Table 1. Distribution of recording mode by type of activity (in %)

TUS type of activity	Buttons	Suggester	Open field	Total	Numbers
Personal care	80,2	16,5	3,3	100	4 886
Professional / study	81,9	13,3	4,8	100	1 198
Household chores	74,4	16,4	9,2	100	1 986
Care for others	80,4	11,5	8,1	100	469
Sociability	71,6	18,9	9,5	100	465
Leisure	61,6	27,9	10,5	100	1 346
Total	76,5	17,5	6,1	100	
Numbers	7 916	1 807	627		10 350

3.2. Navigation paths: what it reveals on user's app appropriation

Following our first strategy described in 2.3, some main results could be highlighted:

- When activities were not selected using buttons (23,5 % of the time): in 19,9% of these, buttons were clicked during navigation. The *"Personal care"* button was the last clicked in 54% of cases. A closer look at the activity finally chosen in this case revealed that users were often seeking leisure activities like *"TV/videos"*, *"reading"*, or *"resting"* – likely due to these being solitary, home-based activities.
- Open field selections (6,1% of cases): in 21,5% of these, users first tried the search field. Common searches like *"prepare"* or *"meal"* didn't generate suggestions, though users were looking for cooking-related activities.
- In the open field, frequently used terms included *"prepare"* (often associated with cooking-related or children care-related activities) and *"watch"* (almost invariably linked to TV or screen-based activities).

Furthermore, the inferential analysis revealed that:

- Button positioning significantly impacted usage patterns: deeper interface placement and lower screen position correlated with reduced selection rates.
- *"Other..."* buttons didn't attract disproportionate clicks (except *"Other personal care activities"*).
- Heavy button users reported fewer distinct activities, suggesting a "cognitive shortcut" where they stick to familiar activities. This contrasted with search/open field users who reported more diverse activities.

3.3. What insights can be drawn to improve the design of the TUS app?

The analysis provided several insights for improvement, including these key enhancements to the application's input methods:

- Button positioning and labelling:
 - Reposition key buttons (*"sleep"*, *"getting ready"*) earlier in interface hierarchy.
 - Enrich labels with frequent activities (*"TV"*) and specific examples (*"car repair"*).
 - Update terminology (*"looking after children"* vs *"care"*...)

- Handling Specific, hard-to-categorize activities: a careful balance is needed between comprehensive coverage and interface complexity. Several options should be considered, like:
 - Duplicating certain buttons to maximize visibility (e.g., *"restaurant dining"*, *"social media"*)
 - Promoting alternative input methods (especially the open field) for specific activities
- Search enhancement: expand search suggestions with common user terms from free-field entries (*"washing clothes"* alongside *"doing laundry"*).

4. MAIN FINDINGS

The study revealed that 76% of activities in the web diaries were recorded using buttons, aligning with the app's design goal. However, button usage varied by activity type, with leisure activities less frequently selected via buttons. For the remaining 24% of activities not recorded via buttons, 74 % were entered using the search field, while 26% were recorded through the open field.

Button positioning (interface depth and screen location) and clarity significantly influenced selection patterns. Examination of search field and open field entries revealed opportunities to enhance button labels and search suggestions.

The study proposed interface improvements: repositioning key buttons, enriching labels with common terms, duplicating strategic buttons across categories, and expanding search suggestions with user-generated terms. While buttons prove highly effective for common activities, alternative input methods (suggester or open field) should also be promoted for rare and complex activities to ensure comprehensive activity recording. These enhancements aim to better capture users' actual activities, thereby reducing measurement errors and improving comparability with paper diaries.

ACKNOWLEDGEMENTS

I would like to thank Anne Pla and Julie Solard, my collaborators, for their valuable feedback on this work.

5. REFERENCES

- [1] E. D. de Leew, 2018, "Mixed-Mode: Past, Present and Future", Survey Research Methods, Vol. 12, No. 2, pp. 75-89.
- [2] Eurostat, 2022, "Position paper on mixed-mode surveys"
- [3] R. Tourangeau, F. G. Conrad, M. P. Couper, 2013, "The Science of Web Surveys", Oxford University Press.

Spatial Mobility and Housing Demand in the Dynamic Microsimulation Model MikroSim

Sarah Bohnensteffen (Federal Statistical Office Germany, Destatis) &
Julian Ernst (Trier University)

1. INTRODUCTION

In times of increasing computational power and availability of microdata micro- based ex-ante simulation analysis is nowadays used extensively around the world for policy analysis and design in different fields. The development and application of microsimulation models is a statutory task of the Federal Statistical Office (Destatis) in Germany since 2016. Destatis responds to this task as part of the research group “Multi-sectoral Regional Microsimulation Model (MikroSim)”, developing a dynamic microsimulation model based on a synthetic population of Germany [1].

Microsimulation models make it possible to analyse complex issues of economic and socio-political importance, considering the future development of a population. Migration movements are particularly important for the development of the regional population structure. At the same time, they are a crucial factor influencing other relevant indicators: In addition to long-term trends in population development and household composition, regional housing demand is closely linked to residential mobility. As residential mobility patterns are themselves strongly linked to different events in the life course and vary considerably from region to region, modelling them poses a particular challenge. We analyse the suitability of different data sources, such as the German microcensus and migration statistics, for estimating the small-scale mobility probabilities of entire households and individual household members.

The German microcensus contains a wide range of information on individuals and households associated with residential mobility of households and individuals. Compared to the relocation dynamics described in the migration statistics, microcensus data might allow a more differentiated view of the situation of households. In order to investigate the influence of events, such as the birth of a child in the household or the completion of an educational stage, we explore the possibility of linking the overlapping sample components of the 2016 to 2019 microcensus to build a data base that allows for longitudinal analysis.

2. METHODOLOGY

An important data source containing information on migration dynamics in Germany is the official migration statistics, which record annual inflows and outflows of residents at municipal level. The advantages of migration statistics lie in the very good coverage of migration across municipal boundaries, the inclusion of migration to and from abroad and the depiction of interdependencies between the origin and destination regions. However, the data source is also associated with various limitations in terms of its usability in the microsimulation model, as it may contain duplications due to multiple moves at the individual level, only a few

socio-demographic characteristics are included and due to the lack of representation of events in the life course that correlate with relocation mobility. Another serious limitation is the fact that no information on household structures is available, as residential mobility often takes

place in the context of entire households moving. Moreover, the reasons for relocation mobility differ greatly depending on the household context. Due to these limitations the microcensus is explored as a possible data source for modelling the relocation mobility of individuals and households in the context of dynamic microsimulation. This official survey covers 1% of households in Germany every year. The advantages of this data source include the availability of information on household composition as well as numerous socio-demographic and socio-economic variables at both household and individual level. Although the microcensus is not designed as a panel, it is based on a rotation sample in which only a quarter of the observations are replaced each year between 2016 and 2019. The rest of the sample remains the same. By combining various approaches that utilise both the unplanned disappearance and unplanned reappearance of household and person IDs in the sample as well as the information available in the microcensus at (meta) data level, we present a rule-based procedure that makes it possible to identify populations moving in and out. A probabilistic matching approach is used to combine these subpopulations into a data set of a moving population.

3. RESULTS AND PRACTICAL APPLICATION

The approach mentioned above allows us to analyse the moving population in terms of socio-demographic characteristics and events in the life course shortly before or after the relocation. By comparing with the remaining immobile population, a statistical model can be developed to estimate small-scale relocation probabilities for entire households and individual household members.

For some variables, values from the previous year, i.e. the value from $t-2$, can also be reconstructed, so that not only the influence of a characteristic and its value in the previous year, but also the influence of changes in this variable between $t-2$ and $t-1$ on the probability of a relocation can be estimated. This applies, for example, to a change of occupation or the attainment of a school-leaving qualification. The main limitation of the approach to identifying out-migration is that no information on the destination of the mobility is included. In addition, the disappearance of persons/households alone cannot always be clearly attributed to relocation mobility, as a person can also leave the sample unplanned due to a death. Supplementing the identification of relocation mobility based on the analysis of IDs with the use of results from the survey itself makes it possible to take the reasons for the disappearance at both household and individual level into account, as they are not limited to residential mobility.

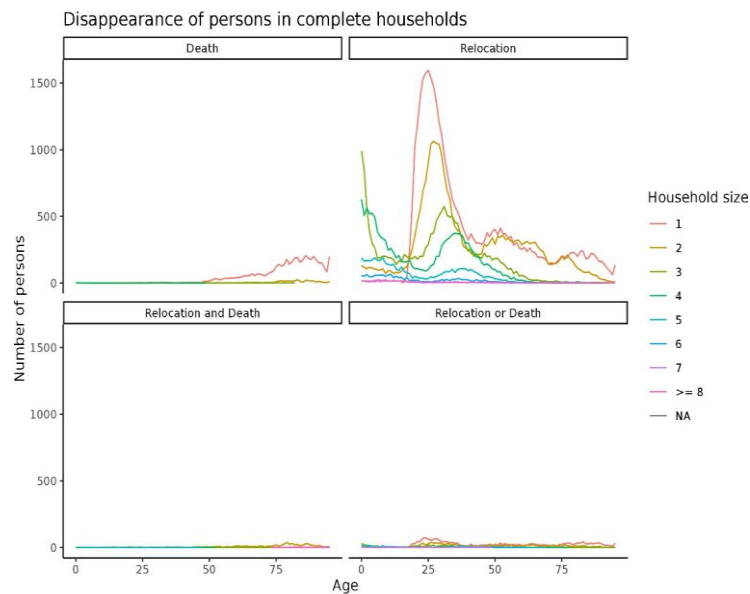


Figure 1. Reasons for disappearance of individuals in households by sex and household size.

First results show the effect of socio-demographic characteristics and events in the life course on the relocation probabilities of individuals and households in Germany. The figure above shows an example of how relocation dynamics differ over the life course depending on the age of the individual and size of the household. While the disappearance of households is generally most frequently caused by relocation mobility, the disappearance due to death also plays a significant role for older residents, mostly in single-person households. The figure shows that mixed types of household disappearance play a minor role in terms of numbers. The integration of the model into the microsimulation model shows the development of migration movements and their influence on the simulated regional evolution of the population structure for different household types up to 2040.

4. MAIN FINDINGS

Coming from the relocation dynamics depicted in the migration statistics, the longitudinal dataset generated based on the overlapping sample parts of the microcensus enables a more differentiated view of the situation of households. We discuss the relevance of modelling migration movements within the dynamic microsimulation model 'MikroSim' and show the limitations of the data base in this context.

5. REFERENCES

- [1] R. Münnich, R. Schnell, H. Brenzel, H. Diekmann, S. Dräger, J. Emmenegger and others A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model. In: methods, data, analyses (2021), 15 (2), p. 24. Available at <https://mda.gesis.org/index.php/mda/article/view/2021.03>.

Speeding up the release of the Total Market Production Index (TMPI)

Mojca Bavdaž (University of Ljubljana, School of Economics and Business), Tjaša Bartolj (Institute for Economic Research, Ljubljana), Aleša Lotrič Dolinar, Sašo Polanec & Tjaša Redek (University of Ljubljana, School of Economics and Business)

1. INTRODUCTION

Today's volatile economic environment requires governments and companies alike to closely monitor what is currently happening in the national and global economy and where it is heading. Official statistics provide the quantitative foundation for effective government policies and sound corporate strategy by ensuring the most comprehensive and comparable set of macroeconomic indicators that inform on the state and performance of an economy and its sectors and branches.

Traditional official statistics production relies on design-based primary data collection that requires time for collection of data and its processing. Even when secondary data is used, it is often available for statistics production with certain time lags. Monthly indicators are published with lags of up to two months. While this is sufficient during periods of stability, turbulent times highlight the need for more timely and accurate information, and the importance of producing official statistics faster [1] [2].

Solutions can be sought in new data sources and model-based approach to estimation besides faster data collection. The proliferation of available data sources presents unprecedented opportunities for early detection of economic trends and anomalies even if fast extraction of actionable insights from this vast trove of information remains a significant challenge. The model-based approach to estimation, while relatively new in official statistics, has long been employed in certain areas of statistics production, such as imputation and seasonal adjustment. Furthermore, there is a growing body of literature on (near) real-time estimation. However, a consensus on the optimal approach has yet to be reached.

The Slovenian NSI (SURs) launched a project to develop and evaluate a methodology for producing the total market production index (TMPI) much faster than with the current 60-day lag. The TMPI, a composite monthly output index, combines four short-term indicators that cover industry, construction, trade and services. By covering a large share of the total gross value added (about three quarters in the EU), it should give a good first indication of the evolution of the total economic performance on a monthly basis compared to GDP growth rates released quarterly [3]. In this extended abstract, we present the project findings regarding the state-of-the-art of real-time estimation in general and at NSIs and what appear to be the most promising avenues for further research.

2. METHODOLOGY

In order to assess the state-of-the art, we first conducted a literature review. We then conducted two group interviews at both Slovenian institutions that are regularly publishing forecasts, but are also preparing nowcasts for their internal purposes, namely the Bank of Slovenia and the Institute of Macroeconomic Analysis and Development (in Slovenian: UMAR, Urad za makroekonomske analize in razvoj). The Bank of Slovenia prepared a formal

presentation and some analyses comparing several methods specifically for this interview. UMAR shared internal documentation listing the variables included in the estimation. Both interviews were semi-structured, using a pre-defined list of topics (timeliness of economic indicators; traditional, new and potential data sources and their characteristics; methods and their assumptions, conditions, results etc.), and were conducted by a project team of three researchers (two specialised in econometrics and one specialised in data collection in official statistics). Altogether, interviews included eight interviewees, experts involved in preparing forecasts and nowcasts.

As a last step, a survey was conducted among national statistical institutes (NSIs), not only to understand the current situation, but mainly to get insights into planned and ongoing developmental work on real-time estimation. The invitation was sent by email to 37 NSI participating in the European Statistical System and 6 other¹ NSIs. Within the European Statistical System, we targeted short-term statistics coordinators, while in other NSIs we asked personal acquaintances to direct us to the most knowledgeable respondent(s). An overview of the questionnaire was attached as a pdf and a link was provided to enter the responses into the web questionnaire. So far, we received response from 26 NSIs, of which 21 from the European Statistical System. We also conducted three follow-up interviews with NSIs that had fast releases and/or worked on nowcasting.

In parallel, a database is being built to include all relevant indicators/variables to be used for modelling. The list of indicators/variables is inspired by lists of indicators/variables previously used in forecasting and nowcasting models for Slovenia and other countries, business micro data from the Slovenian NSI and newly available big data sources, such as electricity consumption. Slovenia has a data rich environment given its strong register orientation, but being a small open economy, indicators/variables beyond its borders must also be taken into account.

3. MAIN FINDINGS

The main findings are presented in two sections. The first section presents the findings from the literature review and interviews with experts. The second section focuses on the survey results.

3.1. Important aspects of modelling

The literature review suggested more than 15 models, including basic ordinary least squares, quantile regression, auto-regressive moving average (ARMA) models, auto-regressive distributive lag (ARDL) models, mixed-data sampling (MIDAS) regression models, vector auto-regression (VAR) models, principal component analysis, dynamic factor models (DFM) [4], Kalman filter, decision trees and machine-learning approaches.

When choosing among the models, data characteristics have to be considered, namely that there are many data series with potentially useful information content and that data comes with different frequencies (most typically with quarterly and monthly frequency but could also have weekly and daily frequency) and has a ragged-edge structure (starting and/or ending at different points in time). Additional complications come from challenges of seasonal and calendar adjustment.

¹ Statistics Canada, US Census Bureau, Australian Bureau of Statistics, Statistics New Zealand, Singapore Department of Statistics, UK Office for National Statistics

Simpler models seem sufficient for stable economic environment, but more advanced models, such as dynamic factor models, tend to perform better in more turbulent times while allowing for all data characteristics [5], [6].

3.2. Situation regarding real-time estimation at NSIs

The survey revealed that the majority of responding NSIs felt at least mild pressures for faster short-term indicators, but most responding NSIs haven't had any plans for producing a real-time monthly output indicator. Among obstacles to real-time estimation (Figure 1), lacking staff is the most critical one as more than half of responding NSIs marked it as an extreme obstacle. Other obstacles are also late availability of relevant data, budget for real-time estimation and – to some extent – also knowledge of methods for real-time estimation. On the positive side, top management support does not seem to be an issue in the responding NSIs.

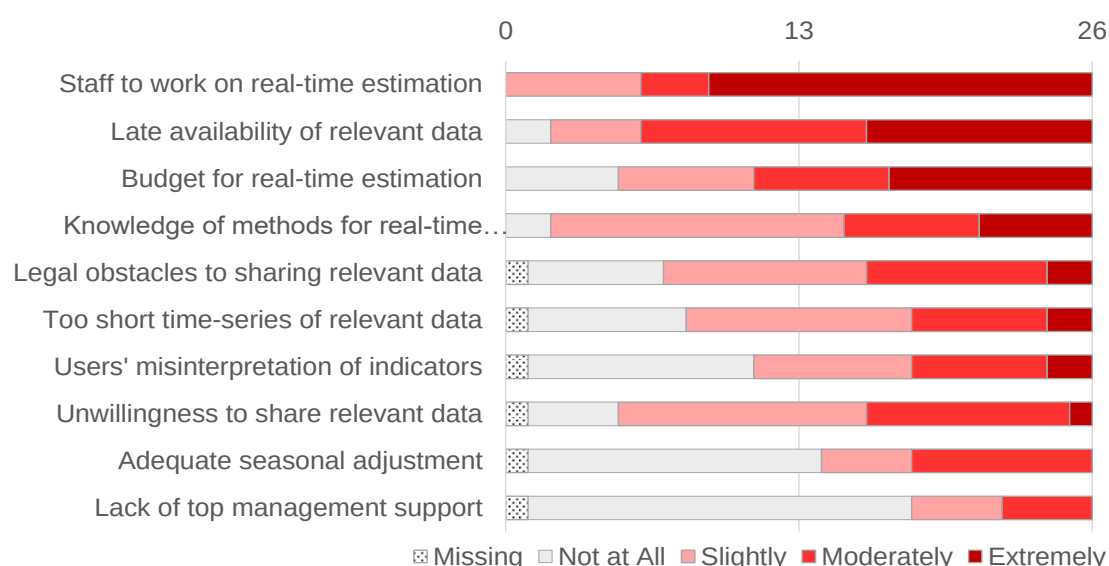


Figure 1. Obstacles to real-time estimation among responding NSIs ($n=26$)

The responding NSIs release the TMPI or a similar index (if they have one) between 21 and 60 days after the observed month. Those few NSIs that have tested any methods for nowcasting, mentioned DFM and Bayesian VAR models.

On the basis of the survey and follow-up interviews, we see three approaches to putting real-time estimation into life at NSIs:

- **Faster data collection:** Although we did not investigate deadlines that NSIs put to businesses for data reporting, we realised that at least one NSI requires businesses to submit data very fast after the observed month, which gives that NSI sufficient time for processing and release the indicator about three weeks after the observed month.
- **New data sources, available in real time:** Some NSIs publish specific indicators based on new data sources that are available in real time or very fast after the observed period, often with a shorter periodicity (e.g. weekly).
- **Modelling:** Some NSIs have been testing methods for nowcasting, but implementation in everyday production remains a challenge.

4. CONCLUSION

Nowcasting methods are nowadays mainly used for quarterly estimation of GDP growth and monthly estimation of inflation, primarily by central banks. An important step forward would be to have a broad output indicator on a monthly basis, such as the TMPI, available very soon after the end of the observed month. Shortening the publication lag can be achieved also by faster data collection, but for largest effects, modelling seems to be a priority option. Future research work should therefore test the set of relevant models. Also, the modelling itself needs good data input, so paying attention to achieving high quality data remains an important NSI task.

ACKNOWLEDGMENT

The work received funding from Statistical Office of the Republic of Slovenia and Slovenian Research and Innovation Agency (V5-2346, P5- 0441).

5. REFERENCES

- [1] M. Bańbura and G. Rünstler, A look into the factor model black box. Publication lags and the role of hard and soft data in forecasting GDP. *Working paper series* (2007), No 751, May 2007, European Central Bank. <https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp751.pdf>
- [2] S. MacFeely. Nowcasting: Data delayed is data denied. *Statistical Journal of the IAOS* (2021), 37, 257–258. DOI 10.3233/SJI-210788
- [3] Eurostat. Total market production (volume) index overview. Statistics Explained, September 2024. [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Total_market_production_\(volume\)_index_overview](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Total_market_production_(volume)_index_overview)
- [4] C. Fulton, Large dynamic factor models, forecasting, and nowcasting (2020, Aug 5). https://www.chadfulton.com/topics/statespace_large_dynamic_factor_models.html
- [5] L. Mosley, 2023: *High-Dimensional Methods for Timely and Interpretable Economic Statistics*. Lancaster University [PhD thesis].
- [6] D. Zult, S. Krieg, B. Schouten, P. Ouwehand, and J. van den Brakel, From Quarterly to Monthly Turnover Figures Using Nowcasting Methods, *Journal of Official Statistics* (2023), 39/2, 253-273.

Statistical Information Communication: past insights, future innovations

Reinoud Stoel, Vivian Meertens, Nynke Krol, Martijn Tennekes, Edwin de Jonge, Marco Puts, Nicole Bongers, Duco Hoogendoorn, Darren den Boer & Ger Snijkers
(Statistics Netherlands)

1. INTRODUCTION

National Statistical Institutes (NSI's) publish statistical information on a daily basis in many forms: press releases, corporate articles on website, reports, and tables. Sometimes infographics are used to illustrate text or in a standalone form. The target audience for all these types of output is broad. It can be an interested citizen, a student searching for data for a thesis, a journalist, a researcher, a government official, or a member of parliament. All of them have their own skills and wishes, and for all of them it's important that they understand and properly interpret the published information (in whatever form), and that they can use the information for the intended purposes.

Just like any NSI, Statistics Netherlands (CBS) puts a lot of effort in the quality of the statistical output to be published. Quality is important at every stage of the production process. Data collection, sampling, processing, and estimation methods are designed to ensure that our statistics are 'unbiased and precise'. Quality frameworks, like the Total Survey Error (TSE) Framework or the GSBPM (Generic Statistical Business Process Model), are used to describe the quality issues at every step in the process. At the output stage quality criteria like timeliness and relevance are used (Eurostat Statistics Code of Practice), but also clarity. CBS does yet not have explicit methods in place to empirically study the quality of our statistical information communication, as is the case for other quality criteria.

Current approaches of communicating official statistics are often based on convenience, user experience and common journalistic principles and practices. Quality aspects with respect to the understanding of the message of statistical communication are less known, and have not received much attention yet. This was the topic of our presentation^[1] at the 2023 Conference on New Techniques and Technologies for Statistics (NTTS). We stressed that it is unclear how published information comes across and how this in turn effects the understanding and interpretation of the published statistics. Naïve statistical information communication may introduce the risk of perception bias in a chart or prime a reader to draw premature conclusions, and may result in an adequate assessment of the (in)accuracy of statistics. We suggested several research directions that can improve the effectiveness of (visual) communication of NSI's. Although it may seem a trivial question, surprisingly little research has been done in this direction. Research has been done in several academic disciplines, such as information visualization, statistics, computer science (information theory), and psychology, but a holistic, interdisciplinary, approach is missing. In our 2023 presentation we encouraged the official statistics community to study how users perceive, interpret, and use their statistical information, and to develop methods and guidelines to elevate the quality of communication above "best practices". We presented our plans to conduct user studies at CBS, both in a controlled environment and in a free environment (for qualitative studies).

1.1. A new theme in the CBS Research Program Methodology

CBS aims for the production of reliable, detailed, timely and phenomenon-oriented statistics, while balancing costs and burden for society. Innovations in the statistical process and business operations should make this possible. Our CBS Research Program Methodology (RPM), run by the R&D Department, paves the way at CBS and delivers knowledge (of processes, methods, and data sources) and skills to facilitate future innovations, transitions and/or improvements in the statistical process. Within our RPM the work is divided into themes, with each theme having a specific niche. Every five years our RPM is calibrated against new societal developments, new production process demands and missing methods, and subsequent consequences for the production of our statistics. Because of the importance stressed in the introduction, our new 2025-2030 RPM will contain a new theme with the title “Statistical Information Communication”.

In the current presentation we will present what we have developed so far, what our plans are for the near future with this new theme, and how the collaboration between the Methodology Department and the Communication Department of CBS was formed. An important goal of this theme is methods development: developing methods to assess the quality of statistical information communication.

2. FUTURE INNOVATIONS: STATISTICAL INFORMATION COMMUNICATION

Research into the quality of statistical information communication is not yet a routine at CBS and faces challenges from many knowledge domains. One of the first challenges is determining unambiguously the relevant quality criteria, as well as the operationalisation of these criteria, i.e. how to assess quality according to these criteria. This is certainly non-trivial. For example, just as our statistical information should be ‘unbiased and precise’, we could also state that our publications should be, and can be, correctly interpreted, valued and used. The communication format affects how the information is interpreted, valued and used. Additionally, for instance, the attractiveness and accessibility of a publication can also be used as quality criteria: an infographic should also be appealing to look at (and thus attract attention); a long block of text is not very attractive to read for all users. There are multiple quality criteria, and certain criteria may be in conflict with each other.

As for correctly interpreting statistical information (one of the above-mentioned quality criteria), we see several challenges. Our underlying assumption is that people (i.e., users of our information) make an effort to take in and process this information systematically. However, psychological research shows the opposite. What plays a role here is that people are reading less well and paying less attention: they have less patience for taking in and processing information. If a piece of text takes up more than one page, many people often find it too long already. Texts are increasingly being supplemented with ‘infographics’. CBS takes these factors into account in all standard communication with the outside world. Additionally, we know that ‘innumeracy’, or a lack of understanding of numbers, is a widespread phenomenon. In general, people are bad at working with numbers and interpreting them. This is also an aspect we need to deal with: we publish statistical data about the Netherlands, and as such, our data requires interpretation. Statistical data are not hard facts, but rather have an uncertainty margin. Here we speak of ‘statistical literacy’. There has been a lot published on innumeracy and statistical literacy. The motivation to process information, and reading and numerical skills will vary per user and user group. Furthermore, cognitive factors, such as how our brain (graphically) processes information, and its limitations (such as colour blindness), also play a role. This requires a multidisciplinary approach. The challenge is to create statistical information that is directly usable by users without them first having to learn how to work with numbers. The focus for the research theme Statistical Information Communication will be on empirical research in which we put the user at the core.

As for operationalizing the quality of statistical information communication for different user groups, we see the following research questions:

- How can we study the quality of our statistical information communication?
- Which quality criteria for texts and graphics can we distil from this?
- Which knowledge domains are supportive here?
- What do these disciplines say about the quality of statistical information communication?
- What knowledge do we already possess, and what do we need to develop?

For measuring the quality of our statistical information communication, we have the following research questions:

- Which research techniques (such as for example user research or 'eye-tracking' techniques) are already available to study the quality of our statistical information communication?
- Which ones work well in practice and which ones don't?
- Are the techniques also methodologically sound?

Finally, depending on the knowledge gained from the first and second research questions, we can determine how our statistical information can be improved.

- What changes to texts, tables and visualisations can improve the readability of our statistical information?

3. PAST INSIGHTS

In recent years, several exploratory studies have been undertaken to evaluate the quality of statistical information communication at CBS. These studies aimed to identify how users interpret and interact with statistical information, laying the groundwork for future research:

- A qualitative Study on Graph Interpretation: This study explored how a broad audience interprets statistical graphs. Participants were shown a graph and asked to describe what they understood from it.
- A study on the Interpretation of the word 'most': In this ad-hoc experiment, CBS colleagues were asked to interpret the phrase, "In a jar of marbles of different colours, most of them are red." Responses varied widely, indicating that even simple terms like "most" can be understood differently.
- An exploratory Study on Colour Perception: This study focused on how users perceive and interpret colours in graphs and infographics. The study highlighted the need for more thoughtful colour selection in data visualizations, particularly in regard to accessibility (e.g., colour blindness).
- Pre-testing Methods for Questionnaire Quality: we reviewed existing pre-test methods used to assess the quality of survey questionnaires and examined their potential application for evaluating statistical information communication. The study categorized and analysed methods, providing insights into how similar approaches could be adapted to test the clarity and effectiveness of published statistical information.

These exploratory efforts underline the importance of moving toward a more structured and systematic research agenda. The insights gained thus far will inform the development of methods and criteria that can be systematically applied to improve the quality of statistical information communication, ensuring they are more easily understood and accurately interpreted by diverse user groups.

The main findings of these exploratory studies will be discussed in detail at the conference, particularly with regard to the development of methods for evaluating statistical information communication. A key takeaway is that methodological rigor is essential in designing and

implementing studies to increase the likelihood of correct interpretation of the statistical information provided.

4. REFERENCES

- [1] Martijn Tennekes, Marco Puts, Edwin de Jonge, Vivian Meertens, and Reinoud Stoel (2023). *Visualization and communication of statistics*. Presented at the 2023 Conference on New Techniques and Technologies for Statistics.

Statistical Insights about Elderly Hospitalization in India: Count Data Model

Kripa Josten & J. Vennila (Statistics, Manipal College of Health Professions,
Manipal Academy of Higher Education)

Objective: The goal of the study was to evaluate various predictors influencing the inpatient admissions/hospitalization among the elderly population of India.

Materials and Methods: The sample of 488 hospitalized patients were taken from the WHO SAGE India wave 2 dataset. The outcome variable of the study was the hospitalization/inpatient admissions of the elderly. The various predictors included were age, gender, behavioral patterns like smoking and alcohol consumption and various self-reported prevalence of NCD's. Poisson regression was done and corresponding IRR and p-value along with the 95% CI were reported.

Results: The predictors of hospitalization included age, gender, arthritis and hypertension. Females had lower incidence to hospitalization compared to their male counterparts (IRR=0.807, 95% CI: 0.66-0.97, $p=0.026$). Additionally, arthritis related hospital admissions were associated with higher age groups (IRR=1.24, 95% CI: 1.02-1.51, $p=0.027$). However, both age and hypertension are inversely related to hospitalizations among the elderly. The use of bootstrap resampling in Poisson regression analysis provided a more robust and accurate understanding of the relationships between variables and the outcome.

Conclusion: The study identifies important predictors that are involved in the hospitalization of India's older population. The study's conclusions show that preventing hospitalizations among the elderly may be achieved via the efficient management of chronic health issues.

1. BACKGROUND

Hospitalization is a vital part of providing healthcare and is frequently linked to medical, social, and economic impacts. Knowing what the predictors are of the hospitalization, can assist the policymakers and healthcare professionals in creating plans to cut down on needless admissions, enhance patient outcomes and better allocate resources. Major determinants of health care utilization include health status and the need for medical treatments to maintain or enhance health. The World Health Organization states that a person's unique characteristics and behaviors, socioeconomic status, and physical environment all affect their overall health [1]. These are contributed by socio-demographic variables such as age, gender, place of residence etc... and determinants of the diseases. Between 1995 and 2014, there was a 2.23-fold increase in the yearly hospitalization rate per 1000 people in India; the increase for NCDs was greater than for CDs, at 3.61 versus 2.25 times. From 38.6% in 1995–1996 to 62.2% in 2014, noncommunicable diseases accounted for a larger portion of all hospital admissions [2].

Regression-type studies are possible using count data models in cases where the dependent variable of interest is a numerical count. This can be used to calculate how a policy action

affects the chance of when there is no event, one event, or several events, as well as the average rate [3]. One such model is Poisson Regression. A comparison of differences can evaluate how the policy changed before and after comparing the pre-post changes in the outcome distributions of treatment and control groups. Comparing treatment and non-treatment units while accounting for confounding factors is an alternate technique [4].

A Poisson regression model can be performed in the same way as a linear regression model, and the results can be interpreted in a similar manner. In fact, many characteristics of the linear regression model are retained by the Poisson model, despite being non-linear and expressing the mean as an exponential function of the explaining factors. In essence it belongs to a family of generalized linear models' class [5].

Rather than restricting responses to a binary category (i.e., hospitalized or not, utilized home services or not, etc.), this model allows for many incidents, or counts of events, within each domain (e.g., allowing for more than one hospitalization per individual) [6].

2. OBJECTIVE

To evaluate the association between various demographic, behavioral, and health-related variables and the incidence of hospitalization for elderly.

3. METHODS & MATERIALS

3.1. Study Population

Study on Global Ageing and Adult Health (SAGE) is an initiative jointly undertaken by International Institute for Population Sciences (IIPS), Mumbai, and the World Health Organization (WHO), Geneva [7]. Data on individuals 50 years of age and above, as well as more limited comparative sample of adult population (18–49), were gathered for the longitudinal SAGE project from nationally representative samples in the Russian Federation, South Africa, Mexico India, Ghana and China. Uttar Pradesh, Karnataka, Assam, Maharashtra, Rajasthan and West Bengal were the six Indian states where SAGE was carried out. Wave 2 in India was conducted in 2015 and covered 9116 and 8152 completed individual and households' interviews respectively [8] [9].

Sample: The number of inpatient hospitalizations with non-communicable diseases in the WHO-SAGE wave 2 is 488.

Variables: The outcome variable is inpatient hospitalization, and the predictors are age, gender, smoking status, alcohol consumption, NCDs such as Arthritis, stroke, Angina, Diabetes, Asthma, Hypertension, Cataracts.

3.2. Statistical Analysis

Poisson regression models were run for checking the various factors affecting the health-care utilization domains using R and Jamovi 2.5.3. Incidence Rate Ratios (IRR) were also provided, and p-values less than 0.001 are considered statistically significant.

4. RESULTS

Variables	N (%)	Hospitalization (n=488)	Robust Estimation

			IRR	[95% CI]	p-value	Bootstrap Std. Err.	Normal-based [95% CI]
Age (median, IQR)		68 (63,73)	0.986	(0.97,0.99)	0.047	0.006	(-.02, -.001)
Gender	Female	271 (55.53)	0.807	(0.66,0.97)	0.026	0.115	(-0.46, -0.01)
	Male	217 (44.47)	-			-	
Smoking	Yes	210 (43.03)	1.029	(0.85, 1.23)	0.762	0.131	(-0.22, 0.28)
	No	278 (56.97)	-			-	
Alcoholic	Yes	74 (15.16)	0.931	(0.72, 1.19)	0.583	0.155	(-0.38, 0.22)
	No	414 (84.84)	-			-	
Arthritis	Yes	118 (24.18)	1.246	(1.02, 1.51)	0.027	0.145	(-0.02, 0.54)
	No	370 (75.82)	-			-	
Stroke	Yes	45 (9.22)	0.908	(0.66, 1.23)	0.546	0.148	(-0.39, 0.18)
	No	443 (90.78)	-			-	
Angina	Yes	50 (10.27)	1.143	(0.87,1.50)	0.334	0.228	(-0.28,0.60)
	No	437 (89.73)	-			-	
Diabetes	Yes	111 (22.75)	1.088	(0.88, 1.33)	0.425	0.178	(--0.24, 0.45)

	No	377 (77.25)	-			-	
Asthma	Yes	53 (10.88)	0.961	(0.73, 1.26)	0.776	0.124	(-0.29, 0.19)
	No	434 (89.12)	-			-	
Hypertension	Yes	179 (36.76)	0.80	(0.66, 0.98)	0.031	0.137	(-0.50,0.03)
	No	308 (63.24)	-			-	
Cataracts	Yes	202 (41.48)	1.06	(0.89,1. 26)	0.489	0.105	(-0.13,0.27)
	No	285 (58.52)	-			-	

-Reference Category, * Statistically Significant

The median age of the elderly population who were hospitalized was 68 with a majority of female population (55.53%). Hospitalization was higher for participants with cataracts (41.48 %) followed by hypertensive patients (36.76 %). For each one-year increase in age, the rate of hospitalization slightly decreases by 2% (p-value=0.047, 95% CI [0.97,0.99]). Females have a 19.3% lower rate of hospitalization compared to males (p-value= 0.026, 95% CI [0.66,0.97]). Individuals with arthritis have a 24% higher rate of hospitalization (p-value= 0.027, 95% CI [1.02, 1.51]). Hypertension is associated with a 20% decrease in hospitalization rates (p-value= 0.031, 95% CI [0.66, 0.98]).

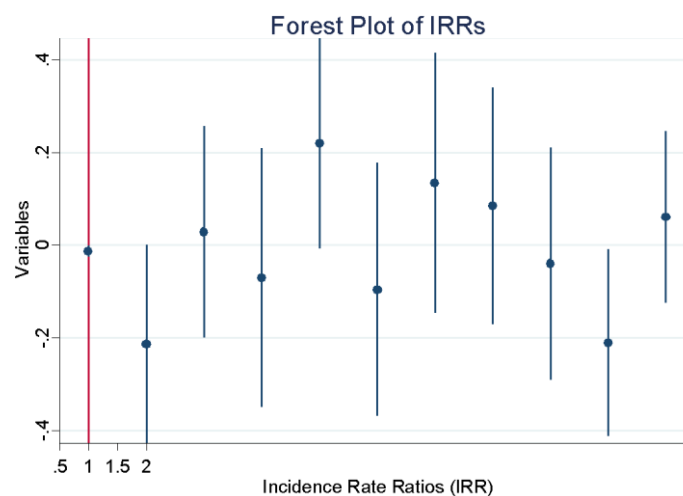


Figure 1. Forest Plot

A reliable indicator of the IRR estimate's variability is the bootstrap standard error. The bootstrap standard error for age, for instance, is 0.006 (95% CI: [-.02, -.001]), indicating that the age-specific IRR estimate is fairly reliable. The forest plots (Fig 1) is graphical representation of multiple models in Poisson regression. This shows the statistical significance of the impact on the outcome, the precision of the estimates and the direction of the effect.

5. DISCUSSION

The results of the study provide insights of the predictors among the 488 samples of elderly population from WHO SAGE wave 2 data during hospital admissions. The results of the study show that hospitalization rates are inversely correlated with age and gender, but they are positively correlated with arthritis and negatively correlated with hypertension.

The results of our investigation supported other studies by confirming that older males had a greater hospitalization rate ([10][11][12]. According to our findings, hospitalization rates for elderly women are lower than those for elderly men. Similar hospitalization patterns have also been noted in other nations, such as Norway [13][14], Italy [15]. Chronic arthritis is known to cause severe pain and limited movement. As a result, patients with arthritis are likely to require more frequent hospital stays for symptom management and therapy. This observation is in alignment with Conesa et al. [16] showing that long-term illnesses like arthritis have a major role in the rise in healthcare use. Studies indicating increased hospitalization rates for hypertensive patients, however, contradict the negative correlation between hypertension and hospitalization. Since the bootstrap approach yields precise standard errors and confidence intervals—especially in scenarios where conventional assumptions might not hold—it strengthens the robustness of these findings.

The limitation of the study might include the smaller sample size and self-reporting by the individuals on NCD's and behavioral habits might add to the bias and the data quality. The study population may not be the representative sample of the general elderly population. Though count models were used, overdispersion of data might affect the fitness of the model.

6. CONCLUSION

The study highlights significant predictors included in the hospitalization of elderly population of India. The findings of the study indicate effective management of chronic health conditions and preventative healthcare practices can significantly reduce the hospitalization rates among elderly. This can help healthcare professionals such as clinicians, researchers etc., create focused interventions that will enhance patient outcomes and make the most use of available resources.

7. REFERENCES

- [1] WHO (World Health Organization), "Health impact assessment (HIA), determinants of health. Geneva, Switzerland: WHO; ."
- [2] A. Pandey, G. B. Ploubidis, L. Clarke, and L. Dandona, "Hospitalisation trends in India from serial cross-sectional nationwide surveys: 1995 to 2014," *BMJ Open*, vol. 7, no. 12, Dec. 2017, doi: 10.1136/bmjopen-2016-014188.
- [3] R. Winkelmann and K. F. Zimmermann, "Count Data Models for Demographic Data," *Math Popul Stud*, vol. 4, no. 3, pp. 205–221, Feb. 1994, doi: 10.1080/08898489409525374.

- [4] R. Winkelmann, "Counting on count data models," *IZA World of Labor*, 2015, doi: 10.15185/izawol.148.
- [5] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Second. 1989.
- [6] F. Skelton, J. M. Hoffman, M. Reyes, and S. P. Burns, "Examining health-care utilization in the first year following spinal cord injury," *Journal of Spinal Cord Medicine*, vol. 38, no. 6, pp. 690–695, Nov. 2015, doi: 10.1179/2045772314Y.0000000269.
- [7] "SAGE Wave-2," IIPS.
- [8] "INDIA Study on global AGEing and adult health (SAGE), Wave 2 WHO SAGE WAVE 2."
- [9] "WHO Study on global AGEing and adult health (SAGE)."
- [10] S. Naaz Akhtar, N. Saikia, and S. N. Akhtar, "Differentials and predictors of hospitalization among the elderly people in India: Evidence from 75 th round of National Sample Survey (2017-18)," *medRxiv*, 2021, doi: 10.1101/2021.08.25.21262606.
- [11] A. Pandey, G. B. Ploubidis, L. Clarke, and L. Dandona, "Hospitalisation trends in India from serial cross-sectional nationwide surveys: 1995 to 2014," *BMJ Open*, vol. 7, no. 12, Dec. 2017, doi: 10.1136/bmjopen-2016-014188.
- [12] M. Olofsson, J. H. Jansson, and K. Boman, "Predictors for hospitalizations in elderly patients with clinical symptoms of heart failure: A 10-year observational primary healthcare study," *Journal of Clinical Gerontology and Geriatrics*, vol. 7, no. 2, pp. 53–59, Jun. 2016, doi: 10.1016/j.jcgg.2015.11.001.
- [13] A. L. Suominen-Taipale, T. Martelin, S. Koskinen, J. Holmen, and R. Johnsen, "Gender differences in health care use among the elderly population in areas of Norway and Finland. A cross-sectional analysis based on the HUNT study and the FINRISK Senior Survey," *BMC Health Serv Res*, vol. 6, Sep. 2006, doi: 10.1186/1472-6963-6-110.
- [14] M. T. Gjessten, K. Brønneck, and I. Testad, "Characteristics and predictors for hospitalizations of home-dwelling older persons receiving community care: A cohort study from Norway," *BMC Geriatr*, vol. 18, no. 1, Sep. 2018, doi: 10.1186/s12877-018-0887-z.
- [15] R. Rozzini, I. Sleiman, S. Maggi, M. Noale, and M. Trabucchi, "Gender Differences and Health Status in Old and Very Old Patients," *J Am Med Dir Assoc*, vol. 10, no. 8, pp. 554–558, 2009, doi: 10.1016/j.jamda.2009.04.005.
- [16] M. Gil-Conesa, J. A. Del-Moral-Luque, R. Gil-Prieto, Á. Gil-De-Miguel, R. Mazzucchelli-Esteban, and G. Rodríguez-Caravaca, "Hospitalization burden and comorbidities of patients with rheumatoid arthritis in Spain during the period 2002-2017," *BMC Health Serv Res*, vol. 20, no. 1, May 2020, doi: 10.1186/s12913-020-05243-0.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. These data can be obtained from <https://iipsindia.ac.in/content/SAGE-data> on request.

ETHICAL CLEARANCE

This study did not include any human participants since the data (WHO-SAGE) used was available in the public domain hence, there were no ethical considerations. The secondary dataset (WHO-SAGE) considered for this study was approved by the World Health Organization's Ethics Review Board in Geneva, Switzerland, as well as the Institutional Review Board of the International Institute of Population Science in Mumbai, India.

Statistical Interoperability Node

Cecilia Albérica García Rodríguez, Jesús Alberto González Yanes (Canary Islands Statistics Institute), Carlos Ballano Fernández (National Statistics Institute), Agustín Tomás Villar Iglesias (Statistical and Cartographic Institute of Andalusia), Jorge Velasco López, Ana María López Lago (National Statistics Institute, INE) & Adrián Juan Verdejo (DG Reform)

1. INTRODUCTION

The European Digital Agenda, a key component of the Europe 2020 Strategy, has driven digital transformation across member states, prompting Spain to adopt its own Digital Agenda in 2013, followed by the Spain Digital 2025 plan. This initiative, which aligns with the Sustainable Development Goals (SDGs), aims to implement structural reforms in response to challenges posed by the COVID-19 pandemic. A primary focus of this plan is the modernisation of Spanish public administration, particularly through the enhancement of statistical data interoperability to support improved decision-making processes.

The Statistical System of the Spanish State Administration is responsible for producing official statistics essential for government decision-making. However, this existing framework encounters significant challenges, such as improvable data exchange and inadequate interoperability among regional statistical systems. These issues have become increasingly pressing considering the COVID-19 pandemic, highlighting the need for enhanced coordination and technical integration among various administrative and statistical entities.

Data interoperability is vital for connecting information from disparate sources, enabling the creation of comprehensive datasets that facilitate analysis and decision-making. Achieving this interoperability necessitates the establishment of standards, methodologies, and tools for data harmonisation, alongside uniform technical interfaces to enable the standardised reuse of statistical information. The advantages of improved statistical data interoperability encompass increased dataset value, enhanced reuse of statistical information, reduced analysis costs, better governmental decision-making, and improved data quality.

To realise this vision, a Statistical Interoperability Node will be developed, aligned with the interSTAT framework¹, facilitating the integration of data and metadata processes. This initiative aims to support strategic decision-making across various administrative levels in Spain, promoting the exchange of statistical data among public administrations and national or regional statistical offices, ultimately benefiting citizens and enterprises through enhanced data accessibility and transparency.

¹ The interSTAT working group, within the Interterritorial Statistics Committee (CITE), has developed a framework to enable data interoperability between national statistical portals, the European Data Portal and the deployment of cross-border services that re-use European statistical open datasets.

2. METHODOLOGY

The methodology is structured into a logical sequence that encompasses the entire project lifecycle, from initial analysis to the final implementation of proofs of concept. By adopting an iterative, agile framework, it allows for flexibility and adaptability as new insights arise. The approach comprises clearly defined phases: current processes analysis, business requirements catalogue, final solution, and model definition, each building on the insights gleaned from the preceding phase.

2.1. Current processes analysis

This phase aims to evaluate the current statistical interoperability landscape among Spanish government bodies. It involves creating a detailed AS-IS Business Process Model using Business Process Model and Notation (BPMN) to visually represent workflows, stakeholders, information systems, and legal frameworks. This foundational document facilitates understanding data flows between entities and identifies bottlenecks and inefficiencies. Additionally, a SWOT analysis assesses strengths, weaknesses, opportunities, and threats, spotlighting areas for enhancement and revealing the drivers and obstacles impacting data exchange.

2.2. Business requirements catalogue

The business requirements catalogue phase is critical for defining needs regarding statistical data interoperability. Building on insights from the current processes analysis, this phase engages stakeholders through workshops and interviews to gather detailed requirements. The resulting catalogue classifies these requirements into strategic, regulatory, functional, technical, and economic categories, ensuring a thorough understanding of interoperability aspects. This documentation provides a clear roadmap for developing alternative solutions and serves as a reference for evaluating proposed models against stakeholder needs.

2.3. Final solution

The final solution phase concentrates on developing alternative solutions to address identified needs. Various alternatives are proposed alongside a baseline scenario that maintains current processes. Each solution is supported by a cost-benefit analysis, considering implications for organisational processes, technology, and legal frameworks. Detailed documentation presents best practice recommendations for technical data management, and a scorecard methodology facilitates the objective evaluation of alternatives, culminating in a comprehensive report outlining the proposed solutions and their advantages.

2.4. Model definition

Following the selection of the final solution, the model definition phase establishes the framework for future statistical interoperability. This phase focuses on co-designing the model, incorporating enhanced data and metadata management policies and the necessary technical and functional architecture for effective implementation. Documentation includes detailed business process models illustrating new workflows and data exchange mechanisms. Additionally, gap analyses identify discrepancies between the current state and future model, resulting in the development of a roadmap outlining implementation steps, timeframes, dependencies, and preliminary risk assessments. This structured approach ensures stakeholder alignment on the vision for interoperability, facilitating effective execution and the long-term sustainability of proposed solutions.

3. RESULTS AND PRACTICAL APPLICATION

Following an exhaustive analysis of potential solutions, several key conclusions have been reached regarding the implementation of interoperability measures. A promising solution involves developing a unified catalogue with services and APIs for automated information ingestion and retrieval, thereby supporting interoperability and streamlining data management for beneficiaries with their own catalogues.

Economic sustainability has emerged as a crucial factor, particularly as two solutions (CP4Data and eDelivery) may incur costs dependent on external licensing. Moreover, while open-source solutions offer free distribution, they typically necessitate substantial development and maintenance resources.

In assessing various models, the Eurostat solution was dismissed due to its limitations in supporting decentralisation and integration with other technologies. The eDelivery solution, though capable, requires each entity to possess its own Domibus agent, rendering it impractical for the Interoperability Node. The custom or open-source solution, while flexible, would entail considerable ongoing development and maintenance, leading to increased complexity and governance challenges, resulting in its rejection.

Ultimately, the Data Platform of Spain was selected as the preferred solution. It effectively meets the established requirements, offering a unified technological framework through IBM's Cloud Pak for Data (CP4D). This integrated approach ensures seamless communication between various components, facilitating the governance and management of the Interoperability Node. Although the primary focus of the Node is not currently on analytical capabilities, the Data Platform of Spain presents future opportunities for advancement towards enhanced data exploitation and artificial intelligence techniques. The platform also establishes a framework for addressing the requirements identified in earlier analyses, ensuring the Node can operate efficiently and securely. Overall, the Data Platform of Spain equips the Interoperability Node with the necessary tools and frameworks to manage data sharing effectively and ensure compliance, transparency, and security, thereby establishing a robust foundation for statistical interoperability.

4. MAIN FINDINGS

The establishment of the Statistical Interoperability Node (SIN) in Spain represents a transformative advancement in public administration and data management. Acting as an intermediation platform, SIN facilitates the seamless exchange of statistical data among various governmental levels while adhering to national and European interoperability standards. Drawing inspiration from modern data governance frameworks such as GAIA-X, the initiative aims to modernise public statistical functions through enhanced interoperability.

Enhancing Public Administration: SIN contributes to the modernisation of public administration by improving the efficiency, transparency, and quality of public services. It promotes seamless data exchange among public entities, reducing redundancy and enhancing decision-making in line with the Interoperable Europe Regulation.

Legal framework alignment: The development of SIN is underpinned by Law 12/1989, ensuring reliability and transparency in public statistics. The platform supports compliance with the National Interoperability Framework and the Data Governance Regulation, positioning itself as a trusted intermediary that upholds legal and security standards.

Innovation through GAIA-X principles: SIN leverages the principles of federated data spaces, allowing public institutions to share data securely while maintaining sovereignty. This

innovative approach fosters collaboration between local and national authorities, coordinated by the National Statistics Institute (INE).

Comprehensive interoperability: The platform incorporates semantic, legal, and technical interoperability layers. By employing metadata standards such as DCAT-AP, SIN ensures a common understanding of data across government levels, thereby enhancing its accessibility and reusability.

Strategic coordination and decentralisation: SIN integrates strategic recommendations from the Interterritorial Statistical Coordination body (CITE), ensuring coordinated statistical data flows that address the needs of both autonomous communities and central government bodies.

In conclusion, SIN represents a significant leap forward in public data governance, setting a precedent for future initiatives that prioritise secure, efficient, and legally compliant data sharing. Its federated model and comprehensive approach to interoperability position it as a leading example in both national and European contexts.

5. REFERENCES

- [1] Ley 12/1989, de la Función Estadística Pública. (1989). Official Journal of Spain (BOE), July 9, 1989, No. 164. Madrid: Government of Spain. Retrieved from <https://www.boe.es/buscar/doc.php?id=BOE-A-1989-15816>
- [2] Data Governance Act (Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance). (2022). Official Journal of the European Union, L 152/1, June 3, 2022. Brussels: European Union. Retrieved from <https://eur-lex.europa.eu/eli/reg/2022/868/oj>
- [3] Regulation on Interoperable Europe (Regulation (EU) 2024/903 of the European Parliament and of the Council of 25 May 2024 on Interoperable Europe). (2024). Official Journal of the European Union, L 144/1, May 27, 2024. Brussels: European Union. Retrieved from <https://eur-lex.europa.eu/eli/reg/2024/903/oj>
- [4] Real Decreto 4/2010, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica. (2010). Official Journal of Spain (BOE), January 29, 2010, No. 25. Madrid: Government of Spain. Retrieved from <https://www.boe.es/buscar/doc.php?id=BOE-A-2010-1331>
- [5] European Data Strategy (COM (2020) 66 final). (2020). A European strategy for data. Brussels: European Commission. Retrieved from <https://ec.europa.eu/digital-strategy/our-policies/european-data-strategy>
- [6] DCAT-AP: Data Catalogue Vocabulary Application Profile for Data Portals in Europe. (2020). European Commission. Retrieved from <https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe>
- [7] GAIA-X: A Federated Data Infrastructure for Europe. (2020). Federal Ministry for Economic Affairs and Climate Action, Germany. Retrieved from <https://www.data-infrastructure.eu/GAIA-X/Navigation/EN/Home/home.html>
- [8] ISO/IEC 21838-2:2021: Information technology - Top-level ontologies (TLO) - Part 2: Basic Formal Ontology (BFO). (2021). International Organization for Standardization (ISO). Retrieved from <https://www.iso.org/standard/74572.html>
- [9] European Interoperability Framework (EIF). (2017). European Commission, Brussels. Retrieved from https://ec.europa.eu/isa2/eif_en

Streamlining the Supply and Analysis of Big Data at Research Data Centres

Oliver Hauke (Destatis)

1. INTRODUCTION

The Research Data Centre (RDC) at Destatis provides a wide variety of large datasets for scientific use. Researchers from diverse fields utilize Destatis' infrastructure to access the provided micro data either onsite or via remote execution.

However, significant technical challenges arise for both researchers and RDC staff when analyzing large and complex datasets, given the limited computing resources available. The execution of research scripts can be extremely time- and resource-consuming, sometimes running for days or weeks and leading to errors. Additionally, variability in researchers' programming expertise across different fields impacts the efficiency of the analysis. Problems and errors in execution can often intensify within the RDC environment, leading to significant delays. In remote execution, extended feedback loops between RDC staff and researchers can potentially last several weeks. Onsite access is constrained by limited reserved time and the potential need for travel.

For typical data manipulation and analysis tasks at RDCs, we demonstrate the effective use of the programming language R to significantly boost efficiency and extend the range of available use cases within the provided infrastructure at the RDC of Destatis.

2. METHODOLOGY

We apply the R-Packages Arrow and DuckDB, in combination with the Apache Parquet storage format, to increase efficiency in realistic use cases. The datasets considered are derived from large datasets offered by the RDC of Destatis, such as the Diagnostic-Related Group Statistics or the Business-Tax-Panel. Analytical tasks at RDCs typically revolve around all parts of the data science project cycle.

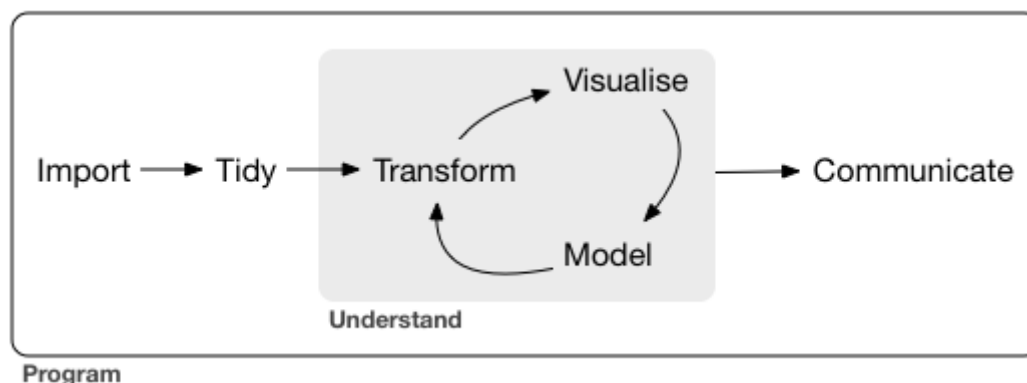


Figure 1. Data science project cycle (Wickham & Grolemund, 2016)

In data wrangling (tidying and transforming together), multiple datasets often need to be joined at our RDC. Common aggregation tasks, such as calculating summary statistics, are typically applied to understand the univariate distribution of each variable. Researchers at the RDC are particularly interested in the interactions between variables. Therefore, they measure covariation, explore patterns and finally derive models. We demonstrate the analytical methods made possible with the considered technologies and highlight the limitations of the considered packages. The presentation provides benchmarks for these analytical tasks applied to large RDC datasets, along with guidelines to significantly enhance efficiency and performance.

To ensure large datasets are fully accessible for analysis, we provide recommendations for RDCs to optimize data provisioning without the need for additional expertise or advanced technologies like databases. Our approach involves distributing data across multiple cohesive files that align with available memory. Common file types, such as CSV, can be enhanced through compression, significantly reducing storage requirements and improving I/O performance when data is stored on a network drive. Switching to Parquet files, which are specifically designed for big data, offers further advantages. Parquet improves storage efficiency and I/O speed through its use of effective encoding and built-in compression. As a column-oriented and chunked format, Parquet allows for more efficient analysis, making it well-suited for large-scale RDC tasks.

These recommendations for storage optimization at RDCs align with Gomolka et al. (2021), who focused on efficient in-memory data analysis in the RDC context. We expand their approach to larger-than-memory datasets, relying solely on R and avoiding the need for additional tools such as databases, Spark or SAS. Since R is widely available and used among researchers, our approach remains widely accessible. Integrated as R packages, we employ Apache Arrow and DuckDB.

The arrow package is designed for efficient analysis and transport of large datasets, enabling fast analysis of larger-than-memory datasets. Arrow datasets can be seamlessly converted to DuckDB databases without any memory copying, enabling even more efficient analysis and queries. All mentioned packages are accessible in R via familiar and user-friendly dplyr syntax.

The integration of the presented methodologies in RDC research projects expands the scope of use cases, including faster data exploration, improved resource efficiency, and the ability to handle more complex analyses within familiar R-based workflows. Our approach ensures that the provided infrastructure can be used not only efficient but also adaptable, scalable, and user-friendly for researchers from various fields.

3. RESULTS AND PRACTICAL APPLICATION

The presentation provides an overview of the current technical capabilities at the RDC of Destatis, informs the audience about planned future developments, and demonstrates how the infrastructure can be used in a user-friendly, stable, and efficient way.

We provide recommendations for RDCs on data preparation techniques for efficient use, with a focus on widely accessible and easy-to-use solutions that offer significant efficiency benefits. One such recommendation is to distribute datasets to multiple cohesive Parquet files.

By considering datasets derived from large RDC datasets, such as the Diagnostic-Related Group Statistics or the Business-Tax-Panel, we present performance benchmarks for typical analytical RDC tasks. This allows us to demonstrate the time and resource benefits of our approach compared to classical tools and techniques.

The presentation includes guidelines and examples addressing RDC staff and researchers, demonstrating how our approach enables new use cases by analysis of larger-than-memory datasets and facilitating interactive exploration of these datasets.

Overall, the presented methodology can be practically applied by RDCs to enhance the availability of large dataset analysis and by researchers to improve their efficiency in utilizing R.

4. MAIN FINDINGS

The evaluation of Arrow and DuckDB in our RDC infrastructure showed huge performance benefits, including reduced computation time, memory usage and storage consumption. Furthermore, the stability has improved as it helps avoid out-of-memory errors.

Alternatively, performing basic analytical tasks in our RDC, such as generating univariate summary statistics or constructing Lorenz curves, on larger-than-memory datasets were only feasible using SAS, and could take several hours to complete. By applying Arrow to a distributed Parquet-based dataset, the same analyses were accomplished within a few seconds. This advancement enables new use cases. Researchers have expressed great interest in interactive exploration of large datasets. This is now feasible within RDCs using the presented tools.

Our approach was easy to deploy within our infrastructure, using only R-integrated packages without the need for additional software or extensive expertise. The syntax is suitable for a wide range of researchers, as it requires only minimal adjustments compared to classical in-memory data analysis in R.

Our presentation will also address the limitations and lessons learned from our evaluation. Initially, RDC staff had to prepare datasets by specifying metadata for datasets with more than 1,500 variables. Arrow does not fully support all data operations available through dplyr verbs. For example, common operations like windowed aggregation, calculating covariation, or generating models are not performed by Arrow when the data exceeds available memory. DuckDB offers solutions for some unsupported data operations. In other instances, queries needed to be reformulated in a specific way for Arrow to process them.

We will present the final results of our evaluation in structured and standardized benchmarks using available large datasets realistic to RDC use cases.

5. REFERENCES

- [1] R Core Team, R: A language and environment for statistical computing (2021), R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>
- [2] Research Data Centre Destatis, Diagnosis-Related Groups Statistics, Metadata Report (Version 2, 2019). Retrieved from <https://www.forschungsdatenzentrum.de/en/health/drg>.
- [3] Research Data Centre Destatis, Business-Tax-Panel, Metadata Report (Version 1, 2024). Retrieved from <https://www.forschungsdatenzentrum.de/de/steuern/btp>.
- [4] H. Wickham, G. Grolemund, R for Data Science, O'Reilly Media, Inc. (2023). Retrieved from <https://r4ds.hadley.nz/>.
- [5] M. Gomolka, J. Blaschke, C. Hirsch, Working with large data at the RDSC, Technical Report, Deutsche Bundesbank (2021). Retrieved from <https://www.bundesbank.de/resource/blob/623988/62b8c17881d63bccc61efd11af0b47db/mL/2021-04-large-data-data.pdf>.
- [6] N. Richardson, I. Cook, N. Crane, D. Dunnington, R. François, J. Keane, D. Moldovan-Grünfeld, J. Ooms and Apache Arrow, arrow: Integration to Apache Arrow, R package (2023). Retrieved from <https://CRAN.R-project.org/package=arrow>.

- [7] H. Mühleisen, M. Raasveldt, duckdb: DBI Package for the DuckDB Database Management System, R package (2023). Retrieved from <https://CRAN.R-project.org/package=duckdb>.
- [8] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, dplyr: A Grammar of Data Manipulation, R package (2023). Retrieved from <https://CRAN.R-project.org/package=dplyr>.

Sustainable open-source projects – key success factors

Eric Anvar (OECD)

1. INTRODUCTION

Figure 1. The Statistical Information System Collaboration Community

SIS-CC (“the Community”) is a reference open-source community for official statistics, focusing on product excellence and delivering concrete solutions to common problems through **co-investment and co-innovation**.



We often hear this statement – including in the context of past NTTS conferences: “Our project is open source... The code is posted on GitHub...”. Great, but can this really be called “open source”? “Posting on GitHub” only contributes to a fraction of the “Open Source” value proposition, as there are many more ingredients to deliver “Open Source” assets.

Drawing on more than ten years of building successfully a reference open-source community in the realm of official statistics, the Statistical Information System Collaboration Community (SIS-CC, see [1], [2]), this presentation introduces key success factors for statistical organisations to build successful open-source (and open-knowledge) communities in the future:

1. The recipe: Formulate a (public) strategy
2. KSF 1: A component based architecture
3. KSF 2: Addressing diverse use cases
4. KSF 3: Community-driven dynamics
5. KSF 4: Open-source delivery
6. KSF 5: Stretching the value with open-knowledge and open-innovation

Keywords: #OpenSource #OpenKnowledge #OpenInnovation #CoInvestment #SDMX #Community

2. METHODOLOGY AND PRACTICAL APPLICATION

2.1. The recipe: Formulate a (public) strategy

Formulating a strategy that formulates a clear *raison d'être*, common long term goals and values, governance and funding model – is fundamental for the open-source initiative to project itself in the long term and ensure its sustainability. If absent – as is clearly the case for the majority of open-source initiatives – misunderstanding will arise from latent divergences on goals or expectations.

The SIS-CC 2020-2025 strategy (see [1]) is one example of such a strategy. Importantly, it results from multiple workshops and ongoing conversations over the 2015-2020 period. The 2025-30 strategy is currently being drafted as a result of the ongoing conversation between community members.

2.2. KSF 1: A component-based architecture

A component-based architecture – or a micro-service architecture – is a fundamental feature of an open-source project in the field of official statistics. Indeed, this is a field which has been active developing open standards and where international collaboration is the norm. Component-based architecture (as opposed to monolithic applications) allows for modular developments, empowering teams to operate in relative autonomy vis-à-vis each other thanks to low coupling; important benefits include agility and quality (lesser risk of regression). In the official statistics ecosystem, component-based architecture enables reusability of software (“plug and play”) allowing cooperative business models between organisations and open-source initiatives.

In the context of SS-CC, the collaboration around the SDMX standard (which is inherently “service oriented”, see [15], [17]) is a powerful enabler of such cooperations (especially, the .Stat Suite largely relies on Eurostat’s developed SDMX-RI libraries, see [10]).

2.3. KSF 2: Addressing diverse use cases

For an open-source project to thrive, it should address a diversity of use cases and/or stakeholders. A project serving a narrow functional scope will likely meet only a narrow audience and will struggle to sustain. In the field of official statistics, as the overall data life-cycle and beyond, supply chain, is relatively well documented and homogenous in principle (with many disparities in implementation), positioning the open-source product to meet the largest possible scope of needs could be advisable.

In the context of SIS-CC, this point is met with the “Full data life-cycle coverage” founding principle (see [1]). As the project is very much SDMX-driven, this principle fits naturally with the expansion in usage of the standard (initially focusing on reporting use cases, see [16], now focusing more on dissemination and data integration, especially with the much better coverage of micro-data and geospatial features in SDMX 3.0 see [17]).

2.4. KSF 3: Community-driven dynamics

Open-source does not mean “free”, it only means a different business model, where resources to innovate and maintain/support the product are mobilised in way that is different from the commercial software endeavour. Open-source should entail transparency in the governance of the product (e.g. key design decisions, priorities, etc.) – something unfortunately rarely met in open-source projects in the field of official statistics (e.g. an organisation, or even an individual, makes unilateral decisions without communicating them). A community-driven dynamics should be put in place and documented to address these issues.

In the context of SIS-CC, a multi-tier governance is in place (Tier 1+, Tier 1, Tier 2 and Tier 3) combining membership (contributing financially to the project) with sophisticated partnerships (with other open-source initiatives, as well as with the private sector). Trust through transparency is illustrated with the multiple open information (see [5], [6], [7], [8]) in addition to the traditional open code repository (see [9]) – which is essential but insufficient to make a project “open-source”.

2.5. KSF 4: Open-source delivery

Open-source delivery starts of course by getting the licensing right – and correctly managing the dependencies on libraries or middleware that the solution builds on. But the delivery must meet the highest standards in terms of technology stack (where most advanced libraries and frameworks are open-source) and mastering of DevSecOps techniques – enabling especially securing the code through automated vulnerability checks, and more generally improving product-quality through automated quality assurance (functional, performance, accessibility, integration), and ensuring (multi) cloud-readiness through containerisation.

In the context of SIS-CC, MIT license was adopted in 2019 allowing for commercial reuse of the software. DevSecOps framework and cloud-nativeness of solutions are central pillars of the project. Switching to this philosophy came with a cost – almost a year of work by the core team – but pays off over time with the mastery of CI/CD techniques and container deployment strategies.

2.6. KSF 5: Stretching the value with open-knowledge and open-innovation

Open-source inherently comes with the notion of continuous innovation and knowledge sharing. In essence, it is about enabling open innovation by sharing knowledge with the widest possible range of players, to build public good. This means bringing value “beyond the software” in the shape of knowledge artefacts (starting with good documentation!), community activities (meetups, workshops...).

In the context of SIS-CC, “stretching the value” through open-knowledge and open-innovation is an explicit strategy (see [1]). This has led over the years to a number of new activities and products: trainings (.Stat Academy, see [11], [12], [13], [14]), data modelling and governance guidelines (see [18], [19]), user research (see [4]), and, more recently, applying AI in official statistics (see 3)). Joining forces with other statistical organisations in order to meet the widest possible audience of data scientists to facilitate their usage of our data is another angle being developed (see [20]).

3. MAIN RESULTS AND FINDINGS

Back in 2011, the SIS-CC featured three founding members. We are now 15, plus an extensive network of close partners in the public and private sector. That’s perhaps the crucial currency to succeed: time. It is over time that an open-source project, driven by public sector organisations to achieve common good, can fully deliver its value. The 5 listed success factors can hopefully inspire future open-source endeavours in the field of official statistics so that they can sustain and deliver value over the long period. Given the relative homogeneity in the requirements, products and customers among statistical organisations, it is likely that a lot more could be achieved by making open-source a true strategic direction. “if you want to go fast, go alone; if you want to go far, go together”.

4. REFERENCES

- [1] SIS-CC [Community website](#)
- [2] SIS-CC [2020-25 Strategy](#)
- [3] SIS-CC “SDMX+AI: Unlocking the potential of NLP to enhance data access” – [public report](#)
- [4]
- [5] SIS-CC “Applying user-centred design approach to official statistics” – [public report](#)
- [6]
- [7] .Stat Suite – [High level product overview and flight planner](#)
- [8] .Stat Suite [Kanban Board](#)
- [9] .Stat Suite [Changelog](#)
- [10] .Stat Suite [Extensive documentation](#)
- [11] .Stat Suite [Code repository](#)
- [12] .Stat Core [Open source dependencies](#)
- [13] .Stat Academy – Learning Paths for [Data Producer](#); [Data Tooler](#)
- [14] .Stat Academy [Courses](#)
- [15] .Stat Academy [Webinars](#)
- [16] .Stat Academy [Become an instructor](#)
- [17] SDMX [2021-25 Strategy](#)
- [18] SDMX [The business case for SDMX](#)
- [19] SDMX [SDMX 3.0 packages under review](#)
- [20] SDMX Guideline [Modelling statistical domains and exchange frameworks in SDMX](#)
- [21] SDMX Guideline [A Reference Framework for SDMX Structural Metadata Governance](#)
- [22] [The SDMX.IO ecosystem](#)

The Effect of Working Conditions on Health and Healthcare Use

Mélina Hillion (Insee)

1. INTRODUCTION

Despite significant progress in preventing occupational hazards, physical and psychosocial exposures at work remain widespread, with continuing implications for public health, social inequality and economic performance.

This study aims to provide a more comprehensive understanding of how working conditions affect health by combining survey data on occupational exposures with administrative health records. This integration allows for a detailed analysis of the short-term effects of working conditions on physical and mental health and health care utilisation.

This approach addresses key limitations of previous research, which has mostly relied on self-reported data [1, 2, 3]. The use of both survey and administrative health data allows for a comparison of self-reported health issues with recorded medical events, reducing potential bias and providing a more comprehensive view of the impact of occupational exposures on health. Our findings provide valuable insights to inform workplace policies and promote better occupational health outcomes.

2. METHODOLOGY

2.1. Data sources

The study is based on two main data sources: the 2013 and 2016 French Working Conditions and Psychosocial Risks Surveys (CT-RPS) and the administrative records of the French National Health Insurance (CNAM).

The CT-RPS surveys provide a detailed overview of occupational exposures, collecting information on physical constraints, psychosocial risks and work organisation, the socio-economic context and several self-reported health measures for a panel of around 22,000 people in employment in 2013. Each participant is surveyed in 2013 and 2016, and his or her career path (staying in or changing jobs, becoming unemployed or inactive) is known between 2013 and 2016.

The administrative health data provides monthly information over the period 2012-2017 on medical consultations, hospital stays, prescribed medication, absences due to illness and recognised accidents at work for more than 80% of the people surveyed. These two sources are matched at the individual level. We investigate their complementarity and differences, particularly in terms of reporting and recognition bias.

2.2. Estimation strategy

Our main estimation strategy exploits the longitudinal design of the CT-RPS surveys matched with administrative health data. It consists of examining changes in health status as a function

of changes in occupational exposure between the first (2013) and second (2016) interviews in the Working Conditions and Psychosocial Risks survey.

To address potential endogeneity (where health status might typically influence job conditions) we employ an instrumental variable strategy. These are external factors correlated with workplace exposure but not directly related to individual health outcomes, helping to establish a causal link between working conditions and health effects. As the instruments used are imperfect, residual biases probably remain in the form of attenuating factors.

As an additional robustness test, we include in the analysis individuals who have exited the labor market between the first and the second interview. The objective is to mitigate the "healthy worker effect," a well known selection mechanism describing the fact that healthier individuals are more likely to remain employed [4].

3. RESULTS AND PRACTICAL APPLICATION

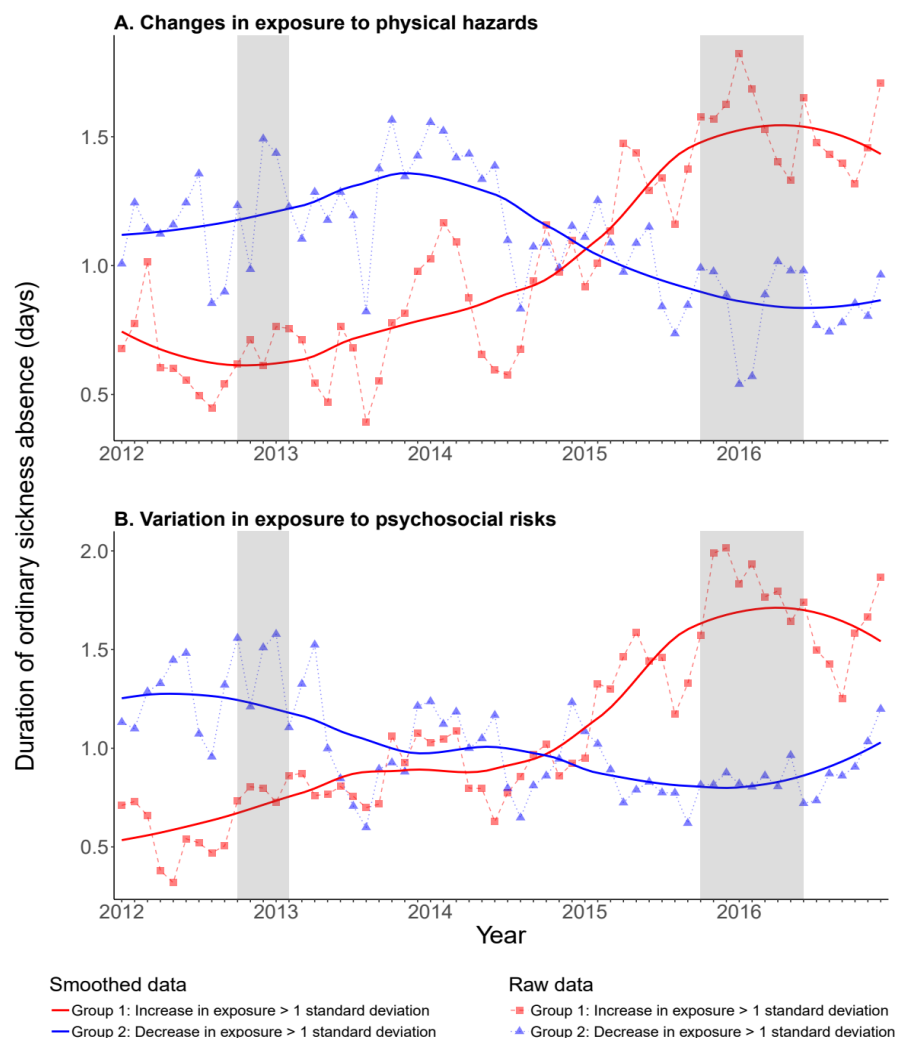


Figure 1. Monthly duration of sickness absence according to changes in occupational exposure among private-sector employees between 2013 and 2016

To illustrate the principle of our estimation strategy, Figure 1 shows the monthly change in the number of days of ordinary sick leave in the private sector as a function of changes in exposure to physical risks (Figure 1.A) and psychosocial risks (Figure 1.B) between the CT 2013 and CT-RPS 2016 surveys. Figure 1.A (resp. Figure 1.B) suggests, based on monthly data, that the number of days of sick leave increases when physical exposures (resp. psychosocial exposures) increase by more than one standard deviation between the two surveys (group 1, in red). Conversely, the number of days of ordinary sick leave decreases when physical exposure (resp. psychosocial exposure) decreases by more than one standard deviation between the two surveys (group 2, in blue).

Our econometric method simply involves generalizing and testing the robustness of this approach. The main results are summarized in the following sections.

3.1. Results

The results indicate that increased exposure to physical and psychosocial risks is significantly associated with a deterioration in health status, including a reduction in psychological well-being, a decline in perceived health, the appearance of functional limitations, and an increased prevalence of chronic illnesses. The study also highlights the significant impact of occupational exposure on medical consultations and the dispensing of medication, as well as on accidents at work, sick leave and presenteeism.

The study also explores the reversibility and cumulative effects of short-term occupational exposures. The results show that occupational exposures, particularly psychosocial exposures, have a partially irreversible short-term effect on the probability of developing a disability or loss of autonomy, the duration of ordinary sick leave and the dispensing of medication to treat musculoskeletal disorders. In addition, poor working conditions have an increasing effect in the short term with the duration of exposure, particularly on activity limitations, ordinary sick leave, and the dispensing of analgesic and psychotropic drugs. The results also suggest a heterogeneous impact of occupational exposure according to sex and age, with women and the over-45s being particularly vulnerable to the effects of physical exposure. These populations have greater activity limitations, are more often off work and use medication more frequently than others when exposed to more difficult working conditions.

The estimates presented in this study may nevertheless underestimate the overall effects of occupational exposure on health. On the one hand, our approach focuses on the short-term impact of working conditions, whereas certain effects, such as cancer and cardiovascular disease, manifest themselves several years after exposure. On the other hand, short-term effects are probably underestimated because of selection bias. In particular, people in better health are more likely to remain in employment and to be exposed to the most difficult working conditions. Our results confirm the existence of this attenuation mechanism and show that the effect of occupational exposures on health is more pronounced in the short term when people who have left employment are included in the analysis.

Despite these limitations, the robustness of the results obtained, regardless of the type of source, health indicator or method used, and the representativeness of our sample, provide a better understanding of the effects of working conditions on health.

3.2. Practical Applications

From a practical point of view, these results underline the need for targeted interventions to improve working conditions. By identifying the specific risk factors that significantly affect health, policy-makers and employers can develop strategies to improve the physical and mental health of employees.

The study highlights the benefits of matching survey data with administrative health data. The use of multiple sources allows results to be cross-validated, providing a more complete and reliable basis for the development of public policies. Administrative data complement surveys by providing new information at lower cost and with unrivalled temporal frequency. The results of this study support initiatives to integrate multi-source data to offer a more accurate picture of societal issues and guide public decision-making.

4. MAIN FINDINGS

This study shows that increased exposure to physical and psychosocial risks at work is significantly associated with poorer health and greater use of health services. Thanks to a unique integration of survey data and administrative health data, the study reveals significant short-term effects of occupational exposures on perceived health, medication prescriptions, sick leave and accidents at work. The results also show that women and people over 45 are particularly vulnerable to these occupational risks. These findings underline the need for targeted interventions to improve working conditions and mitigate their effects on health.

5. REFERENCES

- [1] T. Barnay, *Health, work and working conditions: a review of the European economic literature*, European Journal of Health Economics **17** (2016), 693–709.
- [2] I. Niedhammer, S. Bertrais, and K. Witt, *Psychosocial work exposures and health outcomes: A meta-review of 72 literature reviews with meta-analysis*, Scandinavian Journal of Work, Environment and Health **47** (2021), 489–508.
- [3] E. Cottini and C. Lucifora, *Mental health and working conditions in European countries*, IZA Discussion Papers **(2010)**.
- [4] H. M. Arrighi and I. Hertz-Picciotto, *The evolving concept of the healthy worker survivor effect*, Epidemiology **5** (1994), 189–196.

The ESA "PEOPLE – Ecosystem Accounting" project: experiences from Greece

Ioannis Kokkoris (Department of Sustainable Agriculture, University of Patras),
Bruno Smets (VITO), Lars Hein (Wageningen University) Marcel Buchhorn (VITO),
Stefano Balbi, Villa Ferdinando (Basque Centre for Climate Change), Mathilde De
Vroey, Tim Jacobs (VITO), Bulckaen Alessio (Basque Centre for Climate Change),
Giorgos Mallinis (Laboratory of Photogrammetry and Remote Sensing, PERS Lab,
School of Rural & Surveying Engineering, Aristotle University of Thessaloniki)
Panayotis Dimopoulos (Department of Biology, University of Patras)

1. INTRODUCTION

The European Space Agency (ESA) project "Pioneering Earth Observation Applications for the Environment – Ecosystem Accounting" (PEOPLE-EA) [1] aimed to study and demonstrate the relevance of Earth Observation (EO) for ecosystem accounting in terrestrial and freshwater ecosystems. Within this project tools were developed to use standardised, periodically available EO datasets to develop ecosystem extent, condition and service accounts.

2. METHODOLOGY

Within the project, a series of products have been generated for Greece, integrating national and local scale data, i.e.: (a) the National Ecosystem Type Map of Greece (MAES level 3), recently created for the LIFE-IP 4 NATURA project [2], (b) field survey data for ecosystem condition and ecosystem services from the MAES_GR platform (LIFE-IP 4 NATURA project) [3], and (c) field survey data obtained from the plots of the Habitat Directive monitoring projects. A mapping model has been created that is flexible enough to allow new training data, from ongoing and future observations, to be integrated. From this process, the first ecosystem type extent map (L3) for the region, has been created. Models for ecosystem condition and ecosystem services mapping were also developed to support accounting (e.g. coastal condition, wood provision, nature-based recreation). An evaluation for the thematic accuracy was also conducted from local experts, using relevant spatially explicit datasets for Greece and expert judgment.

3. RESULTS AND PRACTICAL APPLICATION

The main result (product) was the ecosystem type map of Greece (Figure 1), that serves as the baseline for most of the other accounting processes (i.e. for condition and ecosystem services indicators development). The thematic accuracy has been evaluated as of high quality by local experts.

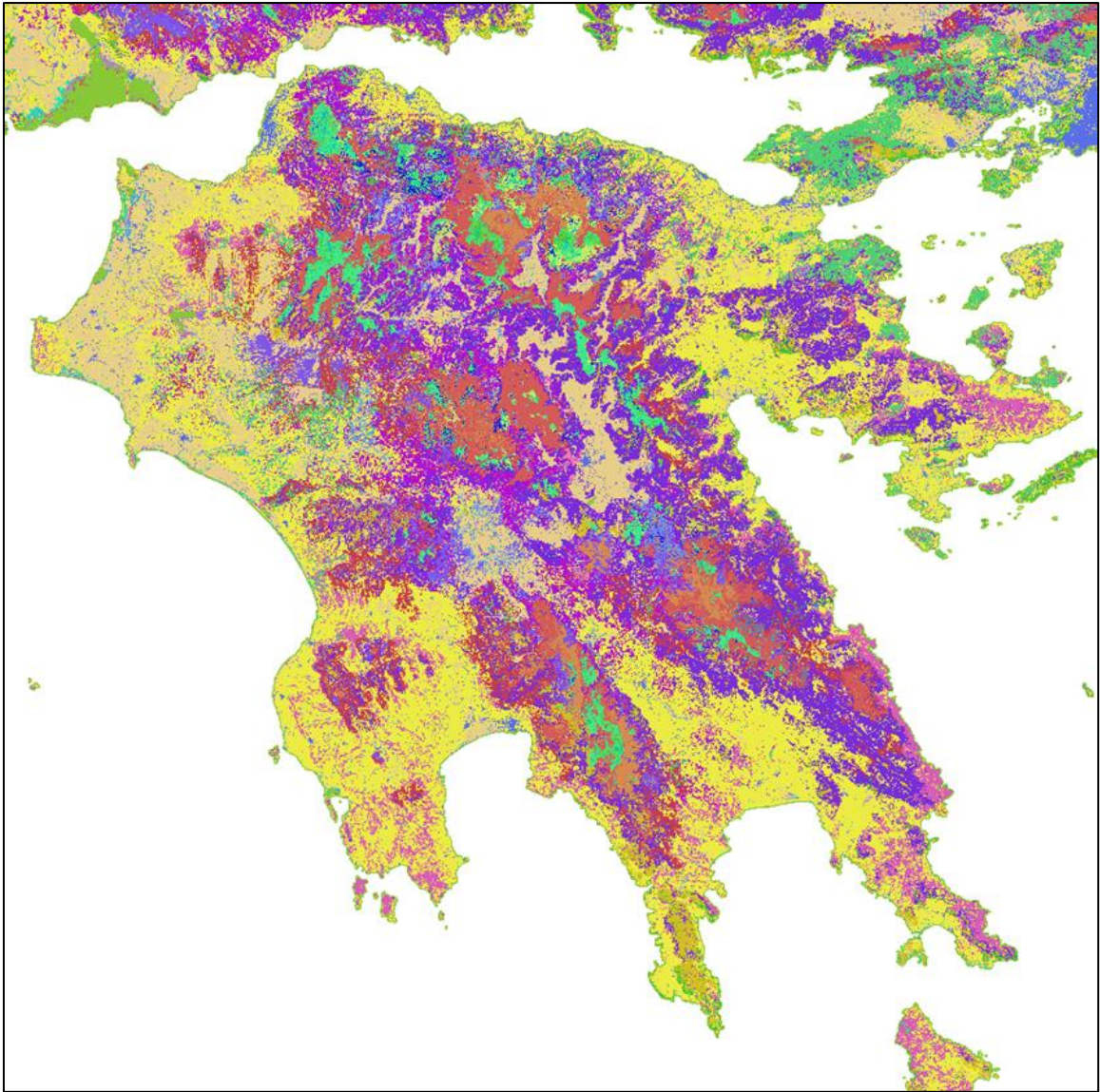


Figure 1. Ecosystem type map (L3) of Peloponnesus, Greece.

The developed ecosystem services condition and services indicators, provide in most cases high quality results, especially for the nature-based recreation potential (Figure 2), with some exceptions due to known issues of the input data (e.g. at the coastal condition map, natural bare rocks and cliffs have been assigned to bad ecosystem condition).

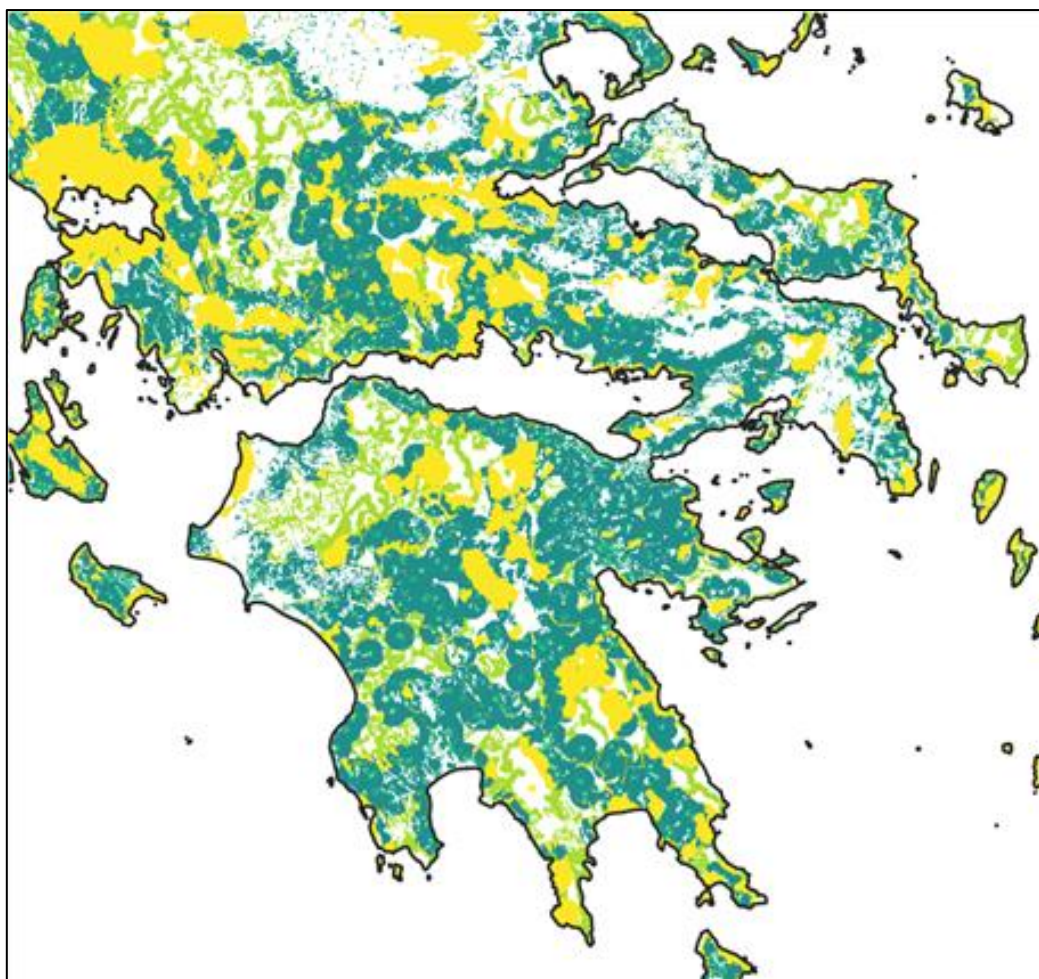


Figure 2. Nature-based recreation potential at Peloponnesus, Greece. Yellow cells represent the highest recreation potential value.

4. MAIN FINDINGS

The Greek case study shows that EO data can support and accelerate ecosystem accounting under the standardised SEEA EA framework providing the most cost-effective way to collect large amounts of data in a standardised form with consistency in space and time. EO contribution was proved to adequately provide the information needed (and missing) for ecosystem type, condition and ecosystem services mapping to represent real world conditions and inform decision and policy making. It is also highlighted that large (local) scale field data are crucial and need to be integrated in the mapping process. However, these data should be also consistently collected with a standardised approach, in order not to bias the accounting process.

5. REFERENCES

- [1] <https://esa-people-ea.org/en>
- [2] Verde, N., Kokkoris, I. P., Georgiadis, C., Kaimaris, D., Dimopoulos, P., Mitsopoulos, I., Mallinis, G. (2020). National scale land cover classification for ecosystem services mapping and assessment, using multitemporal copernicus EO data and google earth engine. *Remote Sensing*, 12(20), 3303.
- [3] Kokkoris, I. P., Kokkinos, V., Michos, E., Kalogeropoulos, R., Charalambides, M., Kounelis, A., Eliadou, E., Damianidis, C.K., Mallinis, G., Bouras, C., Dimopoulos, P. (2021). MAES_GR: A Web-Based, Spatially Enabled Field Survey Platform for the MAES Implementation in Greece. *Land*, 10(4), 381.

The eurostatRTool: a customisable innovative R tool for rapid dissemination of statistical information

Rosa Ruggeri Cannata, Antonio Grosso, Piotr Ronkowski & Anette Sundstroem
(Eurostat)

1. INTRODUCTION

In today's digital landscape, the demand for making statistics more accessible and meaningful to the general public is increasing. This has led to the need for fast and efficient ways to disseminate statistical information, particularly when data visualisations must be produced quickly and updated frequently. In this context, Eurostat launched a project under the ESS Innovation Agenda and developed the eurostatRTool package in R, a tool designed to facilitate statistical data dissemination.

The package provides data producers with a robust and flexible environment for creating customisable HTML dashboards to visualise and analyse economic indicators. This customisation enables data producers to select the indicators to be included and the types of visualisations to be used, tailoring the tool to their specific needs. Customisation options also include the possibility to use a specific logo, insert text in a language different from English, and update the displayed texts using simple input in files.

This visualisation tool combines interactive graphical representation with text, making it possible to convey concise messages to the reader in a storytelling format. Once customised, the EurostatRTool will rapidly generate a set of HTML pages suitable for publication on a website, without additional software requirements. While flexible enough for many applications, this tool is particularly well-suited for displaying frequently updated data, such as monthly monitoring of statistical variables, offering various settings options, including:

- use of specific branding and visual identity
- easy updates of the storyboard with the use of text input files
- use of interactive presentations, such as graphs with interactive charts (time span, selection/deselection of variables, etc.) and table formats
- effortless updates for regular data changes

The tool is freely available on GitHub, with code and wiki documentation, at: <https://github.com/eurostat/eurostatRTool>

2. THE EUROSTATRTOOL MAIN CHARACTERISTICS

The eurostatRTool is based on R Markdown, a powerful package that integrates narrative text, written in plain-text markup language, with embedded R code and outputs, such as charts. This seamlessly generates reports with integrated data, ensuring straightforward updates by simply re-running the R Markdown script. In this way, the production process of a particular visualisation application developed with the eurostatRTool becomes particularly rapid and suitable for content which must be frequently delivered. R Markdown supports multiple output formats, enabling users to produce reports in HTML, PDF, or MS Word format from the same

source document, simplifying the process of web page creation, requiring no prior knowledge of HTML.

The key features and benefits of the eurostatRtool are:

- It provides many alternative customisations. This capability allows any organisation to tailor the outputs to be aligned with its branding and communication strategies, ensuring a consistent visual style.
- It supports several visualisation types, including various charts, maps, and tables. This versatility allows data producers to select the most appropriate graphical representations for their data, catering to specific analysis needs and audience preferences. At the same time, the storyboard layout facilitates an effective storytelling approach to data presentation, structuring data narratives in a coherent flow and making it easier for readers to follow complex analyses and insights.
- It is crafted to streamline the development of interactive visualisations, empowering statistical producers to convey data in an intuitive manner. This feature meets the needs of a broader audience of users, creating an excellent engagement and facilitating a deeper understanding of the statistical information presented.
- Once the visualisation has been settled, updating it with new data will become a very straightforward process. Running the tool needs only basic knowledge of R, making its deployment easy for a broad range of statistical producers. This characteristic becomes especially valuable for projects requiring frequent data refreshes, ensuring that all visualisations are current and accurate without extensive manual intervention.

3. HOW TO SET UP AND CONFIGURE THE TOOL?

3.1. Input your data

Commentary and the storyboard go hand in hand with "frozen" data. It is not possible to comment on live data that may change and not correspond to the data used when writing the related commentary. While this may be seen as a limitation, monthly data rarely change more than once a month, making them an ideal target for the tool.

To create the dashboard, we will rely on different files that should be created and provided as input to the R package for generating the output visualisation. Structural files must be configured once and modified only if changes in the selected indicators occur.

data.csv: This file contains the data which will be displayed. Although it can contain different data and indicators, the structure cannot be changed. This data structure is the one that the user must accurately replicate to build its own data files.

dashboard-structure.xlsx: This file is in relation with the dashboard structure and contains the text snippets to update every time a new visualisation is generated.

dim-labels.csv: This is the master file where the link between indicator codes and labels will be defined. In case they exist, this file will link the levels of the indicator's factor group – e.g., geopolitical entities, gender classes, indicator components – coded in the **data.csv** file with the labels that will be displayed for each group level in the final dashboard.

3.2. Building your own visualisation

After having specified the input files needed for the package, we describe the multiple display modes allowed by the tool.

In the following figures you can see images from an interactive tool generated by the eurostatRtool looking at the ISTAT website. This was done with the only aim of exploring

feasibility aspects and it should not be considered as an official ISTAT output. Those images are of help in understanding the possibilities of the eurostatRtool and we use them to illustrate some display modes.

Timeline Plot: This line plot represents the evolution of the time series for the indicator over time.



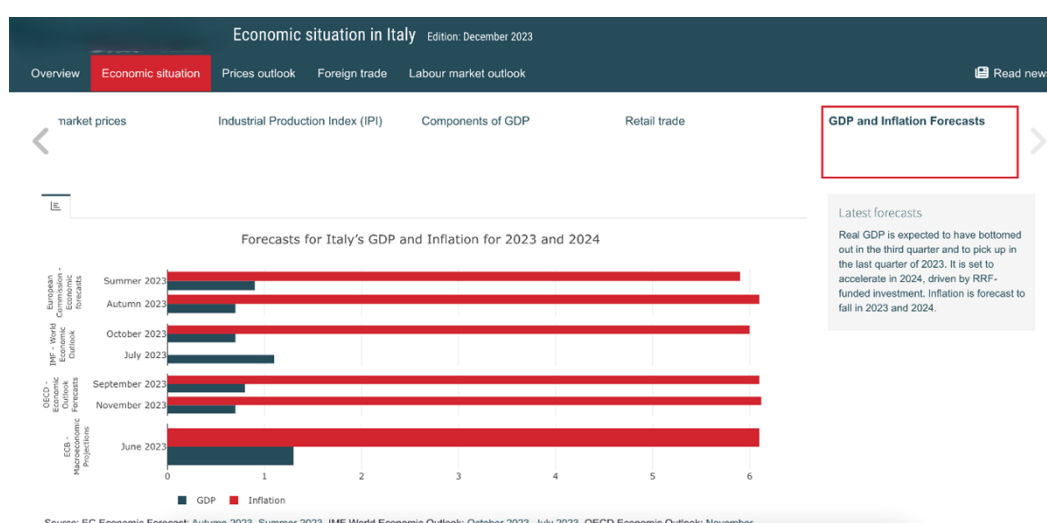
Comp_barchart plot: This bar plot compares the last value of the time series for one indicator against the previous one, considering all the levels in the factor group contained in the column DIM within the data.csv file.



Point_linechart plot: The point_linechart was originally created to display the Euro-coin indicator calculated by the Banca d'Italia as used in [Eurostatistics](#). This display mode can be used with any data specified in the data.csv file that requires to be visualised showing data series both as lines and points.

Forecasts Plot: Although this display mode was originally designed to visualise data on forecasts, the dashboard manager can select this visualisation type to display any indicator that requires a horizontal bar plot as long as the data format requirements are met.

Table 1. Tabular representation of indicators for structured comparison.



3.3. Putting it all together

After having introduced how to handle the input data and the possible graphic visualisation, we will now explore how to customise the tool's layout more in detail. For this purpose, a dedicated scenario has been created with **predefined tabs and headings tailored to the specific needs of National Statistical Institutes (NSIs)**. Although oriented to the needs of a national statistical institute, the same functions could be implemented by whoever wishes to use this freely available tool.

The `update_scenario()` function will enable the data producer to choose the most appropriate layout and selecting the "nsi" function, the overall structure of the dashboard will be modified displaying the specific topics tailored to an NSI. While the menu structure in a primary heading and a secondary one will be fixed (and with a maximum number of tabs), you will be able to customise tabs' heading and all the other elements such as **colours, titles, headers, footers, logos, data, indicators and visualisation modes** are freely customizable.

As you become familiar with the package components, and once the scenario and all the layout customisations are settled, creating your own dashboards will become a **straightforward process**.

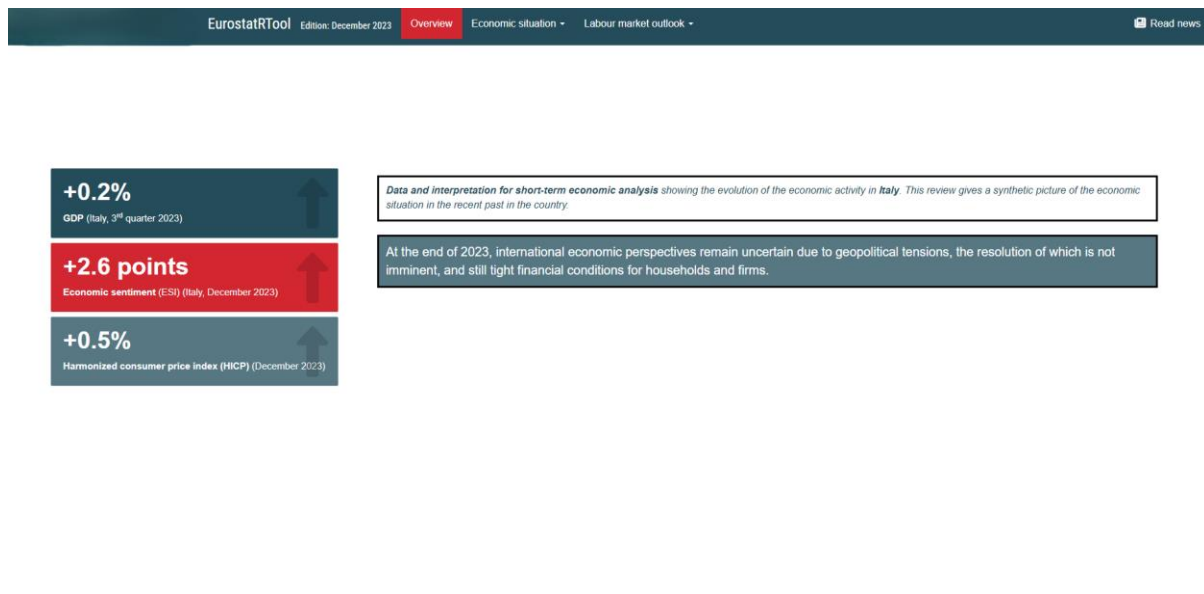
Following the detailed recommendations in the wiki, after having created your own files and uploaded them as input in the package functions, you will finally be able to generate your visualisation according to the data and layout input defined. To generate a new dashboard the main function to use is:

`generate_dashboard` (`output_path`, `pdf_reports`), after setting the input parameters as described in the documentation.

4. CONCLUSIONS

Easy to use and update, the eurostatRTool has been designed to enhance the accessibility and efficiency of statistical dissemination. Its ability to create interactive visualisations rapidly while requiring minimal resources, makes it highly appreciated by those looking to enhance their production of statistical visualisations without investing significant effort in development. Through testing, it has demonstrated its effectiveness in streamlining data visualisation processes, improving user engagement, and supporting flexible customisation options. The ability to integrate various types of visualisations with text ensures that statistical narratives are effectively communicated to a broad audience.

This tool is well-suited for organisations with regular data updates that wish to incorporate a concise storytelling approach into their visualisations with minimal effort. In this regard, the tool's adaptability allows users to modify visual elements, layout structures, and textual content easily, aligning outputs with specific branding and communication strategies. By reducing the complexity associated with data visualisation and dissemination, the eurostatRTool empowers data producers to focus on analysis and interpretation rather than technical implementation. The eurostatRTool serves as a practical and intuitive solution for modern data dissemination needs, offering a structured yet flexible approach to presenting statistical insights in an engaging and user-friendly format.



5. REFERENCES

- [1] Interactive Web Visualisation of Eurostatistics via R: Enhancing the Quality of Data Presentation through Storyboarding, Piotr Ronkowski, Rosa Ruggeri Cannata, Anette Sundstroem, Q2024
- [2] Eurostatistics - from PDF format to interactive web visualisation using R, Piotr Ronkowski, Rosa Ruggeri Cannata, Johannes Buck, NTTS2023
- [3] [Home](#) · [eurostat/eurostatRTool Wiki](#) · [GitHub](#)

The Journey Towards a Governance of Open Source Statistical Tools at Istat

Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Orietta Luzi (Italian National Institute of Statistics, Istat)

1. INTRODUCTION

Modern processes for producing official statistics are highly complex due to the need of integrating different data sources (administrative registers, survey data, etc.). This calls for more efficient and cost-effective statistical IT tools that ensure the implementation of sound methodologies, in line with principles 7 (“Sound Methodologies”¹ and 10 (“cost-effectiveness”)² of the European Statistics Code of Practice. The Italian National Institute of Statistics (Istat) has used and still uses SAS software for various tasks. Anyway, about 20 years ago Istat started to explore Open Source Software Tools (hereafter OSST), and in particular the R environment [1], also in response to a ministerial directive (April 2003) inviting government agencies to adopt OSST and avoid dependence on a single commercial software tool. These efforts have led to the development of a series of tools based, directly or indirectly, on R and distributed to all the members of the Italian National Statistical System and to the general public through Istat’s “Methods and Tools” web repository³. More recently, Istat started to explore Python, a general-purpose language; it is used mainly in the Methodology Directorate, to apply statistical or machine learning methods and to exploit alternative data sources such as big data.

Efforts to develop and disseminate OSST are expected to continue and be strengthened, also in line with the 2019 guidelines issued by the Agency for Digital Italy (AgID)⁴. The EU is also promoting the adoption of OSST in its “Open Source Software Strategy”⁵ (released in 2014 for the period 2014-2017 and subsequently updated in October 2020 for the period 2020-2023). It proposes a “Think Open” approach aimed at promoting the sharing and reuse of OSST. This strategy is an important step towards the goals of the EU’s “Commission’s Digital Strategy”⁶.

A strong push towards the adoption of OSST in the production and dissemination of official statistics is also coming from a number of initiatives launched at the international level. Among others, the expert group on “Shared Tools” set up by Eurostat at the end of 2023 (continuation of the expert group on Open Source for Official Statistics - OS4OS of 2021-22); the UNECE

¹ PRINCIPLE 7 - Sound Methodology: Sound methodology underpins quality statistics. This requires adequate tools, procedures and expertise; in particular, indicators 7.1 (“[...] while constantly striving for innovation”) and 7.7 (“[...] promote better tools when feasible”)

² PRINCIPLE 7 - Sound Methodology: Sound methodology underpins quality statistics. This requires adequate tools, procedures and expertise; in particular, indicators 7.1 (“[...] while constantly striving for innovation”) and 7.7 (“[...] promote better tools when feasible”)

³ <https://www.istat.it/en/classifications-and-tools/methods-and-software-of-the-statistical-process/>

⁴ <https://www.agid.gov.it/it/design-servizi/riuso-open-source/linee-guida-acquisizione-riuso-software-pa>

⁵ https://commission.europa.eu/about-european-commission/departments-and-executive-agencies/digital-services/open-source-software-strategy_en

⁶ https://commission.europa.eu/publications/european-commission-digital-strategy_en

expert group on Open-Source Software launched in early 2024 as part of the ModernStat group (continuation of the Open-Source Adoption Activity proposed by Statistics Canada in 2022); the so called “awesome list of official statistics software”⁷ [2]; the annual conference “Use of R in Official Statistics – uRos” (whose first edition was held in 2013). Finally, it is worth noting that recent ESSnet projects launched by Eurostat require, where appropriate, that new methods be accompanied by OSST implementing them.

In our view, the advantages of statistical OSST outweigh the disadvantages, but this argument alone cannot be used to promote their adoption within a National Statistical Institute (NSI); it is necessary to consider the technical, organizational, and cultural levels. In practice, it is necessary to establish a strategy that should include: (1) procedures for assessment of the potential tools (including code security); (2) guidelines for code development and its documentation; (3) testing procedures; (4) training programs for staff; (5) communication initiatives to share best practices and successful use cases; (6) establishment of internal expertise and support structures; (7) cooperation and partnership with other NSIs or EU statistical offices on key projects/tests or to share knowledge and expertise; (8) engagement with open source communities (encouraging staff to participate in external open source communities, conferences and projects); and, (9) development of corporate policies and governance bodies.

This document presents the work at Istat (Section 2), with a particular focus on the ideas and principles that guide the ongoing activities to define a corporate governance of statistical OSST as presented in Section 3. Section 4 gives some conclusions.

2. ACHIEVEMENTS IN ADOPTION OF OPEN SOURCE STATISTICAL TOOLS AT ISTAT

In the past Istat’s statistical work was carried out with SAS software, which is still a popular tool in statistical production units. The introduction of R took place in the methodology directorate shortly after the year 2000, mainly to test new methods. An informal group of R experts was created and used to meet once a month; today, this group is being formalized and expanded to include experts working also in production Directorates. Regular training in R started in 2007; currently there are two “core” courses (base R and intermediate R) offered twice a year, and a number of short courses on specific topics/packages (e.g. package **ggplot2**) offered once a year. In addition, R tools are used in statistical courses (sampling, data integration, data processing and imputation, etc.) to demonstrate the application of methods. R was chosen as the language to create new packages or packages to replace old tools developed in SAS or as stand-alone applications. The first R packages were released around 2010 (**ReGenesees** for weights calibration and estimation, **SamplingStrata** for stratification and optimum allocation, **SeleMix** for selective editing, **StatMatch** for statistical matching), others followed later (e.g. **FS4** for stratification, **R2BEAT** for determining optimal sample size and its allocation). In other cases, such as the **RELAIS** record linkage system, R is used as the engine for statistical computation. These packages were developed in the Methodological Directorate and are currently used in production units. They are disseminated externally on our corporate website (see footnote 1) and also through other repositories (“awesome official statistics software” list and CRAN in the “Official Statistics” task view⁸); this work places Istat among the major contributors of R packages for official statistics. The tools are promoted through presentations and tutorials given at conferences/workshops, including the uRos conference.

⁷ <https://github.com/SNStatComp/awesome-official-statistics-software>

⁸ <https://cloud.r-project.org/web/views/OfficialStatistics.html>

Taking advantage of the open source philosophy, Istat is also carrying out tests on packages developed in other NSIs: e.g. the R packages for data editing and imputation (**validate**, **validatetools**, **errorlocate**, **simputation** and **VIM**), which are expected to replace our obsolete stand-alone tool CONCORDJava. Similar tests are being performed for disclosure control (R packages **ptable** e **cellKey**). Finally, we are working on extending the functionalities of the **RJDemetra** package, which provides the R interface to JDemetra+, the officially recommended seasonal adjustment software within the European Statistical System.

At Istat we are also using Python, but mainly for research objectives in the methodological department, although it is already adopted to produce some experimental statistics: the SMEI (Social Mood on Economy Index) and TERRA (imporT ExpoRt netwoRk Analysis), a tool to carry out exploratory analysis of Eurostat data on international trade. Other ongoing projects are quite diverse: web mining to integrate and validate information from the Statistical Business Register; estimation of road accidents with big data; estimation of urban greenery via remote sensing images; imputation of education level in the Register of Persons; estimation of shipping routes, etc. A few years ago, it was decided to organize Python courses (basic and advanced) for staff, offered once a year.

3. TOWARDS A CORPORATE GOVERNANCE OF THE OPEN SOURCE STATISTICAL TOOLS AT ISTAT

Section 2 illustrates that we have already achieved some of the objectives of our strategy to promote the adoption of OSST for statistical purposes; what is still at an early stage is the establishment of internal expertise and support structures and the development of corporate policies and governance bodies.

As mentioned in section 2, the informal network of R experts will be enlarged to include methodologists and subject matter experts with a solid knowledge of R, and will become the official support structure for R from 2025 onwards; focusing more on the production of official statistics than on IT infrastructure. This structure will also try to expand the training offered to staff and promote a series of workshops.

A first part of the Istat policy for the governance of statistical OSST is expected to be released by the end of 2024-early 2025. It will mainly deal with statistical-methodological aspects and provide some basic indications on the IT infrastructure, as the two elements are closely linked. The second part, more focused on IT infrastructure and IT requirements, will be released subsequently.

The first part of the governance of statistical OSST will provide a set of guidelines and recommended practices to Istat researchers for (a) the development of new OSST or (b) the adoption (and possible adaptation) of existing OSST and (c) their endorsement, dissemination and maintenance over time.

The guidelines for the development of new statistical OSST will be released at the end of 2024 and include recommendations for writing R code and documenting it according to international standards to facilitate code sharing and re-use. There will also be guidance on some IT tools, e.g. RStudio as reference Integrated Development Environment (IDE; with its tools like Quarto for report preparation) and Git for version control. The adoption and possible adaptation of statistical OSST developed externally should follow a standard procedure from acquisition to testing; the procedure will include: criteria for selection among different potential candidates; preliminary checks to be done (availability and clarity of the documentation; maturity, frequency of update and eventual bugs corrected; limitations and dependencies on other tools; available support, etc.); testing procedures of the selected tools in case studies with increasing levels of complexity. Finally, for both developed and acquired/adapted existing tools, the governance policy will provide guidance on their endorsement, dissemination and promotion

internally and externally. Governance will also give recommendations for user support, maintenance and updating of endorsed tools over time. All defined procedures should be complemented by the identification and establishment of a set of governance bodies with clearly defined roles and tasks.

4. CONCLUSIONS

The complexity of current statistical processes requires statistical software tools that implement innovative methodologies. Statistical OSST have proven to be particularly suitable for this purpose in a cost-effective way. As shown in Section 1, there is an increasing push towards the adoption of OSST at both national and European/international levels, mainly with the aim of promoting standardization, sharing and re-use. Although the advantages outweigh the disadvantages, their implementation in the production of official statistics requires significant efforts at the technical, organizational and cultural levels. A clear strategy should be defined, setting up governance mechanisms and bodies, as well as a support structure. In defining the strategy, we believe that maintenance and support over time is a critical factor if the tools are to be used on a regular basis for the production of official statistics, and therefore these aspects need to be carefully addressed by defining clear guidelines, including the related tasks to be assigned to the support structure.

5. REFERENCES

- [1] R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- [2] Olav ten Bosch. (2022, March 10). "The awesome list of official statistics software & FOSS best practices". Zenodo. <https://doi.org/10.5281/zenodo.7665189>

The path of Earth Observation to official statistics

Carla Martins & Hannes Reuter (Eurostat)

The following abstract collects reflections on the use of Earth Observation (EO) data as an innovation motor for new techniques and technologies in statistics. It describes EO within the ESS's roadmap, first results and further steps into the uptake of EO for Statistics.

1. INTRODUCTION

As part of the big data family, EO data belongs to the ocean of data with the power to attract and/or overwhelm statisticians. The National Statistical Offices and Eurostat have taken the challenge to push for the uptake of EO data for Statistics (EO4S) within the European Statistical System (ESS). The EO4S roadmap is laid out, and some of the first findings are also presented. The future direction of this project is aligned with capacity building, knowledge-sharing and co-creation.

2. THE PATH OF EO DATA TO OFFICIAL STATISTICS

2.1. The seeds- The Warsaw Memorandum

Amid the data revolution¹ statisticians, are faced with a myriad of big data, a need to continue delivering statistical data (including new data requests on new phenomena and data), and high burden of processing administrative registers and surveys.

The Warsaw Memorandum² (WM) was signed by the Presidents and Directors General of the National Statistical Institutes (DGINS) in their 2021 conference on Earth Observation (EO) for official statistics. The Document advocates the gradual incorporation of EO into the Official Statistics. Further steps were outlined in the 48th Meeting of the European Statistical System Committee (ESSC 2022/48/01/EN).

2.2. The plan for EO4S

The EO4S action plan identified a variety of objectives covering the topics Applications, Skills, Methodology, IT, and Data strategy and Governance. As part of this abstract, we outlined the main activities and results until now.

¹ <https://www.undatarevolution.org/>

² <https://dgins2021.stat.gov.pl/warsaw-memorandum>

We performed an inventory of EO4S projects in the ESS, by launching questionnaires, examining existing ESS Net³ and GEOS⁴ grants for EO relevant projects and interviewed ESTAT's Units within the Sectoral and Regional Statistics Directorate⁵. One essential element which became evident was the importance of unbiased statistics^{6 7}. Based on these inventories a draft design of the process from EO to Statistics was created as seen in Figure 1 and Table 1.

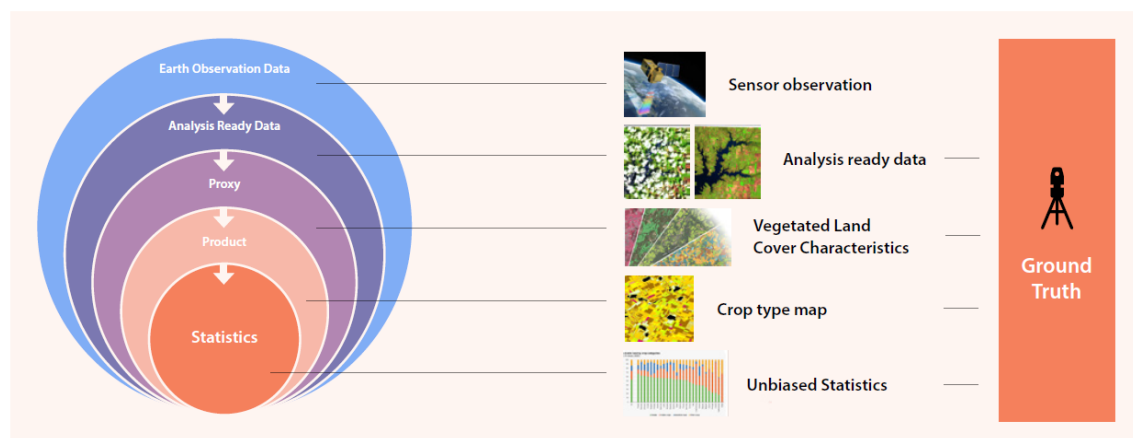


Figure 1. A simplified process diagram which shows the five main steps of deriving a statistical product from raw Earth Observation data.

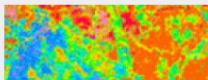
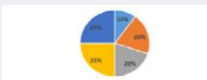
STEP	1	2	3	4
INPUT	Raw data (Satellite image)	Analysis Ready Data	Proxy	Derived product
OUTPUT	Analysis Ready Data	Proxy	Derived product	Statistics
What happens?	Clouds removed Orthorectification	Interprets Analysis Ready Data into a proxy	Conversion between a proxy and a product	Calculation of statistics
Who does it?	EO expert	EO expert Geospatial expert	EO expert Geospatial expert	Geospatial expert Statistician
Examples	Sentinel-2 Mosaic (Apr - Jun 2024) ¹	Phenology stage – End of Season date ²	Crop type classification ³	Statistics ⁴
Visualization				
Colour of the image represents	Red, green and blue (RGB) ratio of satellite image	Calendar dates when leaf discolouration occurs	Different types of crops	% of different types of crops

Table 1. Main steps of deriving a statistical product from raw EO data.

³ https://wayback.archive-it.org/12090/20231227201426/https://cros-legacy.ec.europa.eu/content/essnet-big-data-0_en

⁴ EU action grants in the field of Geospatial, environmental, agricultural, and other sectoral statistics under the European Regional Development Fund and Cohesion Fund. Reports can be found on [CIRCABC](#).

⁵ The Units covered: E1. Agriculture and Fisheries; E2. Environmental Statistics and accounts, sustainable development; E3. Transport; E4. Regional Statistics and Geographical information; and E5. Energy.

⁶ In Eurostat, 2021, Earth Observation for Statistics – methodology and training, Report on process and findings, EFTAS Fernerkundung Technologietransfer GmbH.

⁷ <https://doi.org/10.1016/j.ecoser.2024.101659>

The outer circle represents the raw data directly obtained from the sensor of the satellite and stored in the Ground Station or data distribution centre. The Analysis Ready Data (ARD) is the result of several post processing steps (e.g. removing the clouds, orthorectifying the satellite data, atmospheric correction) and combining them into a space-time data cube ready for further analysis. In the next step the data is transformed into a derived proxy that can use auxiliary data or be an interpretation of the original data (e.g., High Resolution Phenology Layer⁸). The proxy will be further processed into a derived product, which consists of data closer to the statistical demands – such as crop type classes, or land cover classes. The final stage is the statistical table that is obtained not only by counting the pixels in the previous data set, but by an unbiased statistics methodology which measures its accuracy where the value represented in the derived product is compared with the reality on the field, in terms of its geographical measurement and its classification, and where the differences are calculated. In other words: the process is not limited to counting pixels, but the output is a spatial statistical product, entailing a defined statistical methodology. This will guarantee evidence-based results, which can qualify the data as official statistics.

The entire process has distinct roles for EO experts (steps 1,2), geospatial knowledge processor (steps 3, 4) and statisticians (step 5). Essential in the entire process is that the different stakeholders clearly communicate their requirements for their data products and feasibilities. It became already evident that the different stakeholder groups have quite different terminologies. This process has been exemplary tested for crop types and grassland to evaluate the accuracy and draw up a list of necessary requirements.

2.3. Capacity Building

Cooperation and capacity building within the ESS is one of the activities of the EO4S:

- The GEOS grants that allow NSIs and other national authorities to innovate and use EO data in their statistical production⁹.
- The Task Force on Earth Observation that brings together colleagues from NSI's, National Mapping and Cadastral Agencies, and other entities working on EO4S projects.
- Knowledge sharing and capacity building through Eurostat's Collaboration in Research and Methodology (CROS) platform¹⁰.
- Training (ESTP) programme with basic and advanced Earth Observation and Geospatial Analysis courses.
- The AI/ML4OS¹¹ ESS Net project (Work Package 7 on EO).
- The research from the Joint Research Centre (including the Knowledge Centre for EO¹²).
- The Copernicus Data Space Ecosystem (CDSE) platform.
- The EO products supplied by other entities (e.g. DG DEFIS with their Copernicus Entrusted Entities¹³).

⁸ <https://land.copernicus.eu/en/products/vegetation>

⁹ https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Integration_of_statistical_and_geospatial_information

¹⁰ [CROS](https://cros.ec.europa.eu/group/earth-observation-statistics-eo4s/about) - Earth Observation for Statistics (EO4S). <https://cros.ec.europa.eu/group/earth-observation-statistics-eo4s/about>

¹¹ <https://cros.ec.europa.eu/dashboard/aiml4os>

¹² https://knowledge4policy.ec.europa.eu/earthobservation_en

¹³ European Centre for Medium-Range Weather Forecasts (ECMWF), Mercator Ocean, European Environmental Agency (EEA), European Maritime Safety Agency (EMSA), European Border and Coast Guard Agency (Frontex), European Union Satellite Centre (SatCen), European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), European Space Agency (ESA).

3. RESULTS AND PRACTICAL APPLICATION

Eurostat's EO4S incubation projects helped to achieve a better understanding of the need for intermediate data. It became evident that this will be, in the end, essential to calculate the statistical accuracy and usefulness of the EO data for the statistical production process.

For example, with current intermediate data products (crop type maps obtained with EO data – e.g. step 4 above) only pixel counting was possible. The intermediate products that led to the final crop type map (used for classification and masking) were not available. Moreover, probability matrices¹⁴ were missing, not allowing the bias calculation, therefore, only obtaining biased statistics. The EU crop type map produced an EU-wide confusion matrix¹⁵ which was not possible to reuse at other more detailed scales (national or regional).

To account for the effect of bias, statistical results were calculated for limited regions that were processed fully from EO to statistics, where the statistical error was possible to measure with the data from the probability and confusion matrices.

The EO4S projects ensured an initial IT infrastructure within the Copernicus Data Space Ecosystem (CDSE) available to all NSIs with special privileged access. The processing needs for an EU wide operation have been identified.

4. MAIN FINDINGS

Preliminary **lessons learnt**:

- The process from EO to statistics needs to include a built-in design of the metadata and quality assessment. The information on the dataflow and the different steps and algorithms used from one level of information to the next is essential for calculating the unbiased statistics that define the accuracy and value of the final statistical data. To obtain these intermediate steps is quite challenging and requires establishment of the necessary requirements.
- Different stakeholders in the processing pipeline use various terminologies and have different understandings. These definitions need to be aligned for a successful execution of the statistical production process.
- Stakeholders identify the need for a processing at an EU wide operation with local adaptations.

¹⁴ Example for producing a land use map: the probabilities of each pixel (of the EO image) of matching each of the land use classes. The highest probability is used to define the class.

¹⁵ These matrices compare the actual classification (obtained through ground truth data) to the predicted classification (obtained by the algorithm that analyses the EO data).

The role of official statistics in the fast-evolving data ecosystem (focus on European Data Strategy)

Verena Manthe, Judith Czepek (Statistics Germany, Destatis), Barteld Braaksma and Maarten Alders (Statistics Netherlands), Jaana Tael & Liina Osila (Statistics Estonia)

1. INTRODUCTION

Official statistics is facing several opportunities and challenges these days due to a very dynamic data ecosystem. Main drivers are the amount of new (big) data sources, IT developments and the implementation of the European Data Strategy, that will reshape the data ecosystem fundamentally. With this strategy, the European Commission intends to establish a single market for data and thereby ensure Europe's global competitiveness and data sovereignty. Besides the legal foundations created by the Data Act and Data Governance Act (DGA), Interoperable Europe Act and Artificial Intelligence (AI) Act, the technical framework conditions for making data sharing easier are currently created in the form of so-called Common European Data Spaces (CEDS).

Due to legal assignments, official statistics has to constantly improve its statistics and is therefore looking beyond traditional ways of developing, producing and disseminating official statistics e.g. to improve timeliness and reduce the burden on data providers. These new developments have led to intense debates within the European Statistical System (ESS) on the role of official statistics and National Statistical Institutes (NSI) in the changing data ecosystem.

This abstract examines present developments, their impact on official statistics and discusses the future of official statistics in the fast-evolving ecosystem. By practical examples, Statistics Estonia, Statistics Netherlands and Statistics Germany show how successful participation in this major European project can be achieved.

2. CURRENT DEVELOPMENTS IN THE DATA ECOSYSTEM AND THE CHANCES AND CHALLENGES FOR OFFICIAL STATISTICS

For official statistics it is extremely important to monitor and consider new developments and user requirements that may have an impact on future statistical production and products. Those developments mainly occur outside official statistics and are often implemented even without considering official statistics as affected user or stakeholder.

At present, we are observing that a lot of new non-statistical regulations become effective, which also have an impact on the role of NSIs and the production of official statistics. Another observation is the increasing number of new data governance models and technological advancements. Furthermore, the quantity of available data sources is constantly growing e.g. due to the European Data Strategy. As part of the strategy, 14 CEDS are set up in Europe in various areas such as agriculture, energy, green deal, health, mobility, tourism and a lot more to be expected. But not only on European level but also on national and international level the amount of new data initiatives is growing day by day. All CEDS shall be interoperable, so that data sharing will be even easier.

Since the implementation of the European Data Strategy, the DGA and the CEDS has only just begun, there are still plenty of open aspects and unanswered questions, which will be discussed in the following.

2.1. Legislation and ethics

As part of the European Data Strategy several legal acts such as Data Act, DGA and further CEDS specific legal acts, e.g. for European Health Data Spaces (EHDS) have already come into force or are about to be adopted. Although official statistics is primarily based on the Regulation (EC) No. 223/2009 of the European Parliament and of the Council on European Statistics and special statistical legal acts, the legal acts part of the European Data Strategy may also have an impact on official statistics. For example, official statistics is explicitly mentioned in the EHDS Act for the secondary use (reuse) of health data. Furthermore, several countries have appointed their NSI as a competent body or a single contact point that aims to provide support for other public sector bodies to facilitate data sharing according to the DGA Article 7 and 8. Next to the new legal situation, it becomes clear that with the increasing importance of data, we must consider ethical aspects more carefully. In the case of health data this is obvious, but it is much wider. A cornerstone of the EU's AI policy is that all AI used in Europe should be human-centric. This principle is reflected in the AI Act. A new field of Data Ethics is quickly gaining importance and we should reflect on what that means for official statistics.

2.2. Technology and methodology

Beside legal aspects the technical dimension needs to be considered. Besides the development of several digital infrastructures e.g. for CEDS, standardisation is very important for data sharing as well as the compatibility of data formats. Especially for official statistics certain formats have to be available.

Modern statistics often rely on a variety of heterogeneous sources, for example a combination of survey data, administrative records collected by other government bodies and big data sources (e.g. scanner data, satellite data or mobile phone data) may be used. This heterogeneity calls for technical solutions and transmission standards to access the data, as well as knowledge of data formats and metadata. Interoperability is a key notion here, not only at technical, but also at semantic level. Methods to map different sources to harmonise statistical concepts and combine them in a meaningful way need to be developed. Furthermore, the use of AI needs to be elaborated and considered. The ability to integrate data is a unique asset of official statistics, which not many organisations have. In an increasingly complicated world, where we must deal with multi-faceted societal challenges, this ability becomes ever more important.

2.3. Quality

Official statistics prides itself in the high-quality standards it maintains and in a time of alternative facts and fake news this quality focus becomes even more important. European quality standards and quality control procedures are already firmly established but will have to be reconsidered given the increasingly dynamic data situation. New sources come and go, or their nature and contents may change. Continuity at input level is not always guaranteed, while continuity at output level is an absolute must. Users expect more timely statistics, which should be feasible with new high-frequency source data, but this puts pressure on quality procedures.

2.4. Collaboration

Data spaces are an environment where stakeholders with many different backgrounds, interests and capabilities meet. They may come from the private and public sector or academia - even citizens and organised interest groups might participate. Participants in a data space may assume different roles like data user, supplier, processor or steward. Identity and access

management, including authorisation procedures are therefore a central element. Participation for a NSI thus means collaboration with a broad array of other parties. It will become essential to learn how to operate in such a setting while maintaining our unique position and safeguarding our interests. Furthermore, several institutions such as Data Spaces Support Centre or the Joint Research Centre of the European Commission offer consultancy and advice. NSIs should also explore what is happening in this area, to learn and to contribute their insights.

2.5. Financing

The financial aspects of the implementation of DGA and CEDS are also relevant. For example, depending on the decision of the data holder, data can be reused against remuneration or for free. Therefore, CEDS are market oriented to a certain degree. The European Commission sometimes advocates CEDS as a data marketplace to support the private sector, in particular small and medium-sized enterprises, while in other cases the focus is rather on public-private partnerships. Several data spaces are discussing possible ways to finance CEDS and some initiatives have decided to introduce a pricing model to avoid exaggerated prices for data. A relevant question for official statistics is whether it has a sufficient budget to participate in CEDS and what financial terms and conditions would be acceptable.

3. RETHINKING THE ROLES OF OFFICIAL STATISTICS IN THE DATA ECOSYSTEM

Considering the presented developments, opportunities and challenges, official statistics needs to rethink its position and role within this fast-evolving data ecosystem. At the same time, official statistics needs to strengthen its traditional role as an important high-quality data provider for decision-makers in politics and to assert itself in this new environment. Therefore, NSIs need to consider their role from a governance perspective, in the sense of the DGA, and in particular their role as a data authority, data steward and data provider. New opportunities may arise as data user as well, or in contributing to the development of new standards in the data ecosystem. The traditional core values, knowledge and expertise of official statistics provide a sound basis for assuming such roles, but it also takes courage, assertiveness and leadership to act.

Hereafter, some positive examples of NSI participating in the new data ecosystem will be presented:

Statistics Estonia participates in the European Data Innovation Board (EDIB) and shares information with the interested European NSIs. The EDIB is a new body established under the DGA and has a key role to stimulate further evolution of the data ecosystem, for example by promoting interoperability and standards. Statistics Estonia coordinates the national public sector data governance since 2019, e.g. guidance on data quality and governance with a view of becoming one of the competent bodies (Art. 7, DGA). Statistics Estonia is strategically focused on acting as a national, high-quality, metadata driven national data space, enabling data linking and sharing across domains.

Statistics Netherlands has been asked to act as a competent body in the sense of the DGA. Statistics Netherlands has replied positively and is currently preparing the implementation, which includes modification of the applicable Dutch legislation as well as technical and organisational measures. Regarding the EHDS, Statistics Netherlands is joining a preparation program carried out by the Ministry of Health, Welfare and Sport. Together with other partners the technical functionalities for a yet to be designated coordinating Health Data Access Body will be developed. Statistics Netherlands is also involved in other data spaces, like the mobility and green deal data space. Also, on national level data ecosystems are evolving. An example is the participation of Statistics Netherlands in the Federated Data System, an initiative from the Dutch Inter-Governmental Data Strategy that aims to stimulate

and facilitate data exchange and (re)use within Dutch government. A second example is the implementation of 'View of the Netherlands', a multiyear vision of the Dutch Council for Geo-information. A key element of that vision is the development of a data space for the physical environment.

In Germany, the national law for implementing the DGA has passed first legislative steps and is due to come into force at the beginning of 2025. The Act mandates Statistics Germany (Destatis) as the competent body for all sectors (Art. 7) and as the central single information point (Art. 8). The implementation of a service centre for data transparency and the set-up of an overview of all available data resources with information on the conditions for the reuse of protected data in public institutions are in preparation.

Furthermore, Destatis is monitoring the international developments on establishing CEDS with a focus on mobility.

4. OUTLOOK

We need to consider the role of official statistics in the fast-evolving data ecosystem, including possible implications for NSIs and their operations. There are several actions required in order to be prepared for future developments and to avoid losing connection to the data landscape, users and stakeholders. Therefore, now is the right time to shape the future of official statistics and to explore possible new roles in the data ecosystem to remain relevant. We cannot do that alone: we have to collaborate within the ESS and beyond and above all, we need leadership to do it.

5. REFERENCES

- [1] European Commission, Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions - A European strategy for data (2020), <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066>
- [2] Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) (2020), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52020PC0767>
- [3] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities, <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009R0223>

The use of NLP and Machine Learning in official statistics

Ibtissame Sahir Sahir & Mouad Ismaili Alaoui (GOPA Luxembourg)

1. INTRODUCTION

Every year, more than 50,000 jobseekers and beneficiaries, along with 10,000 employers, receive support through an extensive portfolio of services and financial benefits. To ensure efficient use of its limited resources, operations must be precise, delivering targeted services that address the specific needs of both jobseekers and employers. Currently, the database contains approximately 250,000 job offers from the period 2015 - 2023, with an estimated 1% to 10% of these being duplicates.

Proactive measures have been implemented to improve the data quality of job offers. First, job offers are automatically imported from a list of private job portals to enhance coverage. Second, algorithms are developed for automated analysis of the texts of offers to identify the required skills. However, each improvement introduces new challenges that must be addressed.

The main issues related to job posting data are:

- Duplicates in declared job offers primarily result from offers coming from external portals after they had already been submitted by the employers themselves.
- Errors in identifying skills and occupations in job descriptions, especially due to “noise sentences” in the text that are not related to job requirements (e.g., company description or terms of employment).

Thus, this project aims to optimize the management of employment data through two innovative modules:

- **Noise Detector:** This module filters out irrelevant sentences in job descriptions, such as company descriptions, that are unrelated to the job's skill or occupation requirements.
- **Twin Detector:** This module detects and eliminates duplicate job offers from different sources.

These modules leverage advanced natural language processing (NLP) and machine learning methodologies to meet client's specific needs and integrate seamlessly into its operational environment.

2. METHODOLOGY

To address issues with duplicate job offers and irrelevant content in job descriptions, a comprehensive strategy has been developed, focusing on two main modules. These modules are designed to enhance the quality and accuracy of job offer data, ensuring that jobseekers have access to unique and relevant job postings. The following sections provide a detailed overview of each module and their respective functionalities.

The strategy centers on two main modules:

- **Twin Detector:** This tool uses NLP techniques to identify and remove duplicate job offers by evaluating criteria such as employer, location, and job description, ensuring each job offer is unique.
- **Noise Detector:** This module applies NLP techniques to filter out irrelevant sentences from job descriptions that do not meet actual job requirements.

2.1. Implementation Steps

Data Annotation:

- **Noise Detector:** Approximately 6,500 sentences from job descriptions in 300 job ads in French, English, and German were manually annotated. Sentences were pre-segmented according to adjusted rules and classified as:
 - Relevant (noise=0): Directly related to the job or necessary skills.
 - Non-relevant (noise=1): Contains irrelevant information like company descriptions or general conditions.
- **Twin Detector:** Annotated at least 400 pairs of job offers, focusing on edge cases to improve duplicate detection.

Quality Assurance:

- A portion of the annotated data underwent double annotation to ensure accuracy and consistency.

Feature Engineering:

- Used methods like TF-IDF, GloVe, Word2Vec, and BERT to process and analyze textual data effectively.

Model Training and Testing:

- Developed and refined predictive models, including Random Forest, SVM, and Gradient Boosting.
- Models were tested using separate datasets not utilized during the training phase to ensure they generalize well and perform accurately on new, unseen data.

Development of Functions:

- Developed a function that takes a job description and returns a version stripped of irrelevant sentences.
- Developed a function that takes a job offer ID and returns potential duplicate identifiers.

Operational Flow:

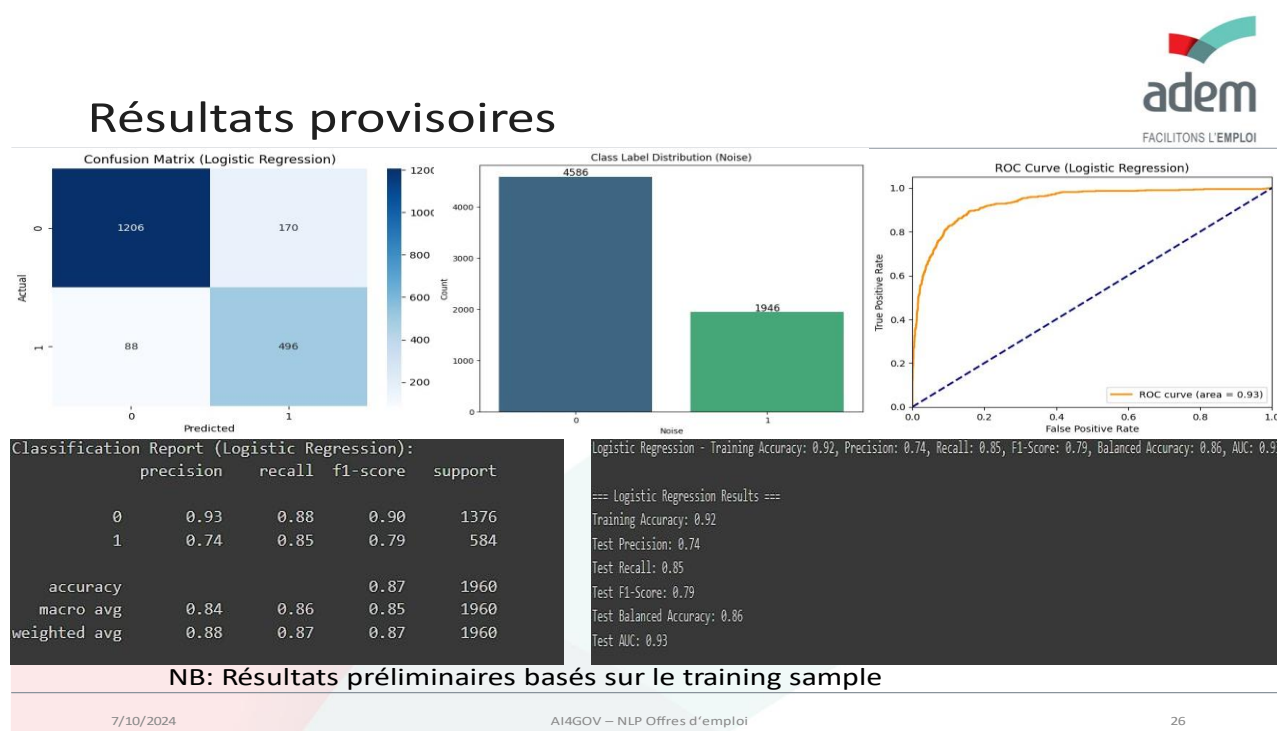
- **Twin Detector:** Quickly checks new submissions for duplicates, removing any that are found to keep the database clean and efficient.
- **Noise Detector:** Cleans up validated job offers by taking out any unrelated content, which helps in correctly identifying job roles and necessary skills.

This approach makes sure that all job offers are thoroughly checked and meet specific job requirements, improving the quality of data management and making employment services more efficient. The use of the Noise Detector and Twin Detector modules provides a clear and effective way to manage job offers.

3. PRELIMINARY RESULTS AND PRACTICAL APPLICATION

a. Noise Detector

The Noise Detector module has demonstrated promising results in filtering out non-essential information from job descriptions. Preliminary results indicate an accuracy of 88% using BERT embeddings and a Voting Classifier. This module significantly enhances the extraction of job-specific requirements and improves the classification of occupations.



b. Twin Detector

The Twin Detector module has made significant progress in identifying duplicate job offers, which helps reduce redundancies in the job offers database. It calculates similarity using job titles, company permits, and descriptions, and flags potential duplicates for manual review.

Preliminary Results:

- **Data Collection:** An initial dataset was collected, including job offers imported in XML format and compared against active offers.

- **Incorporation of Explicit Rules:** Explicit rules were implemented for duplicate detection, such as flagging offers with the same employer and job title as potential duplicates.
- **Annotation:** A dataset is in progress of annotation with at least 400 interesting pairs of offers, focusing on edge cases.
- **Similarity Calculation:** The overall similarity was calculated as a weighted average of scores obtained for job titles, company permits, and descriptions.
- **Integration of Additional Variables:** Additional variables such as contract type, experience level, location, and working hours were proposed to refine detection and reduce false positives.

Practical Application:

- **Workflow Integration:** The modules were designed to operate independently but sequentially, ensuring that job offers are unique and cleaned before further processing. This integration improves data quality and operational efficiency of job offers.
- **Scoring and Manual Review:** The similarity scores are used to flag potential duplicates for manual review, ensuring that only unique and relevant job offers are processed.

Plan for Production Integration:

- **Dashboard Integration:** Develop a dashboard to directly transmit results to the employer service team after processing XML files. This feature is essential for streamlined operations.
- **Output Format Validation:** The enriched XML format, including fields for duplicate identifiers and similarity scores, has been validated for production use.

These steps and agreements support a structured transition aligned with operational needs, ensuring the modules' effectiveness and relevance within Database's systems.

4. BROADER IMPACTS

The broader impacts of these technological advancements extend across various dimensions, contributing to overall improvements in data handling and resource management:

- **Data Accuracy:** Reduced duplications result in cleaner, more reliable data for labor market analysis and informed decision-making.
- **Efficiency:** Automating the detection of duplicates and irrelevant content decreases manual workload, allowing staff to concentrate on higher-value tasks.
- **Modularity:** The flexible design of these modules facilitates adaptation across various datasets and operational contexts.
- **Innovation in Statistics:** These modules demonstrate how NLP and machine learning can revolutionize data management in official statistics, paving the way for future advancements.

This project highlights the transformative potential of NLP and machine learning techniques in improving the quality and efficiency of data processing workflows within official statistics, setting a precedent for future advancements in this field.

The Uses of Administrative Data in the Danish Survey on Income and Living Conditions

Daniel Gustafsson (Statistics Denmark)

1. INTRODUCTION

The European Survey on Income and Living Conditions (EU-SILC) is a comprehensive household survey designed to measure various aspects of everyday life across the European Union. Within this framework, the Danish EU-SILC (DK-SILC) stands out as a prominent user of administrative data. The integration of administrative data into DK-SILC is extensive, encompassing a wide array of topics such as income, labor market conditions, health, housing, and demographics. The applications of this data are diverse, ranging from sampling and calibration to validation.

The availability of administrative data is increasing. Both in countries with established practices and those with limited experience in utilizing such resources. The advantages of incorporating administrative data into statistical production are significant, offering benefits in terms of cost efficiency and enhanced data quality. Administrative data often provide superior coverage compared to purely sample-based surveys, and as response rates continue to decline, the resulting non-response biases introduce further uncertainty into survey outcomes.

However, the use of administrative data is not without its challenges. Issues primarily arise from the definitions employed in these data sources and their timeliness. Since administrative data are generated for specific administrative purposes, their scope can be limited. Moreover, changes in legislation and administrative practices can affect data consistency. Typically, these data are available only after the conclusion of an event or registration period, followed by necessary statistical processing. Whether these limitations impact the usability of the data depends largely on the specific application and subject matter.

This paper aims to outline the structure of registers at Statistics Denmark and explore the various applications of these data within DK-SILC. The different use cases will be categorized according to their roles in the survey production cycle, followed by a discussion on the overall benefits and challenges of utilizing administrative data.

2. METHODOLOGY

A pivotal development in the use of administrative registers in Denmark was the establishment of a central personal registration system in the early 1970s. This system harmonized public administration processes across various bodies by adopting a common personal identifier, which Statistics Denmark also implemented. This common identifier has facilitated the linking of diverse data sources, including those related to individuals, enterprises, and properties.

The early adoption of these practices has allowed numerous official statistical products to be built on administrative data. Consequently, the design of these statistical products often reflects the inherent definitions and availability of the administratively defined data, resulting in a "data first" approach. In contrast, many global statistical practices follow a "needs first" methodology, where user requirements shape the statistical concepts and outputs. While the

"needs first" approach is essential for producing internationally coherent statistics, flexibility in defining concepts is necessary to avoid overlooking the potential of administrative registers due to their limitations.

3. RESULTS AND PRACTICAL APPLICATION

At Statistics Denmark, the production of statistical registers is managed by specialized employees who ensure that these registers are accessible to others for integration into various statistical analyses. These registers consist of versioned microdata for each reference period and are structured within a centralized system that establishes common rules, documentation, and access protocols.

3.1. Sampling frame

One of the primary uses of administrative data in DK-SILC is in the creation of the population and sampling frame. Daily administrative records provide updates on the population, including migration and internal movement, which are critical for maintaining an accurate sampling frame. By utilizing year-end population registers, DK-SILC can generate a precise overview of the population from which samples are drawn, ensuring robust and representative data collection.

3.2. Stratification

Stratification of the sampling frame is another crucial application. Auxiliary data reflecting household income over the past year is utilized for this purpose, focusing on salaries, pensions, and government transfers while excluding income from self-employment and realized capital gains. This targeted stratification enhances sample allocation, allowing for the oversampling of lower-income households. Such an approach is vital for producing more robust results, particularly regarding poverty-related indicators.

3.3. Indirect use

Register data also plays a significant role in pre-filling observations during interviews. This background information can be utilized in two ways: as "hidden use," where respondents are not confronted with specific register values, and as "visible use," where respondents are directly asked about the validity of the information. In the case of DK-SILC, the hidden use of register data helps to avoid raising respondent suspicion, though it can result in missing data if the information is unavailable at the reference date. For example, last known economic activity codes are retrieved from registers, which reduces the need for respondents to verify these details but may lead to gaps in data collection. The visible use of register data in DK-SILC extends to the household composition, where a respondent is presented with everyone living at their address, and asked whether or not they belong to the household.

3.4. Direct use

Direct use of register data is also prevalent in DK-SILC, particularly in instances where statistical definitions align with those in the registers. For example, information such as the year of construction of a dwelling or the number of visits to a doctor can be used without modification. However, when discrepancies arise between statistical definitions and those in administrative records, DK-SILC either employs a model-based approach or directly questions respondents about the differences. Notably, while administrative records capture much of the necessary data for calculating household income, they do not account for inter-household transfers, necessitating self-reported data for those components. Approximately 45% of DK-SILC variables are filled using register data, which significantly reduces the burden on both respondents and data collectors.

3.5. Calibration of weights

Finally, the calibration of weights is becoming increasingly important in light of declining response rates. DK-SILC relies heavily on register data to correct for non-random non-response after data collection. By leveraging the fact that SILC income variables are derived from the entire population, the survey can use various income metrics, such as equivalised disposable household income, mass and poverty status, to calibrate weights. This approach enhances comparability between sample results and the overall population, leading to more reliable poverty indicators.

4. MAIN FINDINGS

The integration of administrative data into the DK-SILC offers substantial advantages, including enhanced data quality, reduced respondent burden, and improved coverage of demographic and economic variables. By leveraging these data sources, DK-SILC can provide a more comprehensive picture of income and living conditions in Denmark, particularly as traditional survey response rates decline.

However, several challenges persist. The reliance on administrative data requires careful consideration of their definitions, potential biases, and the timing of data availability. Moreover, the need for flexibility in defining statistical concepts is crucial to avoid overlooking the insights offered by administrative registers. Addressing these issues necessitates ongoing collaboration between statisticians and administrative bodies to ensure that the data used for surveys accurately reflect the aspects they aim to measure.

Ultimately, the success of integrating administrative data into household surveys like DK-SILC depends on balancing the strengths of these data sources with their inherent limitations, while maintaining a focus on producing reliable, user-centric statistical outputs. As the landscape of data collection continues to evolve, further research and dialogue will be essential in refining these methodologies and maximizing the potential of administrative data in statistical practice.

The Visual Display of Skewed Data: Boxplot, Beeswarm and Beyond

Michael Neutze (Federal Statistical Office of Germany)

1. INTRODUCTION

Skewed data is interesting. Imagine a census resulted in all NUTS-3 regions having the same median population age. Statistical offices would be the first to suffer from budget cuts. Luckily for statisticians, the distribution of most variables is skewed one way or the other and for centuries visualising statistical distributions has been key for understanding what that means. This is especially true for people who are not fluent in the vocabulary of statistical properties. For some time, the boxplot has been the standard graph for displaying quantiles, the span of data and in some variants indicating outliers as well. Driven by Edward Tufte's data-ink-ratio mantra, there have been attempts to give the boxplot an edgier or minimalist design [1] but these attempts failed twofold:

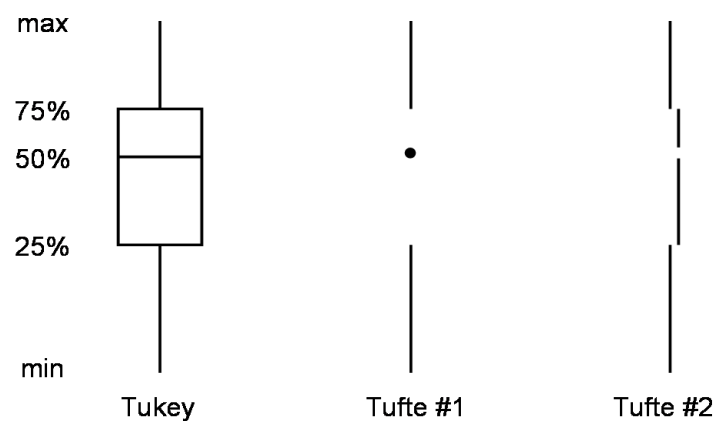


Figure 1. From Tukey's to Tufte's boxplot

For one, they totally kept the level of abstraction of the original, thus pleasing only the already statistically literate [2]. Secondly, they further reduced the poor visual attraction of the boxplot.

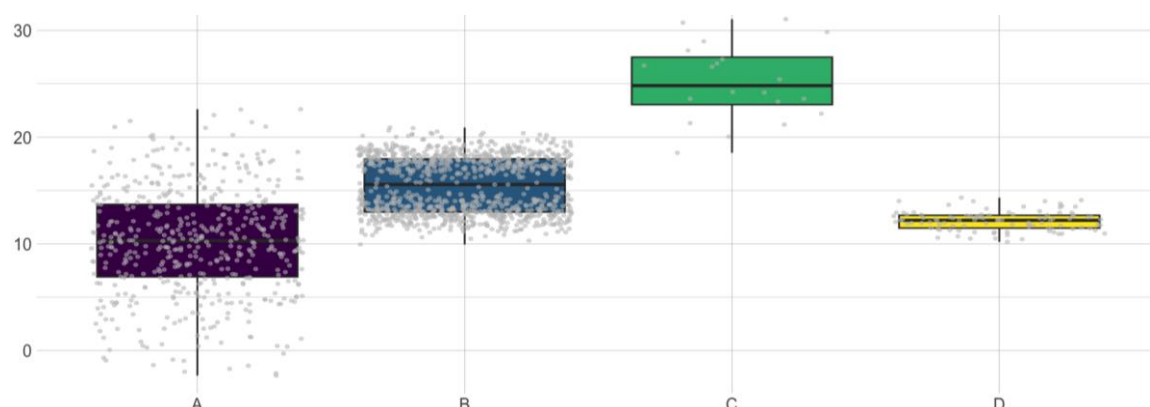


Figure 2. Boxplot with jitter

Overlaying a sample or even all of the data points as jitter on top of a boxplot might improve the understanding [3]. But what then is the explaining factor of the boxplot below the jitter?

2. METHODOLOGY

Nothing explains statistics clearer than a 1:1 mapping of the observed. Within a family or a kindergarten group, we might sort ourselves by height and take a photo of that distribution. Accordingly, in clinical studies each observation may be plotted as seen below [4].

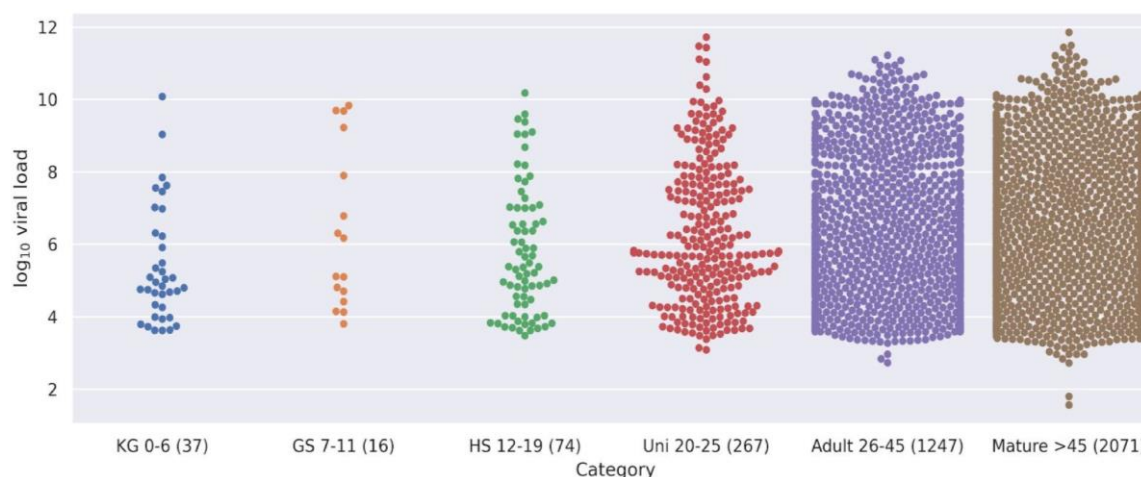


Figure 3. Plotting a dot for every observation

This so called *beeswarm plot* is showing a one-dimensional distribution that locates symbols for each data point in such a manner that they don't overlap. In the example above, every dot has the same meaning (symbolising a person). With this approach your chart can get cluttered pretty quickly, if you want to display large samples, e.g. thousands of observations.

What is the main takeaway here, if this obviously does not scale to the number of inhabitants of even smaller cities? Well, you do not have to plot a single dot for every observation. Clustering your observations is key and applying this to regional statistics can very well offer deeper insights than a choropleth map might do.

3. RESULTS AND PRACTICAL APPLICATION

The average rent of a LAU-region is – among other things – probably a function of its economic prospects, the availability of high paying jobs or the accessibility to those jobs via convenient means of transportation and of the resulting demographic growth. Thus, we expect a highly uneven distribution in most (capitalist) countries for this variable and would not be surprised to find outliers.

In fact, for the German Census 2022 results, we saw significant outliers with strong regional clusters differentiating East from West. How do we know? We plotted a *force directed beeswarm plot*. Instead of using equally sized dots, proportional symbols were used which were colour coded by geographic region. Let us have a look at the final plot before we dissect it:

What People Pay for Renting their Homes in Germany

Municipalities above 5,000 Inhabitants in 2022

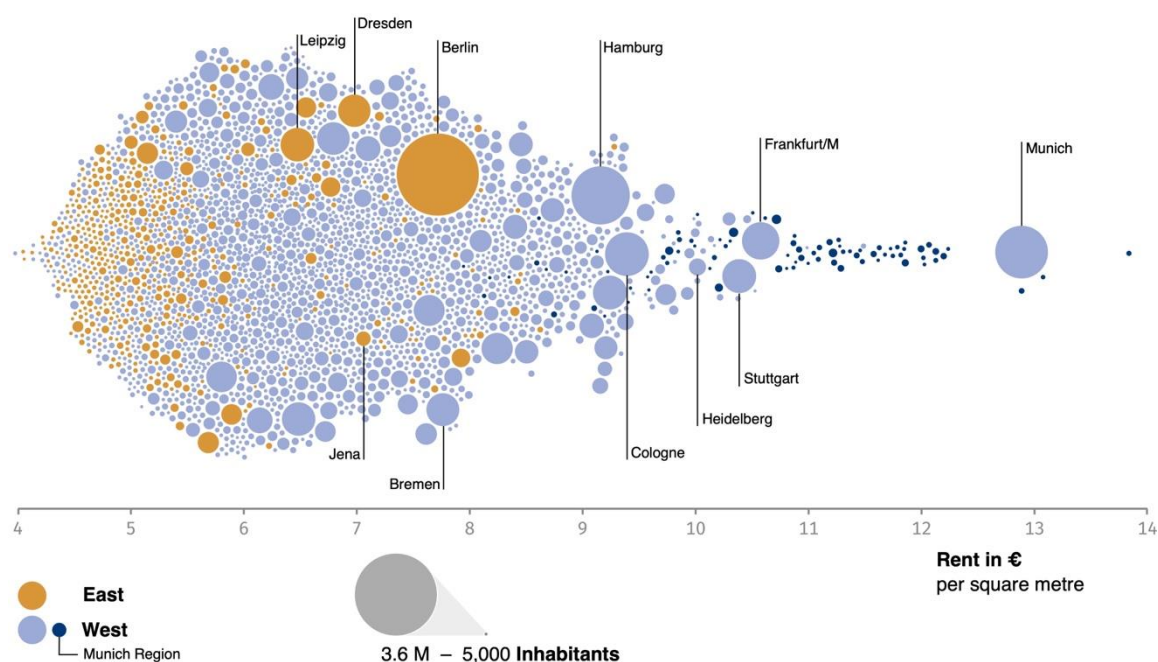


Figure 4. Force-directed beeswarm plot

The x-position of each circle represents the one and only scaling axis, namely rent in Euro per square metre. Circle size is proportional to the number of inhabitants, giving larger populations more visual weight. The y-position of each circle is a non-scaling axis, meaning it does not represent any statistical value. Its sole purpose is for the circles to not overlap and create better readability.

To avoid overlap, circles are spread away from each other by a simulation of physical forces like magnets (or the literal bees) would. Apart from this collide force, position forces push the circles towards a desired position along the x- and y-axis with a configurable strength. In our example the strength was mainly applied along the x-axis so that the circles accurately reflect their position on the rent-scale. Applying forces along the y-axis would result in a more cohesive or streamlined overall shape but at the cost of precision in terms of rent.

Using the JavaScript library *d3.js* has greatly helped creating this chart [5], as each aspect of manipulating the symbols can be individually controlled. The force simulation works as an iterative process, slowly moving circles outwards until an optimal solution is achieved. This can be computationally expensive in the browser depending on the number of symbols. Used as an animation, the iterative process can further help understanding the build-up of the visualisation.

In print media [6] labelling larger circles or outliers with callout-lines is necessary while an interactive version is able to reveal more data on mouseover or touch events. As always, a clearly labelled key is key.

While a thematic map could reveal the most expensive regions faster and locate them accurately in geographic terms, by no means would a map be able to show how far apart the outliers are in terms of rent. And the latter is what many people will probably find interesting.

4. MAIN FINDINGS

Visualisations build on existing familiarity with their concepts. Imagine growing up with bar charts only – your first pie chart will be difficult to understand. Therefore, introducing a new diagram type to your visual vocabulary needs careful consideration.

But building on basic concepts like 1:1 mappings – one circle equals one municipality, together with bigger circles representing bigger numbers, this plot should already be accessible to people with a limited visual vocabulary.

Interestingly, user testing among manager level colleagues resulted in some hesitation. Highly statistically literate people who were not familiar with such plots tried to search for some y-axis labels as they were assuming that positioning along *each* axis must encode some meaning.

We have to assume that first-time readers could struggle with the non-scaling axis: Having learned that visualisations depict data in two dimensions, the main hurdle in understanding is distinguishing the scaled axis with statistical meaning from the non-scaled axis for visual clarity. This is a problem of overthinking that eventually will take care of itself.

The appealing form should not be underestimated. Especially the combination with colour coding an additional variable (in our case the greater geographical region) sparks interest as users want to find out what an exceptionally large or surprisingly coloured bubble represents.

5. REFERENCES

- [1] E. Tufte: The Visual Display of Quantitative Information (1998), 123-125.
- [2] N. Desbarat: I've Stopped Using Box Plots. Should You?, Journal of the Data Visualization Society (2021), <https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/> (received Sept 24, 2024).
- [3] Y. Holtz: *The Boxplot and its pitfalls*, From Data to Viz <https://www.data-to-viz.com/caveat/boxplot.html> (received Sept 24, 2024).
- [4] T. C. Jones, B. Mühlemann, T. Veith, M. Zuchowski, J. Hofmann, A. Stein, A. Edelmann, V. M. Corman, C. Drosten: *An analysis of SARS-CoV-2 viral load by patient age* (2020), https://internationale-gesundheit.charite.de/fileadmin/user_upload/microsites/m_cc11/virologie-ccm/dateien_upload/Weitere_Dateien/analysis-of-SARS-CoV-2-viral-load-by-patient-age.pdf (received Sept 24, 2024).

- [5] H. Stevens: *Force-directed Beeswarm*, Observable Coding Platform (2022), <https://observablehq.com/@harrystevens/force-directed-beeswarm> (received Sept 24, 2024).
- [6] Federal Statistical Office of Germany: *Census 2022 Results*, Press conference hand-out (2024, page 13), https://www.zensus2022.de/DE/Presse/Pressebereich/Zensus2022_PK_Statement.pdf (received Dec 4, 2024).

Telling Stories with Data: Enhancing Museum Accessibility for Blind Visitors Through User Feedback

Lucia De Maio & Andrea D'Andrea (Università degli Studi di Napoli "L'Orientale")

1. INTRODUCTION

Accessibility in museums is a crucial issue that requires special attention to ensure that all people, including those who are blind or partially sighted, can enjoy the cultural and educational resources available. Physical and sensory barriers can severely limit their participation in exhibition understanding. This study aims to explore the experiences and expectations of blind visitors in relation to their enjoyment of museums. A pre-visit questionnaire was used to identify the main barriers and specific needs of users, providing a basis for targeted improvement of the museum experience. Feedback was collected on past experiences and future expectations, identifying the main barriers and practical suggestions for improving museum accessibility.

2. METHODOLOGY

The methodological approach consisted of two main phases:

1. Pre-visit phase: An online questionnaire was developed to collect information about blind visitors' previous experiences in museums, their expectations and wishes regarding accessibility. The survey consisted of 16 questions, of which 13 were closed and 3 open, divided into 3 sections: general information, museum experiences and expectations about museum experiences.

The sample analysed showed a predominance of female participants, with a prevalent age between 46 and 65 years and a medium to high level of education. Most of the participants were living with total blindness, a condition that has been present since birth. Questions covered topics such as the quality of explanations provided by guides, staff training, availability of tactile experiences, use of audio guides and access to multisensory experiences.

The questionnaire was distributed through blind associations and by word of mouth. The data collected was analysed using descriptive statistical methods to identify trends and patterns.

2. A visit to the Museo Scerrato in Naples: on the basis of a preliminary analysis of the results, it was decided that a meaningful experience would be to give the users the opportunity to touch the objects in the Museo Scerrato in Naples and to check their accessibility, pleasure and enjoyment by means of a subsequent questionnaire. The focus was therefore on the selection of original artefacts with highly representative elements, objects with very pronounced characteristics, in order to develop sensitivity, emotion and also imagination.

For blind people, the senses of touch and hearing became essential tools for accessing cultural heritage. The tactile experience is a direct gateway to many works of art and

historical objects. The ability to touch works, to perceive their shape, texture and size, provides a concrete and meaningful experience¹. Of course, not all works can be touched directly, but there are innovative solutions such as the creation of tactile reproductions or three-dimensional models that allow the hands to explore what the eyes cannot see.

In parallel, the use of detailed and narrative audio descriptions is another key tool. Traditional audio guides are not enough: it is necessary to develop specific content that not only describes what is on display, but does so in an engaging way, leading the listener to imagine and understand the essence of the work or object on display. This requires specific storytelling skills and an in-depth knowledge of the blind audience.

3. Post-visit phase: After designing a museum experience specifically adapted for blind visitors, a post-visit questionnaire was administered to evaluate the effectiveness of the measures implemented. This phase included questions on satisfaction, perceived inclusivity and recommendations for future improvements. In addition, qualitative interviews were conducted with some of the participants to explore their impressions and suggestions and to gain a more nuanced understanding of their experiences.

3. RESULTS AND PRACTICAL APPLICATIONS

In the part of the pre-visit questionnaire about user preferences for museum experiences, with particular reference to elements of accessibility and usability for people with visual impairments, the sample was free to give as many answers as they wished.

The results of the pre-visit questionnaire revealed specific preferences and needs for improving accessibility. 75% of respondents said that the most appreciated aspect of a museum visit was the explanation by a guide, highlighting the importance of direct and human interaction. In addition, 69% showed a clear preference for audio guides, confirming that the audio component is fundamental to the full enjoyment of the exhibition content.

66% of respondents highlighted the importance of tactile experiences, calling for greater access to works and exhibits that can be touched. This underlines the need for more structured tactile routes to ensure greater inclusivity. It is also important to note that 48% of respondents expressed an interest in multi-sensory experiences that are accessible to all, while 45% indicated that the ability to move freely within museum spaces is crucial to an accessible and comfortable experience.

These are the main aspects identified by the respondents in the pre-visit questionnaire. In the post-visit questionnaires, the same categories were used to assess the impact of the improvement measures implemented in the participating museums, with the aim of comparing the initial perception with the effectiveness of the proposed solutions. The results were encouraging and in line with what had already been highlighted in the pre-visit questionnaire. For example, when creating the tactile routes, great importance was attached to the choice of objects to be touched, giving the opportunity to experience different materials as well as artefacts from different periods and areas.

4. BARRIERS AND EXPECTATIONS IDENTIFIED

Several participants identified significant barriers to full and equal enjoyment of museum exhibitions. Among these, crowding was identified as a major barrier to mobility and orientation within the museum. Other relevant issues include the lack of tactile aids and the inadequate quality of audio guides and audio descriptions.

¹ Grassini, A. "Per un'estetica della tattilità" (2019)

Expectations expressed by participants included tactile routes, detailed audio guides, descriptions in Braille and specific training for staff to meet the needs of blind visitors. There was a strong desire for empathetic and well-trained staff who could relate to people with visual impairments and make the museum experience enriching and enjoyable.

5. MAIN FINDINGS

The analysis shows how the inclusion of tactile experiences, the use of high quality audio guides and the presence of trained staff can significantly transform the museum experience for blind people. These findings suggest that cultural heritage institutions need to be committed to adopting innovative and participatory solutions to improve accessibility and inclusion, based on the real needs expressed by stakeholders.

6. CONCLUSIONS AND IMPLICATIONS

In conclusion, the study highlights the importance of listening to the voices of blind visitors and responding to their needs in a proactive and structured way. Museums should continue to develop accessible and multi-sensory pathways that encourage the active participation of the visually impaired public. The recommendations made by the participants in this study can serve as a guide for future accessibility initiatives in museums and cultural institutions. It is vital to promote staff training and development to ensure that all visitors can enjoy an engaging and accessible experience.

7. REFERENCES

[1] Grassini, A. "Per un'estetica della tattilità" (2019)

Timeliness criticality and its aggregated economic impact

Frank Pijpers (Statistics Netherlands), Jose Moran (Ecole Polytechnique),
Matthijs Romeijnders (Utrecht University), Pierre Le Doussal (Ecole Normale
Supérieure), Utz Weitzel (Vrije Universiteit Amsterdam), Debrabata Panja (Utrecht
University) & Jean-Philippe Bouchaud (Académie des Sciences)

1. INTRODUCTION

This paper concerns the functionality of any schedule-based system, where timing is of crucial importance, particularly operations in socio-technical systems (STSs), or production in economic ones. The *timely* delivery of goods or services so that they are in the right place at the right time is evidently of value in and of itself. In effect, delays in delivery cause a decline in the value of such goods or services. Any systems or processes involving chains of intermediate use (logistics, emergency response, healthcare, etc.) have interdependent events and breaks at specific time intervals, and are referred to as complex schedule-based systems.

In complex systems, external parameters often determine the phase in which the system operates, that is, its macroscopic behaviour. For nearly a century, statistical physics has been used to extensively study systems' transitions across phases, (universal) critical exponents and related dynamical properties. A recent stylized model [1] of delay propagation on temporal networks is presented, where the magnitude of the delay-mitigating buffer acts as a control parameter. The model exhibits timeliness criticality, a novel form of critical behaviour. We characterize fluctuations near criticality, commonly referred to as avalanches, and identify the corresponding critical exponents. The model exhibits timeliness criticality also when run on real-world temporal systems such as production networks.

2. METHODOLOGY

In its simplest form a schedule-based system takes place with a regular fixed timestep on a network of nodes that are participating in the production of a commodity or a service. At any such node and at every time step the process can incur a delay. It is assumed that there is some modest buffer of time designed into the system so that if the delay is smaller than this, the delay does not affect subsequent steps. The size (in time) of this buffer is a parameter in the model.

It can be shown that a mathematical description of the distribution function of delays in a mean field approximation allows for the formulation of a differential equation for that distribution function. For the solution of this equation it can be shown that there is a finite value for that buffer at which critical phenomena take place. If the buffer size is chosen smaller than this critical value, it can be shown that delays at system level can increase without bound.

Even on less idealised networks, where no analytical solution of that differential equation is available, it can be shown through direct simulation that similar criticality occurs.

3. RESULTS AND PRACTICAL APPLICATION

Direct simulations are carried out, for various values of the number of nodes N and the number of edges per node K to verify the mean field result, as well as for realistic synthetic networks to assess the critical behaviour in situations that are closer to real-world scenarios. In the mean field approximation the network adjacency matrix is reset at every time step, which is not the case for the more realistic scenarios.

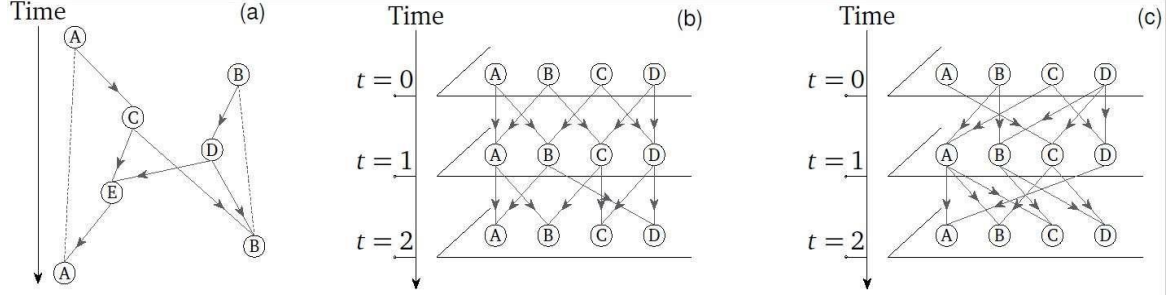


Figure 1. Schematic of our model in terms of temporal networks.

a. How the (time-ordered) dependencies (shown by arrows) among the system components (number of components $N = 5$) in a real-world situation may play out. **b.** mean field (MF) and **c.** synthetic temporal network (STN) representations of our stylized model (number of components $N = 4$) in discretized time steps. For both MF and STN cases, the number of arrows K entering every system component is constant (for example, $K = 2$); for the latter case, the number of arrows leaving every system component is also K .

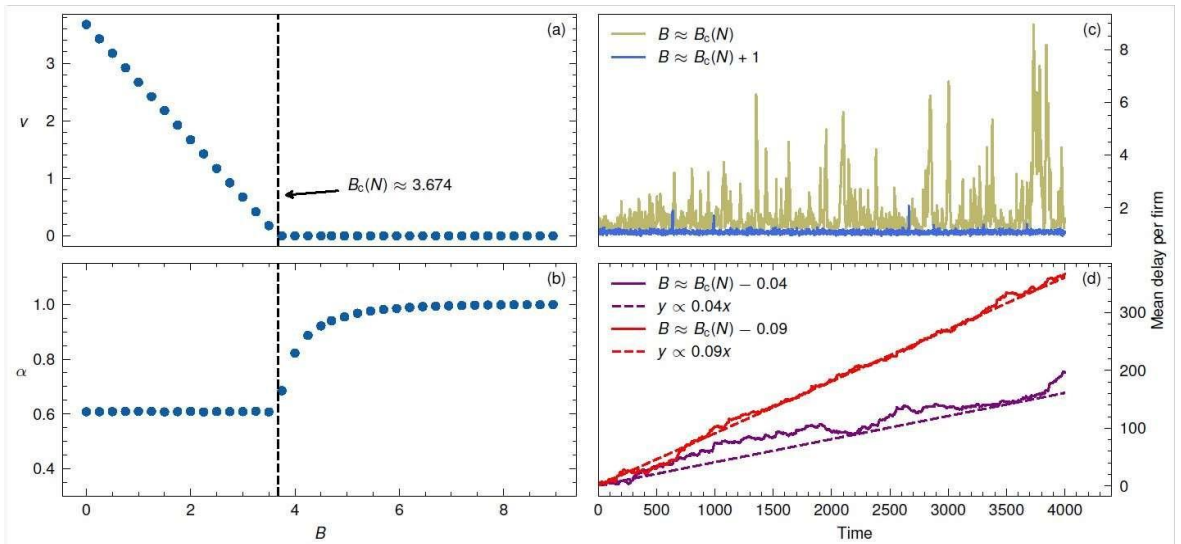


Figure 2. Timeliness criticality. **a, b.** Timeliness criticality in the stylized model obtained from direct simulations with exponentially distributed noise for $N = 10,000$ and $K = 5$: both order parameter ν (**a**) and exponent α (**b**) for the exponentially decaying tail of the delay distribution function $\psi(\tau)$ are shown. The critical value of the temporal buffer $B_c(N)$ is found to be ~ 3.674 . **c, d.** Typical behaviour of the mean delay per node as a function of temporal depth: when the buffer is smaller than the critical buffer $B_c(N)$, the mean delay per node keeps accumulating without bound (**d**); in contrast, large avalanches appear when the buffer is close to $B_c(N)$ from above, whereas a buffer substantially larger than $B_c(N)$ results in

small delays (c). The dashed lines in d show a fit through the data over a long period of time.

In critical phenomena, one expects (power-law) diverging correlation lengths as the control parameter gets closer to the critical value in the disordered phase. For timeliness criticality, we expect diverging correlation lengths in time as $B \rightarrow B^*$ from above. There is, however, a subtlety: correlations in the delays of individual nodes decorrelate within times of $O(1)$. For the MF case this should not come as a surprise since the temporal adjacency matrix A is randomly composed at every time step. Consequently, from Fig. 2 c,d, we expect the signs of diverging correlation lengths to be picked up by the mean delay per node. In autocorrelation functions of mean delays per node and p.d.f.'s for delay avalanches it is therefore useful to correct for scale factors that depend on the buffer size, in order to identify the universality of the behavior of the delays.

4. MAIN FINDINGS

This work concerns the competition between achieving timeliness and striving for efficiency—and more generally between efficiency and resilience—in terms of a stylized model on temporal networks with a delay-mitigating temporal buffer B : more buffer translates to lower efficiency, but more adherence to timeliness and stability. With B as the control parameter, the model exhibits timeliness criticality, a novel form of phase transition in time that occurs at a critical value of B . Above this critical value, delay avalanches of all sizes spontaneously appear. For the stylized model, we have characterized these avalanches and have identified the corresponding critical exponents.

It is also shown that our model runs on real-world temporal networks, which are not regular lattices, and also exhibits timeliness criticality. Nevertheless, several challenges still exist for a meaningful translation of temporal buffers in our model to the real world, particularly for the case of production networks.

Of particular interest for the applicability of our model to real STSs is how relatively small and/or local events can lead to avalanches and system-wide disruptions, such as the cancellation of all train rides to reboot scheduling, a worldwide supply chain blockage due to natural disasters or even a full-blown economic crisis.

Table 1. Statistical measures of period-on-period volatility and their root mean square residues in the National Accounts data for the Netherlands. Noteworthy is that the rMSR values are of the same order of magnitude as the volatility, while the Kurtosis values indicate that the volatility distribution is likely fat-tailed. This is remarkable, because the measures are already highly aggregated from many underlying time series and would be expected to have very close to normally distributed properties. (Data source: Statistics Netherlands (opendata.cbs.nl))

series	cadence	epoch	n	growth rate [yr ⁻¹]	rMSR [yr ⁻¹]	Kurtosis [-]
GDP	annual	'50-'22	72	0.068	0.040	0.33
GDP	quart.	'95-'22	112	0.033	0.018	0.31
Imports	annual	'50-'22	72	0.075	0.074	0.72
Imports	monthly	'15-'22	96	0.055	0.046	0.20

Both financial markets and large economies are known to be volatile (see Table 1 for the Dutch economy), much more than what would be expected from economic equilibrium models based on rational expectations. These observations are usually referred to as the “excess volatility puzzle”, or the “small shocks, large business cycle puzzle” [2]. Several scenarios have been proposed to explain these effects, see e.g. [3]. One possibility, highlighted by Acemoglu, Carvalho and others

(see [4] for a recent review), is the role played by input-output networks. However, their model does not account for temporally cascading delays that are crucial to understand the unfolding of economic crises. Rather, a potent cause of such large scale disruptions may be the critical fragility route developing on production networks: operating close to limiting operational performance of timeliness.

5. REFERENCES

- [1] J. Moran, M. Romeijnders, P. Le Doussal, F.P. Pijpers, U. Weitzel, D. Panja, and J.-P. Bouchaud, “Timeliness criticality in complex systems”, *Nature Physics* (2024) <https://doi.org/10.1038/s41567-024-02525-w>.
- [2] B.S. Bernanke, T.F. Geithner, H.M. Paulson Jr., “Firefighting: the financial crisis and its lessons”, (2019) (Penguin books)
- [3] X. Gabaix, “The granular origins of aggregate fluctuations”, (2011)
- [4] *Econometrica*, 79(3) 733-772
- [5] V.M. Carvalho, A. Tahbaz-Salehi, “Production networks: a primer”, (2019) *Annual review of economics* 11, 635-663

Time Series Exhaustive Automatic Modelling (TEAM)

Carlos Sáez Calvo, Luis Sanguiao-Sande, Félix Aparicio Pérez & José Fernando Arranz Arauzo (Spanish Statistical Office, INE)

1. INTRODUCTION

The TEAM (Time Series Exhaustive Automatic Modelling) software is being developed by Statistics Spain (INE) to overcome some difficulties that arise when using the TRAMO-SEATS methodology [1], namely, the fact that, for some series, the users consider that, for different reasons, the (only) model provided by the automatic model identification of TRAMO-SEATS is not adequate. This issue becomes particularly acute at the annual phase of the revision policy, when the RegARIMA model used during the last year for each series must be evaluated and changed if necessary. In some departments, there are a lot of series to evaluate and very little time and resources to do it. If the model provided by TRAMO-SEATS is unacceptable, the manual search for a good model can be a time-consuming task. Moreover, these difficulties have been exacerbated by the COVID-19 crisis, making it more difficult to find good models for some series.

The TEAM software is an R package that uses the rjdverse ecosystem of R packages to call the core functions in JDemetra+ (version 3) [2].

The idea behind TEAM is to fit many more models than the TRAMO-SEATS program or its JDemetra+ implementation do, and to rank them according to some criteria. Then a list of the best models is provided to the user to make it more likely that the user will find that at least one of those models is adequate.

The programs that implement the TRAMO-SEATS methodology consider also different low orders of the ARIMA model, but are not exhaustive, making some simplifications for the sake of speed of computation.

All these enhancements come at the cost of increasing the computing time needed to fit each time series. For this reason and thanks to the fact that the specifications calculation process is highly parallelizable, the TEAM software has been designed to take advantage of the ability of the R language to run in parallel on the current machine and even on remote machines (by creating a cluster of workers via SSH connection), depending on the capabilities of the available computational systems. The user can also request TEAM whether to use these possibilities or run the process sequentially.

2. METHODOLOGY

The execution of TEAM consists of two independent steps. In the first step, the user establishes the specifications TEAM is going to try, while in the second step each of these specifications is fitted using JDemetra+, some quality indicators are computed and finally all the specifications are ranked according to their quality.

In the first step, the scope of the specifications to be tested can be chosen by the user. For example, the user can decide to fit models both in logs and in levels, and with different

seasonal and regular differencing orders. The user can also try different calendars, different significance levels and kinds of outliers, different trading day and moving holiday treatments and different orders of the ARIMA model.

In the second step, once the specifications to try have been established, TEAM fits a model for each of the given specifications using JDemetra+. Then, the specifications which don't have a canonical decomposition are discarded, since they are unacceptable from the point of view of seasonal adjustment. Also, root cancellation in the AR and MA parts is checked.

Then, for each of the non-discarded specifications, a set of quality indicators are extracted directly from JDemetra+ and a three-phase procedure is used:

1. Ten indicators, considered essential, are checked initially at the 0.1 level. The specifications that don't pass this check are discarded. The ten indicators are normality (Doornik-Hansen), independence (Ljung-Box), Ljung-Box on seasonal lags, linearity (Ljung-Box on the squared residuals), highest correlation between the ARIMA coefficients not exceeding 0.75, number of runs around the mean, length of runs around the mean, seasonal spectral peaks on the seasonally adjusted series, trading day test on the seasonally adjusted series (only for monthly series) and the non-parametric Kruskal-Wallis test for residual seasonality.
2. For the remaining specifications, some additional indicators are checked at the 0.1 level. The level is changed if all specifications are rejected.
3. The remaining specifications are evaluated in five indicators, each of them composed of several lower level indicators. A final score is computed from the five indicators. The five indicators are RegARIMA model adequacy, canonical decomposition, behaviour regarding revisions, annual totals and BIC. These five indicators are standardized and then weighted to compute a final indicator. The user can choose the values of the weights.

If all specifications are rejected in phase 1, a bigger effort is made to model the COVID period properly, since it has been observed that many model inadequacies arise from the difficulties associated with that period. The procedure used consists of the following steps:

- The specifications are again fitted to the series as in phase 1 above, but now including an additive outlier (AO) for each time point in the COVID period (defined as 2020-3 to 2022-8 for monthly series and 2020-I to 2022-IV for quarterly series). The 5 best models according to each of two different criteria are preserved resulting in no more than 10 models.
- For each of these 10 or less models a big effort is made to find adequate interventions in the COVID period. This means that, for quarterly series, all the possible combinations of AO's in each of the COVID quarters and, additionally, for each combination, a level shift (LS) instead of the last AO is also tried, then, the best of them is kept. For monthly series, since we have 30 COVID months it is not possible to try all possible combinations of AO's (the number of combinations would be 2 raised to the 31th power minus 1) and some suboptimization techniques are applied. For example, dividing the 30 COVID months in three 10 month blocks (first 10 months, second 10 months and last 10 months) and optimizing one of the blocks while the two others remain fixed, then optimizing another block with the remaining two fixed and so on until a previously defined number of block optimizations is reached.

3. RESULTS AND PRACTICAL APPLICATION

We show the results of applying this methodology to a set of 24 series in Quarterly National Accounts. Some of these series had a reputation for not being easy to deal with by the TRAMO-SEATS methodology. Phase 1 was successfully completed for 22 series. For the

remaining two, the AO's scheme explained in the previous section was applied, and an acceptable model was found for one of them and no acceptable model was found for the other.

By using TRAMO for automatic model identification, we find 14 series for which the model identified is not acceptable as it failed to meet some of the basic tests of Phase 1 in TEAM. For these series, a manual process would have been required to find a suitable model. In contrast, TEAM provided acceptable models for all but one series without any manual intervention. The table below compares the performance of TRAMO and TEAM, showing the number of Phase 1 indicators not passed in the models identified by each approach.

Table 1. Comparison of TRAMO and TEAM

Series	Phase 1 indicators not passed (TRAMO)	Phase 1 indicators not passed (TEAM)
1	1	0
2	0	0
3	1	0
4	0	0
5	0	0
6	0	0
7	1	0
8	3	0
9	1	0
10	2	0
11	0	0
12	0	0
13	0	0
14	1	0
15	1	0
16	1	0
17	3	0
18	0	0
19	0	0
20	1	0
21	0	0
22	1	0
23	1	0
24	3	2

Using a desktop computer (Intel(R) Core(TM) i5-8500 CPU 3.00GHz with eight cores, 16 GB RAM) the time to analyse the 24 series using the three-phase procedure was 23.85 minutes. For each series, 1297 specifications were evaluated.

Regarding the AO's exhaustive search for the 2 series, an additional time of 1.55 hours was used.

The Quarterly Accounts department has been using TEAM for production with many other series.

4. MAIN FINDINGS

The proposed methodology can be used to complement the traditional TRAMO-SEATS approach, using a reasonable computing time and with little human intervention.

The methodology also offers more flexibility compared to automatic model identification with TRAMO, showing several suitable models and allowing the user to set preferences for specific quality dimensions. Additionally, the exhaustive search ensures that the best possible models are identified, which is not always guaranteed with TRAMO.

More work is needed to address the monthly case, especially for series which have suffered some issues related to the COVID period.

5. REFERENCES

- [1] V. Gomez and A. Maravall, Programs TRAMO and SEATS, Instructions for the User (1997).
- [2] S. Grudkowska, JDemetra+ User Guide Version 2.2, (2017), available at https://ec.europa.eu/eurostat/cros/system/files/jdemetra_user_guide_version_2.2.pdf

Towards buildings demography

Julien Gaffuri (Eurostat)

1. INTRODUCTION

National Mapping and Cadastral Agencies (NMCAs)¹ are national organisations mandated for the maintenance of reference geographical information covering a wide range of thematic areas. Eurostat-GISCO uses this information to address the needs of Eurostat and the European Commission for spatial analysis with the production of geospatial statistics. This article presents recent developments at Eurostat based on NMCA vector topographic databases on buildings. This allows deriving gridded statistical indicators on the building fabric and its evolution at various resolutions, paving the way towards what we propose to name “building demography”.

2. METHODOLOGY

2.1. Input data

NMCAs produce topographic datasets covering various thematic areas such as buildings, transport networks, administrative units, addresses, hydrography, land cover, elevation, utility services, etc. These geographic datasets are often stored as “vector” datasets, representing real world entities as individual objects, with a geometry in vector format (point, line, polygon), various properties (or attributes) describing them and sometimes relations between them. These datasets are described in data specification documents such as [1,2,3]. A vector dataset is thus not a simple image, like a map, but a database of objects whose relations (spatial or not) may be analysed with specific spatial analysis methods.

For the building theme, NMCAs usually provide datasets with each individual building represented as a polygonal vector geometry and information on the building nature, its utilisation (residential, industrial, etc.), its height (ground and roof elevation, or number of floors), lifecycle (construction date, renovation date), energy efficiency, material of construction, etc. There is no standard specification for these datasets in Europe, ensuring common quality requirement and comparability between countries, but similarities exist, which offer promising opportunities for statistical purposes, as presented in this article.

2.2. Methodology 1: Basic statistical indicators

We propose first to simply derive raw quantitative indicators on buildings at 100m resolution grid. This consists in retrieving the buildings that intersect each 100m grid cell, compute the geometrical intersection (for the case when a building is not fully located inside the cell) and compute a statistical indicator based on these building parts. Building counts are not reliable since buildings may have different sizes and the way touching buildings are segmented may vary from country to country. Indicators are thus computed based on both the building **ground area** (without considering building height) and **floor area** (taking into account the building

¹ See: <https://eurogeographics.org/our-members/>

height, when available – the floor area is the ground area multiplied by the number of floors). For both ground and floor areas, the selected indicators are: Areas for all buildings, areas for residential buildings, areas for economic activity buildings only (based on the information on the building utilisation), areas for buildings with a cultural value only (based on the information on the building nature and date of construction).

These indicators are used to show how the urban fabric evolves across space, between cities, showing residential areas, areas with economic activity, and how they interrelate.

2.3. Methodology 2: Combination with resident population

We propose indicators combining residential building area with total resident population. The ratio **residential floor area / population** shows how many residential square meters are available in average per inhabitant. The inverse ratio **population / residential floor area** shows how many inhabitants live in average per residential 100 square meters.

Both indicators are equivalent (since inverse of each other) but offer two relevant interpretation levels. They allow interpreting demographic pressure on the residential buildings stock. Since the population is defined at primary residence place, it also reveals areas with predominance of secondary residences.

For these indicators, the Census 2021 population grid² published by Eurostat at 1km resolution is used.

2.4. Methodology 3: Dasymetric disaggregation

The 100m resolution grid on residential building area can be used to disaggregate 1km resolution gridded statistics on population, assuming that resident population is homogeneously distributed over the residential buildings area. The principle is to use dasymetric disaggregation method [4,5], by distributing the 1km resolution population over the residential area and then re-aggregating the population on the 100m resolution cells.

3. RESULTS

The methodologies were applied on the seven countries listed in table 1, where building datasets are available from NMCA data sources.

Table 1. Building dataset sources

Country	Organisation	Dataset name	Resolution
Poland	GUGiK	BDOT10k	1:10'000k
Luxembourg	ACT	BD-L-TC 2015	1:10'000k
Italy	IGM	DBSN	1:10'000k
France	IGN-France	BD TOPO®	1:10'000k
The Netherlands	Kadaster NL	BAG, TOP10NL	1:10'000k
Austria	BEV	DLM 8000 BAUWERK	1:10'000k
Czechia	CUZK	Inspire BU	1:10'000k

Visualisations of the results is available on:

² <https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat>

- <https://observablehq.com/@eurostat-ws/building-demography>
- <https://observablehq.com/@eurostat-ws/population-europe-grid-100m-resolution>

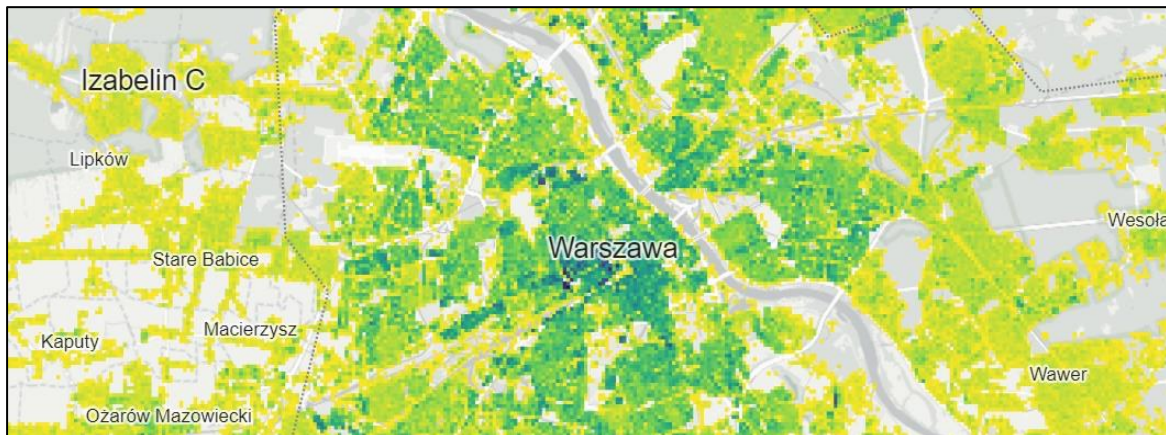


Figure 1. Building residential area around Warsaw

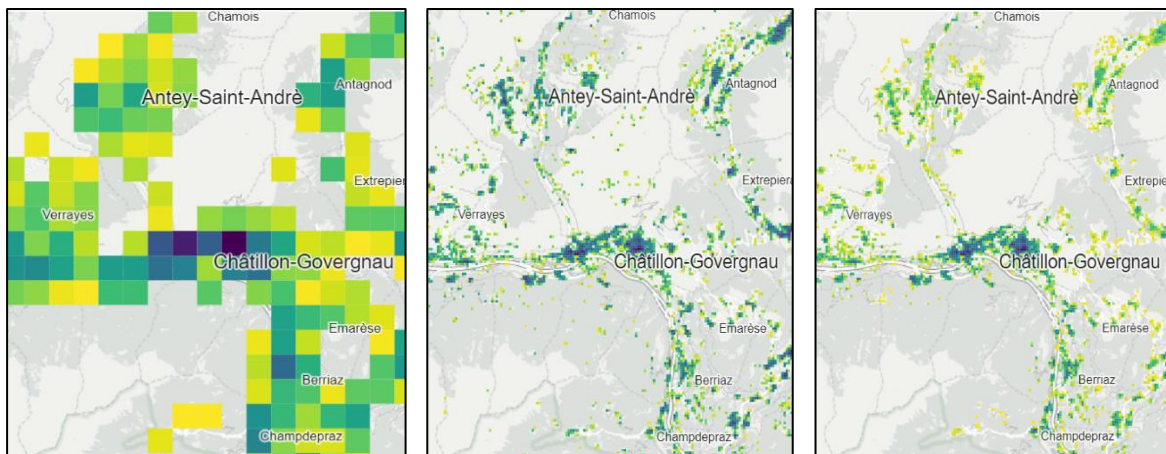


Figure 2. Population 1km resolution (left), residential building area (centre), population disaggregated at 100m resolution (right)

4. MAIN FINDINGS

We presented indicators derived from NMCA building geographical datasets. Some of them result from a combination of geospatial indicators with statistical indicators on population.

We illustrated the utilisation of geographical datasets on buildings for disaggregating statistics [6] – for this purpose, building vector datasets offer more reliable and detailed information than raw satellite or aerial images to measure the urban fabric and its density.

Building statistical indicators may also be an interesting information source to refine statistical area definitions. The degree of urbanisation [7] is currently based on resident population only. It could be combined with building density indicators – some statisticians argue that “buildings do not take the bus”, but people live in buildings. Functional area definitions may also be enriched with activity building clusters.

Buildings are a cornerstone of our society, serving as both the primary consumer of energy and the second-largest source of greenhouse gas (GHG) emissions. They also represent the

largest category of household expenditure. Within the framework of the European Green Deal, which emphasizes the renovation of buildings to enhance energy efficiency, the monitoring of building conditions has become increasingly significant. A critical aspect is understanding the interplay between housing affordability and quality.

To inform policies and track progress, there is a pressing need to consolidate and enrich the data provided by NMCAs on buildings. Detailed information is required to grasp the full scope of building demography, including the location, usage, and occupancy of buildings, as well as their energy sources, consumption patterns, energy performance metrics, construction material, and market value. This requires rich, reliable and quality-controlled information on buildings [8] as proposed for example by the French national observatory on buildings³.

5. REFERENCES

- [1] DBSN (DataBase di Sintesi Nazionale), Italian Military Geographic Institute. URL: <https://www.igmi.org/it/dbsn-database-di-sintesi-nazionale>
- [2] BD TOPO®, descriptive de contenu, Institut de l'Information Géographique et Forestière. URL: https://geoservices.ign.fr/sites/default/files/2021-07/DC_BDTopo_3-0.pdf
- [3] ISO 19131:2022 Geographic information — Data product specifications. URL: <https://www.iso.org/standard/85092.html>
- [4] W. R. Tobler, Smooth pycnophylactic interpolation for geographical regions, 1979, Journal of the American Statistical Association 74(367):519-530.
- [5] H. Kim, X. Yao, Pycnophylactic Interpolation Revisited: Integration with the Dasymetric-mapping Method, 2010, Int. J. Remote Sens. 31(21):5657-5671.
- [6] J. Gaffuri, Improving the quality of official statistics with geographical disaggregation based on dasymetric mapping: Two Eurostat experiments on tourism and population statistics, European Conference on Quality in Official Statistics, 2018, Krakow, Poland.
- [7] V. Angelova-Tosheva, O. Müller (Eds.), Methodological manual on territorial typologies, 2018 edition, Eurostat.
- [8] J. Gaffuri, Geographical data quality for spatial analysis and geospatial statistics, European Conference on Quality in Official Statistics, 2024, Estoril, Portugal.

³ <https://www.observatoire-national-batiments.fr/>

Towards dynamic land cover data for Austria using satellite imagery and time-series classification

Nina Hofer & Nils von Norsinski (Statistics Austria)

1. INTRODUCTION

Land cover commonly describes “the biophysical material over the surface of the Earth and immediate subsurfaces including, among others, grass, shrubs, forests, croplands, barren, waterbodies (including groundwater), and man-made structures” [1]. Since the availability of open-access satellite imagery, the production of large-scale land cover classification maps became more feasible for research, public administration and other stakeholders interested in such information about the surface of the earth. Especially the Copernicus Programme by the European Space Agency offers satellite imagery as well as land information products which have the potential to be integrated in the statistical production process and used to support various surveys in fields of land cover and land use information, such as agriculture, forestry and environment. However, still no continuous national-scale product with a frequent temporal resolution is available. To fill this gap Statistics Austria has recently started to integrate Copernicus Sentinel data for land cover classification. We want to be able to monitor the environment spatially-explicit, timely, consistently and dynamically – meaning to be able to detect changes and to focus on different land cover types depending on different requirements. We therefore continue our activities from previous projects to develop a workflow for the production of annual land cover maps of Austria with a 10 m resolution using Sentinel-1 and Sentinel-2 satellite imagery, a digital elevation model, various open-source land cover data and machine learning models.

So far, we can achieve a classification accuracy of 80 – 90 % and distinguish at least 10 different land cover classes. The results are currently used for calculating various Sustainable Development Goals, Agricultural Land Prices and other land cover related projects such as the Ecosystem Accounts. Besides looking into specific land cover classes and apply tailored models to produce classifications with a higher granularity – for example crop mapping – we are currently also aiming to generalise the 10-class model over time, to be able to predict data from years it has not been trained on and therefore produce respective land cover maps faster.

2. METHODOLOGY

2.1. Remote Sensing Data

Since previous projects have shown that the usage of Synthetic-Aperture Radar (SAR) data could improve classification accuracy when used in combination with optical data especially in cloudy areas, they are an important component in our model. Sentinel-1 backscatter data are being downloaded from the archive of Earth Observation Data Centre (EODC). So far, the monthly mean of the Gamma Nought (MMGMR) product with VV/VH polarization has been used which covers the period from June to September for the years 2019 to 2023.

Sentinel-2 data were downloaded with the R Package “SITS”. The observation period covers data from March to September for the years 2019 to 2023 and images were aggregated over 14-day periods, with a best pixel approach. Within an aggregation period images are sorted

by the proportion of cloud cover. The first pixel which is not masked out by the cloud mask is chosen for the aggregated image [2]. Cloud covered areas are not imputed, whereas a no data flag is assigned to those areas.

For the integration of terrain data, a digital elevation model (DEM) with 10 m resolution was used, which is openly accessible for Austria and produced via Airborne Laser Scanning by the federal states.

2.2. Creation of a ground-truth dataset

A major focus of our previous project has been setting up a classification typology for the land cover classification based on 1. use case requirements, 2. available auxiliary ground-truth data and 3. the investigation of distinguishable classes using cluster analysis. This nomenclature is now used for producing a time series of land cover maps and consists of 10 different classes (bare rocks and screes, cropland, inland wetlands, snow and ice, grassland, heathland and shrubs, sealed surfaces, rivers and lakes, woodland and sparsely vegetated land).

Different sources for the creation of a labelled ground truth dataset were evaluated in the previous project as well, and so far, two data sources which meet the needs of the classification are chosen: The Digital Cadastral Map (DKM) of Austria produced by the Federal Office of Metrology and Surveying and the Land Parcel Identification Data from the Integrated Administration and Control System (IACS). For the latest approach, we created a sample of labelled data by using DKM and IACS data from the years 2019 to 2022 and randomly distributed 4000 point per class and afterwards relabelled them according to the 10 classes needed for classification.

2.3. Model training

For the application of a machine learning model we considered and tested several algorithms in an initial phase. Studies showed that the histogram-based gradient boosting classifier achieves high accuracy when using big datasets, is robust in handling missing values and unbalanced datasets, and is very efficient in model building and prediction [3] [4]. The histogram-based gradient boosting model in scikit-learn was used in version 1.0.2., the masked areas - marked as "no data" - were handled internally by this method. Hyperparameter tuning was performed using grid search. At first a model was trained for each year from 2019 to 2022 and a map was predicted for each year respectively. We then evaluated to train a model for one year and apply it to the remote sensing data of further years to predict a map. This approach didn't lead to good classification results in terms of accuracy - especially in high mountain areas severe misclassifications can be seen. Therefore - to create a model which can generalise to time periods it has not been trained on - we evaluated a time-series classification where one model was trained based on the training data of the years 2019 – 2022. The resulting model was used to predict a land cover map based on satellite data of the year 2023.

3. RESULTS AND PRACTICAL APPLICATION

The time-series classification performs well for most classes (Figure 1), nevertheless for certain classes the classification accuracy is below the objective accuracy of at least 80 %. However, it has to be mentioned that the classes "Sparsely vegetated land" and "Heathland and shrubs" cover only minor parts of the country.

True	Cropland	0.80	0.00	0.00	0.09	0.02	0.01	0.06	0.00	0.01	0.00
	Bare rocks and screes	0.00	0.86	0.00	0.01	0.00	0.05	0.03	0.04	0.00	0.00
	Wetland	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Greenland	0.09	0.02	0.01	0.65	0.03	0.04	0.14	0.00	0.02	0.00
	Urban	0.03	0.01	0.00	0.03	0.88	0.01	0.03	0.00	0.02	0.00
	Sparsely vegetated land	0.04	0.11	0.00	0.09	0.01	0.62	0.10	0.01	0.02	0.00
	Heathland and shrubs	0.07	0.05	0.00	0.11	0.02	0.04	0.67	0.00	0.05	0.00
	Glacier	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
	Forest	0.00	0.01	0.00	0.01	0.01	0.02	0.04	0.00	0.77	0.14
	Water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
		Cropland	Bare rocks and screes	Wetland	Greenland	Urban	Sparsely vegetated land	Heathland and shrubs	Glacier	Forest	Water
		Predicted									

Figure 1. Confusion matrix of the model trained on reference data from 2019 to 2023.

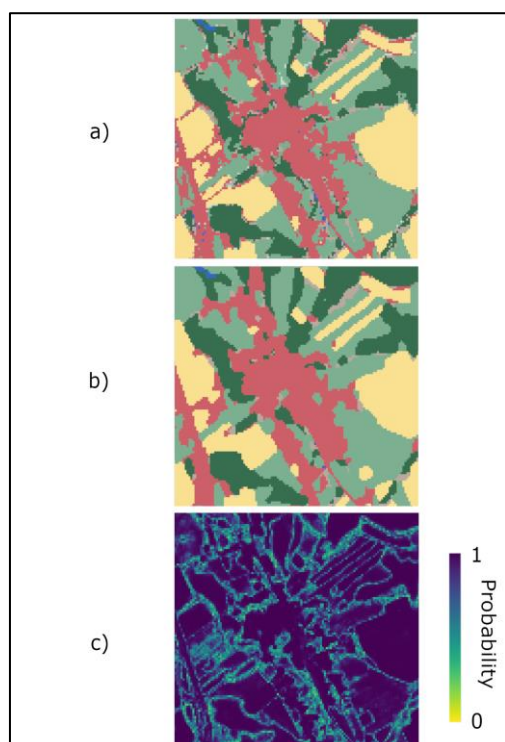


Figure 2. a) Original predictive outcome, b) Result of the bilateral smoothing filter applied, c) Probability of the classified classes.

We used the trained model to predict a result map for 2023 (Figure 2 a). During post-processing it was decided to apply a bilateral smoothing filter to remove noise of misclassifications while preserving edges (Figure 2 b). Usually transition regions between two classes cause misclassifications because of mixed land cover within one pixel. However, a downside of a smoothing filter is the loss of thin structures like roads. To overcome this issue streets were reclassified using the DKM cadastre data.

4. MAIN FINDINGS

To produce land cover maps of Austria faster and dynamically we are testing and implementing the time-series classification approach, with currently acceptable accuracy in most relevant classes. By training the model over several years we have been able to generalise the model over time, so it can predict data from years it has not been trained on. During the procedure it became clear that for high mountains areas SAR imagery and a Digital Elevation Model is necessary, because of the frequent cloud cover and the similarity of sealed areas and bare rock areas in alpine regions.

The stability of the land cover classification over the years is acceptable, though we are conducting an extensive investigation and revision of the training data at the moment to improve the accuracy and stability for challenging classes. While some classes like sealed surfaces are very stable when predicting maps for several years, others like heathland and shrubs are more variable. For now, reliable change detection analysis can be done with the more stable classes like sealed surfaces, woodland, cropland, grassland and rivers and lakes.

By integrating EO data in the production of land cover related statistics, we are not only able to improve the timeliness of data dissemination and statistics, but also increase the coverage of statistical information which can be combined with EO data, reduce missing data and produce spatially-explicit land cover data.

5. REFERENCES

- [1] G. Grekousis, G. Mountrakis, M. Kavouras, An overview of 21 global and 43 regional land-cover mapping products, *International Journal of Remote Sensing* 36 (21) (2015), 1-27.
- [2] R. Simoes, G. Câmara, G. Queiroz, F. Souza, P. Andrade, L. Santos, A. X. Carvalho, K. Ferreira, *Satellite Image Time Series Analysis for Big Earth Observation Data*, *Remote Sensing* 13 (2021), 2428.
- [3] A. Gudmann, L. Mucsi, Pixel and object-based land cover mapping and change detection from 1986 to 2020 for Hungary using histogram-based gradient boosting classification tree classifier, *Geographica Pannonica*, Volume 26, Issue 3 (2022), 165-175.
- [4] J. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics* 29 (2001), 1189-1232.

Towards Statistical Disclosure Control for Complex Networks

Rachel G. de Jong, Mark P.J. van der Loo (Leiden University & Statistics Netherlands) & Frank W. Takes (Leiden University)

1. INTRODUCTION

In recent decennia, research in the field of Network Science where networks are employed as a model for social and economic structures, has surged. National Statistical Institutes, especially those with access to nation-wide administrative data are in the unique position to provide relevant, accurate, and up-to-date data for the construction of population-scale networks. Indeed, Statistics Netherlands has recently constructed a ‘social network’ of the entire Dutch population, where nodes represent people, and edges represent real-world relations including kinship, co-workship, geographical vicinity (neighbours) and shared school [1,2]. This ‘network view’ on society has already yielded applications, including new ways of measuring segregation [3, 4].

The introduction of network microdata poses new challenges in the area of Statistical Disclosure Control (SDC). In contrast to traditional relational microdata where population units are represented by a set of unrelated tuples of attributes, the units represented in network data are interrelated by one or more types of edges. Although this structural information generates valuable insights, it also presents new risks of disclosure as adversaries may use (partial) knowledge of network structure to re-identify nodes or their properties.

Here, we present our ongoing research in the area of Disclosure Control for Complex Networks. Focusing on the risk of node reidentification, the central research questions revolve around attacker scenarios, how to measure and compute risk of disclosure, and approaches to alter a network to mitigate this risk.

To address these questions, we have defined and investigated the notion of d - k -anonymity: a graded variant of k -anonymity that enables one to vary the amount of structural information an adversary can use to reidentify nodes. Since computing this measure is computationally intensive, we have investigated algorithms to efficiently compute it, as well as alternative measures that can be used as approximation. Based on a wide range of computational experiments we draw conclusions on what type of knowledge makes an adversary most dangerous. We have also looked into cascading effects, *i.e.* the extent to which reidentification of one node leads to reidentification of others. Finally, we have started to explore perturbing network data to protect against reidentification.

2. APPROACH AND MAIN RESULTS

In all of the work presented here, we use networks that consist of nodes with at most a single undirected edge between each other. The nodes might represent for example people, households, or businesses. In applications these nodes are typically labelled with non-identifying attributes, such as income or health status. In our scenario, we assume the network is publicly available, and an attacker has some structural information available that can be used to narrow down the possible candidates for a target node. As an example, consider a kinship network of living persons, linked by parent-child relations. If an attacker knows, for

example, that a target node has no children and a single living parent, this information can be used to narrow down the number of nodes representing the target, possibly to a single entity.

2.1. d - k -Anonymity

To quantify the amount of structural information an adversary might have about the network structure surrounding a node, we introduce the notion of d - k -anonymity [5,6]. We say that two nodes are d -equivalent if (1) they have the exact same neighbourhood structure up to distance d and (2) they occupy the same structural location in their respective neighbourhoods. A node is called d - k anonymous, if it is d -equivalent with at most $k-1$ other nodes in the network. In the case of $d=1$, this models the scenario where an attacker knows how many direct neighbours a target node has, as well as all the edges between its direct neighbours. For $d=2$, this is extended to neighbours of neighbours and all edges between them, and so on.

Computing d - k anonymity is computationally expensive, as it requires comparing possibly many nodes, and each comparison possibly requires the computationally intensive operation of determining graph isomorphism. We have therefore designed algorithms that prevent unnecessary comparison of nodes by using heuristics to avoid isomorphism computation. For example, two neighbourhoods cannot be isomorphic if they do not have the same number of nodes and edges, which is easy to determine. The combined optimizations yield a speedup of up to four orders of magnitude as compared to the naïve approach [6].

A computational study on a wide range of model networks as well as on real networks reveals that an adversary is generally capable of reidentifying a large fraction of nodes when they have knowledge of the full neighbourhood structure of a target node, up to and including distance $d=2$. We also demonstrated that a cascading effect, where an attacker knows that a target node is linked to an already identified node, can have a further effect yielding 50% extra reidentification on average [7].

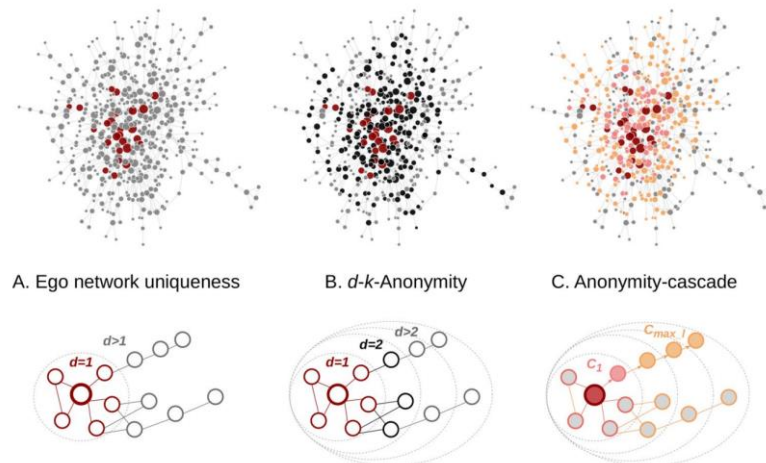


Figure 1. Measuring anonymity assuming an adversary has knowledge of node structure up to and including $d=1$ (A), $d=2$ (B) and accounting for a cascading effect (C). Non-grey nodes are re-identifiable by adversary knowledge [7].

2.2. Comparing measures for k -anonymity

The d - k -anonymity measure assumes that an adversary has a large amount of information available, perhaps an unrealistic amount. In that sense d - k -anonymity may be a ‘too strict’ measure. Moreover, even with all optimisations available, computing d - k -anonymity can be computationally cumbersome for large networks. We have therefore explored and compared various measures of k -anonymity for networks. They differ from d - k -anonymity in the way that equivalence between two nodes is defined.

Amongst the studied measures of equivalence are DEGREE-equivalence where nodes are equivalent when they have the same number of neighbours; COUNT equivalence, that compares the number of nodes and links the neighbourhoods of nodes; degree distribution equivalence (DEGDIST) that compares the distribution of degrees of the neighbourhoods of nodes; vertex refinement query equivalence (VRQ), that compares the degrees of all nodes in a neighbourhood; and HYBRID equivalence, that takes account of full neighbourhood structure and the degrees of the furthest nodes of the neighbourhood. We proved that these measure can be organized in a partial order according to strictness – or the amount of information they assume for the adversary [8].

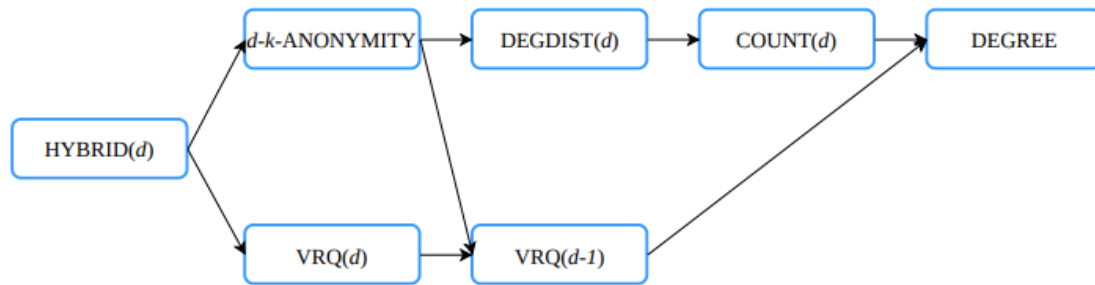


Figure 1. Strictness of anonymity measures. $A \sqsupseteq B$ implies A is stricter (assumes an adversary has more information) than B [8].

A computational study on model networks and a wide range of real networks demonstrated that the COUNT measure is often a good proxy for d-k anonymity while it is computationally much less demanding. Second, we see that it is generally more advantageous for an adversary to have less complete information that includes larger distances from the target node, than very complete information of the direct surroundings of a target node.

2.3. Anonymization of complex networks

As a next step in our research we have started to work on methodology to ‘minimally’ manipulate networks to increase the fraction of non-unique nodes [9]. In existing literature on this topic, a specific anonymization technique is often devised for a certain measure. However, we argue that as in SDC methodology for relational data, the anonymization can be treated as a separate problem from measuring anonymity. We have therefore defined several versions of the anonymization problem that can be investigated regardless of the way anonymity (‘risk of disclosure’) is measured.

Variants of the anonymization problem differ in goal and boundary conditions. One may aim for full anonymization (ensuring all nodes are at least k -anonymous), partial anonymization (ensuring at least a fraction of nodes are at least k -anonymous), or one might find optimal anonymity, given a budget of perturbations. Given those variants, we use computational studies to select a measure for anonymity and found COUNT to be an overall good proxy for many applications.

We also tested a number of general perturbation techniques, including edge deletion, edge swapping and edge addition, and found that edge deletion generally requires less modifications to reach a satisfactory level of anonymity. This is due to the finding that nodes tend to be more anonymous in sparse networks where only a small fraction of all possible edges is actually realized. We used six different approaches of edge deletion that take account of the structure of surrounding nodes in varying degrees. The first and simplest method is random edge deletion (RANDOM). Second, we prioritize edges that connect two high-degree nodes (DEGMIN), or (three) a high-degree with a low-degree node (DEGDIFF). Fourth we prioritize edges that are high-impact in the sense that deleting them affects the structure of

many nodes (AFF). Fifth, we prioritize edges connected to a unique node (AFF-U) or (six) nodes that affect the structure of many unique nodes (U-AFF-U).

Of all tested approaches, the algorithms that target edges that directly or indirectly affect the network structure surrounding unique nodes perform best.

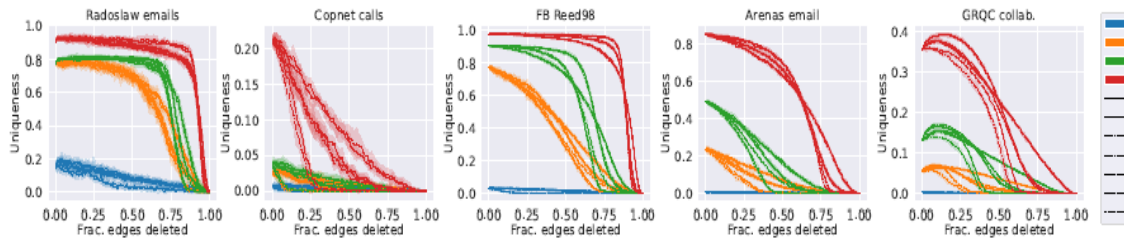


Figure 3. Fraction of unique nodes as a function of fraction of edges deleted, for various algorithms (line type) and anonymity measures (color).

3. SUMMARY AND OUTLOOK

Network science is an extremely interesting field for Statistical Institutes since it promises to investigate social and economic phenomena from the perspective of interconnected units. Moreover, Statistical Institutes are often in the unique position to construct networks from reliable administrative sources that are otherwise unavailable for (network) scientists.

In this work we focus on the problem of Statistical Disclosure Control for network data. We demonstrated, compared and categorized various ways of measuring node anonymity, and showed that (1) adversaries with knowledge of a target's surrounding network structure has a high probability of reidentifying nodes and (2) having incomplete information at a larger distance is more advantageous for an adversary than complete knowledge of the nearby structure of a target. Our research into network anonymization demonstrates that targeted edge deletion methods that aim to affect unique nodes outperform other, simpler tested methodologies. Future work will focus on improving anonymization techniques with better targeting and allowing for explicit account of data utility.

4. REFERENCES

- [1] Van der Laan, J., de Jonge, E., Das, M., Te Riele, S., & Emery, T. (2023). A whole population network and its application for the social sciences. *European Sociological Review*, 39(1), 145-160
- [2] Bokányi, E., Heemskerk, E. M., & Takes, F. W. (2023). The anatomy of a population-scale social network. *Scientific Reports*, 13(1), 9209.
- [3] Van der Laan, J., Das, M., te Riele, S., de Jonge, E., & Emery, T. (2021). Using a network of the whole population of the Netherlands to measure exposure to differing educational backgrounds. <https://doi.org/10.31235/osf.io/7jtb2>
- [4] Kazmina, Yuliia, Heemskerk, E. M., Bokányi, E., & Takes, F. W. (2024). "Socio-economic segregation in a population-scale social network." *Social Networks* 78: 279-291
- [5] Van der Loo, MPJ (2022). Topological Anonymity in Complex Networks. Technical Report, Statistics Netherlands April 21 2022 [PDF](#)
- [6] R.G. de Jong, van der Loo, M.P.J., Takes, F.W. (2023). Algorithms for Efficiently Computing Structural Anonymity in Complex Networks. *ACM Journal of Experimental Algorithmics* 28 1—24

- [7] R.G. de Jong, van der Loo M.P.J., Takes, F.W. (2024). The effect of distant connections on node anonymity. Scientific Reports 14 1156
- [8] R.G. de Jong, van der Loo, M.P.J., Takes, F.W. (2024). A systematic comparison of measures for k-anonymity in networks. <https://arxiv.org/abs/2407.02290>
- [9] R.G. de Jong, van der Loo, M.P.J., Takes, F.W. (2024). The anonymization problem in social networks. <https://arxiv.org/html/2409.16163v1>

Two methods for integrating smart surveys with traditional surveys

Danielle McCool, Bella Struminskaya & Peter Lugtig (Utrecht University)

1. INTRODUCTION

Smart surveys offer a promising advancement in data collection by leveraging features available in modern smartphones -- such as sensors and machine intelligence -- to help survey respondents report on complex or intensive behaviors. This could include tracking their mobility habits, time usage, or expenditures. Automating parts of the data capture can reduce respondent burden and increase accuracy, leading to improved data quality.

The adoption of smart surveys introduces challenges, particularly with respect to integrating the data with data from more traditional survey methods. Measurement differences are likely to arise between the two methods due to the different nature of data collection. Smart surveys typically make use of sensors and machine intelligence to assist respondents in reporting on complex issues, such as detailed behavior (e.g., travel, time use, expenditure) over a number of days. Paper diaries or telephone interviews, in contrast, rely on self-reported measures, which can influence the comparability of the data, making integration non-trivial. The same tasks that respondents are poorly suited for, such as tasks requiring continuous reporting or reliance on recall of detailed behaviors, are well-suited for computers, and the last decade has correspondingly seen an increase in smart surveys which augment existing methodology with the tools available in smartphones [1,2].

Although smart surveys can achieve relatively good response rates [3], they are likely to suffer from problems with representativity. One self-evident example is that a person without a smartphone can't participate in a smart survey if required to download an application to his or her own device. Smartphone ownership and is unevenly distributed within the population [4,5], but even within smartphone owners, privacy concerns [6], age, and educational background [3] may influence participation. While ownership as a concern may become less relevant over time, researchers who decline to offer a non-smart alternative must still address the break in the time-series.

There is a more urgent need to establish best practices for integrating data sources differing both in measurement and representation [2] in smart surveys than has been described in the traditional mixed-mode context where measurement differences have generally been small. The focus has traditionally been on compensating for representational differences, while reducing measurement differences between modes [7,8]. In smart surveys, reducing measurement differences may also substantially reduce their benefit, as the smart features may produce data that are inherently different.

We propose two fundamentally different approaches to the question of data integration: the mixed-mode approach and the multisource approach. The mixed-mode approach prioritizes minimization of mode measurement error, adjusting one mode to accommodate another as a form of pre-processing. The multisource approach prioritizes minimization of overall measurement error, integrating the data at the modeling stage.

2. METHODOLOGY

In smart survey research, the requisite elements of mixed-mode integration (minimal mode measurement effect and a single benchmarking mode), are often missing. For example, in a human mobility study we might seek to reduce the measurement error in estimating self-reported travel distance in a PAPI survey by calculating it on behalf of the user using location services on a smart device, in a time use study, we might reduce categorization errors by suggesting activities on the basis of a search term, or in an expenditure study, address missing data by reducing the burden of entering products by using the device's camera to populate the diary by taking photos of grocery receipts.

Especially where sensors are involved, the resulting raw data may have very different characteristics and require multiple algorithms to arrive at a data set that looks identical between the two modes. For example, consider Figure 1 in which the same underlying travel behavior results in two very different data formats between the raw self-report data and raw sensor data. The path to integrating the timestamped latitudes and longitudes with the diary formatted data is not straightforward, and a researcher who would like to compare or integrate the two data sets is forced to make a decision on how best to address these differences. She may 1) constrain the smart survey data, applying a set of algorithms to get the data into the same shape, or 2) leave the inherently different structures intact and pursue integration after the fact.

3. RESULTS AND PRACTICAL APPLICATION

3.1. Mixed-mode integration

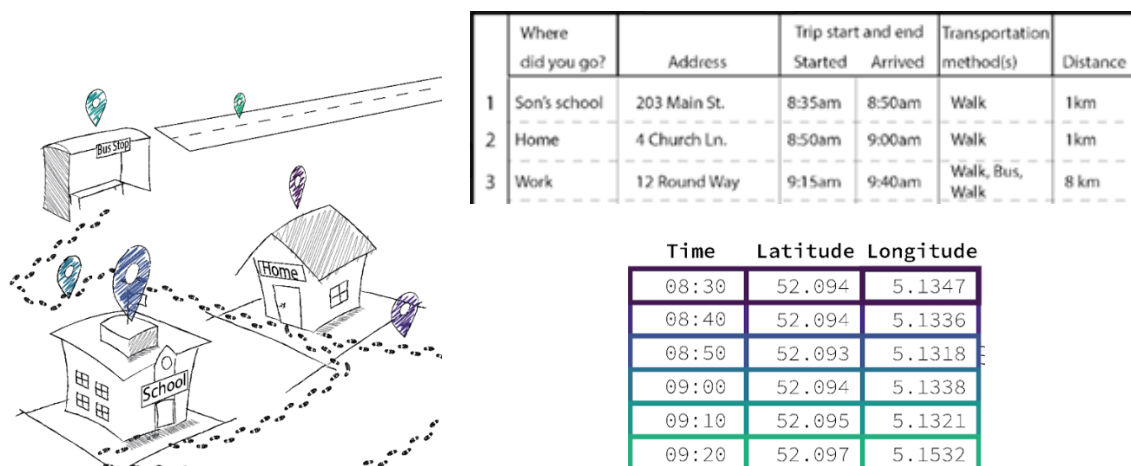


Figure 1. Underlying travel behavior (left) represented in diary format (above) and raw spatiotemporal data (below)

The ultimate goal of mixed-mode-like integration is close alignment of the two data sources, leading to a single source of data that can be analyzed as a single unit. To this end, a researcher will seek to reduce the measurement differences between the smart and non-smart survey. This may require ignoring structural aspects of the smart survey data that improve measurement, aggregating across more coarse intervals of time, or removing incomplete measurements.

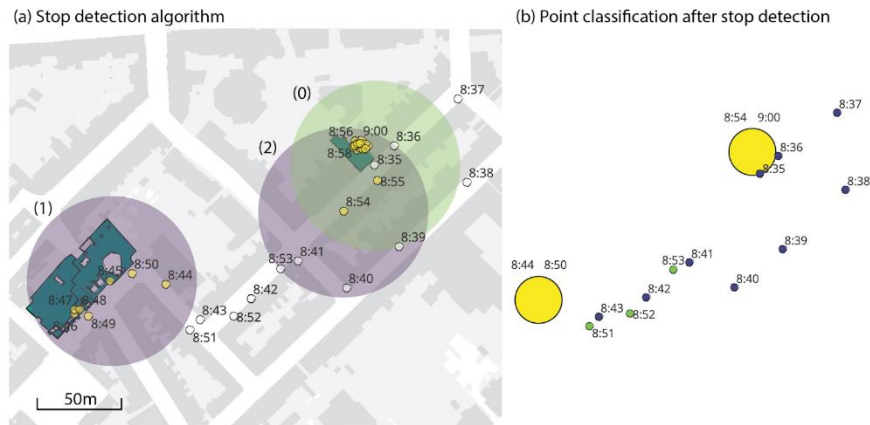


Figure 2. Derivation of diary stop/trip format on the basis of the raw locations using time and radius parameters.

Following the travel diary example, Figure 2 shows the process of aggregating the recorded locations into a consecutive stop/trip diary format. During this process, information on the exact path traveled is lost, and measurement error is induced in an effort to integrate by aligning data formats. Table 1 shows the final result of the diary generation process after processing with stop detection, reverse geocoding, mode detection, and distance calculation on the location data.

Table 2. The diary generated from the smart survey after processing.

Where did you go?	Address	Trip start	Trip end	Transport Method	Distance
(1)	203 Main St.	08:35	08:43	Walk	210m
(2)	4 Church Ln.	08:51	08:53	Walk	210m

3.2. Multisource integration

In multisource integration, each data source maintains its full structure as integration occurs during the analysis stage rather than before. This model-first approach necessitates upfront consideration of the outcome statistics of interest.

As an example, consider the data from Figure 2. Using a multisource approach to integration, a synthetic dataset could be created in which the spatiotemporal data is preserved and forms the basis of analysis for total travel distance, upon which a synthetic data set can be built after matching spatial segments against diary-generated trips to improve labeling and reduce overall measurement error [9].

4. MAIN FINDINGS

Researchers should be attentive to what the raw data measures and whether this represents an underlying conflict with respect to construct validity. This is most evident with sensor data, where differences can be masked when researchers implement existing tools or have only interacted with the resulting data sets prepared with respect to mixed-mode integration. Where the underlying construct being measured differs significantly, the survey design should optimally be administered to allow for a multisource analysis.

Where limited mode measurement effects are predicted, it can be helpful to embed small randomized mode comparison experiments into the field test. Where substantial mode measurement effects are expected, it becomes critical to do so, as it will help provide the necessary information for disentangling measurement and representational differences [10,11].

Deciding on a mechanism for the integration and analysis after the data collection step may depend on the actual measurement differences, whether one survey can serve as a benchmark, whether both modes contribute uniquely to measurement, and differences in aggregation.

5. REFERENCES

- [1] M.P. Couper, G. Gremel, W. Axinn, H. Guyer, J. Wagner, B.T. West, New options for national population surveys: The implications of internet and smartphone coverage, *Soc. Sci. Res.* 73 (2018) 221–235.
- [2] B. Struminskaya, P. Lugtig, F. Keusch, J.K. Höhne, Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges, *Soc. Sci. Comput. Rev.* (2020) 0894439320979951.
- [3] P. Lugtig, K. Roth, B. Schouten, Nonresponse analysis in a longitudinal smartphone-based travel study, *Surv. Res. Methods* 16 (2022) 13–27.
- [4] F. Keusch, S. Bähr, G.-C. Haas, F. Kreuter, M. Trappmann, Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey, *Sociol. Methods Res.* 52 (2023) 841–878.
- [5] J. Klingwort, R. Schnell, Critical Limitations of Digital Epidemiology: Why COVID-19 Apps Are Useless, *Survey Research Methods* 14 (2020) 95–101.
- [6] C. Roberts, J.M.E. Herzing, J. Sobrino Piazza, P. Abbet, D. Gatica-Perez, Data privacy concerns as a source of resistance to complete mobile data collection tasks via a smartphone app, *J. Surv. Stat. Methodol.* 10 (2022) 518–548.
- [7] J. Burton, A. Jäckle, Mode effects, *ISER: Understanding Society Working Paper Series* (2020).
- [8] T. Klausch, J.J. Hox, B. Schouten, Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions, *Sociol. Methods Res.* 42 (2013) 227–263.
- [9] S. Yang, J.K. Kim, Statistical data integration in survey sampling: a review, *Jpn. J. Stat. Data Sci.* 3 (2020) 625–650.
- [10] E.D. de Leeuw, Mixed-mode: Past, present, and future, *Surv. Res. Methods* (2018). <https://ojs.ub.uni-konstanz.de/srm/article/view/7402>.
- [11] J.T.A. Vannieuwenhuyze, G. Loosveldt, G. Molenberghs, Evaluating mode effects in mixed-mode survey data using covariate adjustment models, *J. Off. Stat.* 30 (2014) 1–21.

US Injuries Flood Prediction with Large Language Models Data Augmentation

Claudio Giorgio Giancaterino (Towards Innovation Lab)

1. INTRODUCTION

Climate change poses significant risks to societies worldwide, particularly through extreme events and climate-related hazards.

The research has been focused on the flood event type from the US storm events database, which has data provided by the National Weather Service (NWS) and contains statistics on deaths, injuries and damage estimates.

By leveraging GPT-3.5-turbo (ChatGPT engine) using Scikit-LLM library, the aim has been to retrieve features from narrative data related to climate events, and specifically, has been employed zero-shot text classification and embeddings techniques used as data augmentation.

The injuries forecast performances of the actuarial models like Generalized Linear Models (GLM) and Generalized Additive Models (GAM) have been compared with Gradient Boosting Machine models like LightGBM, XGBoost, and CATBoost.

2. METHODOLOGY

In 2023, the NOAA National Center for Environmental Information (NCEI) released the 2022 U.S. weather and climate disasters report. The U.S. report aims to study climate events, focusing on flood events, as one of the most dangerous situations in Europe. The workflow's research started gathering data from the mentioned report, cleaning the data, using Large Language Models to generate and extract new features from text data columns, building feature engineering for categorical features and other variables, pre-processing the data, and then risk modelling for both single-step and multi-step forecasting. Data was collected from the open-source NCDC Storm Events Database, provided by the National Weather Service (NWS). This database contains statistics on personal injuries, deaths, and damage estimates, covering the United States of America. The data collection began as early as 1950 and continues through to 2022. It includes a chronological listing, by state, of hurricanes, tornadoes, thunderstorms, hail, floods, drought conditions, lightning, high winds, snow, temperature extremes, and other weather phenomena. Data for each year is available on the website. It was first downloaded and then merged into one large dataset. The collected data contained 51 columns including 2 text columns, and 1,794,914 rows.

The task was to study the “flood” event type. The approach followed was to select rows with the “flood” event type, remove all missing values, eliminate redundant columns, and convert monetary values from strings to numbers for features related to the damage. As a result of the cleaning process, the dataset comprises 31 columns and 38,398 records, with timestamps from 2006 to 2022. Given that the dataset includes 2 text columns, the objective was to leverage text data to extract features for data augmentation. For this purpose, we have Scikit-LLM, a Python library that integrates large language models into the Scikit-learn framework. It

serves as a tool for performing Natural Language Processing (NLP) tasks within the Scikit-Learn pipeline and has begun incorporating OpenAI models. In this particular task, GPT-3.5 turbo (ChatGPT engine) was used for data augmentation by creating features through zero-shot text classification and vectorisation.

Starting with zero-shot text classification, the first feature created was the “predicted_sentiment_gpt” with three classes: negative, positive, and neutral. The second categorical feature created was the “predicted_risk_gpt” with three classes: low-risk, medium-risk, and high-risk. Then, I retrieved ten embedding features for each of the text columns. During the feature engineering process, new features were created by extracting year, month, day, and time from the beginning and end dates. The difference between dates was extrapolated in days and hours. Origin and destination names were merged to have an overview of the path and categorical features were encoded, grouping less relevant classes. New target variables were generated. The distance between the starting and ending points of the event was calculated using the haversine distance. Lastly, lag and rolling window mean features were created and used to capture temporal dependencies such as seasonality, where the former represents past values of data points and the latter calculates the average of a specified number of consecutive data points as a sliding window moves over the entire dataset.

The final step, before passing data into the models, involved removing outliers using the Isolation Forest algorithm and addressing large damage costs. Additionally, I’ve eliminated unhelpful features with no variability and correlated predictors to avoid multicollinearity. The prediction study of the flood event type was performed with risk modelling based on the number of direct injuries with the event and all injuries. I’ve compared actuarial models like Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) against Gradient Boosting Machine models like LightGBM, XGBoost, and CATBoost. To fine-tune models, I employed time series cross-validation with Optuna, an automatic hyperparameter optimisation software framework which uses Bayesian optimisation.

In single-step forecasting, the goal has been to predict the next time point ($t+1$) and forecast floods in 2022. In multi-step forecasting, the objective has been to predict multiple future time points, such as $t+1$, $t+2$, $t+3$, etc. One technique for achieving this is Recursive Forecasting, which involves training a single model and using it recursively for each step in the forecast horizon. With multi-step forecasting, I predicted floods for the years 2020, 2021, and 2022.

3. RESULTS AND PRACTICAL APPLICATION

Table 1. Injuries Direct – Mean Poisson Deviance performance comparison

Model	mpd_train	mpd_test
glm	0.000212	0.000092
gam	0.000137	0.000085
lgbm	0.000019	0.000034
xgb	0.000044	0.000033
catboost	0.000067	0.000060

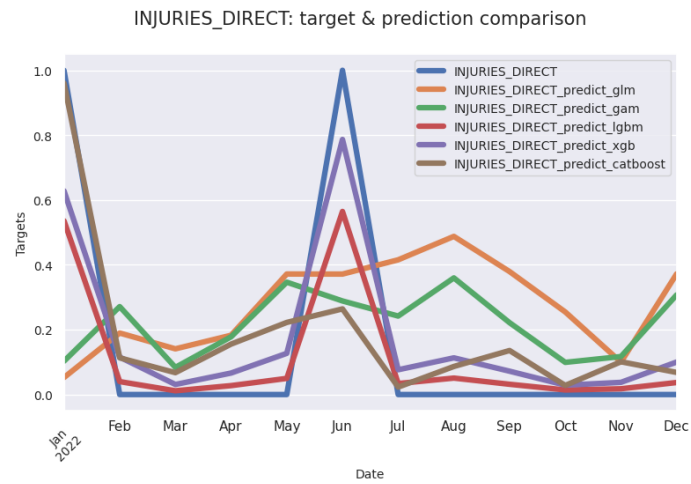


Figure 1. Injuries Direct – plot predictions comparison

Table 2. Whole Injuries – Mean Poisson Deviance performance comparison

Model	mpd_train	mpd_test
glm	0.000238	0.000099
gam	0.000259	0.000096
lgbm	0.000014	0.000024
xgb	0.000017	0.000031
catboost	0.000103	0.000050

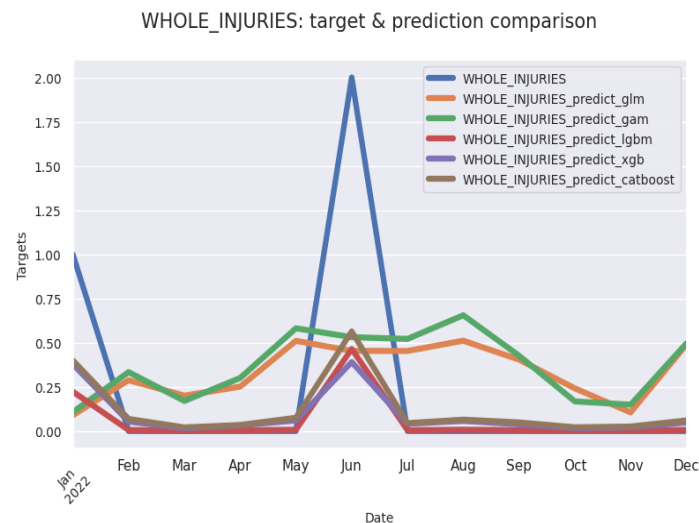


Figure 2. Whole Injuries – plot predictions comparison

The plot results and Mean Poisson Deviance metrics for direct injuries and whole injuries from the single-step prediction indicate that Gradient-Boosting models, particularly LightGBM and XGBoost show better performances.

Addressing climate risk events requires big data to enhance prediction and understand the features involved in these events. This study showcases the potential of integrating modern Machine Learning models with Generative AI to improve climate change prediction. The results highlight the importance of feature engineering in general like the role of features extracted and generated by Large Language Models (LLMs).

Modern Discriminative Machine Learning models and Generative models can be used together to improve the mapping of high-risk zones, thereby providing more accurate estimates of the insurance risk premium.

4. MAIN FINDINGS

With prediction analysis between models with data augmentation and models without data augmentation, Gradient Boosting models can capture information from text features, thereby improving performance by achieving a better fit. In multi-step forecasting, Gradient Boosting models demonstrate greater robustness over the three forecasting periods. The advantage of using multiple data types lies in improving performance and understanding which variables are most impactful. The future challenge is that forecasts pass through the management of multiple types of data, everyone with specific peculiarities in the pre-processing of the data and this study is an example of this. A lesson from this study stands to improve the analysis of the text columns for retrieving new features and better management of generated embedding features from LLMs using a dimensionality reduction instead of retrieving only the first ten embedding columns. Another future challenge is being able to use multiple streaming data types for forecasting.

5. REFERENCES

- [1] Adam B. Smith (2022). "2022 U.S. billion-dollar weather and climate disasters in historical context", NOAA National Centers for Environmental Information.
- [2] P.J. Brockwell, R.A. Davis (2016). "Introduction to Time Series and Forecasting", Springer.
- [3] J. Birkmann, T. Welle (2015). "Assessing the risk of loss and damage: exposure, vulnerability and risk to climate-related hazards for different country classifications", International Journal of Global Warming.
- [4] Dineva Snezhana, (2023). "Applying Artificial Intelligence (AI) for Mitigation Climate Change Consequences of the Natural Disasters", SSRN.
- [5] Database: <https://www.ncdc.noaa.gov/stormevents/>
- [6] Repository: https://github.com/claudio1975/US_Injuries_Flood_Prediction

Use of AI for built-up area land use mapping: A study on construction sites

Athithya Seethalakshmi Loganathan (Statistics Netherlands) &
Saim Muhammad (CBS)

1. INTRODUCTION

Land use mapping is essential for sustainable urban planning, especially in dynamic environments like construction sites. The Land use map dataset, provides detailed classifications of land use. However, the update of such dataset is largely reliant on manual methods which are resource-intensive, time consuming and not scalable for areas prone to rapid land use changes.

Since there is, currently, no single source that provides for the systemic classification of new construction sites, the primary focus of this research is to develop a solution for classifying construction sites, a complex and dynamic category of land use [1]. Given the rapid changes in terrain and structure in these areas, traditional methods struggle to keep up with real-time transitions. The current research aims to integrate **Artificial Intelligence (AI)** and **Earth Observation (EO)** data to develop an automated framework for the classification of construction sites and other built-up areas [2]. The study aims to provide a more efficient, accurate and scalable method for updating land use maps by using high-resolution aerial imagery and advanced deep learning models. Additionally, this study addresses the challenge of differentiating construction sites from other vegetated areas by leveraging **NDVI (Normalized Difference Vegetation Index)** and the **Segment Anything Model (SAM)**.

2. METHODOLOGY

The proposed methodology revolves around a '**ResNet50-based deep learning model**' for classifying construction site polygons. The model was fine-tuned specifically to handle the complex shapes and varying sizes of polygons associated with construction sites. **Pre-trained on ImageNet**, ResNet50 has an efficient feature extraction process, which has been further optimized for classifying construction site polygons by employing additional layers such as **Global Average Pooling** and **Dropout layers**. These layers prevent overfitting and ensure robustness in the classification.

Moreover, during the prediction phase, the methodology also accounts for **grass-covered areas** within these polygons. This is because, construction sites tend to span over a long period of time resulting in grass-cover within the polygon. To handle such complexities within polygons and to accurately segment such variations, NDVI has been used alongside a foundation model (SAM) for segmenting dense grass cover area [3]. This process helps differentiate vegetated regions from active construction zones within a polygon thus improving the overall prediction accuracy. In this research, a definitive prediction score has been designed to account for these modalities in these construction sites and they have been weighted accordingly.

2.1. Data Preparation

The following steps were undertaken for the preparing the data for the current study.

- **Bounding Box Calculation:** For each polygon representing a construction site, a bounding box has been calculated. This is to ensure a standardization in the dimensions of the polygon and ensure that the entire polygon is contained within the extracted image.
- **Aerial Imagery:** High-resolution aerial images (25 cm) from 2017 were obtained via a Web Map Service (WMS). These images provided sufficient details for accurately segmenting and classifying the construction sites. While the use of satellite data was considered, it was concluded that aerial imagery offered higher resolution, thus making it the optimal choice for this phase of the research.

2.2. Model Tuning

Post the preparation of data, the following adjustments were made to the ResNet50 model to ensure accuracy of predictions, improve efficiency and avoid overfitting.

- **Fine-tuning:** The ResNet50 model has been fine-tuned to customize the pre-trained architecture for the specific task of classifying construction sites. The final layers of the model have been replaced with a **Global Average Pooling layer**, a **dense layer** with **dropout**, and a **sigmoid activation** to perform binary classification.
- **Mixed Precision Training:** To optimize the training process of the ResNet50 model, and to reduce the usage of memory, a mixed precision training has been performed. This training helped accelerate the computations on the GPU without sacrificing accuracy.
- **Early Stopping and Model Checkpoints:** To prevent overfitting and ensure optimal performance of the model, early stopping and model checkpointing has been applied, saving the best performing model based on validation results.
- **Threshold Optimization:** After training, a precision-recall curve has been used to determine the optimal decision threshold, ensuring a balance between precision and recall. This threshold was set at **0.4013**, providing the best trade-off for classifying construction sites while minimizing false positives.

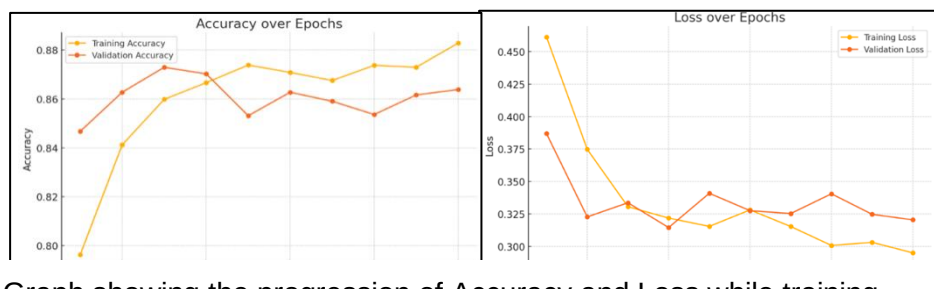


Figure 1. Graph showing the progression of Accuracy and Loss while training

2.3. Integration of NDVI and SAM for Prediction (Inference)

To further enhance the accuracy of the prediction scoring of the construction site polygons, the NDVI and the SAM were integrated into the prediction pipeline. NDVI was employed to evaluate the extent of vegetation within each polygon [4]. This step was crucial in distinguishing between construction sites and areas that contain significant vegetation cover, such as grass or trees, which could otherwise be misclassified as construction zones.

The NDVI values were calculated from CIR (Colour Infrared) imagery. The segments with an NDVI value above 0.5 were flagged as vegetation. These flagged segments were further segmented using SAM, which provided a more granular segmentation of vegetation patches within the polygons. This process allowed for accurately identifying and isolating grass patches from active construction zones.



Figure 2. The first image is an example of a construction site Polygon. The second image represents the NDVI of the construction site. The third image shows the SAM segments of NDVI. It can be clearly observed from the third image that the segments with high grass cover is well separated from the barren land segments.

From the above example, it is clear that the combination of NDVI and SAM ensured that the prediction scoring process not only relied on the base model's classification but also incorporated a more detailed assessment of the land cover within each polygon. By using NDVI to flag vegetated areas and SAM to segment them, the model was able refine its predictions, thus minimizing false positives related to vegetated construction sites.

This process directly impacts the final classification score, ensuring that construction site polygons are evaluated more holistically, considering both built-up and vegetated areas. The integration of NDVI and SAM into the prediction scoring pipeline enhances the reliability of the model, helping to ensure that interpreters receive accurate outputs that require minimal manual correction.

2.4. Prediction

A definitive classification score has been assigned for each polygon using the following formula. This score is designed to predict the likelihood the polygon being a construction site while accounting for factors such as vegetation cover for a more reliable classification.

$$\text{Final Score} = \alpha \cdot \text{Classifier Score} + \beta \cdot \left(\frac{\sum (\text{NDVI Segment Values} > 0.5)}{\text{Number of High NDVI Segments}} \right) + \gamma \cdot \left(\frac{\text{SAM Grass Segments}}{\text{Total SAM Segments}} \right) + \delta \cdot \left(\frac{\text{NDVI Grass Patch Area}}{\text{Total Polygon Area}} \right)$$

Figure 3. Formula for definitive classification score

The **definitive classification score** has been calculated using a weighted combination of factors:

1. **Classifier Score:** The classifier score is the output from the ResNet50 model which indicates the likelihood of the polygon being a construction site.
2. **NDVI Segment Values:** NDVI values greater than 0.5 were included in the classification to determine the extent of vegetation.

3. **SAM Grass Segments:** The number of grass segments identified by SAM, indicating areas of dense vegetation within the polygon.
4. **Grass Patch Area:** The proportion of grass patches relative to the total polygon area was calculated to further refine the classification.

The combination of these metrics allowed for an accurate and comprehensive classification process, minimizing errors and improving the efficiency of the overall land use mapping process. This approach is scalable and offers a potential solution for future iterations of the landuse dataset, reducing the reliance on manual processes and enabling real-time updates for construction site classification.

3. RESULTS

By closely examining the performance of the model and the prediction results, it can be seen that there is a definite improvement in accuracy and increased effectiveness in the classification of construction sites. The following results can be observed:

- **Steady Improvement in Accuracy:** Both training and validation accuracy show consistent improvement across the epochs, reaching close to 88% and 86%, respectively, indicating successful learning.
- **Stabilization of Loss:** The training and validation loss decrease rapidly in the initial epochs and stabilize around the middle of the training, showing that the model is converging well.
- **Close Tracking Between Training and Validation:** Both accuracy and loss for training and validation are closely aligned, indicating that the model is not significantly overfitting and is generalizing well to unseen data.

Metric	Value
Training Accuracy	0.882
Validation Accuracy	0.863
Training Loss	0.295
Validation Loss	0.320
Best threshold	0.4013

Table 1. Training metrics of the classification model

The use of NDVI and SAM played a significant role in enhancing the accuracy, particularly by excluding green areas from being misclassified as construction zones. The model was able to identify construction site polygons with a high degree of accuracy, while the integration of NDVI helped to flag and exclude areas with significant vegetation.



Figure 4. Classified construction sites with scores

4. MAIN FINDINGS

The following key findings highlight the effectiveness of the methodology adopted in this research and its potential for improving land use classification with respect to construction sites.

1. **Enhanced Classification Accuracy:** By integrating NDVI and SAM into the ResNet50 model, the classification of construction sites was significantly improved, especially in distinguishing between green spaces and active construction zones. Grass cover was accounted for, reducing errors related to vegetation within construction polygons.
2. **Robustness and Scalability:** The AI-driven approach, combined with advanced data augmentation techniques, allowed for the model to generalize well across different construction sites. While the current approach focuses on new construction sites, future iterations can extend to model built-up area classes more effectively, providing a scalable solution for future updates to the BBG dataset.
3. **Improved Model Performance:** The use of mixed precision training, early stopping, and threshold optimization ensured that the model maintained high accuracy and robustness without overfitting to the training data. Semantic segmentation could be further fine-tuned with one-shot encoding using SAM, allowing for more precise segmentation of complex land cover types. [5]
4. **Evalaution:** Out of 22 construction site polygons of a sample from 2020, 17 were correctly classified with 75% accuracy. Misclassifications mostly occurred due to buildings and shadow regions, which affected the overall classification accuracy. Although not used for direct map production, these results will be sent to interpreters to expedite feature mapping and verification, accelerating the mapping process and improving decision-making efficiency.

Metric	Value (%)
Completeness	77.27
Correctness	70.83
Precision	70.83
False Positive Rate (FPR)	0.46
F1-Score	73.97

Table 2. Prediction Metrics

5. CONCLUSION

This research provides a novel and effective approach to land use mapping, specifically targeting construction sites. The combination of AI models like ResNet50, NDVI for vegetation detection, and SAM for segmentation allowed for a more accurate and scalable solution than traditional manual methods in the case of construction sites. By automating the classification process, this approach supports faster updates to land use datasets and offers significant improvements in accuracy. The methodology developed here has potential applications for a range of land use categories and offers a pathway towards more dynamic and real-time monitoring of land use changes.

6. REFERENCES

- [1] De Simone, L., Ouellette, W., & Gennari, P. (2022). Operational use of EO data for national land cover official statistics in Lesotho. *Remote Sensing*, 14, 3294. <https://doi.org/10.3390/rs14143294>.
- [2] Sawant, S., Garg, R., Meshram, V., & Mistry, S. (2023). Sen-2 LULC: Land use land cover dataset for deep learning approaches. *Data in Brief*, 51, 109724. <https://doi.org/10.1016/j.dib.2023.109724>.
- [3] Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P., & Li, H. (2023). Personalize segment anything model with one shot. *ArXiv*, abs/2305.03048.
- [4] Thakur, T., Patel, D., Bijalwan, A., Dobriyal, M., Kumar, A., Thakur, A., Bohra, A., & Bhat, J. (2020). Land use land cover change detection through geospatial analysis in an Indian biosphere reserve. *Trees, Forests and People*, 2, 100018. <https://doi.org/10.1016/j.tfp.2020.100018>.
- [5] Osco, L., Wu, Q., Lemos, E., Gonçalves, W., Ramos, A. P., Li, J., & Junior, J. (2023). The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot.

Use of public data sources to increase the accuracy and completeness of the EuroGroups Register

Ioannis Sopranidis, Pau Gayà Riera & Alexandre Depire (Eurostat)

1. INTRODUCTION

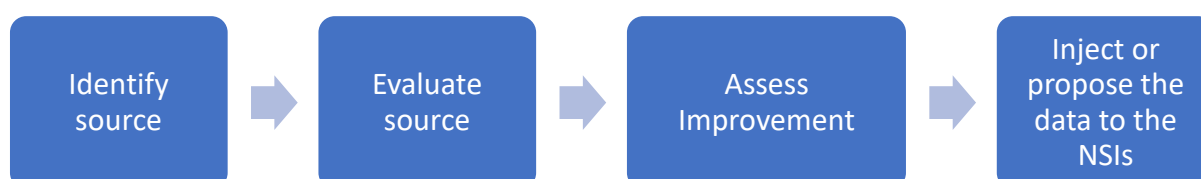
The EuroGroups Register (EGR) is the European statistical business register on Multinational Enterprise (MNE) groups created by the European Statistical System and managed by Eurostat. The main suppliers of data for the EGR are the National Statistical Institutes (NSIs) of the EU and EFTA countries together with a commercial data provider that is used mainly to provide data for non-EU & EFTA legal units.

The EGR quality is already considered good and stable to fulfil its role as the authoritative source for the production of consistent globalisation statistics. However, the statistical domains which are interested to use the EGR as a core infrastructure for their analysis on multinational enterprise groups are increasing, and new experimental statistics are requested to serve policy needs. To move in this direction, there are still some gaps to cover, particularly related to missing economic variables (turnover, assets) at group level, and to improve the non-EU & EFTA part of MNE groups. By improving further its accuracy and completeness, the EGR will strengthen and enlarge even more its role and will become an enriched and trusted infrastructure for the production of new statistics on multinational enterprise groups.

To improve further the EGR data, different activities are ongoing which can help to cover data gaps by filling missing values and improving the delineation of the groups' structure while supporting the monitoring of events and changes that may occur in these MNE groups. During the NTTS 2023, Eurostat presented a preparatory study, which led to an article published in the SJIAOS [1] concluding that public sources can be used in the EGR to support users and improve the quality of data of the MNE groups. In the current abstract we present the process used to qualify the public data sources before they are selected for the EGR, how they are used in the production process of EGR, and the results obtained in terms of increased data quality.

2. METHODOLOGY

The process of qualifying the use of new public data sources in the EGR is split in 4 main steps:



Each step of the process requires specific actions, to finally decide that the new data source is of value and can be used to complement and improve the EGR quality.

Public data sources are assessed in terms of relevance of the available information, as well as by identifying technical limitations imposed by the source on retrieval of the content of interest (e.g., resources required to collect and process the data).

When a potential pool of data sources is identified, this data source must satisfy a list of criteria, to invest in integrating it in the EGR:

1. The new data source is from an official government body or international organization (e.g., <i>Companies House</i> from the UK government).	☐
2. Other NSIs (or international organizations) are using the same data source.	☐
3. The new data source is more frequently updated than another (already used) source by at least one full EGR cycle (e.g., one of the data sources used is only updated once every 2 years, while another is updated every quarter. This means that the one that is updated every quarter should be considered as a better source).	☐
4. The new data source contains units that are not captured by another (already used) source; units which could be potentially part of a multinational enterprise group.	☐
5. The new data source contains more information that is relevant (e.g., employment or turnover) to the EGR production than another (already used) source.	☐
6. The new data source contains the same information than another (already used) commercial data source and, at the same time, reduces the financial cost associated with the collection and processing.	☐

Qualified data for non-EU & EFTA legal entities can be integrated directly in the EGR. Alternatively, for EU-EFTA data, these are proposed to the responsible NSIs. At the moment, the communication and data exchanges with NSIs take place in the form of reports and / or datasets using the secure EDAMIS (the system used by Eurostat to exchange data) channel. A new data visualization and comparison module for the EGR producers and users, the value of which has been demonstrated by a proof of concept, is currently being designed where the NSIs will be able to compare the data coming from the external data sources to decide whether to use it or not. It is expected that trusted data sources will become more and more accepted by NSIs, with a view to regularly and automatically integrate their data directly in the EGR.

3. RESULTS AND PRACTICAL APPLICATION

Following the process to integrate the new data sources, it was decided to select and prioritise five data sources for evaluation: (i) Companies House from the British government; (ii) EDGAR from the Securities Exchange Commission in the US; (iii) GLEIF from the Global Legal Entity Identifier Foundation; (iv) the Multinational Enterprise Information Platform (previously known as ADIMA) from the OECD-UNSD; and (v) web-scraped data from Wikipedia [1].

These open data sources fulfil all or the majority of the criteria and provide several benefits to the EGR by:

- Filling in the missing variables at MNE group level (e.g., consolidated turnover, total assets, and global employment).
- Improving the delineation of the MNE groups (e.g., additional control relationships between legal units).
- Supporting the monitoring of changes or modifications in the MNE groups (e.g., mergers and acquisitions, etc.).

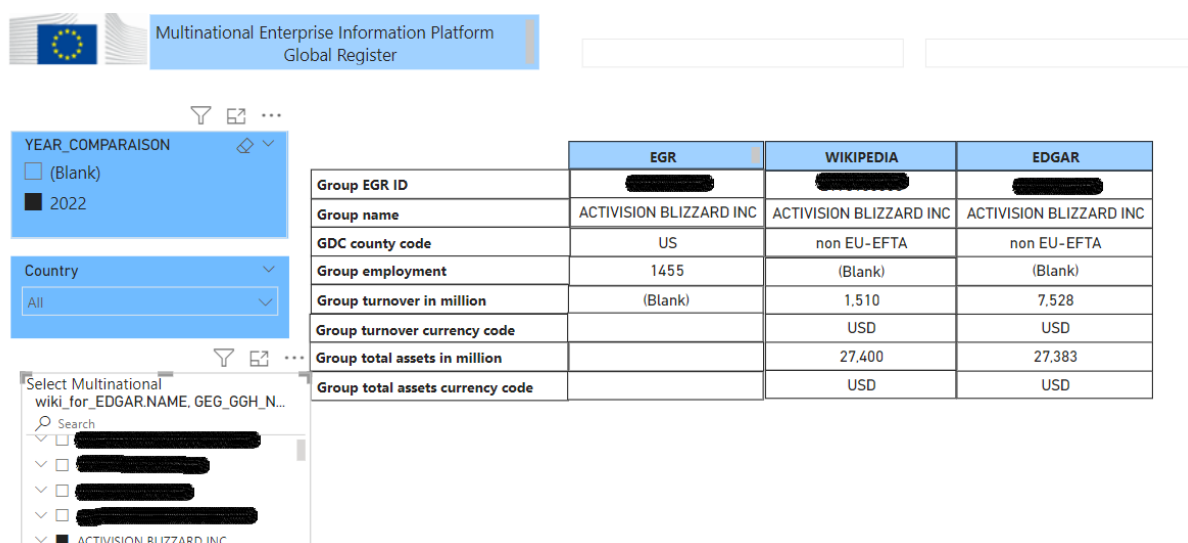
The first public data source which has been already integrated in the EGR, is the Companies House, the UK agency which registers company information. Companies House provides mainly three products:

- a register of around 5 million entities;
- the accounts and monthly updated data for employment, turnover, and activity status; and
- people/legal unit with significant control, including around 8 million relationships.

Once the matching of the public data source with the EGR data is completed, each Member State receives the information for the groups under their responsibility, and the corresponding comparison with the EGR data for their assessment. The database of OECD-UNSD contains 500 groups involving around 125 000 units, 29 000 in EU or EFTA countries. The matching process finds 66% of the units in the EU, and 43% for the whole world. In total it is possible to compare 87% of the groups proposed by this database. The results of the matching exercise between EDGAR and EGR identified a total of 1,062 groups from EDGAR with 100% similarity based on the name of the group, and about 320 candidates with more than 85% similarity. For the Wikipedia data, a total of 1,071 articles in English were identified for the EGR top-tier groups. After filtering the Wikipedia available information for the corresponding 2023 reference year, a total of 699 MNE groups can be collected, which represent around 40% of the top-tier MNE groups.

Using the new data visualization module of the EGR, the data will be summarised in dashboards where the NSIs will be able to see the differences between the EGR data and the public data sources, evaluate the quality and decide whether to use in the EGR final frame or not. Some sample dashboards have been developed in Power BI (see the example in Figure 1).

Figure 1. Power BI, data comparison tool for group information, 2022 data



The next step is to allow the EGR producers to simply select the correct data from the user interface and consolidate their data to the EGR.

4. MAIN FINDINGS

Using this process of evaluation and integration of new data sources in EGR, the accuracy and completeness of the EGR have already significantly improved: Companies House data improved the coverage of the EGR by 5 million legal units for the UK and reduced the cost of data acquisition from the Commercial Data Provider by 25% in 2023.

With the exchange of the matched data between the public data sources and the EGR data, the NSIs will have the opportunity to use data which they are missing or consider of better quality. Finally with the availability of the new EGR data visualization and comparison module, the EGR production process will gain in efficiency, by reducing costs and burden on NSIs and Eurostat.

This solution opens up to the possibility of integrating any additional new and open data source in the EGR core infrastructure in a flexible and cost-effective way (e.g. BEPS CbCR data¹, CSRD data², etc.) to progressively expand the variables available in the EGR beyond what NSIs can provide from their national data sources.

5. REFERENCES

- [1] Palma Laureti A, Bitoulas A, Depire A, Reis F, Gaya Riera P, Sopranidis I. A web intelligence information system to support the production of EuroGroups Register (EGR) statistics. Statistical journal of the IAOS. 2024;40(3):549-562. doi: 10.3233/SJI-240032

¹ Base Erosion and Profit Shift, Country by Country Reports.

² Corporate Sustainability Reporting Directive.

Use of web scraped data for analysing labour market demand for ICT specialists

Magdalena Kaminska & Cristina Fernandez Alvaro (Eurostat)

1. INTRODUCTION

The ongoing digitalisation is profoundly affecting society, businesses and the economy in general. The widespread digital technologies play an increasing role in everyday life; they transform business models, jobs and production and stimulate growth and innovation.

Digitalisation greatly affects the way businesses conduct their operations. It facilitates global outreach for business, their interaction with customers and business partners and it helps them streamline and optimise their operations.

Consequently – by changing many occupations and work-related tasks – digitalisation has an impact on the labour market. As the variety of activities engaging the information and communication technologies (ICT) is increasing, there are more types of jobs that rely hugely on digital technologies. Thus, there is a growing need for qualified personnel who are able to deal with increasingly technology- and data-intensive working environments.

At the same time, digitalisation of various jobs and activities creates a vast and constantly increasing amount of data, which gives opportunities for businesses to capitalise on the data available. This also creates new opportunities for official statistics to use non-traditional data sources for producing statistics. For instance, the vast amount of information contained in online job postings presents prospects for tracking the evolution of labour markets and skill demands [1], “hot technologies” or “in demand” technology skills [2].

ICT occupations are among the jobs with a particularly prevalent persistent labour shortages. The demand for such specialists is foreseen to grow and the increase will primarily be driven by employment growth in these occupations, followed by replacement needs as older cohorts retire [3]. Therefore, there is an urging need to measure the impact of digitalisation on the evolution of labour markets, to identify where are the shortages of specialists and which type of skills are in demand.

The presentation aims to discuss the possibility of using the non-traditional data scraped from the job postings on web portals as a source of data for measuring the labour market demand for ICT specialists.

2. METHODOLOGY

2.1. Data source

The presentation makes use of a large micro-data set of classified Online Job Advertisements (OJA) produced by the Web Intelligence Hub (WIH) of Eurostat (WIH-OJA) [4]. The online postings are web scraped from millions of ads posted in EU countries, collected from hundreds of web sources including job search engines and public employment services' websites. To the extent possible, automatic systems filter out multiple postings for the same vacancy appearing in different job portals (deduplication) to mitigate the risk of overestimation due to

double-counting. These advertisements normally include data on the characteristics of the job (e.g. occupation and location), characteristics of the employer (e.g. economic activity) and job requirements (e.g. education and skills). Part of this information is available only as natural language textual data. The retrieved job postings are classified into several variables by various algorithms (ontologies, Machine Learning models, fuzzy-matching techniques among others) and mapped to different taxonomies and, in particular to the International Standard Classification of Occupations (ISCO) at the 4 digits disaggregation level.

2.2. Indicators

Matching of the occupations listed in the job postings with the ISCO codes allows to analyse the labour market demand for ICT specialists in online job advertisements as defined by OECD and Eurostat [5].

The experimental statistics are derived from the OJAs dataset by calculating the shares of OJAs for ICT specialists in the total number of OJAs. The indicator is published quarterly and refers to a period of the last 4 quarters (i.e. is calculated on four rolling quarters) to smooth seasonal effects. The statistics include also the distribution by sub-groups of ICT occupations as well as the percentage growth of the number of OJAs for ICT specialists in a quarter compared to the same period of the previous year.

For deriving the statistics on the labour market demand for ICT specialists in OJAs, the number of OJAs published during the reference period is used.

It is important to note that the data from OJAs should not be compared to the job vacancies statistics. Firstly, the source dataset of OJAs, which is used for producing indicators on labour market demand for ICT specialists in OJAs, is referring only to the online postings of job positions and does not cover job offers advertised outside of the internet. Secondly, OJAs are not equivalent to job vacancies. The 2 concepts are related, but there may be vacancies that are not advertised, or some employers post job advertisements even if they don't intend to fill their vacancy immediately or within a specific period of time, or it might happen that one online job advertisement refers to several vacancies at the same time.

3. RESULTS

Between the second quarter of 2023 and the first quarter of 2024, 7.9% of online job advertisements in the EU were dedicated to ICT specialists, with the highest demand recorded in Slovakia (15.6%), Cyprus (14.5%) and Austria (14.4%).

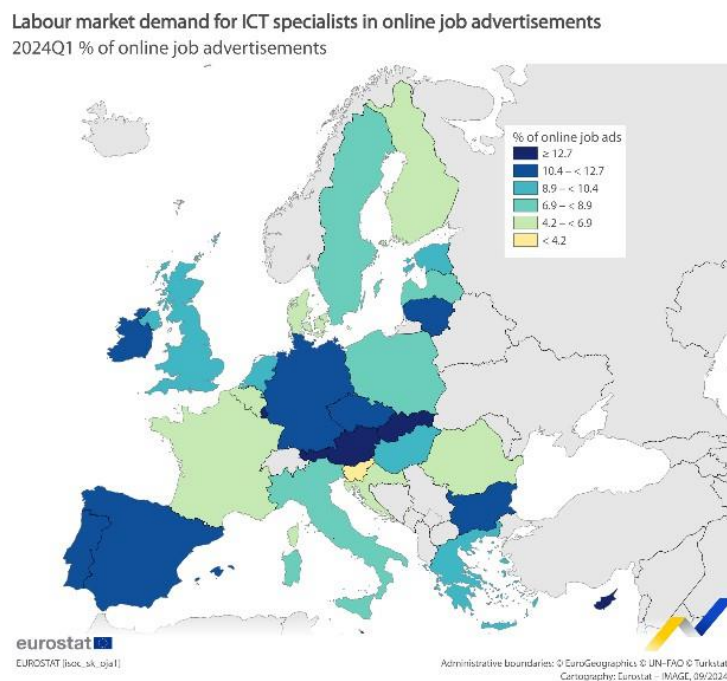


Figure 1. Labour market demand for ICT specialists in online job advertisements, 2024Q1, % of online job advertisements.

The statistics produced based on web scraped data show also a more granular view of the demand for ICT specialists, with the data available at regional level. It allows to identify in which region is the highest concentration of job postings seeking ICT specialists.

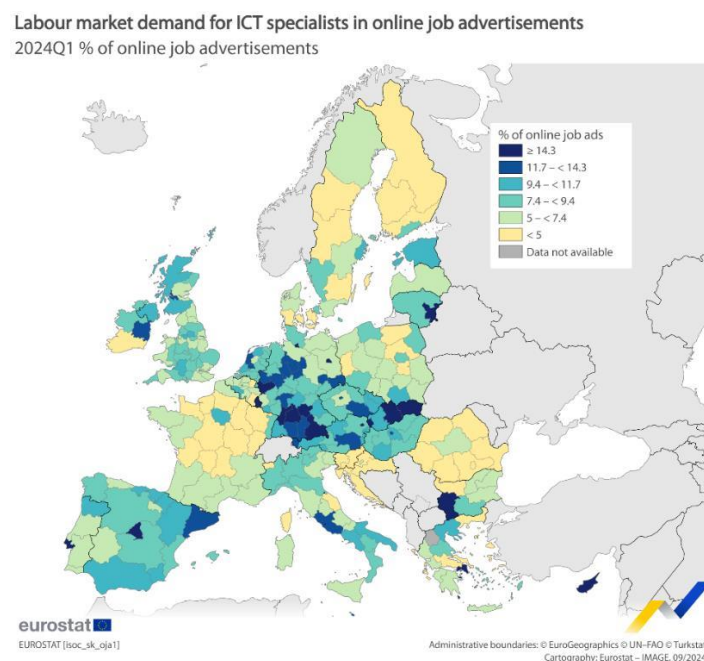


Figure 2. Labour market demand for ICT specialists in online job advertisements by NUTS 2 region, 2024Q1, % of online job advertisements.

3.1. Challenges and opportunities

The use of non-traditional data sources poses some challenges. The coverage of the web portals is not exhaustive, meaning the web scraping does not cover all potential web portals posting job advertisements. Thus, the selection of web portals might have a significant impact on the quality of the data output and the comparability across time and countries. This makes the dataset sensitive to the changes in the list of websites scraped.

Furthermore, deduplication of job advertisements, which involves the identification and removal of duplicate entries within the dataset, may cause potential errors, specifically in the semantic deduplication stage of the pre-processing phase. However, the algorithms are progressively improved to eliminate redundant OJAs.

Moreover, it can be assumed that the OJA source is biased with regard to economic activities and occupations: certain positions are likely to be advertised online more than others. For example, IT positions are more likely to be advertised on the internet than jobs in small shops, such as butchers or bakeries, or in the retail trade sector. Some studies of the representativeness of the WIH-OJA data and the occupation variable have been done [6].

Nevertheless, the use of non-traditional data sources presents also new opportunities. It allows for more granularity, for instance regional level, or information about the type of skills employers look for. That shows potential for the future work to analyse which types of digital skills are in demand, both within the ICT occupations and in other jobs.

4. MAIN FINDINGS

The experimental statistics on the labour market demand for ICT specialists in online job advertisements (OJAs) complement the existing statistics on ICT specialists and digital skills. Compared to many statistics based on official sources, the use of big data allows for increased timeliness and some additional granularity, like for instance regional perspective. However, it still presents many methodological shortcomings which call for their use with caution.

5. REFERENCES

- [1] OECD (2022), "Online job postings as a data source to analyse the impact of digitalisation on labour markets", in Skills for the Digital Transition: Assessing Recent Trends Using Big Data, OECD Publishing, Paris, <https://doi.org/10.1787/7d99dfbe-en>.
- [2] P. Lewis and J. Morris, National Center for O*NET Development, Hot Technologies and In Demand Technology Skills within the O*NET System, (2022), [Hot Technologies and In Demand Technology Skills within the O*NET System \(onetcntr.org\)](https://onetcntr.org/HotTechnologiesandInDemandTechnologySkillswithintheO*NETSystem)
- [3] European Commission, Directorate-General of Employment, Social Affairs and Inclusion, Employment and Social Developments in Europe 2023. Addressing
- [4] labour shortages and skills gaps in the EU, Publications Office of the European Union (2023), [KE-BD-23-002-EN-N.pdf](#), 16.
- [5] Eurostat, Reference Metadata in Euro SDMX Metadata Structure (ESMS) for WIH-OJA [isoc_sk_oja_esmsip2_an_1.pdf \(europa.eu\)](#)
- [6] OECD, EUROSTAT-OECD DEFINITION OF ICT SPECIALISTS (2015),
- [7] [isoc_skslf_esms_an_8.pdf \(europa.eu\)](#), 6-7.

- [8] Cedefop, Napierala, J., Kvetan, V. and Branka, J., Assessing the representativeness of online job advertisements, Publications Office of the European Union, 2022, <https://data.europa.eu/doi/10.2801/807500>.

Using CI/CD methods for continuous benchmarking of statistical methods

Samo Kosík & Marek Cedula

1. INTRODUCTION

Producing methodologically correct and accurate statistics oftentimes involves comparing different models or approaches and selecting the most accurate one. From developers' perspective, comparing all possible options can be time-consuming and not very efficient. Hence why we propose a framework that leverages CI/CD pipelines to automatically test and benchmark all approaches a developer considers.

2. METHODOLOGY

2.1. Model selection

To establish an efficient framework for benchmarking statistical models, we firstly need to select what the output of the model is going to look like. Building a framework that works with different types of outputs (e.g. both text and number-based outputs) would be a very complex task with very few real-world benefits.

For the purposes of this abstract, we decided to work with models that predict time series based on previous data points, namely Autoregression, Moving Average, Exponential Smoothing and ARIMA to predict GDP per capita of certain European countries and United States of America.

Whilst not necessary, we also used the same programming language - Python throughout the entire project.

2.2. Benchmarks to compare the models

Evaluating the models involves running several benchmarks on the results they produced to give relevant insights, namely Mean Absolute Error, Median Absolute Error, Mean Absolute Percentage Error and Root Mean Square Error. Apart from that, producing a top 10 errors matrix for each approach gives information about where the models made their biggest errors.

2.3. GitLab CI/CD pipelines - in general

To automate the entire process, we leveraged GitLab CI/CD pipelines. CI/CD stands for continuous integration and continuous delivery, and it is a set of tools used to automate the process of software development and deployment. Due to pipelines' nature of being able to run a certain pre-defined sequence of commands, we believe they can also be effectively used for evaluating statistical methods. The sequence of executed commands is determined by a YAML file and is run every time the developer commits changes to their code. This approach allows to monitor the effect of every single modification. Lastly, CI/CD pipelines allow creating so-called "artifacts", which are files that remain accessible even when the job finishes.

2.4. GitLab CI/CD pipelines - our specific pipeline

After all repositories are synced and dependencies installed (we use pipreqs for the sake of automatic installation of necessary libraries), models write their results to the predictions table. Once the results are recorded, we conduct the benchmarks that directly utilize the newly generated predictions.

Note that the predictions and benchmarks are re-generated after every commit, even if the models remain unchanged. Additionally, we also run certain helper scripts to ensure coherent style in CSV tables.

3. RESULTS AND PRACTICAL APPLICATION

Through our work, we showcased that the usage of CI/CD pipelines can be extended beyond testing whether the program works to measuring the statistical validity of it.

The outputs of our framework consist of several files, also known as artifacts. These include our metrics file, where all benchmarks are written in JSON format, the top 10 errors table written in CSV format for every single model and the predictions table based on which the benchmarks were run.

The practical applications of our framework mainly include automation of repetitive tasks. When the developer introduces a new approach, they no longer have to manually test it and compare it with the other ones. All this will be automatically executed by CI/CD runners. The only requirement is to make sure that all the models produce outputs in a consistent way.

Figure 1. Artifacts

Name	Size
..	
metrics.json	764 B
predictions.csv	2.22 KiB
top_errors_ar.csv	491 B
top_errors_arima.csv	491 B
top_errors_es.csv	464 B
top_errors_ma.csv	386 B

Figure 1 shows the folder structure and files the framework produces as artifacts. In addition to benchmarking files, the models' predictions are also recorded.

Figure 2. Benchmark results for all models

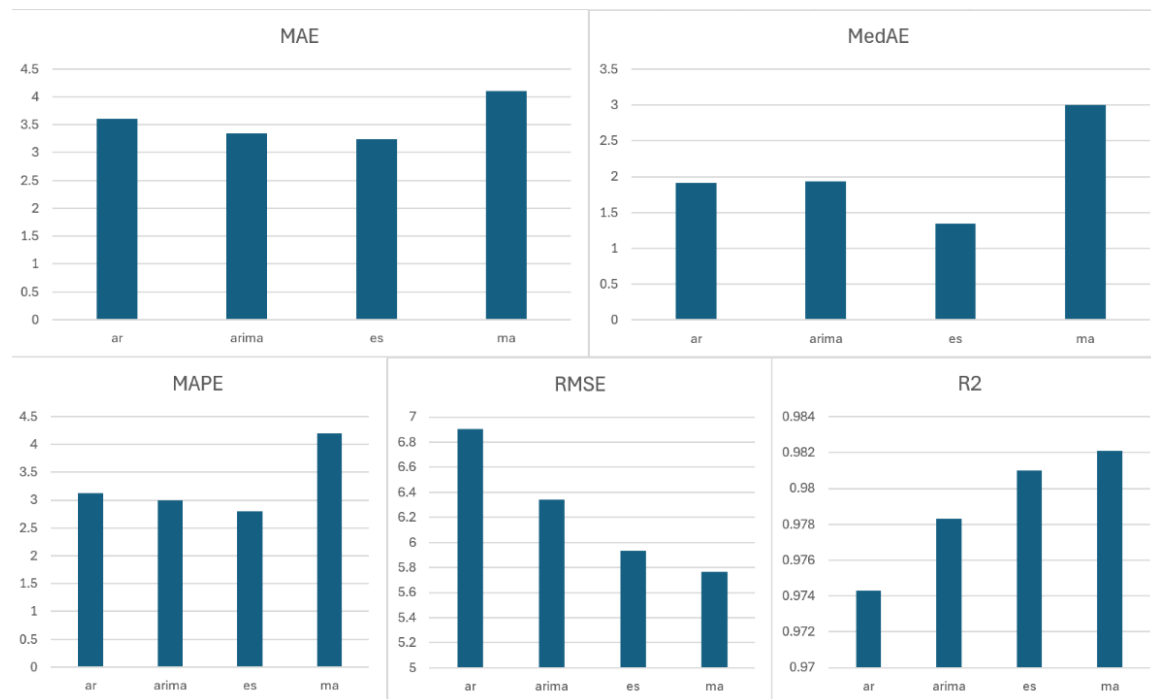


Figure 2 shows results for every single compared model in a graph. These values are written in artifacts as a JSON file.

Table 1. Top 10 errors for ARIMA

The Table 1 shows top 10 errors for ARIMA model; hence it prints the countries' codes where the ARIMA model was most wrong. A similar table is created for every single model.

Country	Error	Real Value	Predicted Value
IE	28.7735246	211	239.7735246
LU	12.7010898	239	251.7010898
DK	9.58938468	127	136.5893847
RO	5.99605385	80	74.00394615
PT	5.05088973	83	77.94911027
LT	4.44857583	86	90.44857583
EE	4.42008578	81	85.42008578
ME	2.956315	52	49.043685
DE	2.81438209	115	117.8143821
BG	2.71567429	64	66.71567429

4. MAIN FINDINGS

In this abstract, we have shown that CI/CD pipelines can be practically used beyond their intended purpose, which is testing software before it goes into production, and thus these pipelines are a useful tool to automatically verify multiple approaches the developer considers. Using them can significantly save time

On the other hand, it is worth noting that this kind of framework works well only for a single type of models, namely those ones that have numerical outputs. Modifying to an extent where it can deal with other types of models, we well, would be possible but also far more difficult to build.

Most of the abstract refers to the code1 that can be found in the References section.

5. REFERENCES

[1] S. Kosík and M. Cedula, NTTS2025, <https://gitlab.com/samokosik1/NTTS2025>

Using Cloud and Open Source for Continuous Innovation in Statistical Production

Trygve Falch (Statistics Norway)

1. INTRODUCTION

National Statistical Offices (NSOs) confront the dual task of delivering reliable data while also encouraging innovation. A framework was created to assure consistency, traceability, and reusability across the data lifetime. This framework identifies five main stable states: source data, input data, processed data, statistics, and output data. Drawing on Dumbill's "The Experimental Enterprise" (2014), the concept uses open-source tools, cloud-based infrastructure, and data science approaches to accelerate innovation while maintaining high statistical production standards.

1.1. Cloud and Innovation

Cloud-based infrastructure is used to increase scalability and flexibility, especially when dealing with non-traditional data sources and complicated statistical algorithms. The cloud enables rapid experimentation while ensuring a stable production environment.

2. METHODOLOGY

Each steady state is defined by specific, outcome-focused expectations. Data transformations between states are handled via version-controlled code (Python or R). The cloud-based platform ensures scalable data processing, is consistent with innovation objectives, and supports privacy-preserving technologies and statistical methods.

2.1. Open Source, Cloud, and Data Science.

By combining open-source frameworks and cloud infrastructure, the approach avoids vendor lock-in while also enabling advanced data science, machine learning, and AI capabilities. This facilitates ongoing modernization by allowing for the seamless integration of new technology and tools without substantial disruptions. The adaptability of cloud-native architectures guarantees scalability and continuing improvements, allowing statistical offices to iteratively modernize workflows while maintaining stable operations. Open-source tools provide a conduit for continuous innovation, guaranteeing that the system progresses alongside advances in statistical approaches.

3. RESULTS AND PRACTICAL APPLICATION

The system has been successfully implemented in a cloud-based environment, simplifying data conversions and allowing for quick iteration. This provides both adaptability and quality in the creation of official data.

3.1. Reusability and Metadata

The integration of metadata and version-controlled processes ensures reusability and traceability across all stages of the data lifecycle, enhanced by the scalability of the cloud infrastructure.

4. MAIN FINDINGS

This framework helps NSOs to use cloud technology to scale their statistical production while incorporating open-source technologies and new approaches. The capacity to swiftly iterate and continuously modernize processes guarantees that statistical outputs remain high-quality and reproducible, even as technology advances.

4.1. Continuous Improvement

By combining open-source tools, cloud infrastructure, and data science, the framework promotes continual improvement. The system supports incremental innovation and modernization, ensuring that procedures are efficient, scalable, and dependable in the face of changing statistical needs.

5. REFERENCES

[1] E. Dumbill, *The Experimental Enterprise*, Forbes, 2014.

Using machine learning to produce census population count from an administrative data system

Antonella Bernardini, Nicoletta Cibella, Giampaolo De Matteis, Antonio Laureti
Palma (Italian National Statistical Institute, Istat)

1. INTRODUCTION

In official statistics production, administrative data are crucial for coping with budget constraints and less willingness on the part of respondents to participate in surveys [1-2]. Administrative data have undoubted advantages, such as: being inexpensive; having a census-like approach to collection and can be viewed as the backbone for the census.

Since 2020, in Italy, ISTAT has produced population and housing census count estimates by relying on the administrative sources organized as 'Signs of Life' (SoL) in an ad hoc integrated database. This involves data processing of more than forty administrative archives, each of them containing basic information on individual SoL covering several years. For each administrative signal, information on the location where it took place is also recorded. Through SoL it is possible to refer to activities carried out by anonymized individuals, being self-employed or working for a company, being a civil servant, attending a school or university are examples of 'direct signs of life'. [3]

To support the census process, since 2018 ISTAT has also been conducting a yearly sample survey by collecting the main characteristics of Italian resident population and its social and economic conditions at national, regional and local levels. The permanent census of Population and Housing does not involve anymore all Italian households, but only a sample of them every year. Through integrating information from statistical sample surveys and data from SoL, the permanent census yearly provides data representing the entire population.

In Italy there is a Population Registry, based mainly on the Municipality Registry Offices, which is typically affected by under- and over-coverage of resident individuals. These mismatches can be relevant and their corrections, or error minimization, is one of the main target of the permanent census process.

The goal of this study is to use SoL to implement a supervised ML-based classification strategy to distinguish between the usual and not usual resident population in Italy in order to evaluate the over- and under-coverage counts in the Population Register (PR).

In contrast to the broad applications of ML in various scientific and statistical fields, the use of ML algorithms for population counting in census statistics is a relatively new field of application and only a few case studies are available in the scientific literature [4].

In our approach, we first pre-processing all data to build an effective feature frame and then, using the census sample survey, we set up a representative training set for a supervised classification strategy based on a Support Vector Machine (SVM) model. Quality assessment of ML results was performed, as well as an evaluation of the importance of the census sample survey data as the training set.

2. METHODOLOGY

The problem of estimating the usual or not usual resident population as a binary classification problem can be based on one or more machine learning algorithms. The selection of the most suitable ML algorithm for scoring is one of the typical steps in any ML approach. We tested several ML models and often found small differences, while small changes in training set construction made big differences. For this reason, at this stage of our study we focused more on the data preparation and identification of training/test sets rather than on applying several ML models. In choosing the ML model, we opted for the Supporting Vector Machines (SVM) model, which can easily handle the class imbalance, the resident population is much larger than the non-resident population, avoiding over- or under-sampling.

2.1. Data preparation

The main SoL considered relevant are shown in Table 1, where each sign measure the presence of individuals in the country. Each combination is 12 monthly dummy signs grouped into a normalized real number variable. The other variables represent presence/absence in the territory or categorical classifications referring to gender (sex), age, citizenship (Italian, any other European or non-EU state) and countries of birth (Italy or not Italy).

Table 1. Variable considered for the classification problem (usual / not usual resident)

Feature	Feature description
COMBI_WS	Combination of work and study signs during the year
FLAG_LAC	Presence or absence from the Municipality Registry Offices
FLAG_TAXFAM	Presence or absence in a household (tax register)
COMBI_ABR	Signs of presence abroad (Income, consular register,...)
COMBI_P2S	Combination of permits to stay from 2012 to 2021
COMBI_CONTR	Combination of signs of contracts (rental, home, car, ...)
CL_BCOUNTR	Country of birth classification (tax register)
CL_AGE	Age classification (tax register)
CL_GENDER	Male or female classification (tax register)

The training set was built using the census survey involved approximately 1.5 million individuals, allowing us to define a target variable of about 45,000 (3%) of not usual residence and 97% of usual residence. Each unit, together with its target variable, was then linked with its SoL to obtain the training set. At this level the data are raw, which means that they may present inconsistencies between different SoL within the same analysis unit. Furthermore, the training set may present a bias in the identification of not usual residents due to the implicit absence of the event.

Therefore, to build a representative training set we performed a double check on the consistency and the representativeness of the collected unit.

2.2. SVM model choice and validation

We checked three different kernel functions of the SVM model: the linear SVM kernel, the SVM-RBF (Radial Basis Function) kernel and the SVM-Polynomial kernel. The SVM-RBF kernel proved to be the best-performing model. The C regularization parameter values were calculated as proportional weights of the class distributions (3%), i.e. with $C_0 / C_1=30$ (zero and one stand for non-resident and resident respectively). The optimized value of C and γ (specific for the RBF kernel) were estimated using grid search and cross-validation algorithms. Meanwhile, the over- and under-fitting of the model was checked using the validation and the learning curves.

To verify the representativeness of collected profiles and possible bias due to sample selection, we implemented the SVM-RBF model by varying the training set. We used several training sets starting from a perfectly consistent set of units and gradually adding inconsistent units up to 50% of the initial training set. Figure 1 shows the SVM-RBF model performance: it is evident that for a training set with about 15% inconsistent units the SVM-RBF model is no longer able to follow the expected class imbalance dynamics, indicating the limit of usability of the prediction model. It is also evident how the balanced accuracy of the model decreases as the quality of the training set deteriorates.

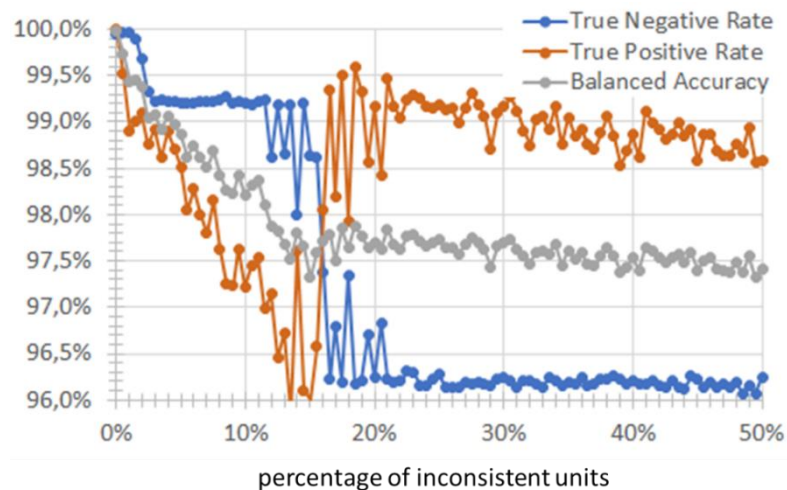


Figure 1. True Positive Rate, the True Negative Rate and the Balanced Accuracy score results of the SVM-RBF model ($C_0=0$, $C_1=30$ $\gamma=0.1$), as function of the percentage of inconsistent units

To further analyse the representativeness of the training sets, Figure 2 shows the comparison of the Balanced Accuracy curves using two test sets. The first set is again from the census survey and the second set is from a sample of the SoL universe, which uses linked values, usual/non-usual residents, from the PR as the target variable. The first test set again varies according to the number of inconsistent units. The second test set is a SoL sample invariant in size but representative of the real population present on the national territory.

From the two plots, it is evident that the difference between the balanced accuracies is larger for a perfectly consistent training set and that the difference between the two curves decreases when adding inconsistent units in the first training set. This means that the cleaning process allows the SVM-RBF model to obtain higher accuracy but at the expense of the representativeness of the training set. In fact, by introducing inconsistencies the model

accuracy decreases (Test set curve) but the representativeness of the model improves, reducing the difference between the two curves (Test set and PR sample).

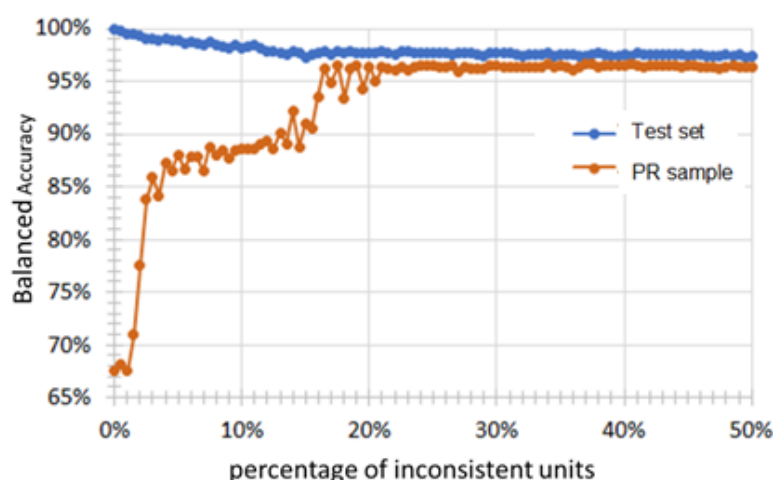


Figure 2. Balanced accuracy discrepancy between Test set and Population Register (PR) sample as function of percentage of inconsistent units.

3. ANALYSIS OF THE RESULTS ACHIEVED

The SVM-RBF prediction model, trained using a training set with about 10% inconsistent units, was applied at the national level to about 64 million individuals to predict usual and not usual resident population in Italy, allowing a complete comparison with information in the PR. In the table 2, the matched usual residents and not usual residents are the two partitions where the SVM-RBF prediction and the PR are consistent. The over-covered population is the not usual resident predicted by SVM-RBF that is resident in the PR. The under-covered population is usual resident predicted by SVM-RBF but not usual residents in the PR.

Table 1. Comparison between the SVM-RBF prediction and the 2021 Population Register on the Italian population

SVM-RBF	PR	Percentage
usual residents	usual residents	95,70%
not usual residents	not usual residents	2,38%
not usual residents	usual residents	1,29%
usual residents	not usual residents	0,64%

As expected, the results show that the usual resident population has a general tendency to be overestimated. The reason is that municipal registers record new arrivals better than outgoing ones because there is administrative pressure to register entry. On the contrary, de-registration largely depends on self-declaration by absent individuals. In addition, there could be an incentive to not de-register if absence from the country might generate a loss of benefits.

4. CONCLUSION

The work has shown a method for counting the census population using ML tools. The method was applied on a complex system of administrative data, organized in SoL, which allowed the

calculation of the over- and under-coverage counts in the Population Register. One of the critical issues addressed was the problem of imbalanced classification, i.e. the rare incidence of not usual resident versus usual resident population. This classification problem has been dealt with using a cost-sensitive approach through the SVM-RBF model.

The ML classification results were compared with the information in the Population Register, highlighting the actual over- and under-counting of the usual resident population. The work also illustrates the critical role of the training set quality in a complex information scenario such as a population census.

5. REFERENCES

- [1] UNECE, 'Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses', United Nations, New York, 2021.
- [2] A. Bernardini et al., 'Evolution of the person census and the estimation of population counts in New Zealand, United Kingdom, Italy and Israel' *Stat. J. IAOS* 38(4) (2022), 1221–1237.
- [3] A. Bernardini, A. Chieppa, N. Cibella, F. Solari, 'Administrative data for population counts estimations in Italian Population Census', "Book of short papers" *SIS* 2021, 274-278.
- [4] M. Zuppardo, V. Calian, Ó. Harðarson, 'Machine learning estimation of the resident Population', *Statistical Journal of the IAOS*, 2022, 1–14.

Using R in statistical observation of entrepreneurial subjects

Mubariz Nuriyev, Saleh Movlamov & Zarifa Naghiyeva (Centre of Scientific Research and Statistical Innovation of the State Statistical Committee of the Republic of Azerbaijan)

1. INTRODUCTION

The article shows a concrete example of using R in conducting sample observations of business entities. Practically, a sampling module was designed to conduct the sampling process according to the sampling plan using the R software system and the business register. At the same time, appropriate modules for verification of observational data, imputation of missing data and estimating of final indicators were developed. The use of R in large-scale data processing automates the production of statistical data, and also leads to an increase in the quality of the output data.

The law on official statistics of the Republic of Azerbaijan prescribes monthly, annual and multi-annual sample surveys with mandatory submission of information in the field of economics. Monthly surveys primarily pursue analytical purposes and provide information on the total volume and number of employees, etc. In contrast, annual sample surveys and follow-up surveys conducted at intervals of several years should provide a broad overview of the most important changes taking place in individual areas in the course of economic development.

Sample frame. At present, the Business Register of Entrepreneurs served as a selection tool for conducting sample surveys in the territories of the Republic of Azerbaijan. It should be noted that the register data is periodically updated with administrative data and other source. Business entities are divided into micro, small, medium and large categories depending on the number of employees. For example, micro-entrepreneurs are those with 10 employees, while small entrepreneurs are those with 50 employees.

Sampling plan. The sample design is greatly influenced by the desired grouping of the survey results and the given sectorial priorities for reflecting certain indicators. In terms of grouping the results, the sample design was focused on presenting for the first time totals in the context of a new classification of economic sectors based on the Classification of Economic Activities (NACE). At the initial stage in drawing up a sample plan, the number of employees is considered the most important statistical indicator for grouping the general population [1].

Selection method. Based on the sampling methodology, the one-stage, stratified random selection of entrepreneurs used so far was chosen. And the general population, that is, the totality of entrepreneurs who could be selected.

Stratification. By stratifying and grouping the population, the accuracy of sample survey results can be greatly improved. According to the goal setting, which says that results should be obtained by region and activity grouping, by size of entrepreneurs with a reasonable degree of accuracy, before the selection, the population was stratified by region, by type of economic activity and by number of employees.

The expediency of using one or another type of sample is predetermined by the degree of representativeness of the data that will be collected during the sample survey in accordance with its program. The representativeness of the sample data is assessed by the extent to which

the average indicators characterizing the sample set of units - n, differ from the corresponding indicators characterizing their general population - N. If these deviations are within possibly acceptable limits, then it is considered that the sample survey data represent the general population.

The absolute value of the error of representativeness depends on the variation (dispersion) of the value of the characteristics of survey units in the number of units selected for the survey, i.e. on the sample size. The mathematical error of representativeness can be expressed by the following equality [2].

$$\Delta_x = \bar{x} - \hat{x}$$

Where,

Δ_x - marginal error of representativeness;

$\bar{x}$ - the size of the trait under study in the general population;

$\hat{x}$ - the average size of the studied trait in the sample population.

In sample surveys, representativeness errors are random in nature and can have both positive and relative values. This means that if several samples were taken from the same general population of units, then in some cases the sample mean in its absolute value was greater than the corresponding value of the general mean $(\hat{x} + \Delta) > \bar{x}$ or $(\hat{x} + \Delta) < \bar{x}$

Due to the random nature of sampling errors, the representativeness of its characteristics varies with the average of representativeness errors.

With proper random selection of units, the average error is calculated by the formula:

$$\mu = \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}$$

Where, μ - mean error of representativeness;

σ^2 – the average square of the deviation, or the indicator of the fluctuation of

The value of the trait of the surveyed units;

n - the size of the sample set of units;

N - the size of the general population of units.

It should be noted that the above formula for calculating the mean error of representativeness in the actual random selection of units can practically be used for the same purpose with other types of samples.

The indicator of fluctuation of the value of this attribute of the surveyed units is calculated by the formula of the unweight average square of the deviation.

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{N}$$

Therefore, the representativeness error is measured not by the average, but by its limiting value. But there is a certain dependence between the average and limiting errors of representativeness, which follows from the law of large numbers. Based on this law, the

probability (p) that the marginal sampling error does not exceed t, a multiple of the mean error, is determined.

The mathematical relationship between the marginal and average representativeness errors is as follows:

$$\Delta = t * \mu$$

Where,

Δ - marginal error of representativeness;

μ - mean error of representativeness;

t — sampling error multiplicity

Coefficient of variation: $CV = \mu / \bar{x}$

Coefficients of sustainability: $CS=1- CV$

Margin of error: $\Delta_{\bar{x}} = t * \mu = \pm t * \frac{\sigma}{\sqrt{n}}$

Limit of average value: $(\bar{x} - t \frac{\sigma}{\sqrt{n}}; \bar{x} + t \frac{\sigma}{\sqrt{n}})$

These indicators allow us to estimate the degree of fluctuation of the data around the average value.

Based on the algorithm given above, the software modules designed in R are as follows[5]:

Program for conducting the selection process

```
registr.stat.unit<-read.table("c:/Canada/registr.csv", header=TRUE, sep=",")
registr.stat.unit
edit(registr.stat.unit)
str(registr.stat.unit)
sample.plan<-read.table("c:/Canada/sample_plan.csv",header=TRUE, sep=",")
sample.plan
str(sample.plan)
n<-length(sample.plan$numberofemployment)
n
number.employment<-as.integer(sample.plan$numberofemployment)
number.employment
number.sample <-as.integer(sample.plan$numberofsample)
number.sample
j=1
while (j<=n)
{
  number.sample_j<-number.sample[j]
  number.sample_j
  number.employment_j<-number.employment[j]
  number.employment_j
  str(number.employment_j)
  registr.stat.object_j<-registr.stat.object[registr.stat.object$numberofemployment ==
  number.employment_j, ]
```

```

registr.stat.object_j
str(registr.stat.object_j)
registr.stat.object_jj<-as.integer(registr.stat.object_j$statcode)
registr.stat.object_jj
str(registr.stat.object_jj)
sample_j<-sample(1:nrow (registr.stat.object_j), number.sample_j)
sample_j
write.table(sample_j, file = "c:/selection.csv", sep="," , append=TRUE,
col.names=FALSE,row.names=FALSE)
j=j+1
}

```

Data verification program

```

library(validate)
sample.data<-read.table("c:/sample_data.csv",header=TRUE, sep=",")
str(sample.data)
#edit(sample.data)
rules1<-validator (
  value.added<-amortization + wage + profit + taxe )
out1<-confront(sample.data, rules1)
out1
edit(out1)
summary(out1)
plot(out1)
rules2<-validator (
  intermediat.cost<- turnover-value.added )
out2<-confront(sample.data, rules2)
out2
summary(out2)
plot(out2)
rules3<-validator (
  profit <- turnover-material-amortization-wage-energy - taxe )
out3<-confront(sample.data, rules3)
out3
summary(out3)
plot(out3)
rules4<-validator(
  value.added<- turnover- intermediat.cost )
out4<-confront(sample.data, rules4)
out4
summary(out4)
plot(out4)
rules5 <- validator (
  taxes.valueadded <- turnover*0.18 )
out5 <- confront(sample.data, rules5)
out5
summary(out5)
plot(out5)
rules6 <- validator (
  profit.taxe <- profit*0.20 )
out5 <- confront(sample.data, rules6)
out6
summary(out6)
plot(out6)
rules7 <- <- validator (

```

```

        aktiv.value <- amortization*1.25 )
        out7 <- confront(sample.data, rules7)
        out7
        summary(out7)
        plot(out7)
rules8 <- validator (
        aktiv.taxe <- active.value*0.20 )
        out8 <- confront(sample.data, rules8)
        out8
        summary(out8)
        plot(out8)

```

Building a regression model for imputation

```

library(readxl)
library(sampling)
library(R2BEAT)
sample.data <- read.csv("C:/sample_data.csv", header=FALSE, sep="," )
str(sample.data)
edit(sample.data)
sample.data.frame<-
data.frame(sample.data$turnover,sample.data$material,sample.data$energy,sample.data$w
age,sample.data$amortization, sample.data$taxes,sample.data$profit)
model.turnover_1 <- lm(sample.data.frame$turnover~sample.data.frame$material)
a0<- model$coefficients[1]
a1<- model$coefficients[2]
model.turnover_1<-a0 + a1 *sample.data.frame$material

model.turnover2 <- lm(sample.data.frame$turnover~sample.data.frame$energy)
a0<- model$coefficients[1]
a1<- model$coefficients[2]
model_turnover<-a0 + a1* sample.data.frame$energy

```

Estimating of observational data

```

j=1
while (j<=11)
{
number.employment<-c(0,1,2,3,4,5,6,7,8,9,10)
i=number.employment[j]
i
sample.plan<-read.csv("c:/sample_plan.csv",header=TRUE, sep="," )
sample.plan
str(sample.plan)
number.employment<-as.integer(sample.plan$numberofemployment)
number.employment
#edit(number.employment)
number.object <- as.integer(sample.plan$numberofobject)
number.object

number.sample <- as.integer(sample.plan$numberofsample)
number.sample

data.frame<-data.frame(number.employment,number.object,number.sample)
data.frame
str(data.frame)

```

```

object0=number.object[1]
object1=number.object[2]
object2=number.object[3]
object3=number.object[4]
object4=number.object[5]
object5=number.object[6]
object6=number.object[7]
object7=number.object[8]
object8=number.object[9]
object9=number.object[10]
object10=number.object[11]

sample.data<-read.csv("c:/sample_data.csv",header=TRUE, sep=",")
sample.data
str(sample.data)

sample.data0 <- sample.data[sample.data$number_of_employment==0,]
n0<-nrow(sample.data0)
n0
mean.turnover0 <- mean(sample.data0$turnover, na.rm=T)
mean.material0 <-mean(sample.data0$material, na.rm=T)
mean.amortization0 <-mean(sample.data0$amortization, na.rm=T)
mean.wage0 <- mean(sample.data0$wage, na.rm=T)
mean.energy0 <- mean(sample.data0$energy, na.rm=T)
mean.taxes0 <- mean(sample.data0$Taxes, na.rm=T)
mean.profit0 <- mean(sample.data0$profit, na.rm=T)

if (n0>0)
{
standart.errorofmean.turnover0 <- sd(sample.data0$turnover)/sqrt(n0)}
standart.errorofmean.turnover0 <- sd(sample.data0$turnover)/sqrt(n0)
standart.errorofmean.material0 <- sd(sample.data0$material)/sqrt(n0)
standart.errorofmean.amortization0 <- sd(sample.data0$amortization)/sqrt(n0)
standart.errorofmean.wage0 <- sd(sample.data0$wage)/sqrt(n0)
standart.errorofmean.energy0 <- sd(sample.data0$energy)/sqrt(n0)
standart.errorofmean.taxes0 <- sd(sample.data0$taxes)/sqrt(n0)
standart.errorofmean.profit0 <- sd(sample.data0$profit)/sqrt(n0)
}
If(object0>0 & turnover0 >0) { total.turnover0<- object0*mean.turnover0}
If(object0>0 & material0>0) {total.material0<- object0*mean.material0}
If (object0>0 & amortization0>0) { total.amortization0<- object0*mean.amortization0}
If (object0>0 & wage0>0) {total.wage0<- object0*mean.wage0}
If (object0>0 & aenergy0>0) {total.energy0<- object0*mean.energy0}
If (object0>0 & taxes0>0) { total.taxes0<- object0*mean.taxes0}
If (object0>0 & profit0>0) { total.taxes0<- object0*mean.profit0}

If (standart.errorofmean.turnover0>0 & object0>0)
{ total.errorofmean.of.turnover0 <- 1.96*standart.errorofmean.turnover0 * object0}
If (standart.errorofmean.material0>0 & object0>0)
{ total.errorofmean.of.material0 <- 1.96*standart.errorofmean.material0 * object0}
If (standart.errorofmean.amortization0>0 & object0>0)
{ total.errorofmean.of.amortization0 <- 1.96*standart.errorofmean.amortization0 * object0}
If (standart.errorofmean.wage0>0 & object0>0)
{ total.errorofmean.of.wage0 <- 1.96*standart.errorofmean.wage0 * object0}
If (standart.errorofmean.energy0>0 & object0>0)

```



```

{total.errorofmean.of.energy0 <- 1.96*standart.errorofmean.energy0 * object0}
If (standart.errorofmean.taxes0>0 & object0>0)
{total.errorofmean.of.taxes0 <- 1.96*standart.errorofmean.taxes0 * object0}
If (standart.errorofmean.profit0>0 & object0>0)
{ total.errorofmean.of.profit0 <- 1.96*standart.errorofmean.profit0 * object0}

sample.data1 <- sample.data[sample.data$number_of_employment==1, ]
sample.data1
n1<-nrow(sample.data1)
mean.turnover1 <- mean(sample.data1$turnover, na.rm=T)
mean.material1 <- mean(sample.data1$material, na.rm=T)
mean.costofmaterial1 <-mean(sample.data1$material, na.rm=T)
mean.amortization1 <-mean(sample.data1$amortization, na.rm=T)
mean.wage1 <- mean(sample.data1$wage, na.rm=T)
mean.energy1 <- mean(sample.data1$energy, na.rm=T)
mean.taxes1 <- mean(sample.data1$Taxes, na.rm=T)
mean.profit1 <- mean(sample.data1$profit, na.rm=T)

if (n1>0 )
{
standart.errorofmean.turnover1 <- sd(sample.data1$turnover)/sqrt(n1)
standart.errorofmean.material1 <- sd(sample.data1$material)/sqrt(n1)
standart.errorofmean.amortization1 <- sd(sample.data1$amortization)/sqrt(n1)
standart.errorofmean.wage1 <- sd(sample.data1$wage)/sqrt(n1)
standart.errorofmean.energy1 <- sd(sample.data1$energy)/sqrt(n1)
standart.errorofmean.taxes1 <- sd(sample.data1$taxes)/sqrt(n1)
standart.errorofmean.profit1 <- sd(sample.data1$profit)/sqrt(n1)
}
If(object1>0 & turnover1 >0) {total.turnover1<- object1*mean.turnover1}
If(object1>0 & material1 >0) { total.material1<- object1*mean.material1}
If(object1>0 & amortization1 >0) {total.amortization1<- object1*mean.amortization1}
If(object1>0 & wage1 >0) { total.wage1<- object1*mean.wage1 }
If(object1>0 & energy1 >0) {total.energy1<- object1*mean.energy1}
If(object1>0 & taxes1 >0) {total.taxes1<- object1*mean.taxes1}
If(object1>0 & profit1 >0) { total.profit1<- object1*mean.profit1}

If (standart.errorofmean.turnover1 >0 & object1>0)
{ total.errorofmean.of.turnover1 <- 1.96*standart.errorofmean.turnover1 * object1}
If( standart.errorofmean.material1>0 & object1>0)
{ total.errorofmean.of.material_1 <- 1.96*standart.errorofmean.material1 * object1}
if( standart.errorofmean.amortization1>0 & object1>0)
{total.errorofmean.of.amortization_1 <- 1.96*standart.errorofmean.amortization1 * object1}
If (standart.errorofmean.wage1>0 & object1>0)
{ total.errorofmean.of.wage_1 <- 1.96*standart.errorofmean.wage1 * object1}
If(standart.errorofmean.energy1>0 & object1>0)
{ total.errorofmean.of.energy_1 <- 1.96*standart.errorofmean.energy1 * object1}
If (standart.errorofmean.energy1 >0 & object1>0)
{ total.errorofmean.of.taxes_1 <- 1.96*standart.errorofmean.taxes1 * object1}
If( standart.errorofmean.profit>0 & object1>0)
{ total.errorofmean.of.profit_1 <- 1.96*standart.errorofmean.profit1 * object1}

total.turnover<- total.turnover0+total.turnover1
total.errorofmean.turnover<- total.errorofmean.turnover0+total.errorofmean.turnover1

total.material<- total.material0+ total.material1

```

```
total.errorofmean.material<- total.errorofmean.material0+total.errorofmean.material1

total.energy<- total.energy0+total.energy1
total.errorofmean.energy<- total.errorofmean.energy0 + total.errorofmean.energy1

total.taxes<-total.taxes_0+total.taxes1
total.errorofmean.taxes<- total.errorofmean.taxes0+total.errorofmean.taxes1

total.profit<-total.profit0+total.profit1
total.errorofmean.profit <- total.errorofmean.profit0+total.errorofmean.profit1
```

2. RESULT

The use of large-scale data (census, statistical register, sample statistical surveys, etc.) in statistical research, the transfer of statistical data processing to digital technology remains an urgent problem of improving data quality. One of the modern solutions to this problem is the digitization and automation of the data processing process.

Currently, various software packages (Access, Excel, SPSS, SAS, STATA, etc.) are used in the process of statistical data processing. However, the level of automation of the data processing process in these software systems is low, and since it does not meet the requirements of digital technology, it is necessary to use the R software package.

Due to the fact that the program codes compiled in the R programming system are open, you can control the quality of the output data.

In view of the foregoing, the R software package was used for statistical observation of business entities, and the corresponding codes were used at the main stages of the data processing process (selection, verification, imputation and evaluation of observation units).

3. USED READING

- [1] I. I. Eliseeva, M. M. Yuzbashev. GENERAL THEORY OF STATISTICS. Moscow. Finance and Statistics 2004.
- [2] Central Statistical Office. Manual on statistics for district and city inspectorates of state statistics. Statistics. Moscow. 1970.
- [3] The Analysis of Household Surveys. A Microeconometric Approach to Development Policy. Reissue Edition with a new preface. Angus Deaton. Winner of the 2015 Nobel Prize in economics. World Bank group.
- [4] 2.Department of Economic and Sosial Affairs. Statistical division. Studies in methods series F. N 96. Households sample surveys in developing and transition countries. United Nations. New York- 2005
- [5] <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>
- [6] https://unstats.un.org/unsd/hhsurveys/pdf/household_surveys.pdf
- [7] https://unece.org/fileadmin/DAM/stats/publications/2015/ECE_CES_39_WEB.pdf
- [8] <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=1105>

Using Shape-Based Time Series Clustering to construct Machine Learning Forecasting Models for Short Panel Data: A Case Study on the Sustainable Development Goals

Manveer Kaur Mangat (UNIDO)

The Sustainable Development Goals (SDGs), adopted by the United Nations in 2015, provide a comprehensive framework to address critical global challenges such as poverty, inequality and environmental sustainability, with the aim of promoting well-being and prosperity for all by 2030. Predicting progress towards SDGs is essential for enabling proactive policy-making and efficient resource allocation, ensuring that interventions are effectively targeted to areas of greatest need. This paper proposes a two-step process for constructing machine learning models to predict SDG indicators. The first step involves a shape-based clustering method that groups countries with similar underlying characteristics, creating more homogeneous clusters for analysis. In the second step, separate machine learning models based on XGBoost and LSTM are built for each cluster, tailored to the specific characteristics of the countries within each of these groups. We apply this workflow to indicator 9.2.1, which measures manufacturing value added as a proportion of GDP. Our comparative study demonstrates that the machine learning models developed using this two-step process significantly outperform the classical ARIMA and exponential smoothing model.

1. INTRODUCTION

At the time of writing, the challenge of short time series data for SDGs, covering only 23 years (2000 - 2023) in the UNSD database, is further exacerbated by a significant number of missing values for certain indicators. Accordingly classical linear statistical time series forecasting models such as ARIMA models are not well suited as they require sufficient data to 1) identify the appropriate order of the model and 2) to estimate the model parameters. An alternative approach would be to consider machine learning based prediction models, which over the years, have established themselves as viable contenders to classical forecasting models [1]. These models are nonparametric and nonlinear and use only historical data to learn the stochastic dependency between the past and the future values [2]. We leverage the panel structure of the SDG data, as well as the structural design of the training sets of machine learning algorithms, to propose a two-step machine learning prediction model for the indicators. The first step involves a shape-based clustering method that groups countries with similar underlying characteristics, creating more homogeneous clusters for analysis. In the second step, machine learning models, based on XGBoost [3] and LSTM [4], are built for each cluster separately. We apply this workflow to the SDG 9 indicator 9.2.1, which measures manufacturing value added as a proportion of GDP and compared their performance to classical time series forecasting models such as ARIMA and exponential smoothing.

2. METHODOLOGY

In this section, we propose a general machine learning pipeline for the prediction of short panel data, which we apply to SDG indicator 9.2.1, which measures the manufacturing value added as a proportion of GDP.

2.1. Data and Pre-Processing

For this comparative study, we focus on SDG indicator 9.2.1, which measures manufacturing value added (MVA) as a proportion of GDP. This data is sourced from the UNIDO database and is available for 213 countries over a span of 34 years, from 1990 to 2023. However, some countries have missing data for certain years. If a country has more than 15 missing observations, it is excluded from the analysis, resulting in the omission of four countries. For the remaining countries, we address the missing data through a straightforward imputation process: we use linear interpolation to estimate the missing values and apply forward and backward filling for the first and last missing observations, respectively.

2.2. Clustering

Since the available time series are very short (consisting of only 34 observations each), constructing a machine learning prediction model for each of the univariate series separately is highly unlikely to result in a useful model, as machine learning algorithms typically require larger datasets to effectively learn patterns and trends. Hence, we conduct a shape-based clustering method that groups countries with similar underlying characteristics together, to obtain a bigger data set for training purposes. Using the R package *dtwlust* [5], we apply partitional and hierarchical shape-based clustering algorithms with various parameter specifications and evaluate the performances using the cluster validity indices (CVI). According to multiple CVIs the best results were obtained when segregating the indicators into five clusters for most of the clustering methods. To obtain the final clusters we used the partitional clustering method with the shape-based distance measure and the shape extraction method (to specify the prototype).

2.3. Feature Extraction

As our aim is to forecast time series, the past values of the time series are natural contenders to be used as features due to their inherent ability to capture relevant patterns and trends. In our study, we consider two types of feature sets consisting of a) lag values of the time series, and b) statistics of the lag values in addition to the lags themselves. By including statistical features, we aim to capture more complex patterns and relationships within the data. One crucial decision must be made regarding the number of lags to be used as features. Using too many lags can introduce unnecessary noise, which may obscure the true signal and degrade the model's performance. Conversely, using too few lags might fail to capture the underlying dynamics of the time series, resulting in unreliable predictions. In our study, we use a rolling window of five years to define the feature set for each time series. We additionally consider the mean and variance to capture the trend and volatility in the rolling windows, respectively. The rolling window approach ensures that the time-varying nature of the time series is adequately captured while older, less relevant information is discarded.

2.4. Training and Test Set

The feature set described in the previous section is computed separately for each country within a cluster and then concatenated to create the final feature set for the regarded cluster. When dividing the data into training and testing sets for each cluster, we first define the training and testing sets for each country individually and then concatenate the resulting sets.

2.5. Tuning

In our empirical study, we use XGBoost and LSTM, to construct our prediction models. To specify the optimal machine learning algorithm, it is imperative to conduct a tuning procedure. This entails selecting appropriate hyperparameters for each algorithm, such as learning rates, tree depths, and the number of boosting rounds for XGBoost, and the number of layers, units per layer, and dropout rates for LSTM. To identify the configuration that yields the best predictive performance, we systematically explore different combinations of these hyperparameters through grid search. This tuning process ensures that each algorithm is optimized for the specific characteristics of the dataset.

2.6. Evaluation

To evaluate the model, we use an out-of-sample (OSS) performance estimation procedure which respects the temporal structure of the data: we use the initial part of the available observations to fit the model, and the last part to evaluate it. We conduct this procedure for multiple train test splits. Thereby we use the scale independent evaluation metrics mean absolute percentage error (MAPE) and mean percentage error (MPE), as they allow to compare and compute averages across heterogeneous time series, as required in our empirical study.

3. RESULTS AND PRACTICAL APPLICATION

Table 1 presents a comprehensive comparison of the time series forecasting models across the different clusters as well as the entire data set (Global) which provides a thorough insight into the accuracy (MAPE) and bias (MPE) of each model's predictions. Overall, the XGBoost models as well as the LSTM unequivocally outperform the classical time series prediction models across all clusters and evaluation metrics. The performance of the machine learning prediction models amongst each other are generally balanced, i.e. overall the number of times that LSTM outperforms XGBoost is similar to the number of times that XGBoost outperforms LSTM.

Table 1. Average performance in terms of MAPE and MPE of each of the models trained on the different clusters (Cluster 1 - Cluster 5) as well as the entire data set (Global) based on 21 repeated out-of-sample holdout iterations.

Metric	Model	Feature s	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Global
MAPE	ARIMA	-	18.37	29.43	52.63	9.93	12.59	19.89
MAPE	Exp. Smooth.	-	21.02	26.07	47.59	10.19	13.1	19.15
MAPE	LSTM	lag	10.45	15.78	21.01	5.67	5.95	13.79
MAPE	xgboost	lag	12.5	15.79	12.49	7.06	8.36	9.41
MAPE	xgboost	lag & stat	12.81	15.49	12.62	7.21	8.42	9.33
MPE	ARIMA	-	-13.11	-16.39	-22.24	2.33	-4.38	-6.51
MPE	Exp. Smooth.	-	-11.95	-12.32	-27.68	3.56	-4.94	-6.24
MPE	LSTM	lag	-6.96	-10.11	-19	3.03	-0.16	-0.18
MPE	xgboost	lag	-6.41	-6.32	-2.89	-1.16	-3.77	-3.66
MPE	xgboost	lag & stat	-6.53	-5.99	-2.91	-1.2	-3.86	-3.62

4. MAIN FINDINGS

In this paper we propose a two-step process for building a machine learning forecasting model for SDG indicators. Leveraging the panel structure of the data, we apply a shape-based clustering method to group countries with similar underlying characteristics together. Subsequently we construct separate machine learning prediction models for each of the clusters, using the LSTM and XGBoost machine learning algorithms. We then conducted an extensive empirical study comparing the performance of 1) the cluster based machine learning prediction models 2) the machine learning prediction models based on the entire data set and 3) the classical prediction models ARIMA and exponential smoothing, using the SDG indicator 9.2.1, which measures manufacturing value added as a proportion of GDP. Overall, the XGBoost models as well as the LSTM unequivocally outperform the classical time series prediction models across all clusters and evaluation metrics. The performance of the machine learning prediction models amongst each other are generally balanced, i.e. overall the number of times that LSTM outperforms XGBoost is similar to the number of times that XGBoost outperforms LSTM.

5. REFERENCES

- [1] Ahmed, N. K., Atiya, A. F., Gayar, N. E., and El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews* 29 (5-6), 594–621
- [2] Bontempi, G., Ben Taieb, S., and Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. *Business Intelligence: Second European Summer School, eBISS 2012*, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2, 62–77.
- [3] Chen, T. and Guestrin, C. (2016). “Xgboost: A scalable tree boosting system”. *Proceedings of the 22nd international conference on knowledge discovery and data mining*, pp. 785–794.
- [4] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. *Neural computation*. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
- [5] Sarda-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette 12*, 41.dtwclust package. *R package vignette 12*, 41.

Visualizing Official Statistics Using Reproducible Analytical Pipelines

Fethi Ozbek (Turkish Statistical Institute)

1. INTRODUCTION

In recent years, the demand for transparent, reproducible, and efficient processes in data analysis and visualization has significantly increased, particularly in the field of official statistics (e.g. SG, 2024, AF, 2024). This article explores the development of a Reproducible Analytical Pipeline (RAP) for visualizing official statistics, focusing on higher education employment indicators. RAPs are automated processes for generating statistical and analytical outputs. By applying software engineering best practices, these pipelines ensure results are reproducible, auditable, efficient, and of high quality (AF, 2024).

Higher Education Employment Indicators (HEEI) are based on the Education-Labor Force Database established by Turkish Statistical Institute¹ (TurkStat) by matching formal education, formal employment, occupational information, earnings and demographic information of individuals in an integrated database in order to reveal the relationship between education and the labor market. With the annual public disclosure of the NEI, the rates of higher education (associate's and bachelor's) graduates' being in formal employment in a reference month by field of education and training, their average monthly earnings, the time it took them to find their first job after graduation, and their employment rates in a field compatible with their graduation are calculated (TurkStat, 2024).

The goal of the study is to demonstrate how to create an automated, repeatable, and flexible system for data visualisation by using RAP.

2. METHODOLOGY

This study utilizes a Reproducible Analytical Pipeline (RAP) to ensure transparency, reproducibility, and efficiency in the analysis and visualization of higher education employment statistics of Türkiye. RAP integrates the entire analytical workflow—from data extraction to final reporting—into a single, automated process. By using RAP, data analysis can be consistently repeated with updated datasets, ensuring a standardized approach to processing and reporting.

The RAP framework was implemented using R programming and hosted on GitHub, enabling version control, collaboration, and code sharing. GitHub serves as a central repository where the R Shiny code and scripts for data wrangling, cleaning, and visualization are stored, allowing for easy tracking of changes and ensuring that updates or corrections are seamlessly integrated into the workflow. The use of GitHub also facilitates transparency, as stakeholders can review the code, providing an audit trail of the entire analytical process. Moreover, this

¹ The opinions and contents of the article remains the responsibility of the author, not of the Turkish Statistical Institute.

approach supports continuous integration, where automated tests validate the correctness of the analyses whenever changes are made.

The core data comes from Türkiye's official higher education employment statistics, focusing on graduates' employment outcomes. Key indicators include the registered employment rate, which measures the percentage of graduates employed in a reference month relative to the total number of graduates, the time taken to find the first job, and average earnings, which are grouped into ordered categories. The alignment between employment and education is assessed using internationally recognized standards (ISCO and ISCED-F), enabling a clear understanding of how educational backgrounds influence labor market outcomes.

Through RAP, the process of generating these insights in visualisation of statistics is fully automated, reducing human error and manual intervention, while maintaining a flexible system that can adapt to new data or analytical needs.

3. RESULTS AND PRACTICAL APPLICATION

The visualization provides an intuitive and user-friendly platform for exploring the relationship between education and employment outcomes. Policymakers, educational institutions, and researchers can use this tool to evaluate how effectively educational programs prepare students for the labor market, track trends in graduate employment, and assess income distribution among graduates. This application highlights the value of interactive tools in making complex data accessible and actionable for decision-makers. The R Shiny application for the visualization can be accessed via the following website: https://fthtrnkr.shinyapps.io/shiny_yig/. And RAP process is available in the GitHub repository (Figure 1.).

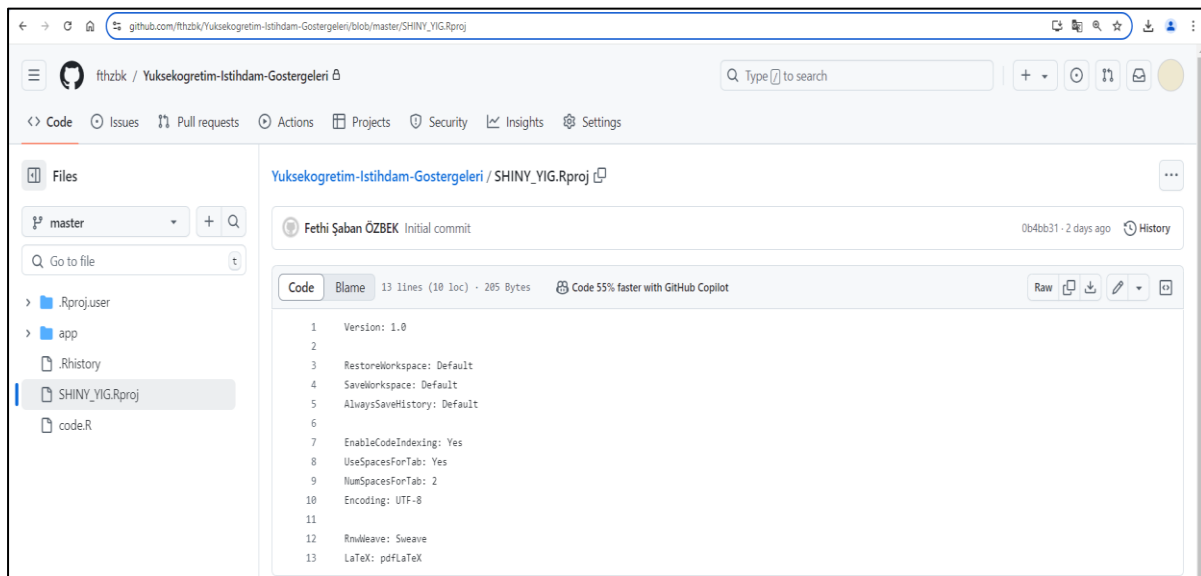


Figure 1. GitHub repository of the project

4. MAIN FINDINGS

The implementation of the RAP for visualizing higher education employment indicators resulted in several key findings. First, RAP significantly enhanced the transparency, reproducibility, and efficiency of the analytical process. By automating key steps in data extraction, analysis, and visualization, the pipeline ensured consistent results while minimizing

manual errors and enabling seamless integration of new datasets. This automated system also facilitated collaboration and version control, ensuring that updates or changes were easily incorporated into the workflow and making the entire process auditable, which is a vital aspect for official statistics.

Moreover, the user-friendly R Shiny visualization proved to be a valuable tool for stakeholders, enabling easy access to detailed data on graduate employment outcomes. This tool has practical applications for policymakers, educational institutions, and researchers seeking to assess and improve the alignment of education with labor market needs. Thus, RAP plays a crucial role in the presentation of official statistics by enhancing transparency, reproducibility, and efficiency in the analytical process, making it an invaluable asset in promoting data-driven decision-making and fostering the modernization of official statistical presentations.

5. REFERENCES

- [1] Turkish Statistical Institute, Statistical literacy (2nd version, in publication), 2024.
- [2] The Scottish Government, 2024. School education statistics.
<https://www.gov.scot/collections/school-education-statistics/#interactivedashboards> F. Lastname and F. Lastname, Article title, Journal title ## (YYYY), page-page.
- [3] The Analysis Function, 2024. (RAP) case studies
<https://analysisfunction.civilservice.gov.uk/support/reproducible-analytical-pipelines/rap-case-studies/>

WebQA Project

Marina Ayestarán Arregi & Ainhoa Vergara López (Basque Statistical Institute, Eustat)

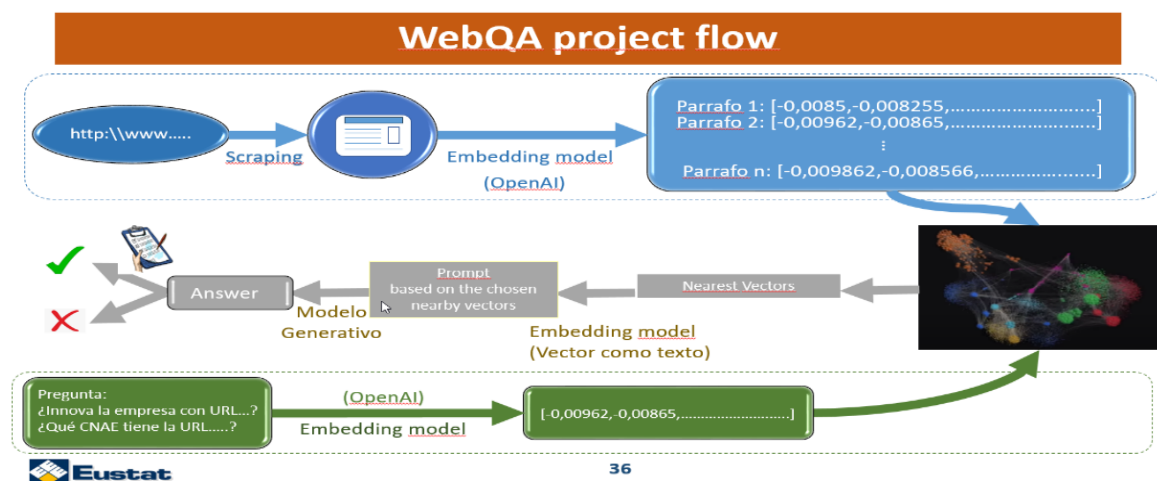
1. INTRODUCTION

Business innovation and accurate classification of economic activity according to the NACE code are key elements in generating high-quality economic statistics. These statistics provide an accurate and up-to-date view of economic realities, serving as a vital tool for informed decision-making in both public and private sectors. They contribute to enhancing business competitiveness, shaping effective public policies, and equipping citizens with the means to understand economic trends and plan for the future. In this context, ensuring the accuracy of NACE codes and evaluating the degree of innovation of companies is a priority, especially for statistical institutes like Eustat, committed to reflecting the economic reality of the Basque Country.

In this project, we combine advanced language models (LLMs) with robust web crawling and data processing tools to address two key objectives: determining whether a company innovates and accurately classifying its economic activity according to the NACE code. Leveraging the increasing web presence of companies and the potential of generative artificial intelligence, we present an innovative framework for extracting, processing, and analyzing business information directly from their websites.

2. METHODOLOGY

The WebQA project flow is as follows:



The project's phases are as follows:

Data collection

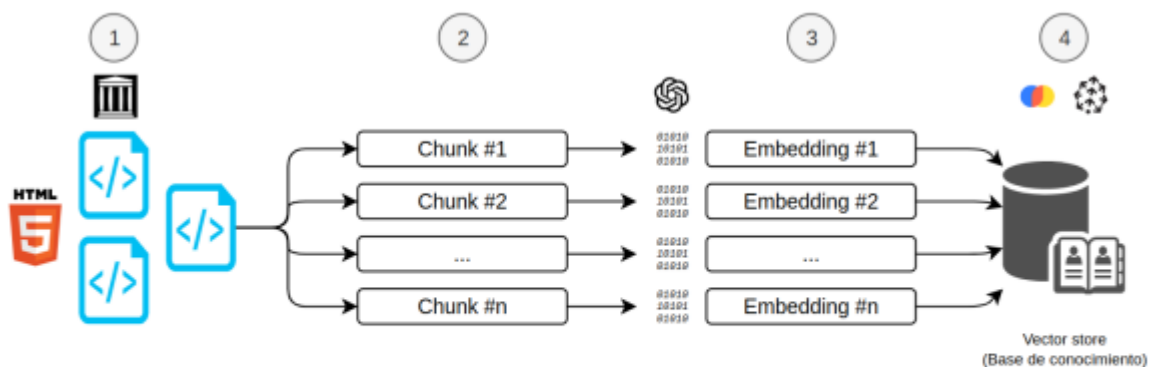
Each run of the crawler results in a WARC file. This file is designed to preserve websites and contains a 'snapshot' of the downloaded pages. This snapshot, which will serve as a source of information for the Knowledge Base, can be reproduced thanks to tools such as ReplayWeb.page.

Once the Warc file is uploaded, it will be available for browsing.

When accessing one of the uploaded files, we will see a list of all the websites it contains. For each crawled website, there will be multiple entries, one for each subpage.

Processing

The aim of this phase is to generate a database with the appropriate structure, content and quality for exploitation. To this end, appropriate workflows shall be performed to cleanse, normalise and transform the data into a structure suitable for analysis: the knowledge base. The flow chart of the processing phase is as follows:



The sub-phases of this phase are represented in the diagram by numbers. The numbers represent the following sub-phases:

Information extraction and normalisation

It extracts the contents of the Warc that are obtained by running the crawler. As it is a file containing a series of records, each of which represents an HTTP response, it must be normalised.

Normalisation consists of getting only the pure text of the HTML, removing the HTML tags and any non-textual content (such as scripts, styles, etc.).

Separation of chunks

Chunks are used to separate the content and allow you to define different lengths, in case you decide to change the model at any time.

Embeddings

Embeddings are numerical representations of words, phrases, or entire documents in a multidimensional space. These representations are fundamental to language models, as they allow the model to process and understand them in a format that the model can handle efficiently.

This high dimensionality helps capture complex language structures, such as meaning, context, or syntax. The model learns to assign a vector to each word or phrase such that words with similar meanings or usages are close in vector space.

Additionally, these embeddings allow mathematical operations to be performed on words or texts, facilitating tasks such as machine translation, text generation, and reading comprehension.

The model used in this project is the OpenAI model.

Storage and creation of the knowledge base.

The database used is FAISS, a vector database. It is an open library developed by Meta AI that allows for efficient semantic searches and vector clustering.

Modelling

The project uses the RAG (Retrieval-augmented Generation) technique, which combines the power of generative models, such as LLMs (Large Language Models), with information retrieval capabilities. This technique allows us to supplement the general knowledge of a model with specific information obtained through the crawling process, where embeddings of the collected data are created.

After creating the database with the embeddings generated from the crawling, we look for the similarity between the embedding of the question posed and those in the database. This allows us to extract only the most relevant information from the knowledge base, enabling the LLM to rely on it when generating a response to the prompt.

The project uses cosine similarity to measure this:

$$\cos(x, y) = \frac{(x, y)}{|x||y|}$$

One of the key steps in this process is choosing the optimal Prompt to ask the LLM. It is crucial to write this prompt correctly, taking into account the question, the desired output instructions, and the context of the question. From this Prompt, we will get the expected answer, supported by both the LLM's inherent knowledge and the information extracted from the database.

3. RESULTS AND PRACTICAL APPLICATION

3.1. NACE predictions

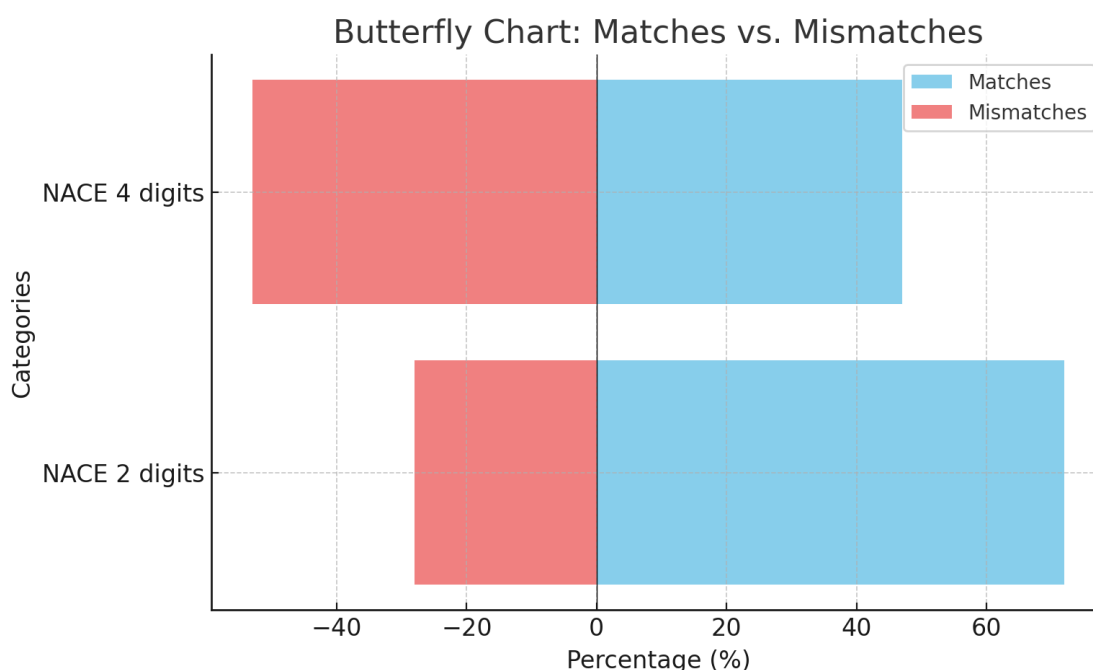


Figure 1. Proportion of matches and mismatches in NACE predictions

The chart compares the accuracy of predictions at 2-digit and 4-digit levels. At the 2-digit level, 72% of predictions match the true NACE code, while this accuracy decreases to 47% at the 4-digit level, reflecting the increased complexity of achieving accurate predictions with more detailed classification levels. These results are part of an ongoing project under testing and refinement, showing encouraging progress toward improved classification accuracy.

3.2. Innovation predictions

The innovation classification model, based on a Large Language Model (LLM), achieved moderate performance metrics, with a precision of 0.50, recall of 0.58, F1-score of 0.54, and an accuracy of 0.55. These results indicate that the LLM correctly classifies approximately half of the cases. Efforts are underway to refine the prompt and improve the model's predictive accuracy, aiming for more reliable and effective results.

4. MAIN FINDINGS

This project represents a promising approach to addressing two key challenges: classifying companies based on their NACE code and identifying whether they engage in innovation, using information extracted from their web pages. Designed as a general-purpose tool, the model builds a knowledge database that can be expanded to answer additional questions of interest in the future. Currently, the project focuses on addressing these two initial questions, with the goal of continuing to evolve and adapt to new needs.

5. REFERENCES

- [1] Heritrix - Web Crawling Tool. Gordon, M., & McCown, F. (2006). Heritrix: A Flexible Web Crawler Based on Extensible Archival Infrastructure. *ACM Journal of Web Science*, 12(4), 35-47.
- [2] OpenAI Embeddings for Language Models. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *NeurIPS Proceedings*, 33, 1877-1901. Retrieved from <https://openai.com/research/language-models-are-few-shot-learners>
- [3] FAISS - Semantic Search. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-548; Johnson, J., Douze, M., & Jégou, H. (2017). Faiss: A library for efficient similarity search and clustering of dense vectors. Meta AI Research. Retrieved from <https://faiss.ai>.
- [4] Retrieval-Augmented Generation (RAG). Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stoyanov, V. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 30, 945-960.

Weekly Nowcasting US Inflation with Enhanced Random Forests

Philipp Wegmüller (State Secretariat for Economic Affairs SECO) & Seton Leonnard (System2)

1. INTRODUCTION

We develop a random forest model which allows for mixed frequencies and missing observations. Further, we generalize the expected value of the target variable at each node of the regression tree to allow for a linear relationship between target and predictors. We apply the proposed framework to provide weekly nowcasts of U.S. inflation using a large set of daily, weekly, and monthly data. The resulting predictions are significantly better than simple univariate benchmarks for US inflation rate indices and match or exceed the accuracy of the nowcasts published by the Cleveland Fed.

Inflation forecasting has long been understood to be practically important and challenging. For example, Bernanke (2007) describes the importance of inflation modeling and forecasting to monetary policy. Many studies, such as Faust and Wright (2013) and Knotek and Zaman (2017), discuss and document challenges in inflation forecasting. The surge of inflation following the COVID-19 pandemic has of course added to practical interest in inflation forecasting.

One component of inflation forecasting is the problem of nowcasting inflation — that is, to predict inflation in the current period (month or quarter), before it is published. These short-horizon forecasts are not only of interest in their own right but also important inputs to the accuracy of forecasts at longer horizons (see, e.g., Faust and Wright (2009), Faust and Wright (2013), and Krüger et al. (2017)). In this paper, we focus on nowcasting inflation in US consumer prices, although our method can be more generally applied. Like much of the literature, our analysis is based on point forecasts. See Knotek and Zaman (2020) and references therein for work on density forecasts of inflation.

2. METHODOLOGY

Building on the success of Medeiros et al. (2021) in forecasting inflation with machine learning approaches, we focus on random forest methods. An important rationale is that we want to be able to include a large number of predictors, motivated partly by continued increases in the availability of data. We extend random forest methods in two dimensions. First, we develop an approach that allows for data with general patterns of missing observations, generated for example by the use of mixed frequency data or alternative indicators only available in the later part of the sample. Second, we generalize the expected value of the target variable at each node of the regression tree to allow for a linear relationship between right-hand side and target variables, which we refer to as a random forest regression node model. The standard model takes the mean of the target variable given the conditions to reach that node as the expected value; we simply generalize this to a slope and intercept term. Thus, the standard model is a subset of this generalization when the slope is constrained to zero. To establish possible benefits of our random forest extensions, we conduct simple Monte Carlo experiments

comparing our approach with alternatives that do not include the missing observation treatment and that take the mean of the target variable given the conditions to reach that node as the expected value (i.e., restrict to 0 the slope coefficients of our general model). As a baseline, we first compare the performance of our missing observation random forest approach with existing RF alternatives using simulated data without missing observations, finding our missing observation approach to work as well as other RF approaches.

Then, still using simulated data, we show that our missing observation random forest model works better than common RF methods for handling missing observations, such as replacing them with averages of the available observations, or dropping the series with missing observations. We then conduct experiments to assess the marginal benefit of adding the linear relationship piece to each node, and we find that, despite the cost of an additional parameter at each node of the model, we are able to reduce out-of-sample mean square error over the missing observation random forest model regardless of missing observations, as our final model combines these two innovations.

3. RESULTS AND PRACTICAL APPLICATION

We then turn to applying our proposed random forest models to nowcast monthly US inflation using a large set of indicators which are either high-frequency or published before the target inflation measures, including about roughly 150 to 200 macro, financial, and (mostly commodity) price indicators. Our analysis looks at nowcasting consumer price inflation directly, as well as breaking it down into its major components of commodities (consumer goods), commodities less food and energy, food, services, and services less energy services. We focus on headline CPI and PCE inflation as well as their ex-food and energy counterparts. In out-of-sample nowcasting, we compare results from random forest models to the nowcasts the Federal Reserve Bank of Cleveland has published since late 2013, based on the approach of Knotek and Zaman (2017).

Figure 1 displays exemplarily the realized and fitted values of CPI inflation starting from 2013:M10 to 2022:M2 (in line with the out-of-sample evaluation period of Section 4.3). We construct these estimates using all the data.

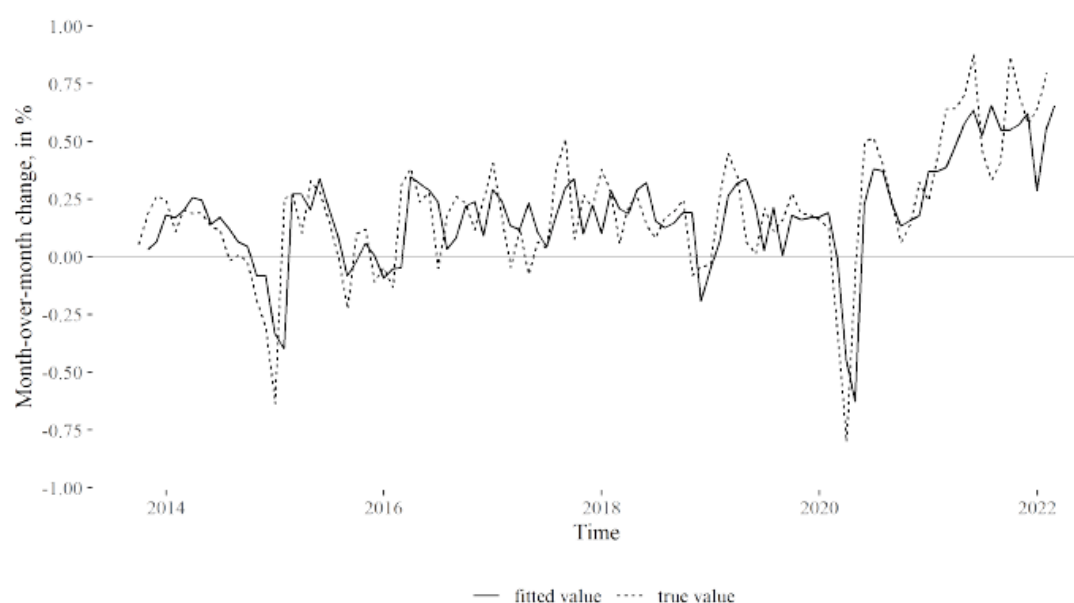


Figure 1. CPI inflation, fitted versus true values, 2013:M10 to 2022:M2 Fitted values from random forest random node model

Our empirical work yields the following main findings. In fitting the data, for the overall CPI, gasoline prices dominate. For categories that do not directly measure energy costs, variables measuring real economic activity such as the weekly economic index of Lewis et al. (2021) or the purchasing managers index often account for the first split in the regression trees. In out-of-sample forecasting, our model performs well, with an approximately 50 percent improvement in root mean square error over the sample 2013:M10-2022:M2 relative to the corresponding nowcast benchmarks from a univariate model and an approximate 10 percent improvement with regards to the Cleveland Fed nowcasts.

Table 1. Forecasting performance, relative RMSE

Forecasting horizon in weeks	Full sample				2013:M10 to 2019:M12			
	1	2	3	4	1	2	3	4
CPI								
AR(1) Benchmark	2.101	2.101	2.101	2.101	1.520	1.520	1.520	1.520
Cleveland Fed Nowcast	0.662***	0.678**	0.742**	0.799**	0.523***	0.532***	0.641***	0.687**
MO-RF	0.455***	0.487***	0.544***	0.564***	0.357***	0.418***	0.506***	0.499***
MO-RFRN	0.434***	0.454***	0.458***	0.492***	0.335***	0.384***	0.421***	0.436***
CPI Core								
AR(1) Benchmark	1.326	1.326	1.326	1.326	0.482	0.482	0.482	0.482
Cleveland Fed Nowcast	1.135	1.135	1.148	1.179	1.046	1.051	1.155	1.163
MO-RF	0.972	0.996	0.998	1.008	0.812***	0.995	1.007	0.986
MO-RFRN	0.961	0.992	1.044	1.051	0.797***	1.118	1.091	1.088
PCE								
AR(1) Benchmark	1.857	1.857	1.857	1.857	1.191	1.191	1.191	1.191
Cleveland Fed Nowcast	0.438***	0.462***	0.596***	0.626***	0.567***	0.571**	0.612**	0.630**
MO-RF	0.349***	0.355***	0.485***	0.482***	0.423***	0.439***	0.527**	0.522**
MO-RFRN	0.352***	0.355***	0.420***	0.415***	0.377***	0.375***	0.455***	0.447**
PCE Core								
AR(1) Benchmark	1.156	1.156	1.156	1.156	0.687	0.687	0.687	0.687
Cleveland Fed Nowcast	0.797***	0.905	1.091	1.129	0.692	0.889	0.975	0.988
MO-RF	0.708***	0.740***	0.938	0.957	0.682***	0.739***	0.793**	0.840
MO-RFRN	0.839	0.882	0.874**	0.881	0.641***	0.702***	0.770**	0.807

RMSE of nowcasting models relative to the benchmark AR(1) model.

Significance levels: *p*-value: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).

Modified Diebold-Mariano tests: the alternative hypothesis states that (i) the nowcasts are significantly more accurate than the benchmark and (ii) the random forest models are significantly better than the Cleveland Fed nowcasts (in bold).

We evaluate the accuracy of our nowcasts by means of root mean square forecast errors (RMSEs) relative to RMSEs from the AR(1) benchmark. Forecast errors are calculated using the final inflation vintage as the measure of actual inflation.⁶ We apply the modified version of the Diebold-Mariano test (Diebold and Mariano, 1995) developed in Harvey et al. (1997). The results are reported in Table 1. The relative RMSEs are shown together with significance levels from the modified Diebold-Mariano tests, in asterisks for the hypothesis that the random forest forecasts significantly improve upon the univariate AR(1) benchmark model and in bold if they significantly outperform the Cleveland Fed nowcasts.

For the full evaluation sample, several results emerge: (1) Forecast accuracy improves in most cases as our forecast horizon shrinks; (2) except for core CPI inflation, predictions from the random forest models exhibit lower root mean square errors than the benchmark at all horizons; (3) for CPI and PCE inflation, the nowcasts are significantly better than the univariate benchmark at all horizons, whereas for the core inflation measures they are significantly better between one and three weeks ahead; (4) the random forest models provide significantly better CPI nowcasts than the Cleveland Fed at the one and two week horizons for the full sample;

for the limited sample at all horizons; (5) the random forest, random node model exhibits somewhat lower forecast errors in most cases than the simple random forest model.

4. MAIN FINDINGS

In this article we present a novel approach to use mixed-frequency data with random forest models. Further, we generalize the models to allow for missing observations and for the expected value of the target variable at each node of the regression tree to be based on a linear relationship between target and predictors. We test these enhanced models both with simulated data and in an empirical application focused on predicting various measures of US inflation. We find that our new random forest models significantly outperform standard random forests when data are missing, and are competitive with if not significantly better than established benchmarks for inflation, such as those of the Cleveland Fed.

When are you ready for a smart survey?

Marc Houben & Remco Paulussen (Statistics Netherlands)

1. INTRODUCTION

Work package 4 of the Smart Survey Implementation (SSI) project focuses on the logistical business process needed to conduct a smart survey. Next to methodology, IT and legal (which are topics of other work packages within SSI) an NSI needs also to be ready to implement a smart survey, from the business process perspective. This should not be a snowed under topic when implementing a smart survey. Thereby, an NSI should be mature enough to perform a smart survey.

2. METHODOLOGY

In WP4 of the SSI project we developed two major topics for an NSI to use to implement a business process for smart surveys.

On the one hand a library of process building blocks. In GSBPM there are two activities to build a business process: 'Design business process' and 'Build business process'. The WP4 building blocks will support these activities. The idea is that an NSI can use these building blocks (process activities) to model their production process. The library holds process activities that are specific for smart surveys and also 'regular' process activities that need to be adapted when using a smart survey. The library can be used to model a generic process or a survey-specific process. That is up to each NSI.

On the other hand a Smart survey Maturity model (SSMM). The model can be used by an NSI to evaluate how mature it is concerning using smart surveys and gives guidance how to grow to a more mature level. The model describes five maturity levels. And the maturity criteria are described from 5 different aspects: Organisation, Methodology, IT, Business process, and Legal.

3. RESULTS AND PRACTICAL APPLICATION

In this presentation, WP4 of the SSI project (CBS, hbits) will show the process building blocks and the maturity model, and will give examples how an NSI can use them.

The aim of the presentation is for an NSI to get an idea of the building blocks and the maturity model. We will show that you need to be at a certain maturity level to implement smart surveys successfully.

4. MAIN FINDINGS

The main finding of WP4 is that a lot of aspects need to be in place within an NSI to use a smart survey, and that the logistical aspect is an important not to forgotten aspect herein.

Where people live and work: a register-based spatial and longitudinal analysis for Italy

Andrea de Panizza & Stefano De Santis (Istat)

1. INTRODUCTION

The paper proposes a fine grain spatial analysis aimed at testing a methodology for identifying spatial agglomerations based on a broad variety of socio-economic indicators. The methodology is then applied to the Functional Urban Area (FUA) of Rome – i.e. a *core-periphery* taxonomy¹ – experimentally profiling sub-communal units within the municipality itself. The classification methodology is also used to perform a longitudinal analysis of societal change for a case-study of an urban district starting from 2011.

The analysis makes extensive use of register-based information. In detail, it considers the base statistical registers on population and business, respectively extended to include information on education and incomes, and an array of economic variables for business units.

2. METHODOLOGY

The first step of the analysis consists integration procedures, carried out at the NUTS-4 level (close to 8 thousand municipalities) of microdata-based indicators. For the case of the three most populated Functional Urban Areas (Naples, Rome and Milan), about 180 territorial units were added within the three core municipalities based on postal codes (bringing the total for the three FUAs to 562, and to 159 for the case of Rome alone).² The variables considered encompass:

- **Demography:** (i) population density; (ii) mean age of residents; (iii) % of females; (iv) family size (number); (v) % of foreign residents; (vi) % of people born outside the local NUTS 3 level area
- **Employment:** (vii) % of persons employed over working age population; (viii) incidence of self-employed; (ix) % of retired individuals;

¹ Large urban centres represent *modernity* from a socio-economic point of view. Indeed, they host command and control functions of enterprise groups, concentrate knowledge-intensive economic activities and, more generally, many different activities interacting in a neighbourhood setting that benefits from relevant economies of urbanisation (Jacobs, 1969). These latter include advantages stemming from the high density of enterprises and from the urban dimension, and include the presence of physical and communication infrastructures, schools and university centres, the availability of skilled labour, specialised know-how, services and intermediate goods, as well as the circulation and dissemination of ideas across different activities. The metropolitan cities of Rome, Milan and Naples are the most populous in Italy and a substantial part of the national economy: together, at the beginning of 2019 they counted 10.7 million residents (17.7% of the national total), and generated 22.7% of Italian added value. In comparison with 2001, their population increased by almost 1 million, or 41% of total population growth recorded in the country. The respective FUAs counted 13 million residents, and hosted about 14 million people on daily basis.

² This approach was chosen as a temporary solution, to allow for the localisation of elementary units (i.e. individuals, local units of enterprises) while the base statistical register of places is still not guaranteeing 100% coverage.

- **Education:** (x) mean years of study of residents; (xi) % of students over residents; (xia) % of students over 6 to 24 years-old;
- **Income:** (xii) income per resident; (xiii) income per person employed
- **Business presence:** (xiv) enterprises per thousand residents; (xv) local units in knowledge intensive activities as a % of total

Indicators are correlated to various degrees. Each municipality can be portrayed for individual variables with respect to parent (NUTS-2 and -3) areas and the whole population of NUTS-4 units, by means of a quick benchmarking tool that could be made public (Figure 1).

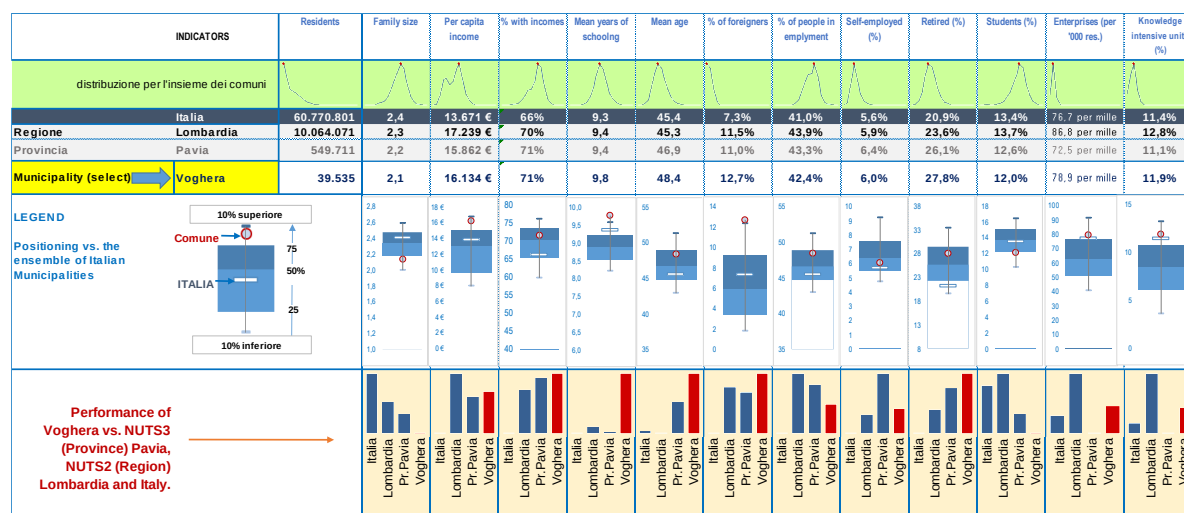


Figure 1. Snapshot of the position of municipalities with respect to individual variables

Operationally, the number of variables is reduced by means of factor analysis, and cluster analysis is undertaken on the ensuing results. This permits to identify a taxonomy of municipalities based on their economic strength joined to societal characteristics.

The same type of analysis is performed separately for the three largest FUAs, to characterise their constituent territories. For the case of Rome, this is complemented by a focus on the urban district of “Pigneto” (eastern quadrant), portraying changes with respect to the development of specific economic activities and the characteristics of residents, in comparison with the dynamics observed for the whole municipality.

3. RESULTS AND PRACTICAL APPLICATION

In this abstract we portray some very preliminary outputs on the three FUAs, based on previous work. In broad terms, we observe that univariate analysis permits to show a clear overall hierarchy between centres and surrounding peripheral municipalities, with some differences among the three FUAs: Rome and Milan are overall more similar, with Naples showing a lesser level of advancement; on the other hand, Rome's municipality alone has 65% of residents of her FUA, as it includes what in the other two cities is the “first belt”, while Milan stands alone in terms of the expulsion of less-than-wealthy residents and lower-end economic activities from the city boundaries (Figure 2).

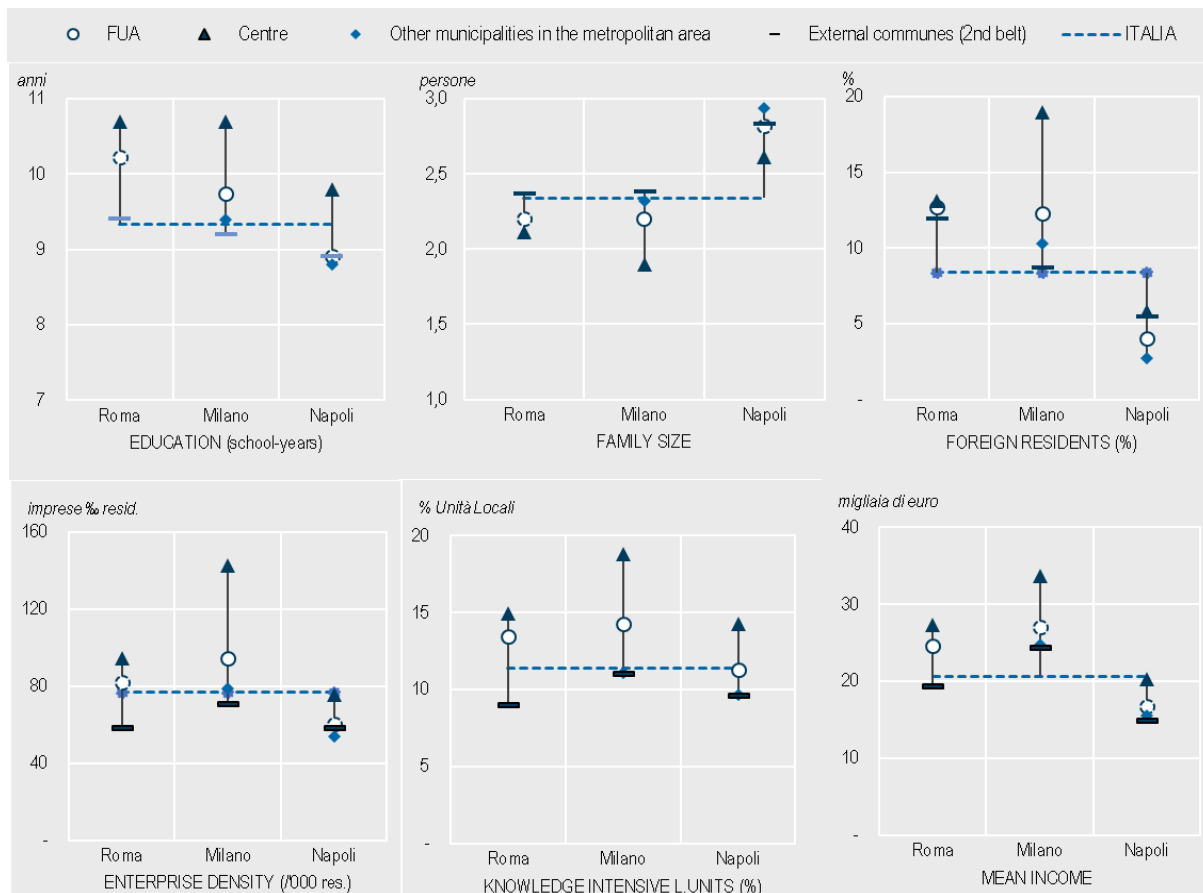


Figure 2. Characteristic features of Rome, Milan and Naples FUAs, by localisation of municipalities (year 2017) Source: Istat, *Rapporto sul Territorio 2020*

Large differences can also be found within the cities themselves. A good example to this respect is provided by the level of education³ in the three FUAs districts. The resulting distribution is highly concentrated, monocentric in Rome and polycentric in the two other FUAs (Figure 3).

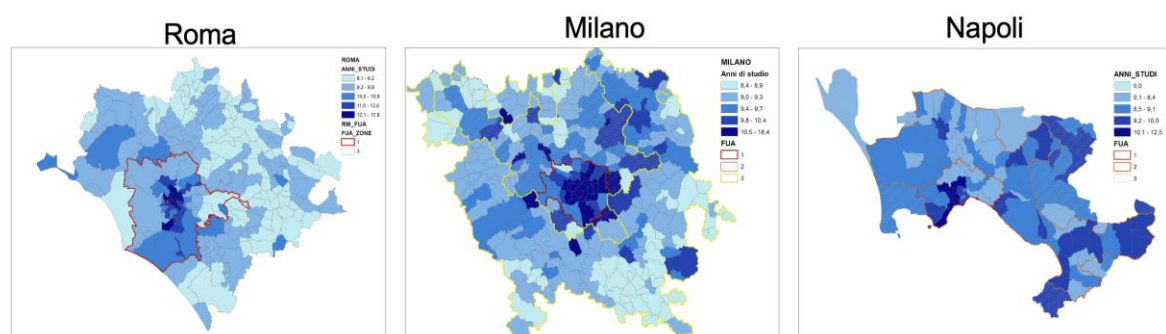


Figure 3. Mean years of formal education of adult residents within the FUAs of Rome, Milan and Naples in 2019 (highlighted boundaries represent Core and first ring of municipalities).

The relevance of spatial correlations among variables (i.e. the presence of multivariate complex phenomena) can be appreciated by looking at a selection of univariates for the case of Rome (Figure 4).

³ Proxied by the average years of study legally needed to obtain the highest educational degree by residents.

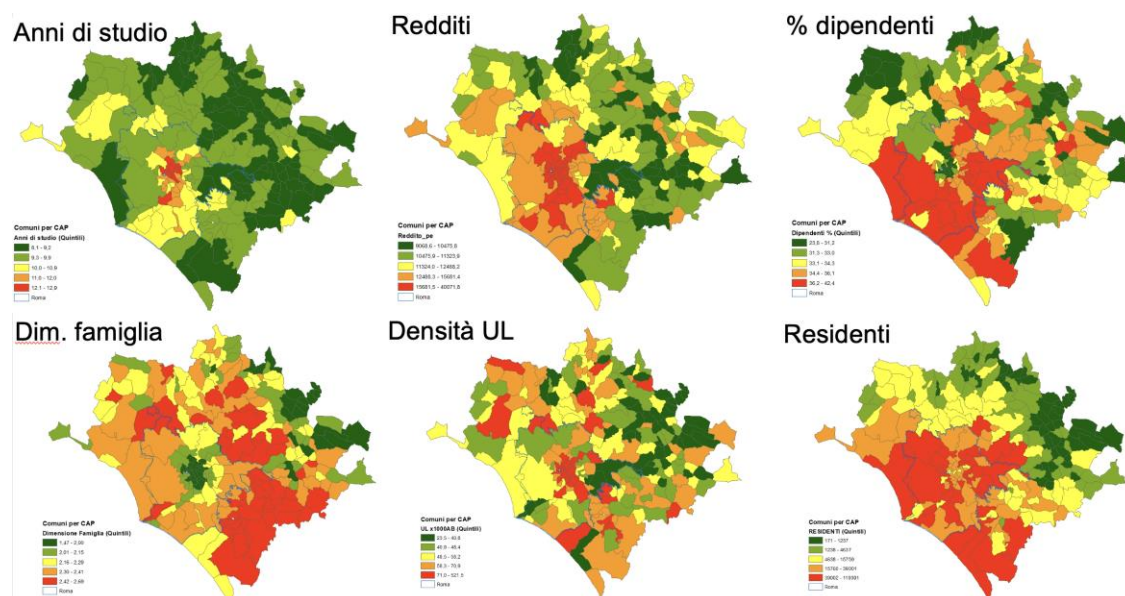


Figure 4. Univariate characterisation of Rome's FUA (postal codes within Rome and surrounding municipalities)

These spatial differences can be summarised in a very raw fashion by means of a multivariate analysis that intentionally attributes a premium to contiguity (presented only for illustrative purposes – Figure 5)

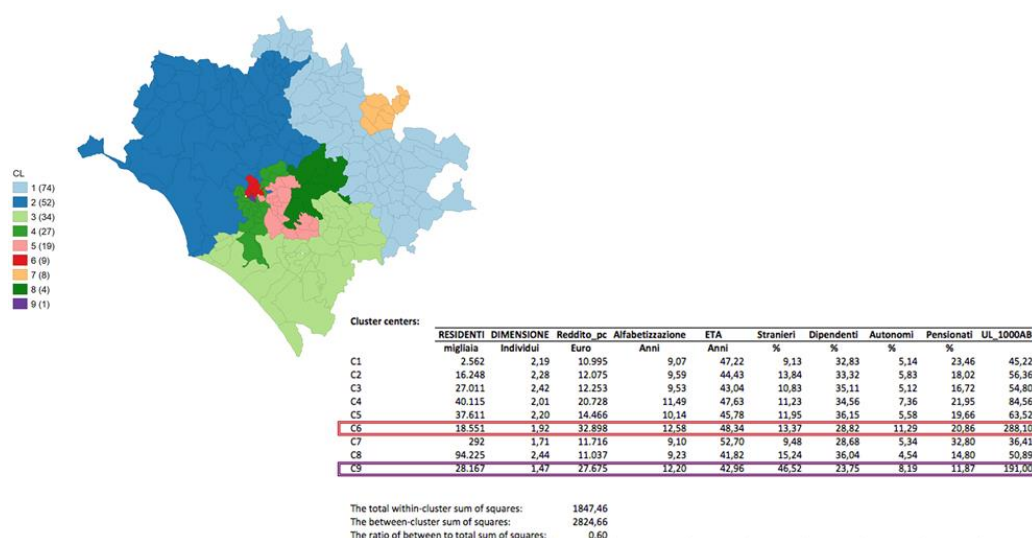


Figure 5. Highly preliminary multivariate characterisation of Rome's FUA

Finally, the paper verifies the potentiality of the same set of indicators to portray features and changes at the sub-communal level, by drilling in to a formerly very popular quarter, which anecdotal evidence showed as undergoing gentrification and increased attractiveness for leisure activities. Hereunder, this is exemplified by looking at the development of distinct economic activities (Table 1).

Table 1. Dynamics of productive activities in Rome and the *Pigneto* quarter. Years 2011-2019 (2011=100)

Sezioni di Attività (ATECO 2007)		Roma	Pigneto
B	ESTRAZIONE DI MINERALI DA CAVE E MINIERE	82,1	
C	ATTIVITÀ MANIFATTURIERE	87,2	80,5
D	FORNITURA DI ENERGIA ELETTRICA, GAS, VAPORE E ARIA CONDIZIONATA	126,5	
E	FORNITURA DI ACQUA; RETI FOGNARIE, ATTIVITÀ DI GESTIONE DEI RIFIUTI E RISANAMENTO	103,5	300,0
F	COSTRUZIONI	91,9	96,3
G	COMMERCIO ALL'INGROSSO E AL DETTAGLIO; RIPARAZIONE DI AUTOVEICOLI E MOTOCICLI	96,9	103,5
H	TRASPORTO E MAGAZZINAGGIO	99,7	110,6
I	ATTIVITÀ DEI SERVIZI DI ALLOGGIO E DI RISTORAZIONE	129,5	144,8
J	SERVIZI DI INFORMAZIONE E COMUNICAZIONE	106,4	114,8
K	ATTIVITÀ FINANZIARIE E ASSICURATIVE	100,0	94,5
L	ATTIVITÀ IMMOBILIARI	109,0	91,7
M	ATTIVITÀ PROFESSIONALI, SCIENTIFICHE E TECNICHE	111,9	136,0
N	NOLEGGIO, AGENZIE DI VIAGGIO, SERVIZI DI SUPPORTO ALLE IMPRESE	121,0	204,5
P	ISTRUZIONE	139,7	150,0
Q	SANITÀ E ASSISTENZA SOCIALE	120,8	119,3
R	ATTIVITÀ ARTISTICHE, SPORTIVE, DI INTRATTENIMENTO E DIVERTIMENTO	111,9	151,4
S	ALTRE ATTIVITÀ DI SERVIZI	112,1	119,4
TOTALE		107,2	118,1

4. MAIN FINDINGS

This highly preliminary exploration shows that – based on the census level information in statistical registers – it is possible to study and synthesise the complexity of the territorial distribution of the multivariate phenomenon represented by the set of indicators. These latter can also be used for purposes of benchmarking, by means of an exploratory data visualisation tool, which could be made available to end-users (Figure 1).

The empirical analysis highlights the presence of a multivariate dependency structure connecting the indicators. This is shown above for the case of the three FUAs, also confirming the expected high variability within each FUA level. This evidence underscores the importance of the availability of information at the sub-communal level, for both cross-sectional and longitudinal analyses.

5. REFERENCES

- [1] Abdi H., Lynne J.W. (2010). "Principal Component Analysis." John Wiley and Sons, Inc. WIREs Comp Stat 2: 433–59. <http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf>.
- [2] Aydin O., Janikas M.V., Assunção R., Lee T. (2018), SKATER-CON: [Unsupervised Regionalization via Stochastic Tree Partitioning within a Consensus Framework Using Random Spanning Trees](#). Conference paper
- [3] Jacobs J. (1969), Economy of cities. Vintage Books. New York.
- [4] ISTAT (2020). Rapporto sul Territorio. Ambiente, economia e società. Roma. <https://doi.org/10.1481/Istat.Rapportoterritorio.2020>

Which data to combine with Mobile Phone Data? Landscaping of the most promising sources and associated challenges

Marie-Pierre Joubert (Insee)

1. INTRODUCTION

Data from mobile network operators (MNOs) is emerging as a promising resource for improving the quality of official statistics, helping to refine public policies and inform public discussion [1]. MNO data can be combined with two categories of data: on the one hand statistical data, gathered by NSIs primarily for statistical purposes (census, surveys...) and on the other hand non-statistical data, collected for alternate purposes and subsequently repurposed for statistical analyses: administrative data but also emerging sources like traffic loop data and public transport ticketing data.

As underlined in the position paper of the ESS Task Force on the use of MNO data for official statistics [2], combining MNO data with other data sources has two major interests. The first one is to improve the general quality of produced statistics. Indeed, MNO data, primarily collected for network maintenance, may lack documentation on coverage deficiencies or network incidents. Adjusting counts through the incorporation of census or administrative data helps enhance the representativeness of these datasets (as illustrated in [3]). The second interest of this combination is to shed a new light on topics addressed by traditional official statistics. For instance, investigating inequalities in exposure to air pollution necessitates the fusion of MNO-derived daytime pollution data with air pollution data from specialized organizations and socio-demographic data collected by NSIs. In essence, every data source has inherent limitations, whether related to information coverage or representativeness, and combining data offers a strategic approach to mitigate these limitations.

This paper is based on the work of Work Package 2 (WP2) within the ESSNet on “Trusted Smart Statistics – Methodological Development Based On New Data Sources”, funded by Eurostat. This research initiative focuses on developing methodologies for the integration of both MNO and non-MNO data. WP2 involves collaboration among NSIs from France, the Netherlands, Germany, Portugal, Romania, Sweden and Italy. Its objective is to establish a short list of the most promising non-MNO sources for integration with MNO data, providing a comprehensive review of their advantages, disadvantages, costs, gains, and sustainability through systematic identification. The paper will present the major findings of WP2 at the time of the NTTS conference (one year and a half after the beginning of the project).

2. METHODOLOGY

2.1. Primary scoring dimensions

In order to analyse the potential of these sources, WP2's members have used as starting point the 'Big Data classification matrix' which was produced by ESSNet BigData 2 [4]. This matrix has then been adapted to the specific question of the combination of MNO and non-MNO sources (see tab below).

CATEGORY OF ANALYSIS	SCORING DIMENSION (low/moderate/high)
Data Type	Technical cost of handling the dataset
Do you know the size of the data set? Will it be a problem to treat it at once? Will you split it for processing ?	
Do you know the structure of the dataset? Are many different files considered a collection? A. Do you have to relate several files to have the entire dataset? B. Are the variables that enable linking of the data already known? If not do you have already a proposal to test the linkability?	
Access	
Who owns the data? Public administration, one company, several companies? Could the multiplicity of actors lead to multiple data formats and therefore potential integration and harmonisation problems?	Easiness of access to a temporally and geographically harmonized data source.
Is it possible to get access with a certain stability ? Does it have to be paid?	
Are there limitations to the amount or aggregation level of data that can be accessed? A. What is the nature of this limitation? Legal, technical, financial, other?	Easiness of access to detailed data
Is there a possibility to access the data to study its relevance?	
Are there potentially competing uses or specific restrictions in the use cases (operator publishing similar statistics, etc.)?	Range of possible use cases
Is this data available in all EU countries?	EU availability of these data
Metadata	
Is the definition of the population accessible? If not do you already have a method to address this issue?	Accuracy and robustness of the information available on these data
Is the reference period of the data available?	
Is the detailed methodology used to build these data available?	
Is the base unit of the dataset accessible?	
Do the units have an identifier?	
Do you have the necessary variables to reach the relevant granularity level for the statistical unit?	
Is there background information that you need to link the base units of the data set to the statistical unit, but that doesn't have the base units of the data set?	
Is there auxiliary information to make the data set useful with auxiliary data (NSI or another source)?	
Does the data contain sensitive variable? (Meaning legal or ethical issues related to its use)	Non-sensitivity of these data

2.2. Secondary scoring dimensions

One important aspect of ESSNet's MNO MINDS is that it aims at producing official statistics, and not only experimental statistics. The short list of the most promising data sources will therefore be established considering the availability of sources in all NSIs and the extent in which they respect the European Statistics Code of Practice ([5]). Naturally, non-MNO sources which already are in the official statistics domain fulfil the Cop requirements, which is not necessary the case of new data sources. This should therefore not be an essential prerequisite for considering a source as promising for being combined with MNO data. Yet going through the adequacy of a source to the CoP principles gives a rough estimation of how costly and complicated it would be to produce official statistics using this source. This adequacy will therefore be considered as a secondary level criterion in the sources scoring process.

3. SOURCE SCORING

Unsurprisingly, the most promising sources at the moment, that is the ones with the best cost-benefit analysis for being combined with MNO data, are the ones which are traditionally used by National Statistical offices. Indeed, they fulfil quasi all the requirements of the assessment matrix and they are also fit with the use cases which aim at improving population's coverage.

However, in addition to these traditional sources, we can't rule out a number of new sources that are being tested in several NSIs, and whose combination with mobile data looks very promising. The pros and cons of these sources, especially their cost of acquisition and treatment and the way in which they could fulfil official statistics' quality criteria are analysed in this paper.

Lastly, some sources, which often correspond to what the common imagination understands by Big Data, have been estimated as less relevant for the combination with MNO data. For instance, Google Maps Popular Time data seems a promising complement to MNO data regarding day-time population. Yet the access to data is complicated, with partnerships to establish with multiple private operators. Moreover handling these data implies high data-science skills because of their huge size. The same analysis holds for connected vehicles for instance.

Most promising data sources to be combined with MNO dat
Census, Population Register (fiscal data, ...), Combination of survey and register, Transportation Surveys
Promising sources but which would require substantial work or which are not yet fully accessible
Vehicle, bicycle and pedestrian sensors, Vessel (boat) traffic data, Pollution data, Satellite data, Electronic invoices, Tourism Household and Border Surveys, Tourism platform data, Credit Card Transaction Data
Less relevant sources
Google Maps Popular Time, Smart Meters, Connected Vehicles, Social Media

4. REFERENCES

- [1] Coudin, E., Poulhes, M., & Suarez-Castillo, M. (2021). The French official statistics strategy : Combining signaling data from various mobile network operators for documenting COVID-19 crisis effects on population movements and economic outlook. Data and Policy.
- [2] MNO, T. F. (2023). Reusing Mobile Network Operator data for Official Statistics: the case for a common methodological framework for the European Statistical System. Eurostat Position Paper.
- [3] Suarez Castillo, M., Sémécurbe, F., Ziemlicki, C., Tao, H. X., & Seimandi, T. (2023). Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics. *Journal of Official Statistics*, 39(4), 535-570.
- [4] Kowarik, A., & members, E. (2020). Typification matrix for big data projects. ESNet Big Data 2 - Grant Agreement Number: 847375-2018-NL-BIGDATA.
- [5] CoP, E. (2017). European Statistics Code of Practice. Eurostat, European Statistical System.

Who are the flood victims? Monitoring the effects of the disaster using Earth Observation, GIS and administrative data

Marek Pieniżek (Statistics Poland)

1. INTRODUCTION

In mid-September 2024, as a result of heavy rainfall in Central Europe, a series of floods occurred. These phenomena were particularly severe in southwestern Poland, the Czech Republic, Austria and Germany. The rising waters in the upper section of the Odra River basin caused the flooding of many towns. Currently, losses are estimated at several billion PLN.

On the 13 September 2024 at 12:50 UTC, The Polish National Centre for Coordination of Rescue Operations and Protection of Population triggered Copernicus EMS Rapid Mapping service to cover the flood event affecting the south-western part of Poland. Copernicus EMS Rapid Mapping was requested to provide flood extent emergency mapping and monitoring.

The EO data allowed to determine the extent of the flood. For which an experimental statistical study was conducted using GIS and combining administrative data. As a result, information was obtained on the characteristics of the population affected by the flood.

2. METHODOLOGY

Remote sensing data obtained from Copernicus EMS Rapid Mapping (EMSR756) were used to create a vector layer of flood-affected areas. On this basis, address points/buildings were selected. This information was combined with data from official registers. As a result, a detailed demographic characteristic of people affected by the flood was obtained. As part of this characteristic, information was obtained, among others, on age, gender, professional activity, level of disability.

3. RESULTS AND PRACTICAL APPLICATION

The data obtained constitute the first detailed estimate of the flood scale. The results can be used to plan the reconstruction of these areas and to plan support for residents. The results will also be published in the context of the Public Services Monitoring System (SMUP.gov.pl)

Working with a mobile network operator (MNO) to create a privacy-conform method for a better access to MNO data

Lorenz Ade & Maurice Brandt (Destatis)

1. INTRODUCTION

Mobile network operator (MNO) data is being analysed worldwide as a potential source for official statistics, as it can be used to analyse a wide range of topics. For example, MNO data could be used to improve statistics on population, mobility or tourism. Additionally, MNO data could lead to the creation of completely new statistics, such as day-time vs. night-time population figures. A major advantage of MNO data is the high level of coverage. For example, 98.1% of households in Germany own a mobile phone. On average, there are 1.87 devices per household (Statistisches Bundesamt, 2022). As a data source it is therefore of great utility for official statistics. But why are there no official statistics based on MNO data yet? MNO data is not primarily produced for official statistics. Instead, it is basically a by-product of the operation of a mobile network. Here the data is mostly used for technical analysis and troubleshooting. Therefore, the data needs to be heavily processed before it can be used in official statistics. These processing steps include the geolocation, deduplication, aggregation and extrapolation of the data. Some MNO's (not only in Germany) offer the usage of anonymized and aggregated data products. In the past, DESTATIS has already produced several promising experimental statistics based on these products, for instance on mobility during the Covid-19 pandemic or population forward projection (see our website at <https://www.destatis.de/EN/Service/EXSTAT/Datensaetze/mobility-indicators-mobilephone.html>). However, the methods by which these products are created are confidential, as they employ information like cell tower positions or local market shares, which the MNOs naturally don't want to share with their competitors. This is however a problem for a National Statistical Institute (NSI), as it has the duty to be transparent about the methods and processing steps it uses in the production of official statistics. To open the black box of the MNO processing, NSI's therefore have to develop their own methods. As the development requires access to the original data, DESTATIS has partnered with T-Systems to gain access to raw MNO data for the first time. Besides providing access to (strongly distorted) data, T-Systems also provides a development environment in their IT-infrastructure and their technical knowledge.

A large part of the data processing to be designed is also ensuring the compliance with data privacy at all steps of the production process. While signalling event data is not as privacy invasive as call details records, the unprotected data would still allow for the tracking of individual movements of cellular device holders. Although the raw data is pseudonymized from the beginning and never to be released, an attacker might be able to identify individuals if insufficient aggregation procedures are employed. To prohibit such reidentification attacks, the second large focus of this research project is therefore the identification of the anonymization needs and the evaluation of suitable methods to prevent reidentification.

At the NTTS 2025, DESTATIS will report on the goals of the project, share the experience of working with MNO operator, as well as present the first preliminary results.

2. PROJECT GOALS

Essentially, this project aims to transfer the usage of MNO data from experimental statistics to official statistics by opening the black box of mobile network signalling processing. This requires designing and implementing a processing and anonymization procedure for the usage of anonymized georeferenced MNO data. To design this process, the conception and set-up of a development environment at the data provider is necessary. Lastly, the project also serves as an attempt to set up a model process for future cooperation's between private data providers, especially MNO's.

Such a cooperation, as it exists with T-Systems as MNO in this project, is vital to reach the planned project goals. Not only is the project dependent on the provided IT-Infrastructure and data but it also benefits heavily of the combination of the technical expertise of T-Systems with the methodical expertise of DESTATIS as NSI. A cooperation with a commercial data provider is not common and comes with its own challenges. At the heart of the problems is the duty of DESTATIS as NSI to employ and document the used methods transparently, whereas all MNO's want keep their algorithms and additional information used, e.g. local market shares, confidential.

To achieve the overall goals of the project, four work packages have been designed. The first work package is designed to develop the necessary IT-Infrastructure for the project. This step is necessary due to the natural differences in IT-infrastructure between DESTATIS and Telekom as MNO. To test the standardized processes developed in this project therefore requires a development environment that is comparable with the T-Systems infrastructure. Additionally, this also makes gaining approval from data privacy regulators easier, as the raw data will not leave the T-Systems infrastructure.

The aim of work package two is to define and implement the necessary interfaces for the transfer of the computed aggregates to DESTATIS. In addition to the technical and content-related results, it is also necessary to comply with IT and data protection regulations

Work package number three is leaning heavily on the work conducted in the ESSnet Big Data 2 and the work currently conducted in the ESSnets Multi-MNO and MNO-MINDS. Its goal is to methodically develop standardized algorithms for processing and preparation that are as flexible as possible for the daily processing of mobile network signal data, particularly for the purposes of official statistics. The methodological solutions that need to be implemented include, for example, the ability to correct distortions, the calculation of location probabilities or the recognition of movement modes. Additionally, data users in the future should be able to control and adjust their data requirements by using standardized parameters, without a need to access the raw data.

The fourth and final work package aims to develop an improved, more flexible anonymization procedure for mobile network signal data in the context of the applicable data protection regulations. This requires a sensible trade-off between flexibility and precision of the data and the protection-compliant processing of the data. To identify the needs in regard to anonymization, dependent on the required scope of the data, the project is part of a German research cluster on the anonymization of georeferenced data called "AnigeD", which itself is part of a large German research network on the anonymization of data.

3. PROJECT STATUS AND NEXT STEPS

The first big steps of the project have already been taken. Both work package 1 and 2 are near completion.

For work package 1, the most important step was the acceptance of the data protection plan for the project. Both the corporate security of our partner T-Systems and the BfDI (Federal Commissioner for Data protection and Freedom of Information) have accepted the concept. This is already a huge hurdle taken, as a failure to coordinate a sufficient data protection plan could have delayed the project timeline severely. As a result, the work on the development environment and the dataset could timely start. The completion of this work, as of writing this abstract, is expected for October 2024.

In work package 2, a preliminary completion is also expected for the very near future. This means that the necessary technical details for the data access have been decided on. Implementation can therefore start when the final authorisations have been obtained. We assume that the knowledge and experience gathered for the access to the development environment helps with creating the exchange interface for regular production.

Work package 3 and 4 are obviously dependent on completion of work package 1. In the meantime, DESTATIS is continuing preparatory work for the start the work packages, chief among them a landscape analysis of anonymization procedures to test on the real data.

4. CONCLUSION

With this project, DESTATIS moves closer to transferring MNO data from the realm of experimental statistics to an incredibly valuable data source for official statistics. To achieve this goal, a standardized processing and anonymization procedure for MNO data is developed. The project benefits immensely from both its cooperation with the MNO T-Systems and the German Research Cluster AnigeD. The cooperation with T-Systems as an MNO allows DESTATIS not only to access (distorted) raw MNO data for the first time, but also to benefit from their technical expertise. AnigeD and the affiliated research network give DESTATIS access to state-of-the-art knowledge on the anonymization of geodata. Preliminary results will be presented at NTTS 2025.

5. REFERENCES

- [1] Statistisches Bundesamt, Laufende Wirtschaftsrechnungen (LWR), Ausstattung privater Haushalte mit Informationstechnik, (2022). Retrieved from: <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsguter/Tabellen/liste-infotechnik-d.html#115470>

