Trusted Smart Statistics:
methodological developments based on new data sources

2022-IT-TSS-METH-TOO
Project n. 101132744

Work Package 3
Methodologies and open source tools for integrating MNO and non-MNO data sources

Deliverable 3.1

Preliminary report on methodologies

May 2024
Partner in charge: SSB (Norway)

Authors[1]: Li-Chun Zhang, SSB (Norway)
Marian Necula, INS (Romania)
Danila Filipponi, Giorgia Simeoni, Tiziana Tuoto, Istat (Italy)
Magdalena Six, STAT (Austria)

MNO-MINDS | Eurostat CROS (europa.eu)

Co-funded by
the European Union

[1] Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Preliminary report on methodologies

MNO-MINDS

25.05.2024

Mobile network operator (MNO) data have great potentials for producing official statistics on population, tourism, mobility and environment. However, MNO data would not suffice on their own whenever the target statistical unit or the measurement unit is not mobile device *per se*.

This preliminary report constitutes Deliverable D3.1 from ESSnet project MNO-MINDS WP3, Methodologies and open source tools for integrating MNO and non-MNO data sources. It consists of three chapters. A reference frame is presented in Chapter One, as a common basis for examining all the methods relevant to utilising MNO data. Chapter Two provides a review of the literature on existing statistical methods and related quality assessment. The methods will be appraised in light of the needs and requirements of official statistics. An outline of further developments for WP3 is given in Chapter Three.

Depending on how the associated uncertainty is defined, one can classify any statistical method under three broad approaches: randomisation, quasi-randomisation and super-population modelling.

- *Randomisation* requires a specialised survey to convert the MNO data into the target statistical outputs, whose uncertainty is considered to be dominated by the sampling error under the known survey sampling design.

- Although MNO data are not observed based on some known probabilities, one may introduce a model of the underlying mechanism *as if* they were, and assess the uncertainty accordingly. Such a *quasi-randomisation* approach can be implemented together with suitable non-MNO population data and, if fit-for-purpose, can remove the need of specialised surveys altogether.

- It is often possible to build a so-called *super-population* model for specific variables from non-MNO sources, using features derived from relevant MNO data. However, different models are needed for different statistics generally, unlike building a quasi-randomisation model that is applicable to all the different variables associated with the same mobile devices.

In Chapters Two and Three we will explain the challenges that exist for all the three approaches and the outlined developments in order to address them. Without specifying the target statistics and the available MNO and non-MNO data, any potential solutions can only be discussed in a preliminary manner at this stage. However, we hope to have anticipated and covered the key elements that are likely to be required in future use-cases and applications.

For those readers who may be interested in examining closely the techniques mentioned in the sequel, we note that the notations are not always consistent across the different parts of this report. This is simply due to the diverse background of the methods and ideas, as well as the conventions that exist in the respective fields of literature. However, we have made an effort to keep the notations self-contained and consistent in each subsection, such as Section 2.3.3, and as far as possible in each section, such as Section 3.2.

Finally, one needs not to read the report page after page. For instance, after Chapter One, it is possible to read first the introductory text in Chapters Two and Three, respectively, before Sections 2.1 and 3.1, and then select among the different sections depending on one's interest. It might also be helpful to read the related sections in tandem, such as Section 2.1 followed by Section 3.1 now that both deal with the randomisation approach.

Contents

1 Introduction

1.1 MNO data

Signal contacts between a mobile *device* and the *mobile network operator (MNO)* infrastructure may have a recorded *time* and a *cell-ID* of the tower (or base station) hosting antennas. Regardless the purposes of contacts, we shall refer to such *(device, time, cell-ID)* records as the *nano MNO data*.

Whilst the time attribute is largely unproblematic, the other two attributes of nano MNO data may cause challenges to secondary statistical uses:

- insofar as the target statistical or measurement unit is *not* mobile device *per se*, a conversion from devices to the target units will be necessary;
- the device's position at the time of contact and movement over time need to be *inferred* (or approximated) from the recorded cell-ID, according to the network infrastructure and various operational contingencies.

Meanwhile, the Official Statistics Agency (OSA) cannot have access to nano MNO data, due to confidentiality, commercial interest and technology reasons. What is being made available to the OSA is (anonymised) *macro* MNO data (Multi-MNO project), which refers to summary measures over multiple devices within a specified time period, such as the number of devices that moved from city A to B during the 24 hours on May 25, 2024, provided it is larger than a specified confidentiality threshold.

However, provided necessary and appropriate computational and regulatory support, it may become possible to process *micro* MNO data in a *multiparty* confidential setting (Ricciato, 2024; Zhang and Haraldsen, 2022), in order to enhance the resulting macro MNO data. Here, micro MNO data refers to summary measures of each distinct device within a specified time period, such as whether or not a particular device moved from city A to B during the 24 hours on May 25, 2024. The multiple parties may include several MNOs, as well as the OSA that contributes data from non-MNO sources. It should be stressed that the final outputs accessible to the parties will remain in the form of (anonymised) macro data. Confidential multiparty computing at the micro level is only a means to enhance the aggregated outputs; but micro MNO data need not to be and will not be revealed to any party, including the owner MNO itself if this is considered desirable.

The methods for combing MNO and non-MNO data for official statistics will assume macro MNO data as outlined above, possibly enhanced by confidential multiparty micro data computing. The only exception will be an *opt-in smart survey*, whereby informed consent is given by sampled individuals to collect their micro or even nano MNO data directly.

1.2 A reference frame for methods

Two structured approaches to data integration have proven to be useful in the past. First, adopting a total error framework allows one to analyse and identify the most important error sources in each situation (Zhang, 2012; Reid et al,

2017; Rocci et al, 2022). Next, specifying a range of generic settings of the data to be combined can provide practical guidance to the relevant methods (ESSnet KOMUSO, 2019; De Waal et al, 2020a). To capture the range of problems pertaining to combining MNO and non-MNO data, we propose a reference frame (Figure 1) that combines the elements from both the approaches, which enables one to place and examine all the relevant methods on a common basis.

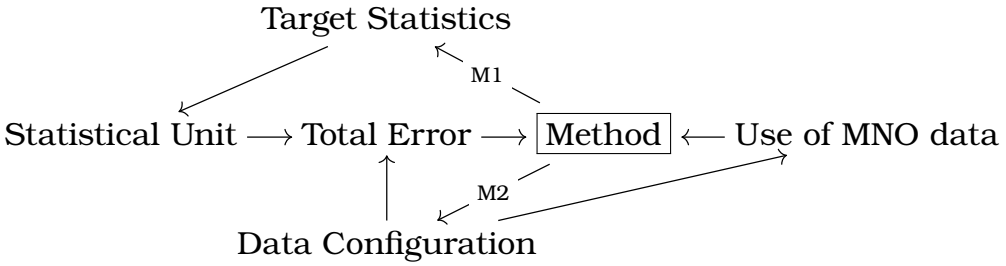

Figure 1: Reference frame for methods to combine MNO and non-MNO data. M1, M-executor methods; M2, M-enabler methods.

Firstly, the statistical unit (and the population) of interest follows from the definition of the target statistics. It is generally the case for official statistics that one is not interested in making statistics of mobile devices but of persons or spatial objects (such as a city centre). Even when each detected device corresponds to a person belonging to the target population, all the detected devices would rarely, if ever, correspond to the target population. It is therefore critical to maintain the distinction between the statistical unit (e.g., person) and the observation unit (device), in order to deal with the potential coverage or selection bias of MNO data.

Secondly, data configuration is characterised by the aggregation process of MNO data and any additional relevant features.

- Whether macro MNO data can be enhanced by micro-data integration will affect the choice of methods. The granularity of location (cell-ID, map grid, administrative area, etc.) or movement (between municipalities, within city, route on a street map, etc.) matters, as well as the reference time period (e.g. within each 24-hour period, or over 12 months).

- Similarly for the non-MNO data. For instance, the choice of methods and the resulting uncertainty of tourism statistics will be affected, depending on whether airline passenger counts are available by the different flights or only as a daily total (e.g. of arrivals at a given airport).

Thirdly, it is also important to clarify whether the MNO data are to be treated as the target measures or as auxiliary information (i.e. covariate, feature) for non-MNO measures, which is referred to as "Use of MNO data" in Figure 1. For instance, if the MNO origin-destination (OD) trip counts are treated directly to as the number of persons making such trips, then bias can be caused by the non-representativity of the detected devices and possibly the errors associated with the OD classification. However, if the same device counts are used as auxiliary information to a Travel Survey, where trip data are collected from the survey respondents directly, then these MNO counts would no longer be a

cause of bias to the resulting statistics, whether the MNO and survey data are combined at the individual or aggregated level.

Now, given the statistical unit (and population), the data configuration and the use of MNO data, it becomes possible to conduct a total-error analysis with respect to the target statistics, in order to identify the most important errors, as well as the corresponding methods that are required to deal with them.

Finally, we shall distinguish two types of methods, referred to as M-executor and M-enabler, respectively. M-executors are methods that are applicable to the available data in the given configuration, and M-enablers are methods that enable alternative improved data configurations and M-executor methods. Let us illustrate with two examples.

- Suppose there are relatively too many employed persons underlying the detected devices compared to that in the target population, whereas the MNOs cannot produce separate macro data according to census or register-based employment status of the device users. Any means of micro processing enhancement (mentioned in Section 1.1) that can enable suitably adjusted macro MNO data would be an M-enabler method in this situation, which would affect the feasible M-executor methods.

- Suppose MNO counts are classified according to device 'Home' municipality assigned by the MNO. Any longitudinal analysis algorithm that produces the MNO-Home classification can be regarded as an M-enabler method in this context, which would affect the properties of MNO-Home classification and the M-executor methods that make use of the resulting macro MNO data. Such matters are within the scope of the ongoing Multi-MNO project.

Although we will be mainly dealing with M-executors, attention will be given to M-enablers when appropriate, such that the development of methods for combining MNO and non-MNO data not only accommodates the existing data configurations but also point to more favourable scenarios in future.

2 Literature review and appraisal

Ahas et al. (2007) illustrate early the potentials of using mobile phone data that are relevant to official statistics. United Nations (2019) provides a first overview in this respect. Nichols et al. (2023) offer recently a comprehensive survey of the literature aimed at the use of mobile phone location data in official statistics, as well as other social, demographic and health studies. The main topics in official statistics are population estimates, mobility, socio-economic indicators, and epidemic (covid-19) tracing-monitoring.

Given that nano MNO data are unavailable, inference of device positions is out of our scope here. To the extent it matters to the target statistics, the errors will have to be dealt with by other methods than directly modelling device position conditional on cell-ID (e.g. Tennekes and Gootzen, 2022).

Below we review the relevant *statistical methods* for using macro MNO data. Many ways of organisation are possible. One can e.g. broadly divide between parametric or non-parametric methods, MNO data used as target or auxiliary

measures, prediction (or regression) vs. other techniques. We shall adopt an *inferential* perspective, which distinguishes *how the associated uncertainty is conceptualised and measured*, whereby all the relevant statistical methods can be classified according to the three broad approaches below.

- *Randomisation* is also commonly referred to as the design-based approach in survey sampling, where a survey is conducted under some known sampling design and the uncertainty of the resulting estimation is considered to be dominated by the associated sampling error.

- *Quasi-randomisation* is a common *model-based* approach to observational studies or nonprobability samples: given observations that are not selected according to some known probabilities, one could postulate a model of the observation mechanism of the MNO data *as if* they had been obtained by designed randomisation, and the same mechanism is applicable to all the attributes associated with detected devices.

- *Super-population* modelling is another common *model-based* approach to observational studies or nonprobability samples. Unlike quasi-randomisation, which e.g. builds a selection model applicable to multiple outcome variables, a super-population model is tailored to specific outcome variables, such that different models are needed for different outcomes generally. The distinction between super-population and quasi-randomisation modelling is convenient and traditional in official or survey statistics (e.g. Zhang, 2019).

In short, we first distinguish whether the basis of inference is a known sampling design or an assumed statistical model and, for model-based methods, whether the assumed model is about the target-agnostic observation mechanism or specific outcome variables.

Notice that although all the relevant statistical methods (or techniques) we have come across can be classified in this manner, different types of methods may be required in a given application to produce the target statistics, in case one needs to deal with multiple important sources of error — as discussed for the total error analysis (Figure 1) previously.

Without attempting to compile an exhaustive reference list of all the relevant methods, we do aim to cover the most typical ideas of the different approaches. Moreover, the existing methods will be appraised to identify the developments relevant to the needs of official statistics.

2.1 Randomisation

Grassini and Dugheri (2022) give tourism statistics in Estonia and Indonesia as currently the only accredited official statistics based on MNO data.

The Central Bank of Estonia has been the official host of the statistics since 2008. The Methodology@Estonia relies on confidential processing of nano MNO data over time, combined with card payment transactions.

The method in Indonesia (Lestari et al. 2018) for producing foreign visitor statistics combines macro MNO counts with a tailored sample survey. One can view a macro MNO count here either as a target measure of devices, or

as a proxy to the target measure of persons (instead of devices), for which the sample survey is used to estimate a device-to-person adjustment factor. The method is a typical example of randomisation approach due to the tailored sample survey and the absence of any statistical model.

To be specific, let m be the number of active-roaming foreign SIM cards, which is a macro MNO count after MNO processing to remove the out-of-scope devices such as carried by fast fliers, seamen, accidental roamers. The total number of foreign visitors *corresponding to* m can be written as

$$N = wm \quad \text{and} \quad w = \xi^{-1}\{P_r(1 - P_w)\}(1 - P_{nr})^{-1}$$

which involves (i) deduplication from foreign devices to travellers, via

$$\xi \text{ = no. active-roaming SIMs per foreign traveller;}$$

(ii) subsetting of tourists among the foreign travellers, via

$$P_r \text{ = proportion of foreign residents among travellers,}$$
$$P_w \text{ = proportion of workers among foreign residents;}$$

and (iii) weighting tourist-roamers to all foreign tourists, via

$$P_{nr} \text{ = proportion of non-roamers among foreign tourists, including e.g. without phone, turn off roaming, switch to local SIMs.}$$

A sample survey is used to estimate w (including all its constituent quantities), which is conceptualised as a finite-population parameter, i.e. the ratio between two population constants m (observed) and N (unknown). In applications, the adjustment is stratified by the border regions, and the estimate of w varies between 0.48 and 2.58 in Table 2 of Lestari et al. (2018).

Two central lessons are worth noting for this randomisation-based method of the only accredited official statistics based on macro MNO data.

- Survey sampling provides a universally valid approach for utilising macro MNO data, just like survey sampling could have been *without* the MNO data. Thus, the added value of MNO data here lies primarily in efficiency gains and reduced sample size (compared to what is necessary otherwise).

- The long-term cost of randomisation approach will *not* be negligible, not least because the adjustment factor w is target-specific, in that it only applies to a particular MNO count m. Different factors are needed across space (e.g. the border regions), time, and topics (e.g. domestic vs. foreign visitors).

2.2 Quasi-randomisation

Expansion of macro MNO counts according to official population sizes and MNO market shares is commonly practiced for the sake of 'representativity'. The underlying idea belongs to the quasi-randomisation approach, which postulates a selection model of the detected devices for the MNO counts.

In terms of Figure 1, the macro MNO counts are treated as target measures, albeit based on a subset of the target population. The non-uniform selection

probabilities across the population are determined by the assumed selection model, the most convenient of which amounts to post-stratification.

The method of Suarez Castillo et al. (2024) is typical. Denote by r the MNO-detected Home (place) for any device $d \in D_r$, and let $D = \cup_r D_r$ contain all the devices. Denote by $j_{d,t}$ the (contact) cell of device d during time t, where $j_{d,t}$ is imputed even if no contacts exist for d during t. We have

$$\sum_j m_{jr,t} \equiv |D_r| \quad \text{where} \quad m_{jr,t} = \sum_{d \in D_r} \mathbb{I}(j_{d,t} = j)$$

is the macro MNO count of Home-r devices with contact cell j during t. Let the weight (or expansion factor) from device to population U_r, $\forall d \in D_r$, be

$$w_d = w(r) = \frac{|U_r|}{|D_r|}$$

Let the device *location* i conditional on $j_{d,t} = j$ be given according to

$$\theta_{ij} = \Pr(i \mid j)$$

Predict Home-r individuals at location i during t by

$$E(N_{ir,t} \mid D_r, [m_{jr,t}]) = w(r) \sum_j \theta_{ij} m_{jr,t}$$

where $[m_{jr,t}]$ denotes the matrix of MNO counts by j and r given t.

Clearly, $w_d = w(r)$ for any $d \in D_r$ amounts to a post-stratification model of selection (by r). To see the problem with this model, imagine one has a perfect location technique such that $\theta_{ij} = 1$ if $i = i_j$ and 0 otherwise, i.e. a location i_j can be assigned without error given j. We would then have

$$E(N_{ir,t} \mid D_r, [m_{jr,t}]) = w(r) m_{ir,t}$$

where

$$m_{ir,t} = \sum_{d \in D_r} \mathbb{I}(i_{j_{d,t}} = i) \quad \text{and} \quad \sum_i m_{ir,t} \equiv |D_r|$$

i.e. a straightforward post-stratification estimator, where the devices D_r are treated as a completely random sample from U_r.

We note that CBS (2020) adopts the same post-stratification model by MNO-Home, but allows for multiple contact cells for each device during a given t and varying active device totals $|D_{r,t}|$ (instead of constant $|D_r|$ by $j_{d,t}$-imputation). These modifications affect only the conditional distribution of location given contact cells and the potential location errors, but not the selection model that characterises the quasi-randomisation approach.

However, the MNO-Home selection model is surely mistaken, because the persons carrying the devices D_r can hardly be a proper subset of U_r. Detecting Home location from device positions is just *not the same* measurement concept underlying the official statistics on $|U_r|$, whether the latter is produced based on population census, sample surveys or administrative registers.

Notice that the Multi-MNO project aims to reduce the spuriousness of MNO-Home classification by leveraging data over a longer period, say, 12 months. This is likely to improve the compatibility between MNO-Home and the usual residence concept, but it cannot remove the definitional discrepancy. External validation will be needed to show if the MNO-Home selection model can become fit-for-purpose, e.g. by auditing as discussed in Section 2.4.

2.3 Super-population modelling

Super-population modelling is perhaps the most common approach to MNO data particularly in applications outside the OSAs, where it is simply known as 'statistical modelling'. As explained earlier, we have adopted 'super-population' to emphasise the distinction to 'quasi-randomisation' modelling. From now on, the shorthand QR may be used for quasi-randomisation and SP for super-population.

2.3.1 Data fusion

The simplest SP modelling approach is to assume that the target distribution (e.g. related to the population of residents) is *the same* as a distribution derived from mobile devices directly. In the literature of using mobile phone position data, this is sometimes referred to as a *data fusion* approach, perhaps because the assumption cannot be empirically established based on the data that are actually available, similarly to statistical matching or data fusion problems.

For example, Batista e Silva et al. (2020) explore temporal changes in EU population density by dasymetric mapping, which is an interpolation technique that disaggregates population counts per administrative areas or census zones to a finer set of spatial units using a 'covariate' distribution of higher spatial resolution, such as macro MNO counts of cellphone contact records between the mobile devices and cell towers at high temporal frequency. The authors call it a data fusion approach, which essentially imputes the distributions required for disaggregation by those derived from a suitable geotagged covariate, such as MNO data or social media posts.

Another example can be found in Koebe et al. (2022), where a large area population size is disaggregated into the small areas therein proportionally to a covariate count with high spatial resolution, while respecting the benchmark constraints at the large-area level. Mobile phone data and satellite imagery are mentioned as possible covariate sources, although only satellite image data are used in the said application.

Such data fusion methods clearly require a high degree of faith and subject-matter judgement. Insofar as the associated error cannot be quantified based on the data actually available, external validation such as auditing would be necessary in order to accept the outputs as official statistics.

2.3.2 MNO features for prediction, time series, etc.

In terms of Figure 1, macro MNO data are used here as features (or covariates) in SP prediction models of some target outcome variable from non-MNO

sources, such as sample survey or census. It is of course possible to include additional features from other non-MNO sources. The term 'prediction models' implies that supervised learning is needed.

Although prediction models may be the most common in practice, other types of models can also make use of features extracted from MNO data. For instance, van den Brakel et al. (2017) study a bivariate time series model, where one series contain sample survey estimates and the other is derived from social media posts. Obviously, the approach remains feasible in principle, if one of the series is compiled based on MNO data.

Such use of 'MNO features' in model-based estimation for official statistics is not unusual conceptually speaking. The question yet is to demonstrate that one actually succeeds in making official statistics in this way.

Some brief illustrations of prediction models using macro MNO features are given below. Due to the popularity of random effects in small area estimation, we make a distinction between fixed effects and mixed effects models.

Fixed effects models Douglass et al. (2015) consider census population counts in Lombardy, denoted by N_i for *sezione* $i = 1, ..., 10506$ in year 2011. MNO counts of call data records (CDRs) can be obtained for spatial grids of size $235 \times 235m^2$ over November and December 2013. The best single covariate is found to be the number of daily callouts during 10-11am, denoted by m_i, in terms of a linear model

$$E(N_i \mid m_i) = \beta m_i$$

In addition, combining CDR and Land Cover covariates, denoted by x_i, a better random forest model is obtained, denoted by

$$E(N_i \mid x_i) = \mu(x_i)$$

To generate useful population estimates, the authors suggest that the census-trained $\mu(x)$ may be "recalibrated over time... using a very small scale stratified population count in key calibration regions" (Douglass et al, 2015). However, the feasibility of this suggestion is neither clarified nor substantiated.

Mixed effects models The targets may be socio-economic indicators.

Steele et al. (2017) report an application of poverty mapping in Bangladesh, where small area estimation incorporating spatial correlations is applied to relevant survey variables using features generated from satellite remote sensing data, MNO CDR counts or in combination of both.

Schmid et al. (2017) apply the Fay-Herriot model with variable transformation and benchmarking to estimate literacy rates in Senegal by gender and commune (431 of them), where a large number of mobile phone covariates are extracted from tower-to-tower CDRs.

Hadam et al. (2023) apply the Fay–Herriot model with variable transformation to small area estimation for North Rhine-Westphalia in Germany based on the Labour Force Survey. The authors explore an alternative definition of unemployment rate, where the unemployed persons are counted at the place of residence while the employed persons are counted at the place of work. MNO

features are based on mobile activities defined as an event caused by a length of stay in a specific geometry without movement (also known as dwell time). The macro MNO counts are associated with the cell towers.

2.3.3 Geographically weighted regression

Gilardi et al. (2022) apply geographically weighted regression (GWR) to combine road sensor vehicle counts with counts derived from TomTom navigation app in vehicles or mobile phones. The approach is the same if the latter is replaced by similar MNO counts. With reference to Figure 1, such macro MNO counts can be considered at proxies to the sensor counts, where two variables are *proxy* of each other if they have similar definition and the same support (Zhang, 2021b). A proxy to the target measure is a special kind of auxiliary variable or feature, because it is often more powerful than all the other auxiliary variables. For instance, the binary register-employed variable is more predictive for the binary ILO-employed variable than age, education, etc. Moreover, some statistical methods only make sense given proxies but not any other auxiliary variables, such as structure preserving estimation (Purcell and Kish, 1980) in small area estimation, or the situation of Gilardi et al. (2022).

To be specific, let $\{y_i : i \in s\}$ denote the sensor vehicle counts at the set of sites s. Let $\{x_j : j \in R\}$ be the TomTom (or MNO) counts for the set of sites R, where $s \subset R$. Let $d_{ij} \equiv d_{ji}$ be the road distance between $j \in R$ and $i \in s$. For any $j \in R$, GWR yields

$$\hat{y}_j = b_j x_j \qquad \text{where} \qquad b_j = \frac{\sum_{i \in s} w(d_{ij}) x_i y_i}{\sum_{i \in s} w(d_{ij}) x_i^2}$$

as the predicted sensor count at site j, for which y_j may be lacking, given a suitable choice of the weights $w(d_{ij})$ that depend on distances d_{ij}.

GWR (Brunsdon et al, 1996) is a special case of nonparametric regression (Stone 1977) or statistical calibration (Osborne, 1991). We shall consider generalisations or adaptions of GWR in Section 3.3. Together, they form a family of nonparametric methods highly relevant for utilising macro MNO data, which can potentially lead to many novel official statistics of interest.

2.4 Quality assessment and guideline

Salgado et al. (2020) outline a probabilistic framework to the uncertainty of statistics propagated from nano MNO data. This covers device location error, device duplication error, selection error of device carriers, and other relevant errors specific to applications. Although the framework is not operational given only macro MNO data, with or without enhancement by micro-data processing, various elements of it are included in the statistical methods reviewed above. For instance, device deduplication and carrier selection are covered under the randomisation approach in Section 2.1. Or, the device location error is handled by geographically weighted regression in a sensible manner.

It is of course possible to assess the quality of specific statistical outputs originated from MNO data, such as a population spatial density derived from

geographically allocated MNO device counts, either by comparisons to external sources or sensitivity analysis; see e.g. Sakarovitch et al. (2018), Vanhoof et al. (2018), Statistisches Bundesamt (2019) and Ricciato et al. (2020). Such ideas are not very different to those employed to assess register-based statistics in their earlier years; see e.g. Myrskylä (1991).

ESSnet KOMUSO (2019) both collected and developed a number of 'quality measures and calculation methods' for multisource statistics, some of which may also be relevant in the context of combing MNO and non-MNO data, given appropriate data configuration and final statistics. However, as remarked by De Waal et al. (2020b), most of these methods are directed at "separate steps, or building blocks, in the statistical production process. We hope that in the, hopefully near, future, an all-encompassing theory or framework to base quality measures for multisource statistics upon will be developed. Such an all-encompassing theory or framework should be able to handle several different types of error sources at the same time and, preferably, use the same statistical theory to treat these error sources."

In this respect, auditing inference (Zhang, 2023) provides a general and valid design-based approach, which can be applied to evaluate the final statistical outputs directly. As formulated by Zhang (2021), "Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, auditing aims not to estimate the target parameter itself but some chosen error measure of any given estimator of the target parameter..." The approach is as universally applicable as survey sampling, given the same inference basis in finite population sampling theory.

Notwithstanding Quality Guideline for Multisource Statistics from the ESSnet KOMUSO project (Brancato and Ascari, 2019), Quality Guidelines for the Acquisition and Usage of Big Data (Kowarik et al., 2020) from the Essnet Big Data II project pay closer attention to new data sources such as MNO data. The statistical production process is divided into Input phase, Throughput phase I (Lower layer), Throughput phase II (Upper layer) and Output phase. The result from the Input phase is so-called raw data (nano or micro data in the case of MNO data), the result from Throughput phase I is so-called statistical data (e.g. macro MNO data), whereas the statistical output is the final product after the Throughput phase II. For each phase quality guidelines are listed. Since the ESSnet Big Data II covered several types of new data sources - among others MNO data - the guidelines listed are partly source-specific.

Currently, the Multi-MNO project (on a reference processing pipeline of MNO event data and network topology data) is developing a comprehensive quality assurance framework. The development so far has analysed how the general quality requirements from ES Code of Practice and ESS Quality Assurance Framework apply to the statistics based on MNO data and the proposed pipeline and considered the quality issues arising in the Input data. While this quality framework pertains to the entire processing pipeline of MNO data, it does not explicitly cover the later phases of combining MNO and non-MNO data, where the present MNO-MINDS project focuses on the methods of utilising macro MNO data at the Throughout phase II (Upper layer).

Ascari et al. (2023) follow a similar approach aimed at defining a structured quality framework for official statistics based on MNO data. They identify the

main components of the quality framework, highlight specific quality aspects related to the institutional environment and input data, and provide reflections on throughput quality.

Finally, Ascari and Simeoni (2024) examine the production process from nano to macro MNO data, regarding the errors that may occur when including MNO data into a statistical process, and propose to split the first phase of the two-phase data life-cycle model (Zhang, 2012) into two phases, one concerning the mobile phone event (or nano) data and the other the device (or micro) data. Such a split is also relevant to the M-enabler methods, e.g. confidential multiparty micro data computing mentioned in Section 1.1, which may involve multiple MNOs and non-MNO data (e.g. from the OSA).

3 Method development: An outline

Based on our review of relevant statistical methods for using MNO data, the WP2 results of landscaping non-MNO sources and the envisaged deliverable MNO data from the Multi-MNO project, we summarise in Table 1 some potential official statistics based on MNO data.

Table 1: Official statistics based on MNO data

Statistics	Possible examples
(Unit: person)	Long-term *de facto* residents
Population	Census zone population updates
Tourism	Foreign visitors to the country
	Residents going abroad
	Multi-destination trips
Mobility	Commuters by origin-destination
	Commuting time/distance to work
(Unit: spatial object)	Green-area utility
Spatial, Environmental	City-centre traffic

Although the four generic types seem reasonable, i.e. population, tourism, mobility and spatial/environmental statistics, one must take the examples in Table 1 as tentative suggestions. In particular, without defining the target statistics and the available MNO and non-MNO data (in terms of the reference frame in Figure 1), one cannot discuss the details of any relevant methods. Nevertheless, let us take a couple of examples from Table 1 to illustrate the possibility and challenges of applying the methodological approaches reviewed earlier and to be developed further.

Consider the first example of long-term *de facto* residents. Suppose the MNOs together can provide m_i as the device count over the 12 months previous to a given time point t, which have municipality i as the usual environment called Home, where $i = 1, ..., n$, and the target statistics is the number Y_i of in-scope persons with municipality i as the *de facto* place of residence.

- To implement the randomisation approach to estimate the factor $w_i = m_i/Y_i$, let a sample be taken from the population of in-scope persons, such that

$w_i = \xi_i \eta_i$ where ξ_i is the number of devices (underlying m_i) per device user and η_i is the proportion of device users among the Y_i in-scope persons. However, there are several potential complications, such as how to correctly identify the devices relevant to m_i, how to cover the entire population including children, elderly and others who may be impractical to survey directly.

- The simplest estimator under the QR approach is given by $\hat{Y}_i = m_i N/m$, where $m = \sum_{i=1}^{n} m_i$ and $N = \sum_{i=1}^{n} N_i$ given the *de jure* population sizes N_i, assuming $\sum_{i=1}^{n} Y_i = N$. But is this simple assumption about the device selection and duplication effect acceptable? The naïve QR estimator reviewed in Section 3.2 yields $\hat{Y}_i = m_i(N_i/m_i) = N_i$, which breaks down completely here. However, does there exist an acceptable QR selection model otherwise?

- To generate observations of the target *de facto* residents for SP modelling, suppose a sample survey is conducted as in the randomisation approach. Let y_i be a corresponding design-based estimator of Y_i. A simple predictor of Y_i is $\mu_i = m_i \hat{\beta}$ under the model $E(y_i) = E(Y_i) = m_i \beta$, which is model-based despite the use of sample survey, because the validity and variance of μ_i are assessed with respect to the model, in contrast to design-based y_i. To alleviate the bias of potential model misspecification, one may apply a small area estimation technique to obtain $\hat{\mu}_i = \gamma_i y_i + (1-\gamma_i)\mu_i$, where γ_i is a shrinkage coefficient to be estimated, as reviewed in Section 2.3.2.

Consider the last example of city-centre traffic. Suppose point-of-interest (POI) $k = 1, ..., n$ in city A, such as the central railway station ($k = 1$), the zoo ($k = 2$), the opera ($k = 3$), etc. Let the target statistics Y_{kt} be the number of motorised vehicle-passings at POI k on day $t = 1, ..., T$ throughout the year.

- For a randomisation approach, one may sample k and hours h during t, count the number of vehicles passing by k during h, obtain design-based estimate y_{kt} of Y_{kt}. However, the cost would be high in order to cover all (k, t).

- The QR approach is not straightforward here, since the statistical unit is POI and the measurement unit is motorised vehicle, neither of which coincides with the mobile device or user, such that a selection model of detected devices or users would not suffice even when these can be related to any (k, t).

- Suppose there exist road sensors of motorised vehicles in city A at reference points (RPoi) $j = 1, ..., m$, for the purpose of traffic control or congestion tax. Let Y_{jt} be the vehicle count at RPoi j on day t. Let x_{it} be the MNO count of devices (travelling on motorised vehicles) passing by RPoi or POI i. This yields a setting of statistical calibration with proxy measures x_{it} and calibration measures Y_{kt}, similar to that for GWR reviewed in Section 2.3.3.

The outline of method development topics below will be organised according to the three broach approaches reviewed above, while allowing for solutions to specific use-cases (e.g. Table 1) to be developed provided these emerge through the relevant works of WP2, WP4 and the Multi-MNO project.

- In principle, the randomisation approach is applicable to all the statistics. But there are some general challenges that require methodological solutions, in order to move beyond the existing use-case of foreign visitors.

- The aim of developing the QR approach is to establish a *basic* selection model, instead of post-stratification by MNO-Home, which can achieve statistical validity that is as general as possible. If successful, this will profoundly scale up the use of MNO data and impact the uptake of other forms of big data such as payment transactions.
- When it comes to SP modelling, we shall focus on two topics that are currently underdeveloped, statistical calibration and origin-destination models, which can have many MNO data applications (as indicated by the examples above).

Moreover, we shall consider the creation of a sandbox environment of synthetic data and open-source tools, which can support the activities outlined above as well as future method developments. The tools for simulating MNO data (nano to macro) will be described, which are capable of supporting comprehensive and realistic simulation environments. The simulation scenarios will be specified in the coming months, taking into account feedbacks from the SPRINT as well as the outputs of WP2 and Multi-MNO project.

Finally, relevant quality guidelines will be part of the final deliverable D3.4 of MNO-MINDS, pertaining to the processing and integrating of MNO data with non-MNO sources. The work will be based on the methodological developments outlined here, as well as the relevant results of WP2 and WP4. The guidelines would also depend on the envisaged outcomes of the Multi-MNO project. A specific outline of the contents and tasks will be formulated in due course.

3.1 Randomisation

As pointed out earlier, to reduce the long-term cost of sample surveys required for adjusting the relevant MNO data, a key challenge is to improve the efficiency of target-specific survey estimation. *Transfer learning over time and domain* will be considered, since official statistics typically need to be repeated over time and disaggregated over space (or other population domains).

Next, *user ambiguity* can be a major challenge for making statistics of the usual residents, in case macro MNO data can only be organised according to the service contractors instead of the users. The activities of actual device carriers will then need to be either observed or inferred from the available MNO data. It is therefore necessary to develop methods that can enable *valid* randomisation approach in the presence of user ambiguity.

Finally, opt-in smart surveys have attracted attention recently, in which mobile devices are heavily involved. Some relevant elements of methodological development will be discussed.

3.1.1 Transfer learning over time and domain

Denote by $\mu(x;\beta)$ a *target* model with unknown parameters β. Suppose there exists a relevant *source* model for a different though similar population, which has been estimated separately, denoted by $\mu(x;\hat{\theta})$, where the two models belong to the same family with different parameter values β and θ. *Transfer learning* in such a setting aims to improve the estimation of β by leveraging $\hat{\theta}$.

For instance, one may estimate β based on the target observations that are associated with the units in s, subject to a chosen penalty of the discrepancy between β and $\hat{\theta}$, such as minimising

$$\Delta(\beta) = \sum_{i \in s} \{y_i - \mu(x_i; \beta)\}^2 + \gamma \|\beta - \hat{\theta}\|_2$$

given $\gamma > 0$. Although the resulting estimator of β is biased due to the penalty term, the variance of estimation can be greatly reduced compared to estimating β only based on s. One can thus view the approach as a form of regularisation, which has shown to be especially helpful in cases with insufficient number of target observations (e.g. Li et al. 2020; Gu et al., 2023a).

Transfer learning for parametric models above can easily be adapted to model-assisted estimation in survey sampling. For instance, to apply transfer learning to design-consistent generalised regression estimation (GREG), one only needs to replace the unweighted loss over s above for transfer learning by a corresponding 'GREG loss'. But we shall investigate more broadly.

Some remarks are necessary given that transfer learning does introduce a bias to the randomisation approach. First, the result may be fit-for-purpose if the induced bias is limited. Next, a design-consistent estimator may have its own finite-sample bias, given the desired sample size reduction. Finally, bias cannot be avoided practically in the randomisation approach to MNO data. As noted by Lestari et al. (2018), seasonality exists in the number of foreign visitors, "since the survey period was limited, it is necessary to repeat the survey and continue observe and, if necessary, correct for seasonality with proper algorithm". Obviously, any such algorithmic correction would introduce bias just like transfer learning. In short, the question is not whether but how to carry out transfer learning for the randomisation approach.

Remark Transfer learning for nonparametric models such as random forest or boosted trees is still an open topic (e.g. Segev et al. 2017; Gu et al. 2023b).

3.1.2 User ambiguity

There are two major complications for device-to-person conversion. The first one is *device duplication*, in case the user can be identified but the MNO cannot deduplicate the multiple devices of a given user, as illustrated to the left in Figure 2. The second one is *user ambiguity*, in case the MNO cannot identify all the users, if a contractor can subscribe for a device without specifying its user, as illustrated to the right in Figure 2, which is the situation in many European countries.

In the use-case of foreign visitors (Lestari et al. 2018), device deduplication is achieved via the coefficient ξ, whereas user ambiguity seems non-existent if one can assume that all the travellers are surveyed directly.

However, for statistics pertaining to the whole domestic population, directly sampling and surveying any in-scope person, including children, teenagers or other specific individuals, may be either infeasible or too costly. It would therefore be necessary and useful to develop sampling and estimation methods

Figure 2: Device-individual connections, device d_1, d_2, d_3, individual k_1, k_2, k_3. Left, device-user (solid). Right, device-contractor in addition (dashed).

in the presence of user ambiguity and device duplication, which allow for any contractor-user connections that may exist in the population.

3.1.3 Opt-in smart survey

An opt-in smart survey is administered through digital instruments, including mobile devices such as phones and tablets. Data from the respondents may be collected actively or passively, which is sometimes known as data donation. There are two ESSnet projects on this topic: ESSNet Smart Surveys 2020-2022 worked to create a common methodological and architectural framework; the ongoing ESSNet Smart Surveys Implementation 2023 will pilot services and solutions for Time Use and Household Budget surveys.

There are unavoidably new problems of representation and measurement associated with smart surveys. For instance, the respondents (or data donors) to an opt-in smart survey are unlikely to be a completely random sample from the target population, censoring effects and measurement errors will arise whether based on active or passive data collection.

Smart surveys can be viewed chiefly as an effort in the evolution of survey methods, which is not situated at the core of this ESSnet project on combining MNO and non-MNO data. But we shall at least consider some methodological elements that may be relevant to the other development topics outlined here, such as user ambiguity, adjustment of selection effects.

3.2 Quasi-randomisation

Let $Y = \sum_{k \in U} y_k$ be the target total over population U. For any given time t, let D_t contain all the *in-scope, active and deduplicated* devices, such that different MNO-measures in $\{y_d : d \in D_t\}$ refer to different units in U and all the carriers of the devices in D_t form a subset of U which varies with t, denoted by

$$P_t = P(D_t) \subset U$$

Let $y_k = y_{k_d}$ be detected (or observed), where k_d is the carrier of $d \in D_t$. Let y_k be missing if $k \notin U_t$. Suppose the QR selection model

$$\Pr(k \in P_t \mid k \in U, x_k, y_k) = \pi(x_k) \tag{1}$$

such that we have *non-informative selection (NIS)* $k \in P_t$ regarding y_k given x_k,

where x_k is generally a vector of features associated with each $k \in U$. We have

$$E(n_x) = \sum_{k \in U} \Pr(k \in P_t \mid x_k = x)\ \mathbb{I}(x_k = x) = \pi(x) N_x$$
$$E(y_x) = \sum_{k \in U} \Pr(k \in P_t \mid x_k = x)\ \mathbb{I}(x_k = x)\ y_k = \pi(x) Y_x$$

where $N_x = \sum_{k \in U} \mathbb{I}(x_k = x)$ and $Y_x = \sum_{k \in U} \mathbb{I}(x_k = x) y_k$ are the population size and total given x, respectively, and $n_x = \sum_{d \in D_t} \mathbb{I}(x_{k_d} = x)$ and $y_x = \sum_{d \in D_t} \mathbb{I}(x_{k_d} = x) y_d$ are the macro MNO counts that exist by definition. Consistent estimation of Y is possible given $\{N_x\}$ from OSA and $\{n_x, y_x\}$ from MNO, since

$$Y = \sum_x N_x \frac{E(y_x)}{E(n_x)} \tag{2}$$

Note that we ignore device duplication and user ambiguity for the moment, but return to them in Section 3.2.2.

Denote by P all the mobile users with active subscription, $P \subset U$, which is treated as stable over any given period of time. The detected mobile phone users P_t form a subset of P, $P_t \subseteq P$. Unlike the randomisation approach, the QR approach is *target-agnostic* since, under the model (1), there exist features $x_U = \{x_i : i \in U\}$ such that, for any $y_U = \{y_i : i \in U\}$ of interest, we have

$$\frac{\sum_{i \in U} y_i \mathbb{I}(x_i = x)}{\sum_{i \in U} \mathbb{I}(x_i = x)} \approx E\Big(\frac{\sum_{i \in P_t} y_i \mathbb{I}(x_i = x)}{\sum_{i \in P_t} \mathbb{I}(x_i = x)}\Big) \tag{3}$$

where the expectation is with respect to the random event $i \in P_t$ given any t.

In the QR approach based on (1) - (3), the timely variation of P_t (e.g. whether a user has detected signals in a given hour) is treated as random given x_U and unrelated to y_U, such that it affects only the size of P_t (i.e. variance of estimation) but does not cause bias of QR-based estimation.

3.2.1 Proof of concept

Only the MNOs know P_t. Neither is P available in its entirety. However, the OSA may have the possibility to automatically search P for mobile phone numbers associated with the sampled individuals in its household surveys. Let s be a simple random cluster sample from U (with household as cluster), and let

$$s_P = s \cap P$$

Whatever imbalance between P and U is expected to be mirrored by s_P and s. Allowing for the extra sampling variation, one can explore (3) by replacing U with s and P_t with s_P, and using any known $Z_U = \{z_i : i \in U\}$ instead of y_U, such as register-based (or census-based) employment status, education level, places of home-work. Notice that one only needs to associate all the specified features z_i to s (but not necessarily to U) in data preparation.

Analysis of such automatic phone number search results has traditionally been conducted Norway. Karlsson et al. (2013, Table 2) show similar results

in Iceland and Norway for their respective EU-SILC samples, with respect to gender, age group, foreign born or not, married or not, and education level. Mobile phone numbers are found automatically for approximately 80% of the sample in Iceland and 81% in Norway.

Table 2: Percentages in a random sample s of persons age 18-79 in 2015, Norway, and subsample s_P of those with mobile phone, or with email address, or with mobile, email, landline. Source: Lagerstrøm and Wangen (2015).

		s_P according to contact information		
Feature	s	Mobile	Email	Mobile, email or landline
Gender				
Male	50.8	51.4	52.2	51.0
Female	49.2	48.6	47.8	49.0
Age				
< 24	13.9	14.2	15.8	14.4
25-34	18.0	18.3	20.0	17.9
35-44	19.5	20.0	21.4	19.6
45-54	17.8	18.1	18.3	18.0
55-64	14.9	14.8	14.3	14.9
> 64	16.0	14.6	10.2	15.3
Education				
Low	24.1	23.6	22.3	23.4
Middle	39.5	39.8	38.8	40.4
High	30.0	31.3	33.3	30.5
Unknown	6.4	5.3	5.6	5.8
Origin				
Native born	82.9	83.9	83.4	84.1
Foreign born	17.1	16.1	16.6	15.9
Total	4000	3750	3366	3868

Table 2 shows the results from an internal report prepared by Lagerstrøm and Wangen (2015). In addition to the mobile phone s_P, we also include s_P of those with email address, as well as s_P with any form of contact including mobile, email and landline phone. Any discrepancy between a percentage in s and s_P is an unbiased estimate of the underlying difference between U and P, the latter of which is the bias of treating P as a simple random sample (SRS) from U with respect to the given feature, i.e. the simplest QR selection model

$$\Pr(i \in P \mid i \in U) \equiv \pi$$

The results are largely compatible with the SRS model, since the discrepancies are quite small, which seems even more promising given that the mobile phone user percentage must have increased since 2015. The coverage of automatic search has increased to 93.75% compared to 2013.

To prove the concept of a broadly valid target-agnostic QR approach to the use of MNO data, we propose to conduct experiments or pseudo experiments as explained below, which considerably extend the scope of Lagerstrøm and

Wangen (2015) and Karlsson et al. (2013). First, the SRS model may not hold for all the *other* features that may be of interest, and it may be inadequate for various disaggregation needs that arise naturally in the use of MNO data. Moreover, one needs to account for the uncertainty of analysis, such as using (s, s_P) for (U, P). Finally, appropriate methods for assessing the uncertainty of the statistics generated by the QR approach need to be developed.

Experiment In countries where mobile phone interview is a standard survey mode in practice, it may be possible to identify s_P given any s without contacting the sample units. There is then no need to actually survey the sample units. The unit of analysis will be persons, but drawing a household sample makes it easier to investigate scenarios of use ambiguity in addition. We refer to this as the setup for a proof-of-concept experiment.

Pseudo experiment For a pseudo experiment, suppose one has a sample s from a completed mixed-mode household survey, where it is possible to label the respondents s_P with mobile phone as the survey mode. In this setup we need to take into account several additional complications below.

- The probability $\Pr(i \in s)$ may not be equal over U. While one may examine (3) in terms of the respondents s_P and the gross sample s, expansion to the population may be necessary to make the analysis relevant.
- The analysis may be unduly affected by inappropriate survey nonresponse adjustment that is necessary in order to generalise from s_P.
- It may be unclear whether mobile phone or landline is used. Possible modes (including mobile phone) are offered as an option after the initial contact with sample units is made, such that the subjective choice of mobile phone may have its own selection effect (in addition to response or not).

3.2.2 Potential complications

Several potential difficulties of implementing the QR approach require a study.

MNO classification The key difficult to implement any selection model here lies in the lack of access to micro MNO data, which makes it impossible to obtain macro MNO data related to proper subsets of arbitrary subpopulations. Any M-enabler in this respect is a method that can generate MNO counts (y_x, n_x) for any given subpopulation $U_x = \{i \in U : x_i = x\}$.

For any $d \in D_t$, let $x^*_{k_d}$ be available to MNO instead of x_{k_d}, where $x^*_{k_d}$ and x_{k_d} are not always be equal to each other. For instance, x_k may be the Population Register home-municipality of person k, and x^*_k the MNO-Home municipality.

Ideally, the QR selection model should depend on features that are known to both the OSA and the MNO, such that the MNO misclassification problem here can be avoided. However, in case the QR-model must use features that are unknown to the MNO, statistical adjustments will be explored.

Sparse cells problem Confidentiality concerns may prevent the MNOs from delivering a large number of macro counts, if there are many sparse cell counts close to 0. Empty sample cells despite non-empty population cells could also cause complications to simple post-stratified estimation.

For instance, let $j_{k,t}$ indicate the present locality of any given $k \in U$ during t. Let the population mobility measure by residence r and locality j be

$$Y_{jr} = \sum_{k \in U} \mathbb{I}(r_k = r) y_{kj} \quad \text{and} \quad y_{kj} = \mathbb{I}(j_{k,t} = j)$$

where r is the residence known to the OSA, e.g. in the Population Register. In case $d \in D_t$, we have $k_d \in U_t$ and $y_{k_d j} = \mathbb{I}(j_{d,t} = j)$ by the device presence locality, whereas y_{kj} is missing if $k \notin U_t$. Let n_{rx} be the number of active device carriers in U_t with $(r_k, x_k) = (r, x)$, among whom y_{jrx} are present at locality j. Under the QR-model (1), we have

$$E(n_{rx}) = \pi(x) N_{rx} \qquad \text{and} \qquad E(y_{jrx}) = \pi(x) Y_{jrx}$$

where N_{rx} is the number of all individuals in U with $(r_k, x_k) = (r, x)$, and Y_{jrx} is the number of those among Y_{jr}. Similarly to (2), we have

$$Y_{jr} = \sum_x N_{rx} \frac{E(y_{jrx})}{E(n_{rx})}$$

This illustrates a potential sparse cells problem, where MNO counts (y_{jrx}, n_{rx}) are needed instead of (y_x, n_x) in (2).

Should one relax the QR-adjustment by using a less detailed classification than x, can one use some marginal MNO counts instead of the cross-classified ones? The pros and cons of such alternatives need to be investigated.

Device duplication Let y_k be the value of interest for any $k \in U$. Let $y_d = y_k$ for any $(dk) \in A$ given device duplication (Figure 2), where A contains the device-user connections. The *observed* total of y_d over D, denoted by Y_D, is given as

$$Y_D = \sum_{d \in D} y_d = \sum_{(dk) \in A} y_k = \sum_{k \in U} y_k \alpha_k$$

where α_k is the number of devices of each $k \in U$, who is a user iff $\alpha_k > 0$. The overall target-agnostic factor

$$\bar{\alpha} = \frac{1}{N} \sum_{k \in U} \alpha_k$$

can be applied to *any* Y_D given above, provided

$$Y_D = \sum_{k \in U} y_k \alpha_k = \Big(\sum_{k \in U} y_k \Big) \Big(\frac{\sum_{k \in U} \alpha_k}{N} \Big) = Y_U \bar{\alpha}$$

i.e. (finite-population) *non-informative device-duplication (NIDD)* if

$$Cov_U(y_k, \alpha_k) = \frac{\sum_{k\in U} y_k\alpha_k}{N} - \Big(\frac{\sum_{k\in U} y_k}{N}\Big)\Big(\frac{\sum_{k\in U} \alpha_k}{N}\Big) = 0$$

This NIDD-assumption is a special case of the condition for valid inference from non-probability samples given by Zhang (2019). Insofar as it is unrealistic for the whole population, one may replace it by a *stratified-NIDD* assumption, whereby NIDD holds in various subpopulations defined according to a feature vector denoted by x, provided which we have

$$\bar{Y}_{D,x} = \frac{Y_{D,x}}{\sum_{k\in U_x} \alpha_k} = \frac{\sum_{k\in U_x} y_k\alpha_k}{\sum_{k\in U_x} \alpha_k} = \frac{\sum_{k\in U_x} y_k}{N_x} = \bar{Y}_x$$

i.e. the subpopulation mean over persons $\bar{Y}_x$ is equal to the subpopulation mean over devices $\bar{Y}_{D,x}$, so that there would be *no need* to estimate the target-agnostic conversion factor $\bar{\alpha}_x = \sum_{k\in U_x} \alpha_k/N_x$ at all. The matter can be included in a proof-of-concept experiment or pseudo-experiment if α_k is available.

User ambiguity Given user ambiguity, $\{\alpha_k : k \in U\}$ are unknown, and Y_D is an *apparent* sum over all the contractors instead of users, i.e.

$$\sum_{(dk)\in C} y_k = \sum_{k\in U} y_k\zeta_k$$

where the edges C are the *device-contractor* connections (dashed in Figure 3), and ζ_k is the number of edges in C of each individual in U, who is a contractor iff $\zeta_k > 0$. For instance, $y_{d_3} = y_{k_3}$ in Figure 3 would be associated with k_2 instead of k_3. The approach above for dealing with device duplication is no longer applicable due to unknown $Y_{D,x}$ and $\bar{Y}_{D,x}$ in the presence of user ambiguity.

Figure 3: Multigraph given user ambiguity, device d_1, d_2, d_3, individual k_1, k_2, k_3. Left, bipartite (as in Figure 2). Right, non-bipartite with user-contractor edges.

Let the number of devices used by j and contracted by k be given by

$$\nu_{kj} = \sum_{d\in D} c_{dk}a_{dj}$$

such as $\nu_{k_2k_3} = 1$ in Figure 3, where $a_{dj} = 1$ if device d is used by individual j and 0 otherwise, and $c_{dk} = 1$ if device d is contracted by individual k and 0 otherwise. Note that ν_{kk} is the number of devices used and contracted by k. It

follows that

$$\alpha_j = \sum_{k \in U} \nu_{kj} \qquad \text{and} \qquad \zeta_k = \sum_{j \in U} \nu_{kj}$$

The MNO total Y_D can then be written as

$$Y_D = \sum_{j,k \in U} \nu_{kj} y_j = \sum_{\substack{k \in U \\ \zeta_k > 0}} y_k \nu_{kk} + \sum_{\substack{k \in U \\ \zeta_k > 0}} \sum_{\substack{j \in U \\ j \neq k}} y_j \nu_{kj}$$

The last expression decomposes Y_D into a 'self-representing' part (involving ν_{kk}) and an 'indirect-representing' part of other users, since

$$y_k^* = \sum_{\substack{j \in U \\ j \neq k}} y_j \nu_{kj}$$

is the apparent total attributed to the contractor k by the MNO which is actually due to the usage and activity of other users than k.

We will investigate if the QR approach can be extended to accommodate user ambiguity. First, in case one introduces a second model in addition to the selection model (1), can the model be estimated from the available data? Next, if sample survey is used in addition to the selection model, can it be aligned with the treatment of user ambiguity in the randomisation approach?

3.3 Statistical calibration

As Osborne (1991) points out, the term "statistical calibration" is perhaps best explained by analogy to the process of scientific calibration, which determines or adjusts the scale of a measuring instrument on the basis of a 'calibration experiment'. For example, let $\{x_j : j \in R\}$ be all the imprecise MNO-measures and $\{y_i : i \in s\}$ the trusted 'calibration measures' from non-MNO sources, where the latter are only available for a subset $s \subset R$. Viewing $\{(x_i, y_i) : i \in s\}$ as a calibration experiment, one may estimate y_j by adjusting x_j for all the rest $j \in R \setminus s$. By *statistical calibration* we shall refer to any method that uses proxy MNO-measurements in such manners resembling scientific calibration.

3.3.1 Spatial statistical calibration

For any $j \in R$, the best predictor of y_j given x_j is

$$E(y_j \mid x_j) = \int y f(y \mid x_j) dy = \frac{\int y f(x_j \mid y) f(y) dy}{\int f(x_j \mid y) f(y) dy}$$

Suppose observations $\{y_i : i \in s\}$ where $s \subset R$. Replacing $f(y)$ by its empirical distribution function arising from s, we can estimate $E(y_j \mid x_j)$ by

$$\hat{y}_j = \sum_{i \in s} w_i(x_j, s) y_i \tag{4}$$

where

$$w_i(x_j, s) = \frac{f(x_j \mid y_i; s)}{\sum_{k \in s} f(x_j \mid y_k; s)}$$

and $f(x \mid y; s)$ is an estimator of the conditional density based on s. An estimator of the conditional cumulative distribution function of y_j given x_j is

$$\hat{F}(y \mid x_j) = \sum_{i \in s} \mathbb{I}(y_i \leq y) w_i(x_j, s)$$

Viewed as a nonparametric regression estimator (Stone, 1977), the weights $w_i(x_j, s)$ in (4) can be given in many other ways. In case the elements of s and R are spatial points, geographically weighted regression (GWR) would yield

$$w_i(x_j, s) = \frac{w(d_{ij}) x_i x_j}{\sum_{k \in s} w(d_{kj}) x_k^2} \qquad \text{and} \qquad \hat{y}_j = x_j \frac{\sum_{i \in s} w(d_{ij}) x_i y_i}{\sum_{k \in s} w(d_{kj}) x_k^2}$$

as reviewed before, where $d_{ij} = d_{ji}$ is some chosen measure of the distance between $i \in s$ and $j \in R$. For instance, $w(d_{ij}) = \mathbb{I}(d_{ij} < d)$ for a given threshold value d, or $w(d_{ij}) = \exp\{-\alpha d_{ij}^2\}$ given the tuning constant α, or the *bisquare*

$$w(d_{ij}) = \mathbb{I}(d_{ij} < d) \left(1 - d_{ij}^2 / d^2\right)^2$$

Remark Spatial statistical calibration is a widely applicable nonparametric SP modelling approach for combining MNO and non-MNO data.

3.3.2 Network statistical calibration

Network statistical calibration is a statistical calibration approach, where the data are given an underlying graph structure (of which spatial data are a special case), and the target measures are subject to network constraints.

For instance, denote by $G = (U, A)$ a *road network* with crossroads U, and $(ij) \in A$ iff road exists in direction i to j for any $i, j \in U$. Figure 4 illustrates a part of road network with 4 crossroads, the incoming and outgoing vehicle directions, as well as two fixed road sensors marked as triangles.

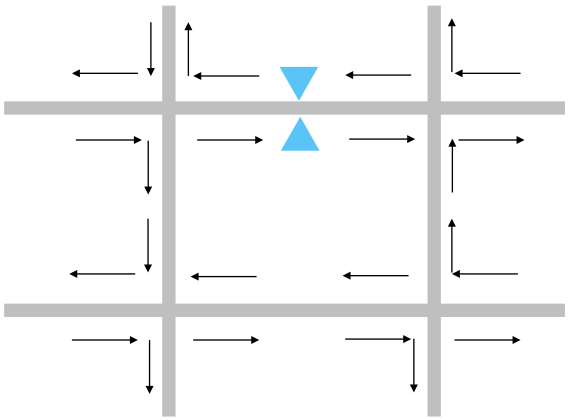

Figure 4: Illustration of road network with two fixed sensors (triangles)

Let $\{y_{ij} : (ij) \in A_s\}$ be the trusted vehicle counts over (ij) in A_s, where $A_s \subset A$, obtained by fixed road sensors or other suitable means. Let $\{x_{ij} : (ij) \in A\}$ be the proxy MNO-counts, where x_{ij} and y_{ij} may differ for various reasons, e.g. x_{ij} may include multiple devices travelling in the same vehicle or devices travelling without vehicles at all.

By network statistical calibration, we shall obtain for any $(ij) \in A$,

$$\begin{aligned} &\hat{y}_{ij} = \tau(x_{ij}; A_s) \doteq E(y_{ij} \mid x_{ij}) \\ &\text{subject to} \quad \sum_{(ji)\in A} \hat{y}_{ji} = \sum_{(ij)\in A} \hat{y}_{ij} \quad \text{for any } i \in U_0 \end{aligned} \tag{5}$$

i.e. the numbers of incoming and outgoing vehicles, called the inflows and outflows, must be equal to each other at any crossroad in U_0, where $U_0 \subseteq U$. Since $E(y_{ij} \mid x_{ij})$ can be modelled without the network constraints, the notation "$\doteq$" in (5) signifies that $\hat{y}_{ij}$ may be close to the conditional expectations of y_{ij} given x_{ij} but not directly given as the estimates of such. Moreover, $U_0 \subseteq U$ generally. For instance, U_0 may consist of the four crossroads in Figure 4 but not the other undepicted ones (that must exist in addition to U_0). One may refer to U_0 as the *interior nodes* of G; non-interior nodes $U \setminus U_0$ exist as long as G is not a closed network. We are not concerned with all the inflows and outflows of any non-interior node except those to and from the nodes U_0.

Remark Network statistical calibration at street level may not be natural given only MNO data, not least because the nano MNO data tend not to be frequent enough for such detailed routing. However, it is as relevant if the nodes are geographic areas instead of crossroads and the connections are not roads *per se* but refer to spatial congruity or origin-destination relationships.

3.3.3 Compositional statistical calibration

Compositional data are proportions of some whole (Aitchison, 1982). Denote by K the fixed number of components. A set of K counts are compositional either given their total or any one of them, i.e. without loss of information the counts can then be transformed to proportions that sum to 1, where we have at most $K-1$ 'freely-varying' counts or proportions. In *compositional statistical calibration* the target and proxy measures are treated as compositional data.

Suppose $K = 2$ to focus on the basic idea and to simplify the notation. Let (x_1, x_2) be the two proxy MNO-counts for the target measurements (y_1, y_2), of which only y_1 is known from non-MNO sources but not y_2. To estimate y_2 given y_1 is the same as estimating the proportion $y_2/(y_1+y_2)$ or ratio y_2/y_1 given y_1. As an example, y_1 may be the known number of cinema goers on a given day and y_2 the unknown restaurant diners, for which MNO-counts (x_1, x_2) are available.

Simple methods of data fusion have been applied to integrate compositional MNO-data (e.g. Batista e Silva et al, 2020). For instance, let

$$E\Big(\frac{y_2}{y_1+y_2}\Big) = E\Big(\frac{x_2}{x_1+x_2}\Big)$$

be the stipulated binomial distribution of (y_1, y_2), such that we obtain

$$\hat{y}_2 = x_2(y_1/x_1)$$

which can as well be given by statistical calibration under the assumption

$$E(y_k \mid x_k) = \beta x_k$$

for any k. However, such data fusion assumptions are too limited generally.

Denote by $i = 1, ..., n$ the different cities or areas, each of which is associated with (x_{1i}, x_{2i}) and (y_{1i}, y_{2i}). To estimate $\{y_{2i}\}$ given $\theta_i = x_{2i}/x_{1i}$ and a tuning constant $\gamma \geq 0$, consider minimising

$$L = \sum_{i=1}^{n}(p_i - \theta_i)^2 + \gamma \sum_{i=1}^{n}(p_i - \mu_i)^2 \tag{6}$$

with respect to p_i and μ_i, where $p_i = y_{2i}/y_{1i}$ and

$$\mu_i = \mu(y_{1i}, z_i)$$

is a predictor of p_i which may depend on additional covariates z_i. For instance, in case y_{2i} is the number of restaurant diners in city i, one may let z_i include the city population size, its number of restaurants, etc.

By (6), the otherwise unconstrained sum of $(p_i - \theta_i)^2$ is regularised (via γ) by a penalty in terms of a model of p_i given the relevant covariates, where $p_i - \mu_i$ would have been the model discrepancy had y_{2i} been observed. It is of course possible to attach weights to each $(p_i - \theta_i)^2$ or $(p_i - \mu_i)^2$ in (6), and so on.

Illustration For a quick illustration of compositional statistical calibration by (6), let us consider the special case without additional z_i and simply use

$$E(y_{2i} \mid y_{1i}) = \beta y_{1i} \quad \Rightarrow \quad \mu_i = E(p_{2i} \mid y_{1i}) = \beta$$

i.e. the penalty is just the smoothness of p_i in the absence of z_i. The estimator $\hat{y}_{2i} = x_{2i}y_{1i}/x_{1i}$ above would follow from minimising the first term of (6) on its own if $\gamma = 0$. More generally, by minimising (6), we obtain

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} p_i = \bar{p} \qquad \text{and} \qquad p_j = \frac{1}{1+\gamma}\theta_j + \frac{\gamma}{1+\gamma}\bar{p}$$

Thus, we recover $p_j = \theta_j$ if $\gamma = 0$. Whereas, if we let $\gamma = 1$, then we would obtain $p_j = \frac{1}{2}(\theta_j + \bar{p})$ instead, which is solved by $p_j = \frac{1}{2}(\theta_j + \bar{\theta})$ given $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_i$.

Remark Identifiability of (6) is a key issue in general terms, although the illustration above has already demonstrated its feasibility in the simplest setup. Notice that identifiability can always be achieved in case necessary external information (or estimates) can be plugged into (6), e.g. by sample surveys.

3.4 Origin-destination estimation

Mobile devices can generate various origin-destination (OD) mobility data. To adjust for inevitable population coverage errors and possible misclassification errors of OD, one may consider OD models that exist in multiple disciplines such as transportation, population migration, spatial econometrics. It is also worth investigating other approaches that can make use of either mathematical models or alternative statistical models of network flows.

3.4.1 Regression estimation

Take e.g. the spatial interaction model considered by LeSage and Fischer (2008), which can be given as

$$\log E(Y_{ij}) = \alpha + h_i^\top \beta + g_j^\top \phi + d_{ij}\theta$$

where $E(Y_{ij})$ is the expected flow from origin i to destination j, h_i is a vector of features measuring the 'push' from i and g_i a vector measuring the 'pull' of destination j, and d_{ij} is a suitable distance measure between i and j. Suitable MNO OD-counts x_{ij} (on the log-scale) provide an additional feature.

Suppose there exists a sample survey which yields a separate design-based estimate $y_{ij} = \log \hat{Y}_{ij}$. One can then improve the efficiency of estimation under the assumed model

$$y_{ij} = \alpha + x_{ij}\xi + h_i^\top \beta + g_j^\top \phi + d_{ij}\theta + e_{ij}$$

where e_{ij} is a sampling error, by replacing y_{ij} with a *synthetic* predictor given the estimated regression coefficients

$$\mu_{ij} = \hat{\alpha} + x_{ij}\hat{\xi} + h_i^\top \hat{\beta} + g_j^\top \hat{\phi} + d_{ij}\hat{\theta}$$

Whereas y_{ij} depends only on the sample observations related to the given (i, j), the predictor μ_{ij} depends on all the other observations as well via the estimated regression coefficients. Hence, one may expect μ_{ij} to be less variable than y_{ij}. Notice that this is SP modelling approach rather than randomisation, although it involves a sample survey, since the validity and variance of μ_{ij} are with respect to the assumed OD model (instead of the sampling design).

To alleviate the bias due to potential model misspecification, one may apply *shrinkage estimation* such as in small area estimation reviewed earlier, Let

$$\hat{\mu}_{ij} = \gamma y_{ij} + (1 - \gamma)\mu_{ij}$$

where $\gamma \in [0, 1)$ is the shrinkage coefficient to be estimated, such as under the mixed effects model given as

$$y = \eta + u + e \qquad \text{and} \qquad \eta_{ij} = \alpha + x_{ij}\xi + h_i^\top \beta + g_j^\top \phi + d_{ij}\theta$$

where y, η, e are N-vectors, $N = n^2$, and u is the N-vector of random effects.

LeSage and Fischer (2008) outline also two possibilities to allow for spatial

autoregressive impacts. First, let

$$(I_N - \rho W_d)(I_N - \lambda W_o)y = \eta + e$$

where $W_d = W \otimes I_n$ and $W_o = I_n \otimes W$, I_N and I_n are identity matrices with the specified dimensions, and W is a row-standardised spatial weight matrix whose non-zero elements allow for spatial impacts of congruous or nearby places and $w_{ij} = 0$ if such impact is absent (including $w_{ii} \equiv 0$). Next, let

$$y = \eta + u \qquad \text{and} \qquad (I_N - \rho W_d)(I_N - \lambda W_o)u = e$$

In other words, spatial autoregressive impact is either introduced for y or u. Similar autoregressive impact has been considered in spatial or social analysis, such as Ord (1975), Friedkin (1990) and Leenders (2002).

However, implementation to OD flows may be challenging computationally given potentially a very large number N. For instance, there are nearly 8000 municipalities in Italy, such that N is about 64 million in this setup. In the meantime, the Permanent Census Survey has a yearly sample of 1.4 million households, such that many flows will not be observed in the sample or have only very few instances. The feasibility and efficacy of synthetic or shrinkage estimation will be studied, with or without spatial autoregressive impacts.

3.4.2 Network flow models

Mobility data can be regarded as flows in a (connected) network of places. Let $G = (U, A)$ be a directed network with a cost c_{ij} on each edge $(ij) \in A$ from node i to j in U. Let B denote the $|U| \times |A|$ node-edge incidence matrix, whose elements are $b_{ia} = 1$ and $b_{ja} = -1$ given each edge $a = (ij) \in A$, and $b_{ia} = b_{ja} = 0$ otherwise. The matrix B sums to 0 by each column.

Let y_{ij} be the flow on edge (ij), which can be given a lower bound l_{ij} and upper bound u_{ij}. Each node $i \in U$ can be assigned an integer number b_i to represent its supply (if $b_i > 0$) or demand (if $b_i < 0$). The minimum-cost flow problem (e.g. Ahuja et al., 1993), i.e. $\min_{(ij)\in A} c_{ij}y_{ij}$, can be dealt with by linear programming, i.e. optimisation of linear objective function given linear equality $\{b_i\}$ and inequality $\{(l_{ij}, u_{ij})\}$ constraints. One may refer to such formulations of *network optimisation* problems as *mathematical* network flow models.

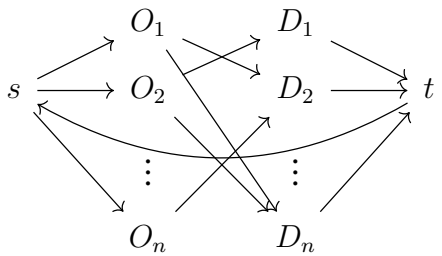

Figure 5: Network flow model of OD mobility.

OD mobility estimation seems possible combining statistical assumptions and network optimisation. Take the setting of Section 3.4.1. As illustrated in

Figure 5, let the nodes in the network G be $U = O \cup D \cup \{s, t\}$, where O consists all the origins and D all the destinations, including when they refer to the same set of places $1, ..., n$. Two additional nodes s and t are introduced, where t accounts for the total inflows to all the destinations and s the total outflows from all the origins, by which the supply and demand are balanced as $b_i \equiv 0$ for any $i \in O \cup D$. The flow from t to s makes $b_s = b_t = 0$ as well.

Let Y_{ij} be the flow from O_i to D_j, given each *permitted* edge (ij), and $Y_{ij} = 0$ by definition if edge (ij) is not permitted (such as O_i to D_i in Figure 5). Moreover, $Y_{jt} = \sum_{i=1}^{n} Y_{ij}$ is the total inflow to destination D_j, and $Y_{si} = \sum_{j=1}^{n} Y_{ij}$ is the total outflow from origin O_i. Finally, $Y_{ts} = \sum_{i=1}^{n} Y_{si} = \sum_{j=1}^{n} Y_{jt}$ is the total flow.

Let the estimates $\{m_{ij}\}$ of $\{Y_{ij}\}$ solve the network optimisation problem

$$\max_{m} \; c^\top m \quad \text{subject to} \quad Bm = 0, \; l \leq m \leq u$$

where c, m, l, u are vectors (over all the edges A in the network). For instance, one can let $c_{ij} = \mu_{ij}$ be the synthetic estimates. It is possible to introduce various constraints, such as $l_{ts} = u_{ts} = \hat{Y}_{ts}$ where $\hat{Y}_{ts}$ is the design-based total flow estimate. Similarly for other selected estimates $\hat{Y}_{si}$, $\hat{Y}_{jt}$ or $\hat{Y}_{ij}$, whether model or design-based. Moreover, the flow bounds (l_{ij}, u_{ij}) can be imposed statistically, e.g. a confidence interval of Y_{ij} or simply $l_{ij} = \min(\hat{Y}_{ij}, \mu_{ij}), u_{ij} = \max(\hat{Y}_{ij}, \mu_{ij})$.

Notice that the solution would be $m \propto c$, if the bound constraints $\{(l_{ij}, u_{ij})\}$ are sufficiently accommodating and $\{c_{ij}\}$ are balanced to start with.

The feasibility of such network flow models will be investigated, including how best to set the costs c, the flow constraints and bounds, as well as other statistical assumptions about the network flows.

3.5 Sandbox of data and tools

In order to tackle data access problems from real-world mobile networks, which is often constrained by privacy concerns, high costs, and logistical challenges, as an alternative, synthetic data generation offers a viable solution to create realistic and representative datasets for developing and testing the proposed methodological. Synthetic mobile network data generation strategy focuses on using the OMNeT++ (OpenSim Ltd. 2024), Simu5G (Nardini et al. 2020), and Veins (Sommer et al. 2018) software stack to generate synthetic mobile network phone data. These tools provide a robust framework for simulating mobile network environments and generating detailed datasets that can be used to enhance the performance and reliability of mobile networks.

OMNeT++ is a discrete event simulation environment that is widely used for simulating communication networks. Its modular and extensible architecture enables one to model a variety of network protocols and technologies. Simu5G is an extension of OMNeT++ that provides a detailed simulation model for 5G networks. It includes features such as enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-type communications (mMTC). Veins is another extension of OMNeT++ that focuses on vehicular network simulations, integrating realistic mobility models with network communication protocols.

One can create comprehensive simulation scenarios that mimic real-world mobile network environments by combining OMNeT++, Simu5G and Veins. These simulations can generate synthetic datasets that include various metrics such as user mobility patterns, call detail records (CDRs), data traffic volumes, and network performance metrics.

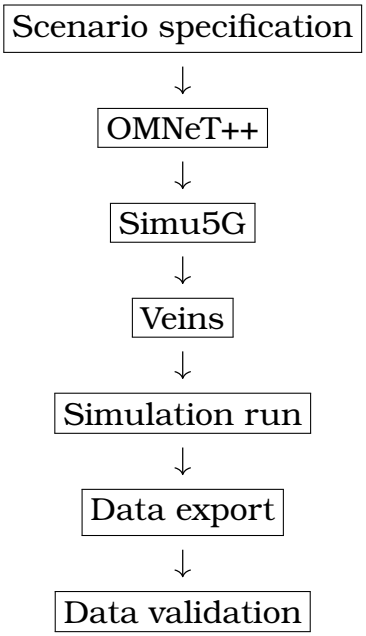

Figure 6: Simulation steps.

The process of generating synthetic MNO data using OMNeT++, Simu5G, and Veins is given in Figure 6. To set up the simulation environment, one needs to define the network topology, specifying the mobility models, and configuring the communication protocols, as illustrated generically by the examples below. In each case the simulation can generate synthetic MNO data at different levels of granularity: nano, micro and macro data.

Example 1: Urban Network In this example, we simulate a mobile network in an urban environment with high population density. The network topology includes a grid of base stations covering a city area, with user devices randomly placed within the coverage area. The mobility model is configured to simulate pedestrian and vehicle movement using random waypoint models, with users moving randomly within the city.

Example 2: Rural Network In this example, we simulate a mobile network in a rural environment with low population density. The network topology includes a few base stations randomly distributed across a large geographic area, with user devices randomly placed within the coverage area. The mobility model is configured to simulate vehicular movement using realistic traffic models, with users moving along roads and highways.

Example 3: Mixed Mobility In this example, we simulate a mobile network in a mixed environment with both urban and rural areas. The network topology includes a mix of base stations covering both city areas and rural regions, with user devices randomly placed within the coverage area. The mobility model is configured to simulate both pedestrian and vehicular movement, with users moving randomly within the city and along roads and highways in rural areas.

The software stack (OMNet++, Simu5G, Veins) is provided as an off-the-shelf virtual machine image, available at `https://simu5g.org/install#download_vm`. Below we provide some more details of the process steps in Figure 6.

Scenario specification The first step in generating synthetic MNO data is to clearly define the objectives and requirements of the simulation. This involves understanding the specific goals of the synthetic data generation effort and detailing the characteristics that the synthetic data should possess. The primary objective is to create realistic and representative datasets.

The synthetic data should exhibit similar statistical properties and patterns as real data. This includes metrics such as call durations, data session frequencies and user mobility patterns. Additionally, the data should be versatile and adaptable to different use cases, which requires specifying the types of data to be generated, such as CDRs and data traffic records. The granularity of the data is also important, as different use cases may require different levels of detail, ranging from high-level summaries (macro) to detailed user-level data (nano). The time interval of the data should be specified, with options for short-term data (minutes or hours) and long-term data (days or months).

The network topology is a critical aspect of the simulation environment to be determined at this stage. The topology defines the structure of the mobile network, including the locations and configurations of base stations, user devices, and other network elements. The topology should be designed to mimic real-world scenarios, taking into account factors such as geographic distribution, population density, and network coverage.

OMNeT++ OMNeT++ is installed and configured on the simulation platform first. OMNeT++ provides a user-friendly graphical interface that simplifies the process of creating and managing simulation projects. Once OMNeT++ is set up, Simu5G can be integrated to provide the necessary models and features for simulating 5G networks.

Simu5G As can be read from its name, Simu5G is an advanced simulation framework that builds on OMNeT++ to model and simulate 5G networks. It incorporates detailed implementations of 5G New Radio features and network elements, allowing one to study and develop 5G technologies in a simulated environment. Understanding its components is crucial for effectively utilising its capabilities.

At the core of Simu5G is the base station, also known as gNodeB, which serves as the primary connection point between user equipment (UE) and the

5G core network. The gNodeB component models the radio interface, implementing physical layer (PHY) and medium access control (MAC) protocols for communication with UEs. It also includes a scheduler that manages radio resource allocation based on various algorithms, and a Radio Resource Control (RRC) module that handles the setup, maintenance, and release of radio bearers. The Packet Data Convergence Protocol (PDCP) manages data transfer, encryption, and compression over the radio interface, while the Service Data Adaptation Protocol (SDAP) manages the mapping between Quality of Service (QoS) flows and data radio bearers.

User Equipment (UE) refers to mobile devices connected to the 5G network. The UE component includes a radio interface that implements PHY and MAC layers for communication with the gNodeB. It features mobility models that simulate user movement across the network, affecting handover and connectivity. The application layer models user applications and traffic generation patterns such as web browsing, video streaming, and VoIP. Similar to gNodeB, the UE component includes RRC, PDCP, and SDAP modules to maintain connections, manage data transfer, and ensure QoS.

The 5G core network is responsible for overall control and data management in the 5G network. Simu5G models key elements of the core network, including a number of functions. The Access and Mobility Management Function (AMF), which manages UE registrations, connection states, mobility, and access control. The Session Management Function (SMF) handles session establishment, modification, and release, managing IP address allocation and QoS parameters. The User Plane Function (UPF) forwards user data packets between the gNodeB and external data networks, while the Network Slice Selection Function (NSSF) manages the selection of network slices, ensuring that UEs are connected to appropriate slices based on their service requirements. The Unified Data Management (UDM) function manages subscriber data and profiles, facilitating authentication and authorisation processes.

In order to create a simulation scenario as close as possible to real world applications, Simu5G provides an application programming interface which implements features such as: network slicing, a fundamental feature of 5G, allows the creation of multiple virtual networks over a common physical infrastructure. Simu5G includes components to model and manage network slices, such as the slice descriptor that defines the characteristics and requirements of a network slice, including resource allocation, QoS parameters, and service types. The slice manager handles the creation, modification, and deletion of network slices, ensuring proper resource allocation and isolation between slices; Quality of Service (QoS) and traffic management are critical for ensuring that different applications and services meet their performance requirements. Simu5G models various QoS mechanisms, including QoS flows that define the QoS requirements for different types of traffic, such as latency, throughput, and reliability. Traffic shaping and policing mechanisms control traffic rates and enforce QoS policies, while priority handling ensures that higher-priority traffic receives preferential treatment in terms of resource allocation and scheduling; Radio Resource Management (RRM) is responsible for efficient utilisation of radio spectrum and resources. Simu5G includes components for resource allocation, implementing algorithms for allocating radio resources to

different UEs based on their QoS requirements and channel conditions.

Moreover, interference management addresses the interference between different cells and UEs, optimising overall network performance. Handover management handles the process of transferring UEs from one cell to another, ensuring seamless connectivity and minimal disruption. Mobility management is essential for maintaining continuous service as UEs move across the network. Simu5G models various aspects of mobility management, including handover procedures that simulate the handover process between gNodeBs, including measurement reporting, decision making, and execution. Tracking area management manages the registration and tracking of UEs within different geographical areas, ensuring efficient paging and location updates. Mobility models provide realistic simulations of UE movement patterns, including pedestrian, vehicular, and aerial mobility.

Simu5G offers comprehensive tools for controlling and configuring simulations, allowing users to define various parameters and scenarios. Simulation configuration files enable users to specify parameters such as network topology, UE density, traffic patterns, and mobility models. Runtime control provides mechanisms for dynamically adjusting simulation parameters and controlling the execution of the simulation. Result collection and analysis tools collect and analyse simulation results, such as performance metrics, QoS indicators, and network statistics.

Veins Veins (Vehicles in Network Simulation) is an open-source simulation framework designed to facilitate the study of vehicular networks by integrating realistic mobility models with communication network protocols. It builds on OMNeT++ and SUMO (Simulation of Urban Mobility) to provide a comprehensive platform for vehicular ad hoc network (VANET) simulations. Veins is integrated to simulate vehicular networks and realistic mobility patterns. Veins includes models for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication, as well as realistic mobility traces that can be used to simulate user movement in urban environments.

Veins integrates two major simulation tools: OMNeT++ for network simulation and SUMO for traffic simulation. OMNeT++ handles the communication aspects, while SUMO simulates vehicle mobility. These two simulation tools are connected through the Traffic Control Interface (TraCI), which allows for real-time exchange of information between the mobility and network simulators. This integration enables the simulation of realistic vehicular communication scenarios where vehicle movements and network communications are closely interlinked.

The OMNeT++ component of Veins provides the environment for simulating network protocols and communication mechanisms. It includes modules that model vehicles as mobile nodes equipped with communication devices, typically IEEE 802.11p or DSRC (Dedicated Short-Range Communications) radios. These modules simulate the communication stack from the physical layer up to the application layer, including channel access, message dissemination, and application-level logic. Vehicles can communicate with each other (vehicle-to-vehicle, V2V) or with roadside infrastructure (vehicle-to-infrastructure, V2I), enabling the study of various VANET applications such as collision avoidance,

traffic management, etc.

The SUMO component simulates the mobility of vehicles, providing realistic traffic patterns and vehicle behaviours. SUMO supports detailed road network modelling, including intersections, traffic lights, and lane changes. It allows for the specification of routes, vehicle types, and driver behaviours, generating realistic vehicle movement patterns based on real-world traffic scenarios. SUMO's flexibility in defining traffic scenarios makes it possible to simulate a wide range of urban and highway environments, contributing to the realism of the vehicular network simulations.

TraCI, the Traffic Control Interface, is the bridge that links OMNeT++ and SUMO. It enables bidirectional communication between the network and traffic simulators. Through TraCI, OMNeT++ can query and control the state of the traffic simulation, such as vehicle positions, speeds, and routes. Conversely, SUMO can receive commands from OMNeT++ to alter vehicle behaviours based on network events, such as rerouting vehicles in response to traffic congestion detected through V2V communication. This real-time interaction ensures that the mobility patterns and network communications influence each other dynamically, providing a more accurate representation of vehicular network scenarios.

The application layer in Veins is where the logic of vehicular applications is implemented. This layer can simulate a wide range of applications, from simple message dissemination to complex cooperative driving scenarios. For instance, safety applications can simulate warning messages broadcasted to nearby vehicles to prevent collisions, while efficiency applications can optimise traffic flow by dynamically adjusting traffic signals based on real-time traffic data. Veins also includes models for simulating the wireless communication environment, such as the propagation of radio waves in urban environments. These models take into account factors like buildings, terrain, and other obstacles that affect signal strength and quality. The accurate modelling of the communication environment is crucial for realistic VANET simulations, as it directly impacts the reliability and performance of vehicular communications.

Mobility models are essential for simulating realistic user movement within the mobile network. Veins provides a range of mobility models that can be used to simulate different types of user movement, including pedestrian, vehicular, and mixed mobility patterns. These models can be configured to match the specific characteristics of the simulation scenario. For example, pedestrian mobility can be simulated using random waypoint models, where users move randomly within a specified area. Vehicular mobility can be simulated using realistic traffic models that take into account factors such as road networks, traffic signals, and vehicle speeds. Mixed mobility models can combine pedestrian and vehicular mobility to simulate scenarios where users move between different modes of transportation. The mobility models should be configured to generate realistic user movement patterns that match the characteristics of the target environment. This includes specifying parameters such as user speeds, movement directions, and stop duration. Veins provides tools for visualizing and analysing mobility traces, which can be used to validate the realism of the simulated movement patterns.

Simulation, data export and validation Once the simulation environment is set up and configured, the next step is to execute the simulation scenarios, generate and collect data on various metrics such as user mobility patterns, call detail records, data traffic volumes aggregated at different levels depending on simulation objectives.

OMNeT++ provides tools for collecting and analyzing simulation data. The data can be exported in various formats, including CSV (Comma-Separated Values) and XML (eXtended Markup Language), for further analysis and processing. The collected data should be organized and structured to facilitate easy access and analysis. This includes creating separate datasets for different types of data, such as CDRs, data traffic records, and performance metrics. To generate realistic mobile network data, the simulation should capture details such as call initiation time, call duration, and the locations of devices. This data can be used to analyze user behavior patterns and network performance. Data traffic records should include information about data sessions, such as the volume of data transferred, session durations, and the types of services accessed. Performance metrics should capture various parameters related to network quality, such as signal strength, latency, and throughput.

The synthetic data generated by the simulation should be validated and evaluated to ensure its ecological validity and accuracy. This involves comparing the synthetic data with real-world data to assess its statistical properties and patterns. Various metrics can be used for validation, including statistical measures such as mean squared error and domain-specific metrics such as call duration distribution and mobility pattern similarity.

4 References

[1] Ahas, R., Aasa, A., Silm, S., Tiru, M. (2007). Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia. *In: Sigala, M., Mich, L., Murphy, J. (eds) Information and Communication Technologies in Tourism 2007*. Springer, Vienna. https://doi.org/10.1007/978-3-211-69566-1_12

[2] Ahuja, R.K., Magnanti, T.L. and Orlin, J.B. (1993). *Network flows: theory, algorithms and applications* Englewood Cliffs (N. J.): Prentice Hall.

[3] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society, Series B*, 44:139-160.

[4] Kowarik et al. (2020). *Deliverable K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data*. Work Package K Methodology and quality, ESSnet Big Data II.

[5] Ascari G., Simeoni G (2024) Applying the extended Total Survey Error approach to statistics based on new data sources: the case of Mobile Network Operators data.*Presented at the European Conference on quality in Official Statistics, Estoril, June 2024*.

[6] Ascari, G., Cerasti, E., Faricelli, C., Mattera, P., Piombo, S., Radini, R., Simeoni, G. and Tuoto, T. (2023). Quality aspects using Mobile Network Operators data for Official Statistics. *Presented at The Second Workshop on methodologies for Official Statistics, Istat, Rome.*

[7] Batista e Silva, F., Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M. A., Ziemba, L., Craglia, M., Koomen, E., and Lavalle, C. (2020). Uncovering temporal changes in europe's population density patterns using a data fusion approach. Nature communications, 11:1-11.

[8] Brancato, G. and Ascari, G. (2019). Quality Guidelines for Multisource Statistics. *ESSnet KOMUSO WP1 Deliverable*.

[9] Brunsdon, C., Fotheringham, A. S. and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28:281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

[10] CBS (2020). Estimating hourly population flows in the Netherlands. *Statistics Netherlands. Report.*

[11] De Waal, T., van Delden, A. and Scholtus, S. (2020a). Multi-source Statistics: Basic Situations and Methods. *International Statistical Review*, 88:203-228.

[12] De Waal, T, van Delden, A. and S. Scholtus (2020b). Commonly used methods for measuring output quality of multisource statistics. *Spanish Journal of Statistics 2*:79-107.

[13] Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D. and Tatem, A.J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), 15888-15893.

[14] Douglass, R., Meyer, D.A., Ram, M., Rideout, D., and Song, D. (2015). High resolution population estimates from telecommunications data. *EPJ Data Science*, 4:1-13.

[15] ESSnet KOMUSO (2019). Complete Overview of Quality Measures and Calculation Methods (QMCMs). Deliverable WP3.

[16] Friedkin, N.E. (1990). Social Networks in Structural Equation Models. *Social Psychology Quarterly*, 53:316-328.

[17] Gilardi, A., Borgoni, R. and J. Mateu (2022). Spatial statistical calibration on linear networks: an application to the analysis of traffic volumes. *Proceedings of the 10th International Workshop on Spatio-Temporal Modelling*, https://boa.unimib.it/handle/10281/400941

[18] Grassini, L. and G. Dugheri (2021). Mobile phone data and tourism statistics: a broken promise? *National Accounting Review*, 3:50-68. http://doi.org/10.3934/NAR.2021002

[19] Gu, T., Han, Y. and Duan, R. (2022). Robust angle-based transfer learning in high dimensions. 10.48550/arXiv.2210.12759

[20] Gu, T., Han, Y. and Duan, R. (2023). A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations. *Pacific Symposium on Biocomputing 2023*, 28:186-197. PMID:36540976.

[21] Hadam, S., N. Würz, Kreutzmann, A.-K. and T. Schmid (2023). Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany. *Statistical Methods & Applications*, https://doi.org/10.1007/s10260-023-00722-0

[22] Kang, C., Liu, Y., Ma, X., and Wu, L. (2012). Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19:3-21.

[23] Karlsson, A., Jónasdóttir, D. and Lagerstrøm, B. (2013). Does telephone number tracing reduce non-response bias in the EU-SILC? A comparison between sample units with and without registered telephone numbers in Iceland and Norway. *Presented at Nordisk statistikermøte, Bergen.*

[24] Koebe, T., Arias-Salazar, A., Rojas-Perilla, N. and Schmid, T. (2022) Intercensal updating using structure-preserving methods and satellite imagery. *Journal of the Royal Statistical Society*, 185(Suppl. 2):S170-S196. https://doi.org/10.1111/rssa.12802

[25] Lagerstrøm, B. O. and G.-E. Wangen (2015). *Erfaringer med Kontakt- og reservasjonsregisteret i Statistisk sentralbyrå*. Internal report, Statistics Norway, in Norwegian.

[26] Leenders, R. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24:21-47.

[27] LeSage, J. and Fischer, M.M. (2010). Spatial econometric methods for modeling origin-destination flows. In *Handbook of Applied Spatial Analysis. Fischer, M.M., Getis, A. (ed.), pp. 409-433*. Berlin, Heidelberg and New York: Springer.

[28] Lestari, T.K., Esko, S., Sarpono, Saluveer, E. and Rufiadi, R. (2018). Indonesia's Experience of using Signaling Mobile Positioning Data for Official Tourism Statistics. *Global Forum on Tourism Statistics*

[29] Li, S., Cai, T.T. and Li, H. (2020). Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84:149-173.

[30] Lopez, P. A. et al.(2018) Microscopic Traffic Simulation using SUMO *The 21st IEEE International Conference on Intelligent Transportation Systems* url = https://elib.dlr.de/124092/

[31] Myrskylä, P. (1991). Census by Questionnaire - Census by Registers and Administrative Records: The Experience of Finland. *Journal of Official Statistics*, 7:457-474.

[32] Nardini, G.; Stea, G.; Virdis, A. and Sabella, D. (2020). Simu5G: A System-level Simulator for 5G Networks. *Proceedings of the 10th International Conference on Simulation and Modeling Methodologies, Technologies and Applications - SIMULTECH* https://doi.org/10.5220/0009826400680080

[33] Nichols, N., O'Brien, A., Feuer, S. and J. Childs (2023). Use of Mobile Phone Location Data in Official Statistics, Social, Demographic and Health Studies. *Research and Methodology Directorate, Center for Behavioral Science Methods*, Research Report Series (Survey Methodology 2023-03). U.S. Census Bureau. https://www.census.gov/library/working-papers/2023/adrm/rsm2023-03.html

[34] OpenSim Ltd. (2024). OMNeT++ Discrete Event Simulator. https://omnetpp.org/

[35] Ord, K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, 70:120-126.

[36] Osborne, C. (1991). Statistical Calibration: A Review. *International Statistical Review*, 59:309–336. https://doi.org/10.2307/1403690

[37] Purcell, N.J. and Kish, L. (1980) Postcensal estimates for local areas (or domains). *International Statistical Review*, 48:3-18.

[38] Reid, G., Zabala, F. and Holmberg, A. (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics*, 33:477-511.

[39] Ricciato, F. (2024). Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics. *Journal of Official Statistics*, 40:3-15. https://doi.org/10.1177/0282423X241235259

[40] Ricciato, F., Lanzieri, G., Wirthmann, A., and Seynaeve, G. (2020). Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, 68:101263. https://www.sciencedirect.com/science/article/pii/S1574119220301097

[41] Rocci, F., Varriale, R. and Luzi, O. (2022). Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes. *Journal of Official Statistics*, 38:533-556.

[42] Sakarovitch, B., Bellefon, M.-P. d., Givord, P., and Vanhoof, M. (2018). Estimating the residential population from mobile phone data, an initial exploration. *Economie et Statistique*, 505:109-132.

[43] Salgado, D., Sanguiao, L., Bogdan, O., Barragán, S., and Suarez-Castillo, M. (2020). A proposed production framework with mobile network data. *Workpackage I Mobile Network Data Deliverable I.3 (Methodology)*.

[44] Segev, N., Harel, M., Mannor, S., Crammer, K., El-Yaniv, R. (2017). Learn on Source, Refine on Target: A Model Transfer Learning Framework with Random Forests. *IEEE Trans Pattern Anal Mach Intell*, 39:1811-1824. doi:10.1109/tpami.2016.2618118

[45] Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing socio demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A*, 180:1163-1190. https://doi.org/10.1111/rssa.12305

[46] Sommer, C. German, R. and Dressler, F. (2018) Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis *IEEE Transactions on Mobile Computing (TMC)*, 10:3-15. https://doi.org/10.1109/TMC.2010.133

[47] Statistisches Bundesamt (2019). Mobile phone data representing the population. https://www.destatis.de/EN/Service/EXSTAT/Datensaetze/mobile-phone-data.html

[48] Steele JE et al. (2017). Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface*, 14: 20160690. http://dx.doi.org/10.1098/rsif.2016.0690

[49] Stone, C.J. (1991) Consistent Nonparametric Regression. *Annals of Statistics*, 5:595-620. https://doi.org/10.1214/aos/1176343886

[50] Suarez Castillo, M., Sémécurbe, F., Ziemlicki, C., Tao, H. X., and Seimandi, T. (2023). Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics. *Journal of Official Statistics*, 39:535-570. https://doi.org/10.2478/jos-2023-0025

[51] Tennekes, M. and Y. A. Gootzen (2022). Bayesian location estimation of mobile devices using a signal strength model. *Journal of Spatial Information Science*, 25:29-66.

[52] United Nations. (2019). *Handbook on The Use of Mobile Phone Data for Official Statistics.* https://unstats.un.org/bigdata/task-teams/mobile-phone/MPD%20Handbook%2020191004.pdf

[53] van den Brakel, J., Söhler, E., Daas, P. and Buelens, B. (2017). Social media as a data source for official statistics; the Dutch consumer confidence index. *Survey Methodology*, 43:183-210. https://doi.org/10.13140/RG.2.2.19294.64326

[54] Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34:935-960.

[55] Zhang, L.-C. (2023). Audit sampling as a quality standard for multisource official statistics. *Spanish Journal of Statistics*, 5:67-83. doi:https://doi.org/10.37830/SJS.2023.1.05

[56] Zhang, L.-C. (2021a). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, 184:571-588.

[57] Zhang, L.-C. (2021b). On the Use of Proxy Variables in Combining Register and Survey Data. In *Administrative Records for Survey Methodology*, eds. A. Y. Chun, M. Larsen, G. Durrant and J.P. Reiter. John Wiley & Sons, Inc.

[58] 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

[59] Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66: 41-63.

[60] Zhang L.-C., Haraldsen G. (2022). Secure Big Data Collection and Processing: Framework, Means and Opportunities. *Journal of the Royal Statistical Society Series A*, 185:1541-59. https://doi.org/10.1111/rssa.12836