ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Work package K

Methodology and quality

Deliverable K12: Revised literature overview

Final version, 30.11.2020

Prepared by:
Jacek Maślankowski (GUS, PL)
David Salgado (INE, ES)
Sónia Quaresma (INE, PT)
Gabriele Ascari, Giovanna Brancato, Loredana Di Consiglio, Paolo Righi and Tiziana Tuoto (ISTAT, IT)
Piet Daas (CBS, NL)
Marlene Weinauer, Magdalena Six, Alexander Kowarik (STAT, AT)

Work package Leader:

Alexander Kowarik (STAT, AT)
alexander.kowarik@statistik.gv.at
telephone : +43 1 71128 7513

Contents

Contents 2

1.1. Cereal Yield Modeling in Finland Using Optical and Radar Remote Sensing 4

1.2. Integrating Vegetation Indices Models and Phenological Classification with Composite SAR and Optical Data for Cereal Yield Estimation in Finland 6

1.3. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories (in Canada) 8

1.4. Radar Vegetation Index as an Alternative to NDVI for Monitoring of Soyabean and Cotton (in India) 11

1.5. Pre-crop Values From Satellite Images for Various Previous and Subsequent Crop Combinations in Finland 13

1.6. Pre-Crop Values from Satellite Images to Support Diversification of Agriculture 15

1.7. Sentinel EO Browser for viewing Sentinel SAR and optical satellite images 17

1.8. NDSI review: Generate wet snow maps using NDSI and optical and SAR data 19

1.9. VWC & NDWI bridge: Inferring Vegetation Water Content From C- and L-Band SAR Images 21

1.10. A new method for crop classification combining time series of radar images and crop phenology information 23

1.11. Multi-data approach for crop classification using multitemporal, dual-polarimetric TerraSAR-X data, and official geodata 25

1.12. A review of assessing the accuracy of classifications of remotely sensed data 27

1.13. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data 29

1.14. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification 31

1.15. Classifying websites by industry sector: a study in feature design 33

1.16. Classifying firms with text mining 35

1.17. A Heuristic Approach for Website Classification with Mixed Feature Extractors 37

1.18. Three Methods for Occupation coding Based on Statistical Learning 39

1.19. Learning from Imbalanced Data 41

1.20. Measuring the UK's Digital Economy with Big Data 43

1.21. A knowledge-based approach to unstructured data 45

1.22. Measuring the internet economy in The Netherlands: a big data analysis 47

1.23. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation 49

1.24. Data quality management, data usage experience and acquisition intention of big data analytics 51

1.25. Data quality management, data usage experience and acquisition intention of big data analytics 53

1.1. Cereal Yield Modeling in Finland Using Optical and Radar Remote Sensing

SPECIFICATION	DESCRIPTION
Bibliographic data	Remote Sens. 2010, 2(9), 2185-2239; https://doi.org/10.3390/rs2092185 Cereal Yield Modeling in Finland Using Optical and Radar Remote Sensing Heikki Laurila , Mika Karjalainen , Jouko Kleemola and Juha Hyyppä
Link	https://www.mdpi.com/2072-4292/2/9/2185
Short overview (strengths, weaknesses)	During 1996–2006, the Ministry of Agriculture and Forestry in Finland (MAFF), MTT Agrifood Research and the Finnish Geodetic Institute performed a joint remote sensing satellite research project. It evaluated the applicability of optical satellite (Landsat, SPOT) data for cereal yield estimations in the annual crop inventory program. Four Optical Vegetation Indices models (I:Infrared polynomial, II: NDVI, III: GEMI, IV: PARND/FAPAR) were validated to estimate cereal baseline yield levels (yb) using solely optical harmonized satellite data (Optical Minimum Dataset). Optical VGI yield estimates were validated with CropWatN crop model yield estimates using SPOT and NOAA data (mean R2 0.71, RMSE 436 kg/ha) and with composite SAR/ASAR and NDVI models (mean R2 0.61, RMSE 402 kg/ha) using both reflectance and backscattering data. CropWatN and Composite SAR/ASAR & NDVI model mean yields were 4,754/4,170 kg/ha for wheat, 4,192/3,848 kg/ha for barley and 4,992/2,935 kg/ha for oats. Keywords: optical vegetation Indices models; classification; NDVI; GEMI; FAPAR; PARND; SAR/ASAR; CropWatN; LAI-bridge; Finland; CAP; Kalman Filter; data fusion; harmonized data; phenological SAR and optical spectral profiling
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society

	[ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.2. Integrating Vegetation Indices Models and Phenological Classification with Composite SAR and Optical Data for Cereal Yield Estimation in Finland

SPECIFICATION	DESCRIPTION
Bibliographic data	Remote Sens. 2010, 2(1), 76-114; https://doi.org/10.3390/rs2010076 Integrating Vegetation Indices Models and Phenological Classification with Composite SAR and Optical Data for Cereal Yield Estimation in Finland (Part I) Heikki Laurila, Mika Karjalainen , Juha Hyyppä and Jouko Kleemola
Link	https://www.mdpi.com/2072-4292/2/1/76
Short overview (strengths, weaknesses)	During 1996–2006 the Ministry of Agriculture and Forestry in Finland, MTT Agrifood Research Finland and the Finnish Geodetic Institute carried out a joint remote sensing satellite research project. It evaluated the applicability of composite multispectral SAR and optical satellite data for cereal yield estimations in the annual crop inventory program. Three Vegetation Indices models (VGI, Infrared polynomial, NDVI and Composite multispetral SAR and NDVI) were validated to estimate cereal yield levels using solely optical and SAR satellite data (Composite Minimum Dataset). The average R2 for cereal yield (yb) was 0.627. The averaged composite SAR modeled grain yield level was 3,750 kg/ha (RMSE = 10.3%, 387 kg/ha) for high latitude spring cereals (4,018 kg/ha for spring wheat, 4,037 kg/ha for barley and 3,151 kg/ha for oats). Keywords: Composite multispectral modeling; SAR; classification; SatPhenClass algorithm; minimum dataset; cereal yield; phenology; LAI-bridge; CAP; IACS; FLPIS; phenological SAR and optical spectral profiling
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....

Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.3. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories (in Canada)

SPECIFICATION	DESCRIPTION
Bibliographic data	McNairn, H.; Champagnea, C.; Shanga, J.; Holmstromb, D.; Reichert, G. 2008. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. ISPRS Journal of Photogrammetric Remote Sensing 64: 434-449. September 2009ISPRS Journal of Photogrammetry and Remote Sensing 64(5):434-449 DOI: 10.1016/j.isprsjprs.2008.07.006 Heather McNairnCatherine ChampagneCatherine ChampagneJiali ShangJiali Shang, Gordon Reichert
Link	https://www.researchgate.net/publication/222042356_Integration_of_optical_and_Synthetic_Aperture_Radar_SAR_imagery_for_delivering_operational_annual_crop_inventories
Short overview (strengths, weaknesses)	Agriculture plays a critical role within Canada's economy and, as such, sustainability of this sector is of high importance. Targeting and monitoring programs designed to promote economic and environmental sustainability are a vital component within Canada's agricultural policy. A hierarchy of land information, including up to date information on cropping practices, is needed to measure the impacts of programs on land use decision-making and to gauge the environmental and economic benefits of these investments. A multi-year, multi-site research activity was completed to develop a robust methodology to inventory crops across Canada's large and diverse agricultural landscapes. To move towards operational implementation the methodology must deliver accurate crop inventories, with consistency and reliability. In order to meet these operational requirements and to mitigate risk associated with reliance on a single data source, the methodology integrated both optical and Synthetic Aperture Radar (SAR) imagery. The results clearly demonstrated that multi-temporal satellite data can successfully classify crops for a variety of cropping systems present across Canada. Overall accuracies of at least 85% were achieved, and most major crops were also classified to this level of accuracy. Although multi-temporal optical data would be the preferred data source for crop classification, a SAR-optical dataset (two Envisat ASAR images and one optical image) provided acceptable accuracies and will mitigate risk associated with

	operational implementation. The preferred dual-polarization mode would be VV–VH. Not only were these promising classification results repeated year after year, but the target accuracies were met consistently for multiple sites across Canada, all with varying cropping systems. -- Recently McNairn et al. (2008) reported results obtained in Canadian prairie growing conditions that the composite VV-VH SAR combined with the optical data to be the most suitable for red hard wheat, maize and soybean (Glycine max L.) classification with over 85% overall accuracy (Kappa range 0.47–0.89). McNairn et al. (2008) applied three primary classification methodologies Neural Networks, Gaussian Maximum-Likelihood Classifier and Decision trees for crop classification using composite SAR and optical data McNairn, H.; Champagnea, C.; Shanga, J.; Holmstromb, D.; Reichert, G. 2008. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. ISPRS Journal of Photogrammetric Remote Sensing 64: 434-449.
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....

Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.4. Radar Vegetation Index as an Alternative to NDVI for Monitoring of Soyabean and Cotton (in India)

SPECIFICATION	DESCRIPTION
Bibliographic data	Radar Vegetation Index as an Alternative to NDVI for Monitoring of Soyabean and Cotton September 2013 Conference: XXXIII INCA International Congress At: Jodhpur Provides a new technology by providing basic methodology for using SAR RVI algorithm with X,Y and cross polarization XY,Y levels into crop and soil covariance modeling.
Link	https://www.researchgate.net/publication/267020154_Radar_Vegetation_Index_as_an_Alternative_to_NDVI_for_Monitoring_of_Soyabean_and_Cotton
Short overview (strengths, weaknesses)	The study explores the possibility of Radar Vegetation Index (RVI) for vegetation monitoring in Cotton and Soya bean fields as an alternative to monitoring with Normalized Difference Vegetation Index (NDVI). The RVI is the measure randomness of scattering has been proposed as a method for monitoring the level of vegetation growth, particularly when time series of data are available. The SAR can penetrate the clouds thus it has the potential to monitor the crop growth in all the seasons. This point is very important in the context of Indian subcontinent since monsoon clouds hampering the monitoring of crops in such season. The present study carried out in Vidharba region of Maharashtra, India using four Radarsat-2 data sets acquired in the monsoon season of 2011 (ie: 13thJuly 2011, 06thAugust 2011, 30 August 2011, 23rdSeptember 2011). The derived RVI was compared with the MODIS NDVI product of same date. The research shows some significant improvements in the RVI technique than NDVI in some context. The RVI is linearly increasing as crop grows unlike the NDVI becomes saturated after level of growth in the crop. The Cultivation started in the first week of July 2011, the NDVI becomes saturated mostly in second week of August, 2011 however RVI shows further increase in 30th August 2011 with the growth in the vegetation. These particular findings attributed to the fact that maximum randomness of scattering in microwave signal occurs where plant spread fully whereas peak greenness (NDVI) occurs before that. The present study found that RVI can be utilised in place of NDVI for vegetation monitoring thus monitoring of crops is possible even in the monsoon season of India. This method can be better used with RISAT descending mode since its acquiring image on optimum incident angle of around 360 look angle thus it is highly suitable for vegetation monitoring. The Radar Vegetation Index can also have the prospect of using in soil moisture models in vegetated fields where NDVI is a critical one (Kim and J. Van Zyl ., 2009). The result of this study can be effectively used for monitoring soyabean and cotton and also can be used as a vegetation mask for soil moisture monitoring. The further study is required to derive biophysical parameters from RVI and application of RVI in soil moisture estimation in vegetated fields. Keywords : RVI, Radar Vegetation Index, NDVI
Data sources	[ ] Web data

	[ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……...., ……...,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.5. Pre-crop Values From Satellite Images for Various Previous and Subsequent Crop Combinations in Finland

SPECIFICATION	DESCRIPTION
Bibliographic data	Master Thesis (Aalto University)
Link	https://www.maanmittauslaitos.fi/tietoa-maanmittauslaitoksesta/ajankohtaista/lehdet-ja-julkaisut/positio/viljelykasvien-tunnistaminen-sentinel-kuvilta-suomessa https://www.maanmittauslaitos.fi/sites/maanmittauslaitos.fi/files/attachments/2019/03/Viljelykasvien-tunnistaminen-sentinel-kuvilta-suomessa.pdf
Short overview (strengths, weaknesses)	Sentinel Supervised classification for crops currently cultivated in Finland. Different classification results indicate and average accuracy of 89 %, currently EU requires over 95% classification accuracy for automated satellite monitoring. Keywords: Supervised classification, Copernicus Sentinel, Neural Networks, Multilayer Perceptron (MLP), Convolutional Recurrent Neural Network (ConvRNN) , Support Vector Machine (SVM), ConvRNN algorithm
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation

	[ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.6. Pre-Crop Values from Satellite Images to Support Diversification of Agriculture

SPECIFICATION	DESCRIPTION
Bibliographic data	• Front. Plant Sci., 09 April 2019 | https://doi.org/10.3389/fpls.2019.00462 • Pre-crop Values From Satellite Images for Various Previous and Subsequent Crop Combinations • Pirjo Peltonen-Sainio, Lauri Jauhiainen, Eija Honkavaara, Samantha Wittke, Mika Karjalainen and Eetu Puttonen
Link	https://www.luke.fi/en/news/pre-crop-values-from-satellite-images-to-support-diversification-of-agriculture/ https://www.frontiersin.org/articles/10.3389/fpls.2019.00462/full
Short overview (strengths, weaknesses)	Monocultural land use challenges sustainability of agriculture. Pre-crop value indicates the benefits of a previous crop for a subsequent crop in crop sequencing and facilitates diversification of agricultural systems. Traditional field experiments are resource intensive and evaluate pre-crop values only for a limited number of previous and subsequent crops. We developed a dynamic method based on Sentinel-2 derived Normalized Difference Vegetation Index (NDVI) values to estimate pre-crop values on a field parcel scale. The NDVI-values were compared to the region specific 90th percentile of each crop and year and thereby, an NDVI-gap was determined. The NDVI-gaps for each subsequent crop in the case of monocultural crop sequencing were compared to that for other previous crops in rotation and thereby, pre-crop values for a high number of previous and subsequent crop combinations were estimated. The pre-crop values ranged from +16% to -16%. Especially grain legumes and rapeseed were valuable as pre-crops, which is well in line with results from field experiments. Such data on pre-crop values can be updated and expanded every year. For the first time, a high number of previous and following crop combinations, originating from farmer's fields, is available to support diversification of currently monocultural crop sequencing patterns in agriculture. Keywords: Finland, NDVI, climate change adaptation and mitigation
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade

	[ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.7. Sentinel EO Browser for viewing Sentinel SAR and optical satellite images

SPECIFICATION	DESCRIPTION
Bibliographic data	• Sentinel EO Browser for viewing and overlaying Sentinel and other SAR and optical satellite images (Sentinel, Landsat etc.)
Link	https://apps.sentinel-hub.com/eo-browser/?lat=41.9002&lng=12.4997&zoom=10 Tutorial with different Indices introduction: https://apps.sentinel-hub.com/eo-browser/?lat=41.9002&lng=12.4997&zoom=10
Short overview (strengths, weaknesses)	Global image coverage provide pre-calculated Indices (e.g. for snow cover, Water potential, Vegetation cover & phenological development): Some of the common options: • True Color - Visual interpretation of land cover. • False Color - Visual interpretation of vegetation. • NDVI - Vegetation index. • Moisture index - Moisture index • SWIR - Shortwave-infrared index. • NDWI - Normalized Difference Water Index. • NDSI - Normalized Difference Snow Index. NDWI https://eos.com/ndwi/ NDSI https://eos.com/ndsi/ Snow and cloud https://eos.com/snow-and-cloud/ Kewords: NDVI, Moisture Index, NDWI, Normalized Difference Water Index, NDSI Normalized Difference Snow Index
Data sources	[X] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [X] Sensors [X] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade

	[ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……..., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.8. NDSI review: Generate wet snow maps using NDSI and optical and SAR data

SPECIFICATION	DESCRIPTION
Bibliographic data	• Wet snow maps using optical and SAR data
Link	https://earth.esa.int/documents/10174/3166008/ESA_Training_Vilnius_07072017_SAR_Optical_Snow_Ice_Exercises.pdf
Short overview (strengths, weaknesses)	Normalized Difference Snow Index (NDSI) Practical tutorial to use NSDI and SAR data to generate wet snow maps NDSI https://eos.com/ndsi/ Snow and cloud https://eos.com/snow-and-cloud/ Kewords: NDVI, Moisture Index, NDWI, Normalized Difference Water Index, NDSI Normalized Difference Snow Index
Data sources	[X] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [X] Sensors [X] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data

	[ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……...., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.9. VWC & NDWI bridge: Inferring Vegetation Water Content From C- and L-Band SAR Images

SPECIFICATION	DESCRIPTION
Bibliographic data	Inferring Vegetation Water Content From C- and L-Band SAR Images
Link	https://ieeexplore.ieee.org/document/4305371
Short overview (strengths, weaknesses)	VWC, NDWI with Inferring Vegetation Water Content From C- and L-Band SAR Images A bridge algorithm between SAR C,L band (VWC) and optical NDWI data is proposed bare soils and the other one has been modified for vegetated fields. VWC - Vegetation Water Content NDWI https://eos.com/ndwi/ Snow and cloud https://eos.com/snow-and-cloud/ Kewords: VWC - Vegetation Water Content, NDVI, Moisture Index, NDWI, Normalized Difference Water Index, NDSI Normalized Difference Snow Index
Data sources	[X] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [X] Sensors [X] Images, Videos [ ] E-mail [ ] Not specified/not applicable [ ] Others .....
Domains	[ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [X] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics

	[ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [X] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [X] Data quality [X] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [X] Data noise [ ] Machine learning [X] Data visualization [ ] Others ....., ……...., …….,
Classification (A – very relevant, B – relevant, C – less relevant)	

1.10. A new method for crop classification combining time series of radar images and crop phenology information

SPECIFICATION	DESCRIPTION
Bibliographic data	Bargiel, D. (2017). A new method for crop classification combining time series of radar images and crop phenology information. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.06.022
Link	https://www.sciencedirect.com/science/article/abs/pii/S0034425717302821
Short overview (strengths, weaknesses)	• General aim: The main aim of this paper is ensuring a full implementation of knowledge about crops' phenology into classification approach to improve classification results. • Strengths: The paper confirms that multi-temporal classification based on dense time series of Sentinel-1 images enable for very good results with regard to non-cereal crop type. The paper includes quality aspects of performed image classification. The improved accuracies for different time types of cereals can be achieved using phenology information. • Weaknesses: The phenology information is not easy and not generally available information.
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [x] Images, Videos [ ] E-mail [ ] Not specified/not applicable [x] Others: satellite images; ground true data from fields visits
Domains	[x] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [x] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics

	[ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [x]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [x] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others: remote sensing, type crops identification
Classification (A – very relevant, B – relevant, C – less relevant)	A

1.11. Multi-data approach for crop classification using multitemporal, dual-polarimetric TerraSAR-X data, and official geodata

SPECIFICATION	DESCRIPTION
Bibliographic data	Hütt, C., & Waldhoff, G. (2018). Multi-data approach for crop classification using multitemporal, dual-polarimetric TerraSAR-X data, and official geodata. *European Journal of Remote Sensing*. https://doi.org/10.1080/22797254.2017.1401909
Link	https://doi.org/10.1080/22797254.2017.1401909
Short overview (strengths, weaknesses)	• General aim: The goal of the paper is using multi-temporal SAR data combined with official geodata for crop identification using machine learning algorithms. • Strengths: The paper is the next example of successful crop classification using SAR images. The used methodology allows to recognize 5 classes of grouped crops (winter cereals, maize, sugar beet, rapeseed, potato). The papers shows the possibility of pre-processing and classification using open source software (SNAP). Additionally, the paper includes quality aspects of performed image classification. • Weaknesses: The advanced method of classification and accuracy evaluation was performed using commercial software (ArcGIS).
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [x] Images, Videos [ ] E-mail [ ] Not specified/not applicable [x] Others: satellite images; ground true data from fields visits
Domains	[x] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [x] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....

Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [x]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [x] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others: remote sensing, type crops identification
Classification (A – very relevant, B – relevant, C – less relevant)	A

1.12. A review of assessing the accuracy of classifications of remotely sensed data

SPECIFICATION	DESCRIPTION
Bibliographic data	Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*. https://doi.org/10.1016/0034-4257(91)90048-B
Link	https://doi.org/10.1016/0034-4257(91)90048-B
Short overview (strengths, weaknesses)	• General aim: The paper reviews the necessary considerations and available techniques for assessing the accuracy of remotely sensed data. • Strengths: The paper is well-described principles of assessing the accuracy of remotely sensed data with examples. • Weaknesses: The basic experience with remotely sensed is needed to understand the process of accuracy assessment.
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [x] Images, Videos [ ] E-mail [ ] Not specified/not applicable [x] Others: satellite images; ground true data from fields visits
Domains	[x] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [x] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [x]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics

	[ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [x] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others: remote sensing, type crops identification
Classification (A – very relevant, B – relevant, C – less relevant)	A

1.13. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data

SPECIFICATION	DESCRIPTION
Bibliographic data	Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*. https://doi.org/10.1109/LGRS.2017.2681128
Link	https://doi.org/10.1109/LGRS.2017.2681128
Short overview (strengths, weaknesses)	• General aim: The paper describes a multilevel deep learning architecture that targets land cover and crop type based on multi-temporal and –source remote sensed satellite data. • Strengths: The use of the deep learning algorithms for crop recognition is detailed described. The presented method uses open source software (Orfeo Toolbox). The paper includes quality aspects of performed image classification. • Weaknesses: The classification process is detailed described while the pre-processing is described in general terms
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [x] Images, Videos [ ] E-mail [ ] Not specified/not applicable [x] Others: satellite images; ground true data from fields visits
Domains	[x] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [x] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture

	[ ]Financial transaction data [x]Earth observation [ ]Mobile networks data [ ]Innovative tourism statistics [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [x] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others: remote sensing, type crops identification
Classification (A – very relevant, B – relevant, C – less relevant)	A

1.14. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification

SPECIFICATION	DESCRIPTION
Bibliographic data	Elber P.., Bischke, B., Dengel A.& Borth D. (2017). EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. arXiv preprint arXiv:1709.00029v2, 2019.
Link	https://arxiv.org/abs/1709.00029
Short overview (strengths, weaknesses)	- The challenge of the paper is classifying the land cover and, possibly, land use by adopting remote sensing satellite imagery. - Proof of the extent to which machine learning is capable to detect satellite imagery areas in order to identify the land use or land cover changes. This can lead to improve geographical maps updates etc. • Weaknesses of the paper - All the paper relies on a hand-crafted classification standard called "EuroSAT" which is not really compliant with Eurostat's Lucas or Corinne Land Cover classification standards. This can determine transcoding tables amongst the diverse standards and can affect the final Land Cover estimation.
Data sources	[ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [X] Images, Videos [ ] E-mail [ ] Not specified/not applicable [X] Land Cover/Use Maps
Domains	[X] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [X] Environment and energy [X] Science, technology, digital society [ ] Not applicable Please specify the domain: Environment
Possible applications	[ ]Online job vacancies [ ]Enterprise characteristics [ ]Smart energy [ ]Tracking ships [ ]Process and architecture [ ]Financial transaction data [X]Earth observation [ ]Mobile networks data

	[ ]Innovative tourism statistics [X] Other Official Statistics indicators such as Land Cover, Land Use, GDP, etc.
Quality aspects	Are quality indicators mentioned? [ ] yes [X] no
Keywords	[ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [X] Machine learning [ ] Data visualization [X] Others Remote Sensing, Land Cover Classification, Convolutional Neural Networks
Classification (A – very relevant, B – relevant, C – less relevant)	A

1.15. Classifying websites by industry sector: a study in feature design

SPECIFICATION	DESCRIPTION
Bibliographic data	Berardi, G., Esuli, A., Fagni, T., & Sebastiani, F. (2015, April). Classifying websites by industry sector: a study in feature design. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (pp. 1053-1059). ACM.
Link	https://www.esuli.it/publications/SAC2015a.pdf http://dx.doi.org/10.1145/2695664.2695722
Short overview (strengths, weaknesses)	The aim of the paper was to classify websites according to industry codes (custom classification from market research) using scraped text. A Support Vector Machine was used for classification. Strengths are the extensive feature engineering and selection procedure (text to data) as well as the hierarchical classification scheme. The efficiency of different feature engineering and selection approaches is compared. A high model performance could be achieved. Whether and how the very imbalanced data was handled is not clear from the text. Only one machine learning algorithm was used.
Data sources	Please check all the data sources used in the paper: [x] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [ ] Others .....
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [x] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable

	Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: text mining
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), A

1.16. Classifying firms with text mining

SPECIFICATION	DESCRIPTION
Bibliographic data	Caterini, G. (2018). Classifying firms with text mining. Department of Economics and Management.
Link	https://www.researchgate.net/publication/331590502_Classifying_Firms_with_Text_Mining_29th_EC2_on_Big_Data_Econometrics_with_Applications_Rome_2018
Short overview (strengths, weaknesses)	The aim of this paper was to create statistics on birth, deaths and survival rates of enterprises using the general business register and textual descriptions of economic activity which were written by enterprises at the time of registration. Missing values and errors for the economic activity (NACE) of enterprises were predicted using machine learning. A pairwise comparison of enterprises that registered in the same area was then conducted to identify the reactivation of enterprises. Furthermore, random forest was used to predict the reactivation of enterprises. This study shows that textual data sources from administrative data can also be used for text mining and machine learning. It has the advantage that the text has the purpose of describing the enterprise characteristic of interest (economic activity in this case). Weaknesses of the study were a rather simplistic feature engineering procedure. Additionally, the available data was not well documented (e.g. class imbalance).
Data sources	Please check all the data sources used in the paper: [ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [x] Others: Administrative data, textual descriptions of economic activity
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [x] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade

	[ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: text mining, missing data handling
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B

1.17. A Heuristic Approach for Website Classification with Mixed Feature Extractors

SPECIFICATION	DESCRIPTION
Bibliographic data	Du, M., Han, Y., & Zhao, L. (2018, December). A Heuristic Approach for Website Classification with Mixed Feature Extractors. In 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS) (pp. 134-141). IEEE.
Link	https://doi.org/10.1109/PADSW.2018.8645042
Short overview (strengths, weaknesses)	The authors present a website classification schema based on deep neural networks. The algorithm showed a better performance than traditional machine learning models. The strength of the paper are the very advanced feature extraction methods. The paper is not written with the application of official statistics in mind, but it could be used for website classification and the derivation of enterprise characteristics.
Data sources	Please check all the data sources used in the paper: [x] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [x] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies

	[x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: text mining, deep learning
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B

1.18. Three Methods for Occupation coding Based on Statistical Learning

SPECIFICATION	DESCRIPTION
Bibliographic data	Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. Journal of Official Statistics, 33(1), 101-122.
Link	https://www.degruyter.com/downloadpdf/j/jos.2017.33.issue-1/jos-2017-0006/jos-2017-0006.pdf
Short overview (strengths, weaknesses)	Machine learning methods are applied to the automatic encoding of open-ended questions for the example of occupation. The methods used could be applied to other text mining tasks and to scraped data. A strength of the study are production rate plots that show the trade-off between accuracy and automatic coding. Additionally, both deterministic and machine learning methods are combined into the same algorithm. Since very short, relatively clean texts are used for classification in this article, the methods might have to be adapted for text data from internet sources.
Data sources	Please check all the data sources used in the paper: [ ] Web data [ ] Social media [x] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [x] Others: answers to open-ended questions
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [x] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable Please specify the domain: .....

Possible applications	Please check all possible applications: [ ] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: text mining
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B

1.19. Learning from Imbalanced Data

SPECIFICATION	DESCRIPTION
Bibliographic data	He, H., & Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, (9), 1263-1284.
Link	https://doi.ieeecomputersociety.org/10.1109/TKDE.2008.239
Short overview (strengths, weaknesses)	The article gives an overview over methods for dealing with imbalanced data, which is a frequent problem for classification tasks and is therefore also relevant to official statistics (e.g. NACE classification). Both methodologies to deal with imbalanced data as well as evaluation metrics are discussed.
Data sources	Please check all the data sources used in the paper: [ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [x] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [x] Not applicable Please specify the domain: .....

Possible applications	Please check all possible applications: [x] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [x] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [x] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [x] Data noise [x] Machine learning [ ] Data visualization [x] Others: imbalanced data
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), A

1.20. Measuring the UK's Digital Economy with Big Data

SPECIFICATION	DESCRIPTION
Bibliographic data	Nathan, M., Rosso, A., Gatten, T., Majmudar, P., & Mitchell, A. (2013). *Measuring the UK's Digital Economy with Big Data*. London: National Institute of Economic and Social Research.
Link	http://www.niesr.ac.uk/sites/default/files/publications/SI024_GI_NIESR_Google_Report12.pdf
Short overview (strengths, weaknesses)	Nathan et al. use the economic activity of enterprises to determine whether they belong to the digital economy. They use data from the commercial partner Growth Intelligence which includes collected and preprocessed internet data from various sources and has information about the economic activity of every enterprise in the UK. Therefore, no matching of official statistics data and big data is needed, which is an advantage of that very rich data set. No official classification for economic activity is used, but an internal classification system of Growth Intelligence. The data collection and processing procedures seem to be not very transparent and are done nearly exclusively by Growth Intelligence.
Data sources	Please check all the data sources used in the paper: [x] Web data [x] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [x] Others: Preprocessed big data by commercial partner
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [x] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable

	Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [ ] yes [x] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others:
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B

1.21. A knowledge-based approach to unstructured data

SPECIFICATION	DESCRIPTION
Bibliographic data	Oostdijk, Nelleke (2018). A knowledge-based approach to unstructured data. CBS, Heerlen.
Link	https://www.aanmelder.nl/102324/wiki/355576/sessions-abstracts-papers-and-presentations https://www.aanmelder.nl/i/doc/447c46bbb0574872242085faab003537?forcedownload=True
Short overview (strengths, weaknesses)	The paper compares machine learning and knowledge-based (or deterministic) approaches to text classification. It highlights in which scenarios a knowledge-based approach is favorable compared to machine learning. Examples from text classification in social media are used for illustration.
Data sources	Please check all the data sources used in the paper: [ ] Web data [x] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [x] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [x] Online job vacancies [x] Enterprise characteristics

	[ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [x] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: text mining, deterministic approaches
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B

1.22. Measuring the internet economy in The Netherlands: a big data analysis

SPECIFICATION	DESCRIPTION
Bibliographic data	Oostrom, L., Walker, A. N., Staats, B., Slootbeek-Van Laar, M., Azurduy, S. O., & Rooijakkers, B. (2016). Measuring the internet economy in The Netherlands: a big data analysis. *The Hague: Statistics Netherlands*.
Link	https://www.nederlandict.nl/wp-content/uploads/2016/10/measuring-the-internet-economy.pdf
Short overview (strengths, weaknesses)	This paper uses the preprocessed data of a commercial partner (Dataprovider) to measure the internet economy in the Netherlands. The Dataprovider data contains a list of all enterprise websites in the Netherlands and some characteristics of these websites. That data was matched to existing CBS data (eg. the GBR) to derive characteristics of enterprises based on their web presence. The article offers a pragmatic definition of the internet economy based on how enterprises use the internet. Weaknesses of the study are that the matching process could only identify websites for 35% of all Dutch enterprises while most likely a higher proportion of enterprises have a website. Also, no social media profiles were taken into account.
Data sources	Please check all the data sources used in the paper: [x] Web data [ ] Social media [x] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [x] Others: Preprocessed data of commercial partner/Trusted Smart Statistics
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [x] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society [ ] Not applicable

	Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [x] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [ ] Others:
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), A

1.23. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation

SPECIFICATION	DESCRIPTION
Bibliographic data	Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1).
Link	https://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/27165/Powers%20Evaluation.pdf?sequence=1&isAllowed=y
Short overview (strengths, weaknesses)	The article gives an overview over evaluation measures for machine learning algorithms. It argues that common evaluation metrics like Precision, Recall and F-Measure are biased and other metrics should be preferred. The concepts of Informedness and Markedness are introduced. These quality indicators for machine learning models seem useful for official statistics, too.
Data sources	Please check all the data sources used in the paper: [ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [x] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy [ ] Science, technology, digital society

	[x] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [x] Online job vacancies [x] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [ ] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [x] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [x] Machine learning [ ] Data visualization [x] Others: evaluation measures
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), A

1.24. Data quality management, data usage experience and acquisition intention of big data analytics

SPECIFICATION	DESCRIPTION
Bibliographic data	Kwon, H., Lee, N., Shin, B., Data quality management, data usage experience and acquisition intention of big data analytics. [w:] International Journal of Information Management. Jun2014, Vol. 34 Issue 3, p387-394.
Link	https://www.semanticscholar.org/paper/Data-quality-management%2C-data-usage-experience-and-Kwon-Lee/ea35336a9151b41d54f2d0adef7e9bece40548c0
Short overview (strengths, weaknesses)	This paper show the relevance of consistency and completeness in using the big data analytics. Thus, data quality is represented by two attributes – consistency and completeness – to show the usefulness of the data used by enterprises. The strength of the paper are the tested hypotheses of the relevance of this two data quality attributes.
Data sources	Please check all the data sources used in the paper: [ ] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [X] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport [ ] Environment and energy

	[ ] Science, technology, digital society [x] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [ ] Online job vacancies [ ] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [x] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [ ] Innovative tourism statistics [x] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [x] Data quality [ ] Data integration [ ] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [ ] Machine learning [ ] Data visualization [x] Others: data quality evaluation
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), C

1.25. Data quality management, data usage experience and acquisition intention of big data analytics

SPECIFICATION	DESCRIPTION
Bibliographic data	Matsudaira K., Capturing and Structuring Data Mined from the Web, Communications of the ACM, March 2014, Vol. 57, No. 3.
Link	https://dl.acm.org/doi/abs/10.1145/2567664
Short overview (strengths, weaknesses)	Web data mining is defined as a method to take data from web. URLs are classified into three hierarchical elements: • URLs that exist but are unknown • URL frontier • URLs crawled so far The process of web mining is as follow: 1. DNS lookup 2. Fetch robots.txt 3. Fetch URL 4. Parse content 5. Store results It is useful for web scraping activities as it is oriented to web data.
Data sources	Please check all the data sources used in the paper: [X] Web data [ ] Social media [ ] Traditional surveys [ ] Web searches data (incl. Google Trends) [ ] Census data [ ] AIS data [ ] Sensors [ ] Images, Videos [ ] E-mail [ ] Financial transaction data [ ] Not specified/not applicable [ ] Others:
Domains	Please check the statistical theme of the paper: [ ] General and regional statistics [ ] Economy and finance [ ] Population and social conditions [ ] Industry, trade and services [ ] Agriculture and fisheries [ ] International trade [ ] Transport

	[ ] Environment and energy [ ] Science, technology, digital society [x] Not applicable Please specify the domain: .....
Possible applications	Please check all possible applications: [X] Online job vacancies [X] Enterprise characteristics [ ] Smart energy [ ] Tracking ships [X] Process and architecture [ ] Financial transaction data [ ] Earth observation [ ] Mobile networks data [X] Innovative tourism statistics [ ] Shared economy [ ] Other ……..
Quality aspects	Are quality indicators mentioned? [x] yes [ ] no
Keywords	Please check all that apply (3-5 keywords) or add others: [ ] Mobile phones (incl. Call detail records, Mobile call records) [ ] Non-probability sampling [ ] Sentiment analysis [ ] Data quality [ ] Data integration [X] Big Data definition [ ] Big Data infrastructure [ ] Big Data architecture [ ] Data noise [ ] Machine learning [ ] Data visualization [ ] Others: ……..
Classification (A – very relevant, B – relevant, C – less relevant)	Assess the relevance of the paper to apply in official statistics: (A – very relevant, B – relevant, C – less relevant), B