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 3.2

Report on methodologies

April 2025
Partner in charge: SSB (Norway)

Authors[1]: L.-C. Zhang, J. Haug, J. Fosen, SSB (Norway)
L. Di Consiglio, M. D'Orazio, C. Faricelli, T. Pichiorri, A. Piovani, T. Tuoto, Istat (Italy)
L. Sanguiao Sande, S. Barragán Andrés, C. Sáez Calvo, M. Novás Filgueira, INE (Spain)
R. Cărtuță, B. Oancea, INS (Romania)

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 not the grating authority can be held responsible for them.

Acknowledgement

We would like to thank the colleagues at SSB, Istat and INE as well as our MNO partners for their help with preparing the data.

We thank also the colleagues in the MNO-MINDS consortium as well as the participants of the MNO-MINDS SPRINT in Vienna in June 2024 for discussions and suggestions.

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 report constitutes Deliverable D3.2 from ESSnet project MNO-MINDS WP3, Methodologies and open source tools for integrating MNO and non-MNO data sources. It consists of three parts, 15 chapters.

Part I provides a general introduction to the later parts. The first chapter explains the MNO data to be considered in this project. A reference frame is presented as a common basis for examining all the methods relevant to utilising MNO data. Chapter Two reviews the 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 the application scenarios to be considered in this workpackage is given in Chapter Three, which covers all the four topic areas mentioned above.

Part II contains five chapters of method development. Chapter Four deals with sample surveys where mobile device usage data are collected specifically, which may enable the use of separately processed MNO data at the aggregate levels. Chapter Five develops several transfer learning approaches given MNO features, in situations where only supervised learning has been considered previously. Chapter Six outlines and extends statistical calibration methods, where imperfect MNO measures are combined with non-MNO target measures, such as road sensor counts of vehicles, in order to predict the targets where only MNO data are available. Chapter Seven considers the problem of origin-destination flow estimation, which is natural for many mobility and population statistics. Chapter Eight studies post-stratification estimation for adjusting the selection errors inherent of any MNO data in a target-agnostic manner, where attributive post-stratification is developed in the presence of mis-identification of mobile device users which is a common problem in many countries.

Part III presents seven chapters on the application scenarios preluded in Chapter Three, where the methods in Part II can be applied to combine MNO and non-MNO data for producing official statistics. However, note that we do not have purposely engineered MNO data in any of the studies. In most cases, the MNO macro data have been obtained through other pilot projects, whose aim was to explore how MNO data may be used for official statistics purposes. It is safe to say that further improvements of the MNO data are easily identified in every case, including when the empirical results already seem acceptable compared to the alternatives that do not use the MNO data. Thus, one should treat them as illustrations of the relevant methods, rather than proven applications to be copied directly.

For those readers who may be interested in examining closely the techniques mentioned in the sequel, we note that the notations are self-contained and consistent in each section in Part I, and in each chapter in Parts II and III. But they are not always consistent across the different parts of this report, simply due to the diverse background of the methods and ideas, as well as the conventions that exist in the respective fields of literature.

Finally, one needs not to read the report page after page. For instance, one may move from Chapter Two in Part I to any specific application scenario in

Part III directly. Or, it is possible to read first Chapter One and then Part II, if one is more interested in methods than specific applications. It might also be helpful to read the related chapters in tandem, such as Section 2.2 followed by Chapters 8, 14 and 15, which all deal with the Quasi-randomisation approach.

Contents

Part I

General introduction

Chapter 1

Concepts of data and methods

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.

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 March 8, 2025, provided it is larger than a specified confidentiality threshold.

However, with 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), so as to enhance the resulting macro MNO data. Here, micro MNO data refer to summary measures associated with distinct devices, such as whether or not a particular device moved from city A to B during the 24 hours on March 8, 2025. 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 in combining MNO and non-MNO data, we propose a reference frame (Figure 1.1) that combines the elements from both the approaches, which enables one to place and examine all the relevant methods on a common basis.

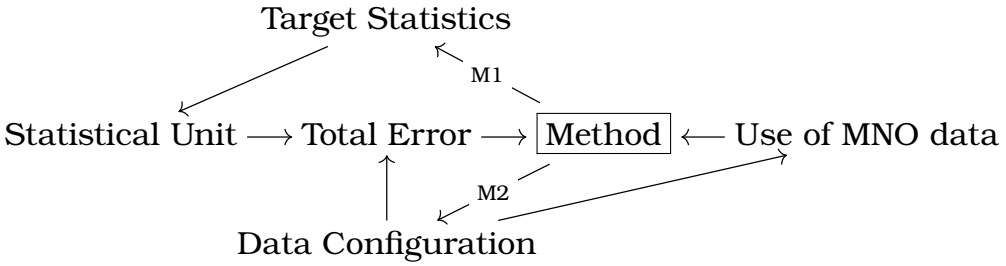

Figure 1.1: Reference frame for methods to combine MNO and non-MNO data. M1, M-executor methods; M2, M-enabler methods.

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 1.1. 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, as long as the MNO and survey data are coherent with each other 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 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.

Chapter 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 modelling device position conditional on cell-ID (e.g. Tennekes and Gootzen, 2022; Salgado et al., 2021).

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 the 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 1.1) 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

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

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

P_{nr} = 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 factors), 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 statistics and MNO market shares is commonly practiced for the sake of 'representativity'. The underlying idea belongs to the quasi-randomisation approach, based on a postulated selection model of the detected devices for the MNO counts.

In terms of Figure 1.1, the macro MNO counts are treated as the 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 refers to past census, demographic accounting or population register.

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 to the official population statistics, if the latter refers to residence *de jure* rather than *de facto*. See appendix A for an illustration of the resulting bias.

Finally, even when the MNO-Home classification is perfectly aligned with the official population statistics used to calculate the expansion factor, it is unclear that weighting by geography is most effective for adjusting the MNO selection effect, compared to other demographic characteristics such as the user age, i.e. if the selection probability varies more with the latter.

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 'modelling'. As explained, 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.

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.

MNO features for prediction, time series, etc.

In terms of Figure 1.1, macro MNO data are used 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 the social-media series is replaced by an MNO series.

Such use of 'MNO features' in model-based estimation for official statistics is straightforward 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 model of Fay and Herriot (1979) 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.

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 1.1, 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). Together, they form a family of nonparametric methods highly relevant for utilising macro MNO data, which can potentially lead to many novel official statistics. The key and difficult challenge, however, is to obtain macro MNO data by appropriate processing, which most likely needs to be tailored to the given application.

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, geographically weighted regression may be used to handle the device location error.

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 qual-

ity 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 (2021a), "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.

The Multi-MNO project (on a reference processing pipeline of MNO event data and network topology data) has published a comprehensive Business Processes and Quality Framework. The development has analysed how the quality requirements from ES Code of Practice and ESS Quality Assurance Framework apply to 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).

Chapter 3

Application scenarios

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 3.1 potentially some official statistics based on MNO data.

Table 3.1: Official statistics based on MNO data

Statistics	Possible examples
(Unit: person)	Long-term *de facto* residents
Population	Census zone population size updates
Tourism	Foreign visitors
	Residents going abroad
	Domestic tourist visits
Mobility	Commuters by origin-destination
	Local trips
(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 3.1 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 1.1), one cannot discuss the details of any relevant methods. Nevertheless, let us take a couple of examples from Table 3.1 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? Note that the naïve QR estimator reviewed in Section 2.2 yields $\hat{Y}_i = m_i(N_i/m_i) = N_i$, which is useless.

- 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.

Consider the last example of city-centre traffic in Table 3.1. Suppose point-of-interest (POI) $k = 1, ..., n$ in a city, such as the central railway station ($k = 1$), the zoo ($k = 2$), the opera ($k = 3$), and so on. Let the target Y_k be the number of motorised vehicle-passings at POI k over a given time period.

- For a randomisation approach to estimate the factor $w_k = m_k/Y_k$, one needs to survey a sample of the people in the passing vehicles, which is hardly practical. Moreover, the MNO count m_k needs to target the number of devices in the passing vehicles, which can be very challenging.

- The QR approach is inadequate on its own here, since the statistical unit is POI and the measurement unit is vehicle, neither of which coincides with the mobile device or user, such that a selection model of detected devices or users would not suffice in any case.

- 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_j be the vehicle count at RPoi j. Let X_i be the MNO count of devices passing by RPoi or POI i. This gives a statistical calibration setting with proxy measures X_i and calibration measures Y_j, similar to that for GWR reviewed in Section 2.3. The challenges now are the computation of X_i, and whether the relationship (Y_j, X_j) at the RPois j can be effectively extrapolated to predict Y_k given X_k at the POIs k.

We have summarised the application scenarios to be explored in Part III in Figure 3.1 where, in addition to the two examples above, we shall consider foreign tourist accommodation statistics referred to as nights-spent, tourist visits to specific areas such as Gotland, as well as resident population mobility statistics in terms of commuters and trips.

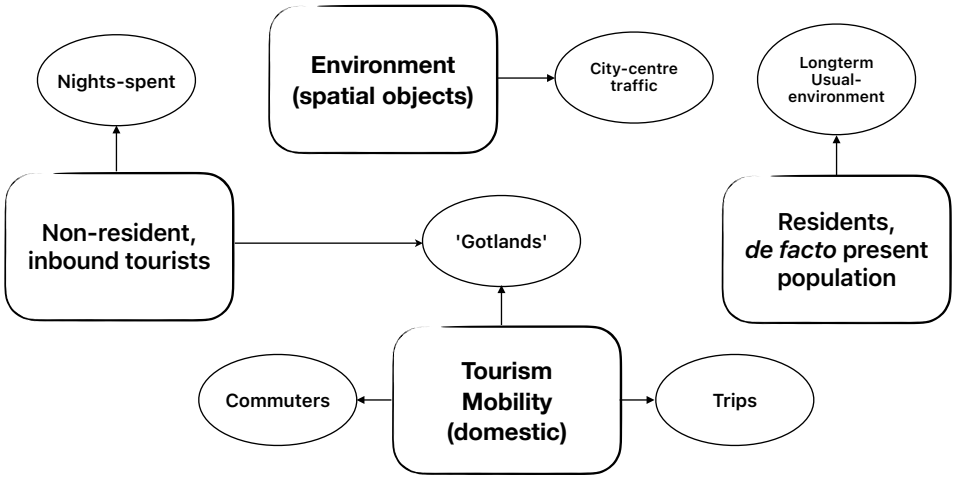

Figure 3.1: Application scenarios to be explored in this report

However, note that we do not have purposely engineered MNO data for any of these application scenarios. In most cases, the MNO macro data have been obtained through other pilot projects, whose aim was to explore how MNO data may be processed to potentially lead to acceptable official statistics. It is safe to say from experience that in *all* such projects, the initially prepared MNO data would lack the required properties and reengineering of the data would be necessary. Even in those cases, where the empirical results already seem acceptable compared to the alternatives that do not use the MNO data, further improvements of the MNO data are easily identified. Thus, one should treat the application scenarios in Part III as illustrations of the relevant methods, rather than proven applications to be copied directly.

Part II

Methods

Chapter 4

Mobile device usage survey

Three types of sample surveys collecting mobile device usage data are relevant for combination with MNO macro data for producing official statistics.

- Provided MNO macro count M can be related to the target total Y via a *conversion factor* W, one can design a sample survey to estimate W specifically. An example is foreign tourist statistics in Indonesia reviewed earlier.
- In a standard sample survey for estimating Y, target outcomes are collected from the sample units, denoted by y, such as $y = 1$ for commuter and $y = 0$ otherwise. One may collect additional device usage data *assoicated* with the sample units, so that sample-based estimation of Y can then incorporate the relevant MNO macro data as well.
- In an *opt-in smart survey*, informed consent may be given by the sampled individuals to collect their micro or even nano MNO data directly, by which the target outcome can be derived directly, such as the number of trips of a device user. It may be possible to estimate Y given additional relevant MNO macro data that include the opt-in sample device data, where the sample device users and non-users are appropriately differentiated.

As a matter of fact all the three types of survey generate both y-outcomes and device usage data. The difference is the intent. The first type of survey is installed because one plans to estimate Y via W. The second type exists to estimate Y even without the device usage data, but one hopes to improve the estimation with the help of device usage data. The third type is conceived as an additional survey mode to collect the target y-outcome, which may reduce response burden and measurement error otherwise.

A template of survey questionnaire is developed by WP4 of MNO-MINDS, which can be considered for all three types of survey. However, the opt-in smart survey involves several challenges beyond the scope of MNO-MINDS, such as the collection and processing of micro or nano MNO data, how they are related to the corresponding MNO macro data, and the missing observations of a device user if the device is not active all the time.

In this chapter, we focus on the second type of survey, where the methods we develop shall encompass that for the first type of survey that aims to estimate Y via W. As we shall explain, a particular challenge is *user ambiguity*, if MNO macro data can only be organised according to service contractors instead of

users. We notice that the user ambiguity problem is also relevant for the other two types of survey if one is interested in subpopulation y-totals beyond the overall population total Y.

4.1 Estimation of population total

Let U denote the population, such that the target population total is given as

$$Y = \sum_{i \in U} y_i$$

Given the sample s and the sample inclusion probabilities $\pi_i = \Pr(i \in s)$, the basic Horvitz-Thompson (HT) estimator of Y is

$$\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i \tag{4.1}$$

However, the estimator (4.1) *ignores* the associated device data altogether.

Let α_i be the number of *in-scope* devices of each sample unit $i \in s$, including $\alpha_i = 0$ for a non-user. For a device of user i, denoted by d, suppose the MNO can derive an associated value which equals to the outcome of the user, denoted by $y_d = y_i$. For instance, suppose $y_i = 1$ if the user i makes an inter-city trip on a given day, and $y_i = 0$ otherwise; and $y_d = 1$ if the MNO detects an inter-city trip of the device d on the same day, and $y_d = 0$ otherwise.

It should be stressed that α_i is not the number of devices each user may or may not have, but the number of devices relevant to the y-outcome specifically. In the above example, a user may have two devices but takes only one of them on the inter-city trip, in which case $y_i = 1$ and $\alpha_i = 1$ not $\alpha_i = 2$, because the device d that is left home would not yield $y_d = y_i$. Whereas in another application one may have $\alpha_i = 2$, because both the devices are relevant.

We now proceed under the assumption $y_d = y_i$ given α_i for $i \in s$. Let the MNO device total be

$$Z = \sum_{d \in D} y_d = \sum_{i \in U} \alpha_i y_i$$

where D denotes all the devices associated with U. By means of calibration estimation (e.g. Deville and Särndal, 1992), one may now adjust the sampling weight $1/\pi_i$ to obtain a calibrated weight w_i satisfying

$$\sum_{i \in s} w_i \alpha_i y_i = Z \tag{4.2}$$

and the corresponding calibration estimator of Y as

$$\hat{Y} = \sum_{i \in s} w_i y_i \tag{4.3}$$

Insofar as the observed MNO device total Z is a close proxy to the unknown target Y, calibration estimation by (4.3) could reduce the variance of $\hat{Y}_{HT}$.

Remark It is worth noting that the estimator intended by the first type of survey is a kind of ratio estimator in the class (4.3), which is given as

$$\hat{Y}_W = \hat{W} Z \quad \text{given} \quad \hat{W} = \frac{\sum_{i \in s} y_i/\pi_i}{\sum_{i \in s} \alpha_i y_i/\pi_i} = \frac{\hat{Y}_{HT}}{\hat{Z}_{HT}}$$

Since the calibration estimator (4.3) accommodates other choices of calibrated weights with respect to Z, as well as additional known population totals beyond Z, it is clear that the approach of estimating Y via the conversion-factor W is encompassed by the more general approach expounded here.

4.2 User ambiguity

One is often interested in subpopulation y-totals in addition to Y. Denote by Y_x the subpopulation totals classified by x. For calibration estimation, one would naturally like to use the corresponding subpopulation device totals. Here arises the complication of user ambiguity.

There are two relevant aspects to device-to-person conversion. The first one is *device duplication*, where some users have multiple devices, as illustrated to the left in Figure 4.1. The second one is *user ambiguity*, if only the service contractors are known to the MNOs but not the actual users, as illustrated to the right in Figure 4.1, which is the situation in many European countries.

Figure 4.1: 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).

We assume that each device d in D has one and only one user in U, as well as one and only one contractor in U. As before, let α_k be the number of edges in A of each individual $k \in U$, who is a user iff $\alpha_k > 0$, where A contains all the device-user edges (dk) as illustrated in Figure 4.1, denoted by $a_{dk} = 1$ or 0 otherwise. Moreover, let ζ_k be the number of edges in C of each individual $k \in U$, who is a contractor iff $\zeta_k > 0$, where C contains all the device-contractor edges (dk) as illustrated in Figure 4.1, denoted by $c_{dk} = 1$ or 0 otherwise. We have

$$\sum_{k,j \in U} c_{dk} a_{dj} \equiv 1 \quad \forall d \in D$$

Moreover, let the number of devices used by j and contracted by k be

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

where ν_{kk} is the number of devices used and contracted by k. It follows that

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

Given user ambiguity, the MNO device total Z is an *apparent* sum over all the contractors instead of users, i.e

$$Z = \sum_{d \in D} y_d \Big(\sum_{j,k \in U} c_{dk} a_{dj} \Big) = \sum_{j,k \in U} \nu_{kj} y_j = \sum_{\substack{k \in U \\ \zeta_k > 0}} y_k \nu_{kk} + \sum_{\substack{k \in U \\ \zeta_k > 0}} \sum_{\substack{j \in U \\ j \neq k}} y_j \nu_{kj} \tag{4.4}$$

The last expression decomposes Z 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. For instance, $y_{d_3} = y_{k_3}$ in Figure 4.1 would be associated with k_2 instead of k_3.

It follows that the partition of Z to $\{Z_x\}$ according to the contractors will not be the same as that according to the users.

Remark One may refer to the mapping from $\{y_d : d \in D\}$ to the user-partition $\{Z_x\}$ as 'device-deduplication', in analogy to record deduplication or linkage, where each device is considered a 'record' and each user an entity, and each entity may be associated with multiple records. Disentangling the matter into two issues, user ambiguity and device duplication of given user, we clarify that the fundamental challenge is user ambiguity (even without device duplication), without which duplicated devices would not have caused additional problems to the breakdown of Z to $\{Z_x\}$.

4.3 Estimation of subpopulation total

The apparent MNO device total of contractor-subpopulation x is

$$Z_x^* = \sum_{\substack{k \in U \\ \zeta_k > 0}} \mathbb{I}(k \in U_x) \sum_{j \in U} \nu_{kj} y_j \tag{4.5}$$

We can construct a vector for each user $j \in s$, whose xth component is

$$z_x(j) = \sum_{\substack{k \in U \\ \zeta_k > 0}} \mathbb{I}(k \in U_x) \nu_{kj} y_j$$

which is the device total of j attributed to the contractor-subpopulation x.

Note that $z_x(j) \equiv 0$ for all x when j is not a device user. Note that only the xth component is $y_j = y_d$ and 0 otherwise, if $\nu_{jj} = 1$ and $\sum_{d \in D} a_{dj} = 1$, i.e. the user j

has a single in-scope device contracted by herself. The contractor information $\mathbb{I}(k \in U_x)$ is needed for each device contractor k otherwise.

Let $[Z_x^*]$ be the vector of totals and $[z_x(i)]$ the vector for each $i \in s$. We have

$$[Z_x^*] = \sum_{i \in U} [z_x(i)]$$

The calibration estimator (4.3) of Y can now use the weights satisfying

$$\sum_{i \in s} w_i [z_x(i)] = [Z_x^*] \tag{4.6}$$

It is of course also possible to include additional calibration totals beyond (4.6). In any case, the calibration estimator of subpopulation total Y_x is

$$\hat{Y}_x = \sum_{i \in s \cap U_x} w_i y_i$$

whether or not $\hat{Y}$ by the weights (4.6) is more efficient than that by (4.2).

4.4 Methodological remarks

It is important to remind the reader of the key assumptions underlying the calibration estimation methods developed above.

First, we assume $y_d = y_i$ given α_i for $i \in s$. In survey, one needs to obtain α_i as the number of *in-scope* devices associated with $i \in s$, which may differ to how many devices a user i has otherwise, and α_i may vary from one target outcome to another. For valid calibration by the MNO device total Z, one needs to make sure that $y_d = y_i$ for all the devices D associated with U.

Notice that the assumption is the same when estimating Y via W. In the example of foreign tourists to Indonesia, α_i refers to the active roaming devices not all the mobile devices the traveller i carries otherwise, and considerable effort is required to process the MNO data, e.g. in order to remove the spurious devices associated with crew members, fishermen in the coast seas, etc.

Second, to make use of the subpopulation MNO device totals $[Z_x^*]$, one needs to obtain the contractor information of any sample device user i in case her devices are contracted by someone else, in order to construct $[z_x(i)]$, whether or not the contractor is in the sample.

Chapter 5

Transfer learning

Pan and Yang (2019) classify different settings of transfer learning in their widely cited survey. We shall take as our starting point *inductive* learning that is perhaps the most obvious for survey sampling, where some labelled data in the *target domain* (or population) are required as the training data to 'induce' the target predictive function, given there are a lot more labelled data in the *source domain* (or a similar but different population).

For instance, suppose one would like to build a regression model of certain heart condition by relevant risk factors for the German population, i.e. the target domain, based on a sample survey. Meanwhile, suppose similar surveys have been carried out in many other European countries previously, and the same regression model can be estimated based on those data, i.e. the source domain. Inductive learning aims to leverage the large amount of source domain data to improve model-fitting for the German population, compared to only using the sample from the German population.

To be specific, denote by $\mu(x;\beta)$ a *target* model with unknown parameters β. Suppose there exists a *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 but with different parameter values β and θ. Transfer learning aims then to improve the estimation of β by leveraging $\hat{\theta}$.

A usual technique is to estimate β based on the observations in a sample s from the target population, subject to a chosen penalty of the discrepancy between β and $\hat{\theta}$, such as minimising

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

given $\gamma > 0$. Although the resulting estimator of β is biased due to the penalty term, its variance can be greatly reduced than otherwise. One can thus view this as regularisation towards $\hat{\theta}$ instead of 0, which is helpful given insufficient number of target observations (e.g. Li et al. 2020; Gu et al., 2024).

Meanwhile, insofar as $\theta \neq \beta$ correspond to two different distributions of y_i given x_i, no matter if the distribution is fully specified or not given θ or β, one may consider $\hat{\theta}$ as the *prior information* on the target distribution of y_i, such that $\hat{\beta}$ by transfer learning via $\Delta(\beta;\gamma)$ combines this prior information with the data associated with s in a non-Bayesian manner.

This insight can enable combining sample surveys with MNO data in many applications, where MNO data supply source domain information. In terms of the reference frame (Figure 1.1), such transfer learning typically uses the MNO macro data as proxy to the target parameters, in a manner which does not require mobile device usage data at the micro level.

In this chapter, we shall outline several transfer learning methods that are relevant to the various application scenarios in Part III.

5.1 Transfer learning for HT estimator

Let N be the target population size, and n the size of a sample s. Zhang et al. (2025) express the Horvitz-Thompson (HT) estimator as a prediction estimator

$$\hat{Y}_{HT} = \sum_{k \in s} y_k + \sum_{j \neq s} \mu(x_j, b_s)$$

where $x_i = \pi_i N/n$ is the 'design' covariate, and $b_s = \frac{1}{n}\sum_{k\in s} \frac{y_k}{x_k}$ is an estimate of the regression coefficient in the linear predictor

$$\mu(x_j, b_s) = x_j b_s + \frac{1}{N-n}\sum_{k\in s}(x_k b_s - y_k)$$

For our purpose, we now rewrite $\hat{Y}_{HT}$ as a simple function of b_s, i.e.

$$\hat{Y}_{HT} = \sum_{k\in s} y_k + \sum_{j\neq s} x_j b_s + \sum_{k\in s} x_k b_s - \sum_{k\in s} y_k \equiv \sum_{i\in U} x_i b_s \equiv N b_s$$

where

$$b_s = \arg\min_{\beta} \sum_{k\in s} (y_k - x_k\beta)^2/x_k^2$$

Now, suppose we have a proxy to the target parameter $p = Y/N$ from relevant MNO macro data, denoted by $\tilde{p}$. Given the prior information of $\tilde{p}$ and $Y \approx N\tilde{p}$, consider regularising b_s by minimising

$$\Delta(p;\gamma) = \frac{1}{n}\sum_{k\in s}(y_k - x_k p)^2/x_k^2 + \gamma(p-\tilde{p})^2 \tag{5.1}$$

which yields the transfer-learning estimator as a convex combination of the source parameter $\tilde{p}$ and the survey estimator $\hat{p}_{HT}$:

$$p(\gamma) = \frac{1}{1+\gamma} b_s + \frac{\gamma}{1+\gamma}\tilde{p} \qquad \overset{w^{-1}=1+\gamma}{\Leftrightarrow} \qquad p(w) = w\hat{p}_{HT} + (1-w)\tilde{p}$$

One can think of different ways to choose the tuning constant. For instance,

$$\gamma = \arg\min_{\gamma} \sum_{k\in s} w_k\{y_k - x_k p_k(\gamma)\}^2$$

where w_k is a chosen weight for $k \in s$, and $p_k(\gamma)$ is the delete-k estimator of p

given by (5.1) based on on $s \setminus \{k\}$. Or, a more direct approach based on

$$\text{MSE}\big(p(w)\big) = w^2 V(\hat{p}_{HT}) + (1-w)^2(\tilde{p}-p)^2$$

which is minimised at

$$w = \frac{(\tilde{p}-p)^2}{V(\hat{p}_{HT}) + (\tilde{p}-p)^2}$$

By transfer learning of HT estimator one can simply use

$$\hat{Y}_{TL} = N\big(\hat{w}\hat{p}_{HT} + (1-\hat{w})\tilde{p}\big)$$

given $\hat{V}(\hat{p}_{HT})$ as unbiased of the HT-variance $V(\hat{p}_{HT})$, and

$$\hat{w} = \max\Big(0,\ \frac{(\tilde{p}-\hat{p}_{HT})^2 - \hat{V}(\hat{p}_{HT})}{(\tilde{p}-\hat{p}_{HT})^2}\Big)$$

For an application scenario, let the total inbound tourist visits be the target of interest. Denote by $k = 1, ..., N$ the inbound passengers at a given airport. Let $y_k = 1$ if passenger k is a tourist, let $y_k = 0$ otherwise, such that

$$Y = \sum_{k=1}^{N} y_k \quad \Leftrightarrow \quad p = \frac{Y}{N}$$

is the target parameter at this airport. Meanwhile, let $m_k = 1$ if passenger k is a roaming customer of the MNOs, let $m_k = 0$ otherwise, such that

$$M = \sum_{k=1}^{N} m_k \quad \Rightarrow \quad \tilde{p} = \frac{M}{N}$$

is a proxy to p from the source domain of mobile phone data.

In case $\hat{Y}_{HT}$ is available by passenger survey but suffers a large variance, transfer learning for the HT-estimator can be considered.

5.2 Transfer learning for ensemble estimation

Suppose a sample survey yields an estimator of the total Y with acceptable precision, but not the subtotals Y_t satisfying

$$Y = \sum_{t=1}^{T} Y_t$$

e.g. as in small area estimation, or, equivalently, the population proportions

$$p_t = \frac{Y_t}{Y}$$

Consider the estimation of the *ensemble parameter* $\{p_t : t = 1, ..., T\}$, given a known proxy breakdown from the source domain

$$X = \sum_{t=1}^{T} X_t \quad \Leftrightarrow \quad q_t = \frac{X_t}{X}$$

For transfer learning given $\{q_t : t = 1, ..., T\}$, consider minimising

$$\Delta(p; \gamma) = -\sum_{t=1}^{T} \hat{Y}_t \log p_t + \gamma \sum_{t=1}^{T} X_t (\log q_t - \log p_t) + \lambda \Big(\sum_{t=1}^{T} p_t - 1 \Big) \tag{5.2}$$

where the penalty with multiplier γ is related to the Kullback-Leibler divergence from the target distribution $\{p_t\}$ to the source distribution $\{q_t\}$, and the last term with multiplier λ is due to the ensemble parameter restriction $\sum_{t=1}^{T} p_t = 1$.

Clearly, the solution is $\hat{p}_t$ if $\gamma = 0$ in (5.2), whereas it tends to q_t as $\gamma \to \infty$. Given non-trivial γ, setting the partial derivatives of Δ to 0, we obtain

$$\begin{aligned} \dot{p}_t &= \frac{\hat{Y}_t + \gamma X_t}{\hat{Y} + \gamma X} = \frac{(\hat{Y}_t / \hat{Y})(\hat{Y}/X) + \gamma X_t / X}{\hat{Y}/X + \gamma} \\ &= \psi(\gamma) \hat{p}_t + \{1 - \psi(\gamma)\} q_t \end{aligned} \tag{5.3}$$

where

$$\psi(\gamma) = \frac{\hat{Y}/X}{\gamma + \hat{Y}/X}$$

and $\sum_{t=1}^{T} \dot{p}_t = 1$ holds automatically. Notice the resemblance to the shrinkage estimator of James-Stein (1961).

For an application scenario, let Y_t be the monthly number of domestic trips, where $t = 1, ..., 12$, and Y is the yearly total. Let q_t be derived from mobile device data, which is biased due to the coverage and measurement errors but has a negligible variance compared to a travel survey estimator

$$\hat{p}_t = \hat{Y}_t / \hat{Y}$$

Notice that it is a standard technique in small area estimation to apply a linear mixed model (LMM)

$$\hat{Y}_t = Y_t + e_t = \beta_0 + \beta_1 X_t + v_t + e_t$$

where v_t is a random effect under the model and e_t the sampling error of $\hat{Y}_t$. Now, in the extreme case of $X_t/X = Y_t/Y$, transfer-learning should intuitively yield $\dot{p}_t = X_t/X$, i.e. no error at all, but not the LMM-predictor since one needs to estimate $(\beta_0, \beta_1) = (0, Y/X)$ based on only 12 observations. In short, transfer learning is particularly worth considering given good proxy X_t/X.

5.2.1 Tuning parameter ψ

In practice, we need to choose the tuning parameter $\psi(\gamma)$, or ψ directly. Let

$$E\Big(\sum_{t=1}^{T}(\dot{p}_t - p_t)^2\Big) = \psi^2 \sum_{t=1}^{T} V(\hat{p}_t) + (1-\psi)^2 \sum_{t=1}^{T} u_t^2$$

be the total mean squared error (MSE) of $\{\dot{p}_t\}$ over repeated sampling, where

$$u_t = q_t - p_t$$

for $t = 1, ..., T$ are treated as constants rather than random variables. The total MSE above is minimised given

$$\psi = \frac{\tau_u}{\tau_u + \tau_e} \quad \text{and} \quad \tau_u = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \quad \text{and} \quad \tau_e = \frac{1}{T}\sum_{t=1}^{T} V(\hat{p}_t)$$

A transfer-learning estimator $\hat{p}_t^{TL}$ follows from replacing $\psi(\gamma)$ in (5.3) by

$$\hat{\psi} = \frac{\hat{\tau}_u}{\hat{\tau}_u + \hat{\tau}_e}$$

given

$$\hat{\tau}_u = \frac{1}{T}\sum_{t=1}^{T}\Big((\hat{p}_t - q_t)^2 - \hat{V}(\hat{p}_t)\Big) \quad \text{and} \quad \hat{\tau}_e = \frac{1}{T}\sum_{t=1}^{T}\hat{V}(\hat{p}_t)$$

5.2.2 Out-of-sample domains

While ensemble estimation via (5.2) suffices if survey estimates are available for all $t = 1..., T$, what if this is not the case?

Let $\mathcal{D}_1 = \{t : \exists \hat{Y}_t\}$ contain the in-sample components, and let $\mathcal{D}_0 = \{t : \nexists \hat{Y}_t\}$ contain the out-of-sample components.

Let the LMM parameter estimates $(\hat{\beta}, \hat{\sigma}_v^2)$ be obtained from $\mathcal{D}_1$, which yields the so-called synthetic estimate $\hat{\beta}_0 + \hat{\beta}_1 X_t$ for $t \in \mathcal{D}_0$. One can view the synthetic estimates as the trivial solution to

$$\min_{v} \sum_{t \in \mathcal{D}_0} v_t^2$$

To leverage the source domain $\mathcal{D}_1$ for the target domain $\mathcal{D}_0$, consider transfer learning to estimate $\{v_k : k \in \mathcal{D}_0\}$ by minimising

$$L(\mathcal{D}_0) = \sum_{k \in \mathcal{D}_0} v_k^2 + \alpha \sum_{k \in \mathcal{D}_0} (\hat{\beta}_0 + \hat{\beta}_1 X_k + v_k - Y_k^*)^2 \tag{5.4}$$

given tuning constant $\alpha \geq 0$ and a proxy Y_k^* derived from X_t, e.g. $Y_k^* = \hat{Y} X_k / X$. Clearly, we would recover the synthetic estimates if $\alpha = 0$; whereas $\hat{v}_k$ will be pulled towards $Y_k^* - \hat{\beta}_0 - \hat{\beta}_1 X_k$ as α increases. It is possible to choose α as follows.

- Given α, let $\{\tilde{v}_k(\alpha) : k \in \mathcal{D}_1\}$ be obtained from minimising $L(\mathcal{D}_1)$, given by (5.4)

in terms of $\mathcal{D}_1$, just like one would have done with $L(\mathcal{D}_0)$.

- Choose the value of α, such that the corresponding $\{\tilde{v}_k(\alpha) : k \in \mathcal{D}_1\}$ are closest to the LMM-based $\{\hat{v}_k : k \in \mathcal{D}_1\}$ according to some chosen metric.

By transfer learning via (5.4) we have moved from the inductive setting to the *transductive* setting (Pan and Yang, 2019), where labels $\hat{Y}_t$ exist only in the source domain $\mathcal{D}_1$ but not the target domain $\mathcal{D}_0$. Notice that uncertainty assessment of such transductive learning estimates

$$\hat{Y}_t^{TL} = \hat{\beta}_0 + \hat{\beta}_1 X_t + \hat{v}_t$$

will have to be based on some additional model assumptions, now that we do not observe any survey estimate $\hat{Y}_t$ for $t \in \mathcal{D}_0$. For instance, one can model

$$\text{mse}(\hat{Y}_t^{LMM}) = \eta(X_t, c_t)$$

over $\mathcal{D}_1$ given known covariates c_t in addition, and apply $\widehat{\text{mse}}(X_t, c_t)$ to $\mathcal{D}_0$.

5.3 Transfer learning for flash estimation

Zhang and Haug (2025) propose two transfer learning approaches in a generic setup for flash estimation, where the model for the available (or early) data is known *a priori* to be biased for the rest of the population.

Denote by $\mu(x, s)$ a predictor of y for any unit with features x, which is learned from $\{(y_i, x_i) : i \in s\}$, where s is a labelled sample (of early observations). Denote by R a *target* set of units with known x_j, $\forall j \in R$ and $s \cap R = \emptyset$, for which no labelled data are yet available and y-prediction is needed. However, it is known that $\mu(x, s)$ is biased for $\{y_j : j \in R\}$, because y_j for $j \in R$ and y_i for $i \in s$ do not have the same distribution conditional on x_j and x_i.

An application scenario to MNO macro data is flash estimation, where y_i is nights-spent by tourists and s consists of early reports from some but not all the units, while x_i is nights-spent based on MNO device counts for $i \in s \cup R$, where R consists of the late reporting units.

5.3.1 Augmented learning

In case of linear prediction $x^\top \beta$, there exists the following equivalence:

$$\min_\beta \sum_{i \in s} (x_i^\top \beta - y_i)^2 + \gamma \sum_{j \in R^*} (x_i^\top \beta - y_j^*)^2 \tag{5.5}$$

$$\Leftrightarrow \quad \min_\beta \sum_{i \in s^*} \{x_i^\top \beta - \mathbb{I}(i \in s) y_i - \mathbb{I}(i \in R^*) y_i^*\}^2 \ \text{ given } \ s^* = R^* \cup s \cup \cdots \cup s \tag{5.6}$$

where R^* denotes a set of units that are similar to or even overlap with R, and y_j^* is a proxy to the target observation y_j including when $j \in R^* \cap R$, and s is duplicated γ^{-1} times in s^* (if practically possible). Unlike regularising $\hat{\beta}$ towards some given $\hat{\theta}$ from the source domain, estimation of $\hat{\beta}$ by (5.5) is the same as based on the *augmented sample* s^* in (5.6).

Generally, to exploit the conceptual equivalence between working with the augmented loss (5.7) and the augmented sample (5.6), one can obtain $\mu(x, s^*)$ based on an *augmented sample* s^* given an arbitrary model or algorithm μ, which is referred to as *augmented learning*:

$$\min L(s^*) \quad \text{given} \quad s^* = s \cup R^*, \quad \{(x_i, y_i) : i \in s\}, \quad \{(x_j, y_j^*) : j \in R^*\} \tag{5.7}$$

One can experiment with the choice of R^* instead of the tuning parameter γ.

A typical source of $\{(x_j, y_j^*) : j \in R^*\}$ is historic data, where a unit $j \in R$ may exist in R^* for which only a past observation y_j^* is available but not the target y_j. The idea is that incorporating these data can be helpful if y_i relates to x_i for $i \in R$ in a similar manner as y_j^* relates to x_j for $j \in R^*$, certainly if i and j refer to the same unit or possibly otherwise.

Insofar as $R^* \cap R \neq \emptyset$, augmented learning by (5.7) yields a hybrid setting of inductive and transductive learning, where s^* is the source domain, and some labelled but non-representative data are available in the target domain R.

5.3.2 Quasi transfer learning

Let the learning for flash estimation at time point t target the model for

$$q_t = s_t \cup R_t$$

which is partially covered by the available s_t although by stipulation s_t is not 'representative' of q_t. The term "quasi" indicates that we aim at something that is not the target model for R_t directly but close to it.

$$\begin{array}{ccc} \mu(x, s_t^*) & \xleftarrow{\hat{g}(\cdot)} & \mu(x, s_b^*) \\ & \Downarrow & \\ \mu(x, q_t) & \xleftarrow{\hat{g}(\cdot)} & \mu(x, q_b) \end{array}$$

Figure 5.1: A scheme of quasi transfer learning

Let a source model $\mu(x, s_t^*)$ be fitted to s_t^*, as given by (5.7). To leverage it for $\mu(x, q_t)$, choose additionally two source models $\mu(x, s_b^*)$ and $\mu(x, q_b)$ in the same setup as $\mu(x, s_t^*)$ and $\mu(x, q_t)$ but for some time point b in the past, and a *transfer scheme* such as the one in Figure 5.1.

According to this scheme, the estimated relationship $g(\cdot)$ between $\mu(x, s_t^*)$ and $\mu(x, s_b^*)$ over the time points (t, b) is transferred to that between $\mu(x, q_t)$ and $\mu(x, q_b)$ over the same (t, b). One can consider

$$E\{\mu(x, s_t^*)\} = \mu(x, s_b^*) + g(x) \tag{5.8a}$$

$$E\{\mu(x, s_t^*)\} = g\big(x, \mu(x, s_b^*)\big) \tag{5.8b}$$

where $g(x)$ is either an offset to $\mu(x, s_t^*)$ in (5.8a) or a feature in (5.8b).

Quasi transfer learning is also a hybrid setting of inductive and transductive learning. The scheme in Figure 5.1 requires the relationship between s_t^* and s_b^*

to be similar to that between q_t and q_b over the same time lag, where s_t^* and s_b^* have the same sample composition and likewise for q_t and q_b. Moreover, at any given time point t, s_t^* and q_t overlap in terms of s_t, while the units $R^* = s_t^* \setminus s_t$ and $R_t = q_t \setminus s_t$ are comparable and likely overlapping as well.

Remark It is possible to combine naïve learning $\mu(x, s)$, augmented learning $\mu(x, s^*)$ and quasi transfer learning $g\big(x, \mu(x, s_b^*)\big)$ by ensemble learning methods, such as voting, averaging or stacking.

Remark See Zhang and Haug (2025) for associated methods of uncertainty estimation and an application to the Norwegian Retail Turnover Index.

Chapter 6

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 the scale of a measuring instrument on the basis of a 'calibration experiment'. For example, let $\{x_j : j \in R\}$ be the imprecise MNO-measures and $\{y_i : i \in s\}$ the trusted 'calibration measures' from non-MNO sources, where

$$s \subset R$$

Viewing $\{(x_i, y_i) : i \in s\}$ as a calibration experiment, one may estimate y_j by adjusting x_j for $j \in R \setminus s$.

By *statistical calibration* we refer to any method that uses proxy MNO macro measurements in such manners resembling scientific calibration. The trusted calibration measures y_i are obtained from non-MNO sources, such as traffic sensor counts, passenger or customer ticket numbers. A major challenge will be to obtain good enough proxy measures from the MNOs.

6.1 Spatial statistical calibration

Let R contain the units, for which one has available proxy measures x_j. For any $j \in R$, the best predictor of the target measure y_j given x_j is

$$\begin{aligned} E(y_j \mid x_j) &= \int y f(y \mid x_j) dy = \frac{\int y f(y, x_j) dy}{f(x_j)} \\ &= \frac{\int y f(x_j \mid y) f(y) dy}{\int f(x_j \mid y) f(y) dy} \end{aligned}$$

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

$$\hat{y}_j = \sum_{i \in s} w_i(x_j, s) y_i \tag{6.1}$$

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 (6.1) can be given in many other ways. In case the elements of s and R are spatial points, geographically weighted regression (GWR) would yield

$$w_i(x_j, s) = \frac{w(d_{ij}) x_i x_j}{\sum_{k \in s} w(d_{kj}) x_k^2}$$

and

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

(Brunsdon et al, 1996), where $d_{ij} = d_{ji}$ is some chosen measure of the distance between $i \in s$ and $j \in R$. For instance,

$$w(d_{ij}) \propto \mathbb{I}(d_{ij} < d)$$

for a given threshold value d, or

$$w(d_{ij}) \propto \exp\{-\alpha d_{ij}^2\}$$

given the tuning constant α, or the *bisquare*

$$w(d_{ij}) \propto \mathbb{I}(d_{ij} < d) \, (1 - d_{ij}^2 / d^2)^2$$

Like any spatial prediction method, the associated uncertainty evaluation must depend on assumptions that are hard to verify, because ultimately any point depends on any other point and every point is spatially unique. A basic cross-validation mean absolute error (MAE) estimator is given by

$$\text{mae} = \frac{1}{n} \sum_{i \in s} |y_i - \hat{y}_{(i)}|$$

where $\hat{y}_{(i)}$ is the GWR prediction of y_i based on the delete-i sample $s_{(i)}$.

6.2 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 be transformed to proportions that sum to 1, with 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 the idea and 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 is the same as estimating the proportion $y_2/(y_1 + y_2)$ or ratio y_2/y_1 given y_1.

6.2.1 Simplistic approach

For example, y_1 may be the known number of cinema goers on a given day and y_2 the unknown restaurant diners, where (x_1, x_2) are the corresponding MNO counts. For another example, let y_1 be the number of passenger tickets travelling into a city centre, let y_2 be the number of other people travelling into the city centre, provided corresponding MNO counts (x_1, x_2). In both cases, if the device-person ratio is the same of y_1 and y_2, *and* the device-detection ratio is the same for x_1 and x_2, then we would simply have

$$\frac{x_2}{y_2} = \frac{x_1}{y_1}$$

Whilst the device-person ratio might not vary much from y_1 to y_2 in any case, it will be more demanding–if possible at all–for the MNO to accurately distinguish the devices going to cinemas vs. restaurants, or those travelling on public transport vs. otherwise.

Still, such simple methods have been applied to compositional MNO-data, as a data fusion approach (e.g. Batista e Silva et al, 2020) by directly assuming

$$E\Big(\frac{y_k}{y_1 + \cdots + y_K}\Big) = E\Big(\frac{x_k}{x_1 + \cdots + x_K}\Big)$$

such that in the case of $K = 2$ we obtain

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

Note that this can also be given by statistical calibration under the assumption

$$E(y_k \mid x_k) = \beta x_k \quad \forall k \tag{6.2}$$

or transductive learning from source domain (x_1, y_1) to target domain (x_2, y_2). However, regardless the origins of idea, the approach is too limited generally.

6.2.2 A modelling approach

The simple model (6.2) can be extended in two respects: first, one can consider ensemble estimation of $\{y_{2i}\}$, where y_{2i} is defined in the same way over a set of units $i = 1, ..., n$; next, one can introduce other relevant features z_i to y_{2i}.

For example, let y_{2i} be the number of people travelling by other means than public transport into the cite centre i, and z_i may include the relevant numbers of vehicles, card payment transactions, visitors to culture institutions or events, etc. Or, 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.

Now, given (x_{1i}, x_{2i}) associated with (y_{1i}, y_{2i}) for $i = 1, ..., n$, let

$$p_i = \frac{y_{2i}}{y_{1i}} \qquad \text{and} \qquad q_i = \frac{x_{2i}}{x_{1i}}$$

be the target and proxy compositions, respectively, with known (x_{1i}, x_{2i}) and y_{1i}. To estimate $\{y_{2i}\}$, consider minimising

$$L = \sum_{i=1}^{n} (p_i - q_i)^2 + \gamma \sum_{i=1}^{n} (p_i - \mu_i)^2 \tag{6.3}$$

with respect to p_i and μ_i, given the tuning parameter $\gamma > 0$, where

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

is a predictor of p_i which generally may depend on additional covariates z_i.

By (6.3), the otherwise unconstrained sum of $(p_i - q_i)^2$ 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 $(p_i - q_i)^2$ or $(p_i - \mu_i)^2$ in (6.3), and so on.

Illustration For a quick illustration of compositional statistical calibration by (6.3), let us consider the special case without additional z_i and simply use

$$\mu_i = E(p_i \mid y_{1i}) = \beta$$

i.e. the penalty is just the smoothness of p_i in the absence of z_i. The simplistic estimator $\hat{y}_{2i} = x_{2i} y_{1i} / x_{1i}$ would follow from minimising the first term of (6.3) on its own if $\gamma = 0$. More generally, by minimising (6.3), we obtain

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

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

Remark Identifiability of (6.3) is an issue generally speaking, although the above has already demonstrated its feasibility in the simplest setup. Notice that identifiability can always be achieved as long as necessary external information (or estimates) can be plugged into (6.3), e.g. by sample surveys.

Remark Apart from identifiability for well-defined compositional statistical calibration, successful applications will depend heavily on how close the proxy MNO compositions q_i are to p_i. Notice that this does not necessarily require $x_{1i} \approx y_{1i}$ *and* $x_{2i} \approx y_{2i}$ but rather $x_{2i}/x_{1i} \approx y_{2i}/y_{1i}$. Whereas success is guaranteed provided effective proxy MNO compositions q_i, since setting $\gamma = 0$ in (6.3) would simply yield $p_i = q_i$.

6.3 Network statistical calibration

In *network statistical calibration*, the data are given a graph structure of which spatial data are a special case, and the target measures are subject to network constraints induced by the graph structure.

For instance, denote by $G = (U, A)$ a *road network* with crossroads U, and $(ij) \in A$ iff road exists in the direction i to j for $i, j \in U$. Figure 6.1 illustrates a part of road network with crossroads 1 to 4, the permitted directions of vehicles, as well as two fixed road sensors marked as triangles.

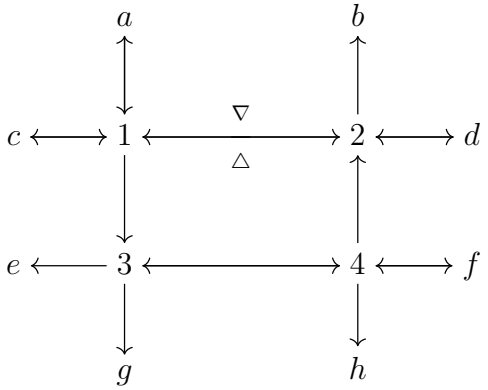

Figure 6.1: Illustration of road network with two fixed sensors (triangles)

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 counts from another source, where x_{ij} and y_{ij} may differ for various reasons, such as coverage or measurement errors.

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

$$\begin{gathered} \hat{y}_{ij} = \tau(x_{ij}; A_s) \doteq E(y_{ij} \mid x_{ij}) \\ \text{subject to} \quad \sum_{(ji)\in A} \hat{y}_{ji} = \sum_{(ij)\in A} \hat{y}_{ij} \quad \text{for any } i \in U_0 \end{gathered} \tag{6.4}$$

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 (6.4) signifies that $\hat{y}_{ij}$ may be close to the conditional expectations of y_{ij} given x_{ij} but not directly given as such estimates. Moreover, $U_0 \subseteq U$ generally. For instance, U_0 may consist of the four crossroads 1, 2, 3, 4 in Figure 6.1 but not the others a, ..., h that exist as long as G is not a closed network.

Traffic network

It may be practically impossible to anchor mobile device signals to the road network directly, because the nano MNO data tend not to be frequent enough and the physical location of a device needs to be inferred. However, the network formulation applied equally to a *traffic network*, in which connections are not roads but refer to congruity among geographic areas.

- Let U refer to geographic areas, and $(ij) \in A$ iff traffic is possible in direction i to j for any two congruous areas $i, j \in U$.
- Let $\tilde{U} = U \cup \{0\}$ and $\tilde{A} = A \cup \{(0i), (i0) : i \in U\}$, such that $G = (\tilde{U}, \tilde{A})$ includes an imaginary node 0 that is connected to every $i \in U$ in both directions.
- Within a specified time period, say, every 24 hours, let x_{ij} be the MNO-count of devices that moved from i into j, let x_{0i} be those that started moving from within area i, and let x_{i0} be those that stopped moving in area i.

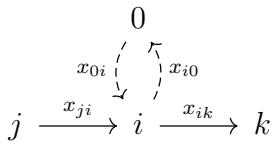

Figure 6.2: Illustration of between-area traffic with stopover-state 0

That is, for each device that have been active during the given time period, there exists a detected initial area, such that x_{0i} is the total of those starting in area i, as illustrated in Figure 6.2. Following any device starting from outside i, some of them may move into area i from a congruous area j, which are included in x_{ji}, and some of them may then move from area i into a congruous area k, which are included in x_{ik}. Note that the same device may contribute multiple times to the traffic in and out of an area during the whole period. By the end, x_{i0} is the total of devices that are last detected in area i.

Let $\{y_{ij} : (ij) \in A_s\}$ be some trusted non-MNO vehicle or traveller counts from i to j, for every $(ij) \in A_s$. One can obtain $\hat{y}_{ij}$ for any $(ij) \in \tilde{A}$ by network statistical calibration (6.4) that is applied to $\tilde{A}$ instead of A.

A solution

Let $\mu_A = \{\mu_{ij} : (ij) \in A\}$ be the predicted values of $\{y_{ij} : (ij) \in A\}$ by any chosen model which do not satisfy the constraint of (6.4). To obtain

$$\hat{y}_{ij} = \mu_{ij} + z_{ij}$$

with minimum changes to μ_A, consider the Lagrangian

$$L = \frac{1}{2} z^\top W z - \lambda^\top H(\mu + z) \tag{6.5}$$

where μ is the R-vector of μ_A given $R = |A|$, similarly for vector z, and W is a diagonal $R \times R$-matrix of chosen weights, and H is the $N \times R$-matrix of constants given $N = |U|$ such that the constraints of (6.4) can be expressed as $H(\mu + z) = 0$ (Pannekoek and Zhang, 2015). The solution is

$$\lambda = -(HW^{-1}H^\top)^{-1} H\mu$$
$$z = W^{-1}H^\top \lambda = -W^{-1}H^\top (HW^{-1}H^\top)^{-1} H\mu$$

An algorithm

A potential numerical difficulty with the above solution arises when (N, R) are so large that the required matrix inverse is infeasible directly. Below we devise an iterative algorithm for (6.4) which is applicable in such cases.

Let ℓ_i be the *load* of node i, which is its total inflow or outflow when they are equal to each other, i.e. satisfying the constraints of (6.4). If

$$\mu_{i+} = \sum_{(ij)\in A} \mu_{ij} \neq \sum_{(ji)\in A} \mu_{ji} = \mu_{+i}$$

then one can define the minimum-change load to be

$$\ell_i = \frac{1}{2}(\mu_{i+} + \mu_{+i})$$

The required adjustments of the outflows from i can be given as

$$\frac{\mu_{ij}}{\mu_{i+}}(\ell_i - \mu_{i+}) = \frac{\mu_{ij}}{\mu_{i+}}\ell_i - \mu_{ij} \qquad \text{which yields} \qquad \frac{\mu_{ij}}{\mu_{i+}}\ell_i$$

as the adjusted flow from i to j. Similarly for the required adjustments of the inflows. However, μ_{ij} would be adjusted twice due to $\ell_i - \mu_{i+}$ and $\ell_j - \mu_{+j}$, respectively, and the two adjustments may not agree unless

$$\ell_i \equiv \mu_{i+} \equiv \mu_{+i}$$

An iterative algorithm is therefore needed.

Load-diffusion algorithm Let $\mu_{ij}^{(0)} = \mu_{ij}$, $\forall (ij) \in A$. Let $t = 0$. Set tol > 0.

i. Increase t by 1, and compute

$$\ell_i^{(t)} = \frac{1}{2}\Big(\mu_{i+}^{(t-1)} + \mu_{+i}^{(t-1)}\Big) \quad \text{and} \quad y_{ij}^{(t)} = \frac{\mu_{ij}^{(t-1)}}{\mu_{i+}^{(t-1)}}\ell_i^{(t)} \quad \text{and} \quad z_{ij}^{(t)} = \frac{\mu_{ij}^{(t-1)}}{\mu_{+j}^{(t-1)}}\ell_j^{(t)}$$

ii. Stop if $|y_{ij}^{(t)} - z_{ij}^{(t)}| \leq$ tol, $\forall (ij) \in A$; otherwise, let $\mu_{ij}^{(t)} = \frac{1}{2}\Big(y_{ij}^{(t)} + z_{ij}^{(t)}\Big)$, go to (i).

Convergence is the case because the absolute difference $\ell_i^{(t)} - \mu_{i+}^{(t)}$ or $\ell_i^{(t)} - \mu_{+i}^{(t)}$ is reduced after each iteration, such that

$$\hat{y}_{ij} = \lim_{t\to\infty} y_{ij}^{(t)} = \lim_{t\to\infty} z_{ij}^{(t)}$$

This *load-diffusion* algorithm solves (6.4) by making minimum load changes. Notice that it is possible to fix

$$\hat{y}_{ij} = y_{ij}$$

for any $(ij) \in A_s$ by imposing these $\mu_{ij}^{(t)}$ at the end of each iteration.

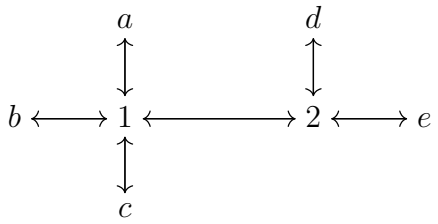

Figure 6.3: Network statistical calibration with $U_0 = \{1, 2\}$

An example

Figure 6.3 illustrates a simple example, where network constraints apply to the nodes $U_0 = \{1, 2\}$, while the other nodes $\{a, b, c, d, e\}$ may have additional inflows and outflows that are not of concern here. Let

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 \end{pmatrix}$$

$$\mu = (\mu_{12}, \mu_{1a}, \mu_{1b}, \mu_{1c}, \mu_{2d}, \mu_{2e}, \mu_{21}, \mu_{a1}, \mu_{b1}, \mu_{c1}, \mu_{d2}, \mu_{e2})^\top$$

such that we seek $H(\mu+z) = 0$ if $H\mu \neq 0$ with respect to the network constraints. Notice that μ_{12} and μ_{21} are involved in both the constraints, which is the reason that generally z cannot be obtained separately one node after another. For the weights, one may e.g. let $\text{Diag}(W) = 1/\mu$, such that relatively greater changes are allowed for z_{ij} given larger μ_{ij} initially.

To apply the load-diffusion algorithm to Figure 6.3, let m_{1+} and m_{2+} contain the outflows from 1 and 2, respectively, where

$$\begin{pmatrix} m_{1+} \\ m_{2+} \end{pmatrix} = \begin{pmatrix} 0 & \mu_{1a} & \mu_{1b} & \mu_{1c} & \mu_{12} & 0 & 0 \\ \mu_{21} & 0 & 0 & 0 & 0 & \mu_{2d} & \mu_{2e} \end{pmatrix}$$

i.e. arranged according to the adjacency to $\{1, a, b, c, 2, d, e\}$, and let m_{+1} and m_{+2} contain the inflows to 1 and 2, where

$$\begin{pmatrix} m_{+1} \\ m_{+2} \end{pmatrix} = \begin{pmatrix} 0 & \mu_{a1} & \mu_{b1} & \mu_{c1} & \mu_{21} & 0 & 0 \\ \mu_{12} & 0 & 0 & 0 & 0 & \mu_{d2} & \mu_{e2} \end{pmatrix}$$

At the tth iteration, $y_{ij}^{(t)}$ is obtained from adjusting the terms of $m_{i+}^{(t-1)}$ for $i = 1, 2$, and $z_{ji}^{(t)}$ is obtained from adjusting the terms of $m_{+j}^{(t-1)}$ for $j = 1, 2$. Notice that μ_{12} would be adjusted once due to $\ell_1 - \mu_{1+}$ and once due to $\ell_2 - \mu_{+2}$; similarly, μ_{21} would be adjusted twice due to $\ell_2 - \mu_{2+}$ and $\ell_1 - \mu_{+1}$. Finally, it is possible to impose, say, $\mu_{12}^{(t)} = y_{12}$ at the end of each iteration given $(12) \in A_s$.

Chapter 7

Origin-destination flow estimation

Mobile devices can generate various origin-destination (OD) flow data. *De facto* population and commuters will be used as two motivating application scenarios in this chapter. To adjust for MNO population coverage errors and possible OD measurement errors, one may consider OD regression models from multiple disciplines such as transportation, migration, spatial econometrics, where OD device counts are treated as features, while the target OD observations and additional features may be obtained from other survey or non-survey sources. It is also worth investigating other approaches that can make use of either mathematical or statistical models of network flows.

7.1 Regression estimation

It is natural to consider OD-specific extensions of regression models. 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 for $i, j = 1, ..., n$, 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-count x_{ij} (on the log-scale) provides an additional feature.

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

$$y_{ij} = \alpha + x_{ij}\xi + h_i^\top \beta + g_j^\top \phi + d_{ij}\theta + e_{ij} \tag{7.1}$$

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 pertaining 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 μ_{ij} is valid with respect to the OD regression 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

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

where $\gamma_{ij} \in [0, 1)$ 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, the implementation to OD flows will be challenging given 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 in Italy 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 shrinkage estimation need to be demonstrated in practice, with or without spatial autoregressive impacts.

Note that one can reduce the dimension from $N \times N$-matrix to N-vector, by only considering the out-flows from all the origins or the inflows to all the destinations. The MNO counts x_{ij} can be aggregated accordingly, the irrelevant pull or push effects can be dropped, as is the distance effect $d_{ij}\theta$ in (7.1).

7.2 Multisource estimation

Let *home place* be origin and *work place* be destination, denoted by $i, j = 1, ..., n$. For each individual in the target population, $k \in U$, let

$$u_k = \begin{cases} 1 & \text{if in-work} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \delta_k = \begin{cases} 1 & \text{regular traveller} \\ 0 & \text{otherwise} \end{cases}$$

such that the population *commuter* total is given by

$$Y = \sum_{k \in U} u_k \delta_k$$

Let commuter k contribute to OD commuter flows according to the indicators $\delta_{k,ij} = 1, 0$, such that the target *commuter flows* are given by

$$Y_{ij} = \sum_{k \in U} u_k \delta_{k,ij} \delta_k$$

where one can let $\delta_{k,ii} \equiv 0$ by definition if Y_{ii} is not of interest. Meanwhile, let

$$r_k = \begin{cases} 1 & \text{if mobile device user} \\ 0 & \text{otherwise} \end{cases}$$

such that the MNO count of *traveller flows* can be given as

$$M_{ij} = \sum_{k \in U} r_k \delta_{k,ij} \delta_k$$

where we disregard any potential user ambiguity problem for now to keep focus on OD flow estimation here.

Consider two conditional independence assumptions of the distribution of $(u, \delta, \delta_{ij}, r)$ to account for the MNO-user selection effect:

$$\Pr(u\delta_{ij}\delta r = 1 \mid x) = \Pr(u\delta_{ij}\delta = 1 \mid x) \Pr(r = 1 \mid x) \tag{7.2a}$$

$$\Pr(u\delta_{ij} = 1 \mid r\delta = 1, x) = \Pr(\delta_{ij} = 1 \mid r\delta = 1, x) \Pr(u = 1 \mid r\delta = 1, x) \tag{7.2b}$$

where $x_k = x$ denotes the appropriate individual features, given which the OD commuter status $u\delta_{ij}\delta = 1$ and user status $r = 1$ are conditionally independent, and the OD flow indicator $\delta_{ij} = 1$ and the in-work status $u = 1$ are independent conditional on the MNO traveller status $r\delta = 1$ in addition.

We have then

$$\begin{aligned}
\Pr(u\delta_{ij}\delta = 1 \mid x) &\overset{(7.2a)}{=} \Pr(u\delta_{ij}\delta = 1 \mid r = 1, x) \\
&= \frac{\Pr(u\delta_{ij} \mid r\delta = 1, x) \Pr(r\delta = 1 \mid x)}{\Pr(r = 1 \mid x)} \\
&\overset{(7.2b)}{=} \Pr(\delta_{ij} \mid r\delta = 1, x) \Pr(u = 1 \mid r\delta = 1, x) \Pr(\delta = 1 \mid r = 1, x) \\
&= \Pr(\delta_{ij} \mid r\delta = 1, x) \Pr(u\delta = 1 \mid r = 1, x) \\
&\overset{(7.2a)}{=} \Pr(\delta_{ij} \mid r\delta = 1, x) \Pr(u\delta = 1 \mid x)
\end{aligned}$$

Let N_x be the subpopulation size given x. We obtain

$$\hat{Y}_{x,ij} = N_x \widehat{\Pr}(u = 1 \mid x) \widehat{\Pr}(\delta = 1 \mid u = 1, x) \frac{M_{x,ij}}{M_x} \tag{7.3}$$

as an estimator of subpopulation OD commuter flows, where $M_{x,ij}/M_x$ is an

estimator of $\Pr(\delta_{ij} \mid r\delta = 1, x)$ based on the subpopulation MNO traveller flows directly. An estimator of the population OD commuter flows is then

$$\hat{Y}_{ij} = \sum_x \hat{Y}_{x,ij}$$

Apart from MNO data yielding $M_{x,ij}/M_x$, different sources can be utilised for the other terms involved in $\hat{Y}_{x,ij}$.

- The subpopulation totals N_x are given by the population statistics.
- Population proportion $\Pr(u = 1 \mid x)$ may be given by administrative registers; otherwise, $\widehat{\Pr}(u = 1 \mid x)$ can e.g. be obtained from the Labour Force Survey.
- Population proportion of traveller among people in-work, $\Pr(\delta = 1 \mid u = 1, x)$, may be available in administrative registers or estimated from sample data of commuters. Note that in the absence of other sources of $\Pr(u = 1 \mid x)$, one would need to estimate $\Pr(u\delta = 1 \mid x)$ from a survey of commuters directly.

As outlined above, it is possible to derive a *multisource estimator* of Y_{ij} under the assumptions (7.2a) and (7.2b). The variance of $\hat{Y}_{x,ij}$ or $\hat{Y}_{ij}$ rests largely on any sample estimate required for $\Pr(u\delta = 1 \mid x)$. Apart from the assumptions (7.2a) and (7.2b), a key challenge is to ensure the validity of the estimator based on MNO data,

$$\widehat{\Pr}(\delta_{ij} \mid r\delta = 1, x) = \frac{M_{x,ij}}{M_x}$$

7.3 Network flow models

OD flows can be presented using a connected network of places (as nodes). Let $G = (\mathcal{V}, A)$ be a digraph where each directed edge $(ij) \in A$ points from node i to j in $\mathcal{V}$. Let B denote the $|\mathcal{V}| \times |A|$ 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 f_{ij} be the flow on edge (ij), which can be given a lower bound l_{ij} and an upper bound u_{ij},

$$l_{ij} \leq f_{ij} \leq u_{ij}$$

Each node $i \in \mathcal{V}$ 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.

$$\min_{(ij)\in A} c_{ij} f_{ij}$$

is defined by assigning a cost c_{ij} to each edge $(ij) \in A$, which 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 *network optimisation* problems as *mathematical* network flow models.

For OD flow estimation one may consider combining statistical assumptions and network optimisation. As illustrated in Figure 7.1, let the nodes in the network G be

$$\mathcal{V} = O \cup D \cup \{s, t\}$$

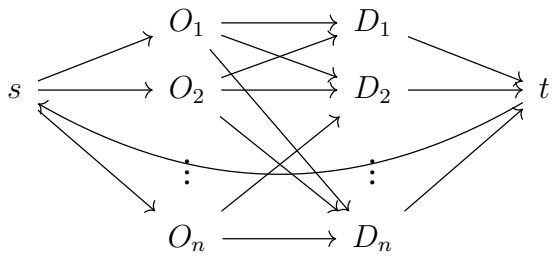

Figure 7.1: OD network flow model.

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. The flow from t to s makes $b_s = b_t = 0$, such that the supply and demand are balanced as

$$b_v \equiv 0 \quad \forall v \in \mathcal{V}$$

Let Y_{ij} be the flow from O_i to D_j, given the set of *permitted* edges $(ij) \in A$, and $Y_{ij} = 0$ if edge (ij) is not permitted, e.g. $(ii) \notin A$ if $Y_{ii} = 0$ by definition. 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, the total flow is

$$Y_{ts} = \sum_{i=1}^{n} Y_{si} = \sum_{j=1}^{n} Y_{jt}$$

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

$$\min_{f} \Delta(f, c) \quad \text{subject to} \quad Bf = 0,\ l \leq f \leq u \tag{7.4}$$

where c, f, l, u are vectors (over all the edges A in the network), and $\Delta(f, c)$ is a chosen objective function. For instance, one can let $\{c_{ij}\}$ be a set of baseline estimates, such as the OD regression estimates in Section 7.1, from which the final estimates f should not deviate too much, in case the estimates c do not satisfy the constraints by themselves. The solution to (7.4) then aims to make minimum changes to c in order to satisfy all the constraints.

It is possible to introduce various constraints, such as

$$l_{ts} = u_{ts} = \hat{Y}_{ts}$$

where $\hat{Y}_{ts}$ is an unbiased total flow estimate, or an observed total based on a trusted source. Similarly for other selected estimates $\hat{Y}_{si}$, $\hat{Y}_{jt}$ or $\hat{Y}_{ij}$. Moreover, the flow bounds (l_{ij}, u_{ij}) can be imposed statistically, e.g. a confidence interval of Y_{ij}, or simply $l_{ij} = \min(\hat{Y}_{ij}, \mu_{ij}), u_{ij} = \max(\hat{Y}_{ij}, \mu_{ij})$ where $\hat{Y}_{ij}$ is a direct estimate subject to large variance and μ_{ij} a synthetic estimate with a small variance but possibly biased. The intuition is similar to shrinkage estimation.

The above outline applies obviously to OD commuter flows estimation. Let us finish by considering the estimation of *de facto* population sizes at local

places as a problem of network flows estimation.

Denote by U the *de jure* population according to the population register. Let *de jure home place* be origin and *de facto home place* be destination, denoted by $i, j = 1, ..., n$. Let Y_{ij} be the population total of people with origin i and destination j, i.e. OD flows from *de jure* places to *de facto* places. Let

$$N_i = \sum_{j=1}^{n} Y_{ij} \quad \text{and} \quad Y_{\cdot j} = \sum_{i=1}^{n} Y_{ij}$$

be the known *de jure* population size at origin i and the unknown *de facto* population size at destination j, respectively. The total population size is

$$N = |U| = \sum_{i=1}^{n} N_i = \sum_{j=1}^{n} Y_{\cdot j}$$

Let M_{ij} be the MNO count of mobile device users with *de jure* home place i and *de facto* destination j, where we ignore the potential problem of user ambiguity for now to focus on OD flows estimation. The counts M_{ij} are subject to missing due to the non-users, i.e. $r_k = 0$ for some $k \in U$, such that

$$M_{i\cdot} = \sum_{j=1}^{n} M_{ij} < N_i \quad \text{and} \quad M_{\cdot j} = \sum_{i=1}^{n} M_{ij} < Y_{\cdot j}$$

Consider three alternative models of MNO user selection:

$$\Pr(r = 1 \mid x, i, j) = \Pr(r = 1 \mid x) \tag{7.5a}$$

$$\Pr(r = 1 \mid x, i, j) = \Pr(r = 1 \mid x, i) \tag{7.5b}$$

$$\Pr(r = 1 \mid x, i, j) = \Pr(r = 1 \mid x, j) \tag{7.5c}$$

given chosen individual features x, origin i and destination j, with known

$$U = \bigcup_{x} U_x \quad \text{and} \quad N_i = \sum_{x} N_{x,i} \quad \text{and} \quad N = \sum_{x} N_x$$

The model (7.5a) yields the subpopulation OD flow estimate given x as

$$\hat{Y}_{x,ij} = M_{x,ij} + \hat{c}_{ij} \quad \text{and} \quad \hat{c}_{ij} = (N_x - M_x)\frac{M_{x,ij}}{M_x}$$

where M_x is the total of users given x and $M_{x,ij}$ the OD flows among them. However, these $\hat{Y}_{x,ij}$ do not sum to the *de jure* population sizes $N_{x,i}$ directly. Let $M_{x,i} = \sum_{j=1}^{n} M_{x,ij}$. The model (7.5b) yields

$$\dot{Y}_{x,ij} = M_{x,ij} + \dot{c}_{ij} \quad \text{and} \quad \dot{c}_{ij} = (N_{x,i} - M_{x,i})\frac{M_{x,ij}}{M_{x,i}}$$

which sum to $N_{x,i}$ over j directly. Under the model (7.5c) an estimator of $\hat{Y}_{x,ij}$

which do not sum to $N_{x,i}$ can be given as

$$\tilde{Y}_{x,ij} = M_{x,ij} + \tilde{c}_{ij} \quad \text{and} \quad \tilde{c}_{ij} = \hat{p}_j(N_x - M_x)\frac{M_{x,ij}}{\sum_{i=1}^{n} M_{x,ij}}$$

given

$$\Big[\frac{M_{x,ij}}{\sum_{i=1}^{n} M_{x,ij}}\Big][\hat{p}_j] = \Big[\frac{N_{x,i} - M_{x,i}}{N_x - M_x}\Big]$$

Using the network flows representation (Figure 7.1) with edges $O \times D \subset A$, one can solve (7.4) for the $N_x - M_x$ non-users. Apart from $Bf = 0$, the bound constraints are

$$\begin{cases} l_{sO_i} = u_{sO_i} = N_{x,i} - M_{x,i} \\ l_{ts} = u_{ts} = N_x - M_x \end{cases}$$

One can define $\Delta(f, c)$ to have minimum changes to c, where c is given by $\hat{c}$ for the model (7.5a), or $\dot{c}$ for the model (7.5b), or $\tilde{c}_{ij}$ for the model (7.5c).

Finally, instead of relying on one of the three models, one can combine them to enhance robustness again single-model failure. For instance, one can let $c = \hat{c}$ by (7.5a), and impose the bound constraints derived from (7.5b) and (7.5c), i.e. for any $(ij) \in O \times D$, let

$$\begin{cases} l_{ij} = \min(\dot{c}_{ij}, \tilde{c}_{ij}) \\ u_{ij} = \max(\dot{c}_{ij}, \tilde{c}_{ij}) \end{cases}$$

Remark The effect of model misspecification by (7.5a) - (7.5c) is limited by the high population MNO-user percentage now-a-days. The key to any successful application will be the availability and quality of MNO counts $M_{x,ij}$.

Chapter 8

Attributive post-stratification

When it is possible to derive the target outcome from device signals directly, quasi-randomisation (QR) estimation using MNO macro data aims to adjust the *selection error* of the detected device users as a subset of the entire target population. However, several complications may need to be considered, such as lack of suitable features for QR estimation, multiple devices of given mobile phone users, ambiguity (or mis-identification) of device users.

8.1 Post-stratification estimator

Denote by $U = \{1, ..., N\}$ the target population units. Let the population mean of some associated $\{y_k : k \in U\}$ be the parameter of interest, denoted by

$$\theta = N^{-1} \sum_{k \in U} y_k$$

Let $\delta_k = 1$ if y_k is observed and 0 otherwise. Let

$$P = \{k \in U : \delta_k = 1\}$$

be a proper subset of U. The *selection* error of P is defined as $\hat{\theta} - \theta$, where

$$\hat{\theta} = \frac{\sum_{k \in U} \delta_k y_k}{\sum_{k \in U} \delta_k}$$

By treating δ_k as a random variable, for any $k \in U$, one can introduce a QR model of the selection error. In particular, suppose

$$\Pr(\delta_k = 1 \mid k \in U_g) = \pi_g \quad \text{given} \quad U = U_1 \cup \cdots \cup U_G \tag{8.1}$$

as a partition of U, with $N_g = |U_g|$ and $N = \sum_{g=1}^{G} N_g$, such that

$$\hat{\theta}_{qr} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\theta}_g \qquad \text{where} \quad \hat{\theta}_g = \frac{\sum_{k \in U_g} \delta_k y_k}{\sum_{k \in U_g} \delta_k} \tag{8.2}$$

is a *post-stratified* QR-estimator. Note that although conceptually one may

allow the selection probability $\Pr(\delta_k = 1)$ to vary from one individual to another more freely than the group-specific probability π_g, such models are impractical without access to $\{\delta_k : k \in U\}$ at the individual level.

The QR-estimator (8.2) is approximately unbiased of θ under the QR model. There is no need to estimate the probabilities $\{\pi_g\}$ explicitly, as long as $\{(N_g, \hat{\theta}_g)\}$ are given directly. This is convenient because, although the indicator δ_k will vary over time and across different y-outcomes of interest in applications, the QR model and estimator can be given generically as above in a manner that is agnostic to the chosen time period and outcome.

To apply the QR-estimator to MNO data, we assume that $\{N_g\}$ are known to the Official Statistics Agency (OSA), and that each MNO can provide the aggregated data yielding $\{\hat{\theta}_g\}$ for the post-strata $\{U_g\}$. Note that this is possible *only if* the relevant features are known to the MNO, such as age, registered address, which may be available for the service users, but not features such as activity or social status of the users which are unknown to the MNO. Note that given multiple MNOs, each of them would need to provide

$$\{(\sum_{k \in U_g} \delta_k, \sum_{k \in U_g} \delta_k y_k) : g = 1, ..., G\}$$

as aggregated counts, so that an overall $\hat{\theta}_g$ can be calculated for each $g = 1, ..., G$. We shall assume that $\{\hat{\theta}_g\}$ are available in a multi-MNO setting below.

Lack of features In the absence of device duplication or user ambiguity, we have $\delta_k = 1$ for the users of detected devices and 0 otherwise. Lack of suitable features is the case if the desired post-stratification requires some features that are unknown to the MNOs. Note that the desired QR-estimator could still have been implemented in an environment of secure multiparty computing, where all the relevant features are combined at the user level but only the group-specific aggregated data are calculated and made available to the OSA.

Device duplication Let $D = \{1, ..., M\}$ consist of all the detected devices, each with associated y_d, such that

$$\hat{\tau} = M^{-1} \sum_{d \in D} y_d$$

calculated over all the devices (instead of users) is subject to device duplication error in addition to the selection error of $P = \{k_d \in U : d \in D\}$, where k_d denotes the user associated with the device d.

However, by QR estimation, each MNO needs to calculate $\{\hat{\theta}_g\}$ instead of the naïve $\hat{\tau}$, for which y_j of user j needs to be derived from all the associated $\{y_d : k_d = j\}$. The exact mapping may depend on the outcome of interest.

First, the value y_j is obvious if the y_d's are all the same, such as it is may be the case if $y =$ place of *de facto* residence. Next, the value y_j can be inferred. For example, let $y_d = 1$ if the device 'made' an inter-city trip on a given day, and 0 otherwise, then $y_j = 1$ iff there is at least one $y_d = 1$ among $\{y_d : k_d = j\}$.

In short, y_j can be derived logically as along as $\{y_d : k_d = j\}$ are 'coherent'. An example of incoherence arises, if two devices of the same user are travelling in two parts of the country at the same time; however, this would not be an issue of device duplication but rather user ambiguity. Another example may be when a user has multiple devices serviced by different MNOs, which cannot be correctly processed without multiparty confidential computing; however, this would be another problem rooted in user ambiguity, in that the same person appears as distinct users to the MNOs.

User ambiguity User ambiguity has been described in Section 4.2. While the overall device average $\hat{\tau}$ is unaffected by user ambiguity, the apparent group-specific average for QR estimation is now given as

$$\hat{\theta}^*_g = \sum_{\substack{k \in U_g \\ \zeta_k > 0}} (y_k \nu_{kk} + y^*_k) / \sum_{\substack{k \in U_g \\ \zeta_k > 0}} \zeta_k$$

in the same notation as (4.4), which is affected by user ambiguity error as long as the post-stratum of the contractor of a device can differ to that of the actual user of the device.

8.2 Proof-of-concept study

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 the QR-estimator by replacing U with s and P 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 8.1 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)

Table 8.1: 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

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 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 macro data, one may 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, potential remedies of the complications such as device duplication or user ambiguity may be explored.

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 QR-estimation 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).

8.3 Analysis of selection error

The selection error $\hat{\theta} - \theta$ vanishes if the population covariance $Cov_N(\delta_k, y_k)$ of δ_k and y_k is zero, where Cov_N is defined with respect to the finite-population distribution function with mass $1/N$ on each $k \in U$, since we have

$$\hat{\theta} - \theta = \frac{N}{n}\Big\{\frac{\sum_{k\in U}\delta_k y_k}{N} - \Big(\frac{\sum_{k\in U}\delta_k}{N}\Big)\Big(\frac{\sum_{k\in U} y_k}{N}\Big)\Big\} = \frac{N}{n}Cov_N(\delta_k, y_k) \tag{8.3}$$

given $n = \sum_{k\in U}\delta_k$. Moreover, let $\hat{\theta}^c$ be the unobserved non-user mean,

$$\hat{\theta}^c = \frac{1}{N-n}\sum_{k\in U}(1-\delta_k)y_k$$

Let $V_N(\delta_k) = \bar{\delta}(1-\bar{\delta})$, and $\bar{\delta} = n/N$. The difference between the user and non-user means is

$$\hat{\theta} - \hat{\theta}^c = \frac{Cov_N(\delta_k, y_k)}{V_N(\delta_k)}\ .$$

8.3.1 QR adjustment

For the QR-estimator (8.2), we have, similarly to (8.3),

$$\hat{\theta}_g - \theta_g = \frac{N_g}{n_g}Cov_g(\delta_k, y_k)$$

$$Cov_g(\delta_k, y_k) \coloneqq Cov_N(\delta_k, y_k \mid k \in U_g) = \frac{\sum_{k\in U_g}\delta_k y_k}{N_g} - \Big(\frac{\sum_{k\in U_g}\delta_k}{N_g}\Big)\Big(\frac{\sum_{k\in U_g} y_k}{N_g}\Big)$$

such that

$$\hat{\theta}_{qr} - \theta = Cov_N(\bar{\delta}_g^{-1}, Cov_g) + E_N(\bar{\delta}_g^{-1})\big(Cov_N - Cov_N(\bar{\delta}_g, \theta_g)\big)$$

where $\bar{\delta}_g = \sum_{k \in U_g} \delta_k / N_g = n_g / N_g$, Cov_g is the shorthand of $Cov_g(\delta_k, y_k)$ and Cov_N that of $Cov_N(\delta_k, y_k)$. Since $\bar{\delta} = E_N(\bar{\delta}_g)$, we obtain

$$\hat{\theta}_{qr} - \theta = E_N(\bar{\delta}_g^{-1})E_N(\bar{\delta}_g)(\hat{\theta} - \theta) - E_N(\bar{\delta}_g^{-1})Cov_N(\bar{\delta}_g, \theta_g) + Cov_N(\bar{\delta}_g^{-1}, Cov_g) \qquad (8.4)$$

It follows from (8.4) that we have $\hat{\theta}_{qr} - \theta = \hat{\theta} - \theta$ if $\bar{\delta}_g = \bar{\delta}$, i.e. the selection error of $\hat{\theta}$ is unadjusted by QR-estimation in this case. Whereas, if $\bar{\delta}_g \neq \bar{\delta}$ and $Cov_g \equiv 0$, i.e. $\hat{\theta}_g$ has no selection error, then the selection error of $\hat{\theta}$ exists only due to the covariance between $\bar{\delta}_g$ and θ_g, i.e.

$$\hat{\theta} - \theta \overset{Cov_g = 0}{=} \bar{\delta}^{-1} Cov_N(\bar{\delta}_g, \theta_g)$$

It seems from (8.4) that post-stratification for QR-estimation may be guided by the resulting $Cov_N(\bar{\delta}_g, \theta_g)$, which is unknown but can be approximated by the observed $Cov_N(\bar{\delta}_g, \hat{\theta}_g)$. Of course, in reality, the choice of post-stratification may be limited by data accessibility or other practical considerations, and one may be able to reasonably gauge the effect of QR-estimation by comparing $\hat{\theta}_{qr}$ to $\hat{\theta}$ directly in light of one's intuitive grasp of the selection error of $\hat{\theta}$.

8.3.2 Sample analysis

For proof-of-concept experiments, let s be a simple random household sample of persons from U, and let $s_P = s \cap P$ be the subsample of MNO users. One can construct $\big(\bar{y}(s), \bar{y}(s_P), \bar{y}_{qr}(s_P)\big)$ as sample analogy to $(\theta, \hat{\theta}, \hat{\theta}_{qr})$, in order to analyse various errors and effects subject to cluster sampling variances.

Selection error Let H (or H_s) be the number of households in the population (or sample). Let s_i be the subsample of persons in each household $h = 1, ..., H_s$. The difference between the mean of y_k over s_P and s is given as

$$\begin{aligned} b = \bar{y}(s_P) - \bar{y}(s) &= \frac{\sum_{h=1}^{H_s} \sum_{k \in s_i} \delta_k y_k}{\sum_{h=1}^{H_s} \sum_{k \in s_i} \delta_k} - \frac{\sum_{h=1}^{H_s} \sum_{k \in s_i} y_k}{\sum_{h=1}^{H_s} \sum_{k \in s_i} 1} \\ &= \frac{\sum_{h=1}^{H_s} t_i}{n(s_P)} - \frac{\sum_{h=1}^{H_s} Y_i}{n(s)} \end{aligned}$$

Over repeated cluster sampling of s, with fixed $\{\delta_k : k \in U\}$, we have $E(b) \doteq \hat{\theta} - \theta$.

By Taylor expansion of $n(s_P)$ and $n(s)$ around their expectation, we obtain an approximate variance estimator of b, which is given as

$$\hat{V}(b) = \big(1 - \frac{H_s}{H}\big) \frac{H_s s_z^2}{n(s)^2}$$

where

$$s_z^2 = \frac{1}{H_s - 1}\sum_{h=1}^{H_s}\big(z_i - \frac{1}{H_s}\sum_{h=1}^{H_s} z_i\big)^2 \quad \text{and} \quad z_i = \frac{n(s)}{n(s_P)}t_i - Y_i$$

Moreover, for the sample estimator of $Cov_N(\delta_k, y_k)$, denoted by

$$cov_N = \frac{\sum_{k\in s}\delta_k y_k}{n(s_P)} - \frac{n(s_P)}{n(s)}\Big(\frac{\sum_{k\in s} y_k}{n(s)}\Big) = \frac{n(s_P)}{n(s)}b$$

we have

$$\hat{V}(cov_N) = \big(\frac{n(s_P)}{n(s)}\big)^2\hat{V}(b)$$

QR error Let $n_g(s) = |U_g \cap s|$, such that $n(s) = \sum_{g=1}^G n_g(s)$. Let

$$\bar{y}_{qr}(s_P) = \sum_{g=1}^{G}\frac{n_g(s)}{n(s)}\cdot\frac{\sum_{i\in s_P}\mathbb{I}(k\in U_g)y_k}{\sum_{k\in s_P}\mathbb{I}(k\in U_g)}$$

be the *sample* QR-estimator, such that $b_{qr} = \bar{y}_{qr}(s_P) - \bar{y}(s)$ is given by

$$\begin{aligned}
b_{qr} &= \sum_{g=1}^{G}\frac{n_g(s)}{n(s)}\Big(\frac{\sum_{h=1}^{H_s}\sum_{k\in s_i}\delta_k y_k\mathbb{I}(k\in U_g)}{\sum_{h=1}^{H_s}\sum_{k\in s_i}\delta_k\mathbb{I}(k\in U_g)} - \frac{\sum_{h=1}^{H_s}\sum_{k\in s_i} y_k\mathbb{I}(k\in U_g)}{n_g(s)}\Big)\\
&= \sum_{g=1}^{G}\frac{n_g(s)}{n(s)}\Big(\frac{\sum_{h=1}^{H_s} t_{h,g}}{n_g(s_P)} - \frac{\sum_{h=1}^{H_s} Y_{h,g}}{n_g(s)}\Big)\\
&= \frac{1}{H_s}\sum_{h=1}^{H_s}\Big(H_s\sum_{g=1}^{G}\frac{n_g(s)}{n(s)}\big(\frac{t_{h,g}}{n_g(s_P)} - \frac{Y_{h,g}}{n_g(s)}\big)\Big) = \frac{1}{H_s}\sum_{h=1}^{H_s}\dot{z}_i
\end{aligned}$$

where $t_i = \sum_g t_{h,g}$ and $Y_i = \sum_g Y_{h,g}$ in each household. We have $E(b_{qr}) \doteq \hat{\theta}_{qr} - \theta$ and, similarly to $V(b)$ above,

$$\hat{V}(b_{qr}) = \big(1 - \frac{H_s}{H}\big)^2\frac{s_{\dot{z}}^2}{H_s} \quad \text{and} \quad s_{\dot{z}}^2 = \frac{1}{H_s - 1}\sum_{h=1}^{H_s}\big(\dot{z}_i - \frac{1}{H_s}\sum_{h=1}^{H_s}\dot{z}_i\big)^2$$

QR effect For the effect of QR-adjustment on the selection error, let

$$\begin{aligned}
e_{qr} = \bar{y}_{qr}(s_P) - \bar{y}(s_P) &= \sum_{g=1}^{G}\Big(\frac{n_g(s)}{n(s)} - \frac{n_g(s_P)}{n(s_P)}\Big)\frac{\sum_{h=1}^{H_s} t_{h,g}}{n_g(s_P)}\\
&= \frac{1}{H_s}\sum_{h=1}^{H_s}\Big(\sum_{g=1}^{G}\big(\frac{n_g(s)}{n(s)} - \frac{n_g(s_P)}{n(s_P)}\big)\frac{H_s}{n_g(s_P)}t_{h,g}\Big) = \frac{1}{H_s}\sum_{h=1}^{H_s}\dot{t}_i
\end{aligned}$$

We have $E(e_{qr}) \doteq \hat{\theta}_{qr} - \hat{\theta}$ and

$$\hat{V}(e_{qr}) = \big(1 - \frac{H_s}{H}\big)^2 \frac{s_t^2}{H_s} \quad \text{and} \quad s_t^2 = \frac{1}{H_s - 1} \sum_{h=1}^{H_s} \big(\dot{t}_i - \frac{1}{H_s} \sum_{h=1}^{H_s} \dot{t}_i\big)^2$$

8.4 SUD estimator

Consider an approach for dealing with user ambiguity and device duplication in addition to selection error, which combines the MNO data computed for the contractor groups $g = 1, ..., G$ and the device-user-contractor connections obtained *separately* for the specific y-outcomes at various time points (such as by sample surveys of resident mobile device usage). Notice that our approach will allow the devices of a user to have different service contractors.

Using the notations introduced in Section 4.2, where $\sum_{j \in U} a_{dj} \equiv 1$ and $\sum_{k \in U} c_{dk} \equiv 1$ for any $d \in D$, we have

$$Z = \sum_{d \in D} y_d = \sum_{d \in D} \Big(\sum_{j,k \in U} c_{dk} d_{dj} \Big) y_d = \sum_{\substack{k \in U \\ \zeta_k > 0}} \sum_{\substack{j \in U \\ \alpha_j > 0}} \Big(\sum_{d \in D} c_{dk} d_{dj} y_d \Big)$$

i.e. the total of y_d over the devices given as the sum over contractors k and users j. The breakdown of Z by $\{U_g\}$ according to the users is

$$Z = \sum_{g=1}^{G} Z_g \qquad \text{where} \quad Z_g = \sum_{\substack{j \in U_g \\ \alpha_j > 0}} \Big(\sum_{d \in D} a_{dj} y_d \Big) = \sum_{\substack{j \in U \\ \alpha_j > 0}} z_j$$

is the total of z_j over the users in U_g. However, the observed breakdown of Z according to the contractors would be

$$Z = \sum_{l=1}^{G} Z_l^* \qquad \text{where} \quad Z_l^* = \sum_{g=1}^{G} \sum_{\substack{j \in U_g \\ \alpha_j > 0}} \sum_{d \in D} C_l(d,j) a_{dj} y_d \quad \text{and} \quad C_l(d,j) = \sum_{\substack{k \in U_l \\ \zeta_k > 0}} c_{dk}$$

Notice that $C_l(d,j) = 1$ occurs if the given device d of user j is contracted by someone in U_l, no matter who that person is, and $C_l(d,j) = 0$ otherwise.

For a model of independent $C_l(d,j)$ and $a_{dj} y_d$ conditional on $j \in U_g$, let

$$\phi_{lg} = \Pr\big(C_l(d,j) = 1 \mid a_{dj} = 1, j \in U_g\big) \tag{8.5}$$

be the probability that a device used by someone in U_g is contracted by someone in U_l. Given the model (8.5), we have

$$E_\phi(Z_l^*) = \sum_{g=1}^{G} \phi_{lg} Z_g$$

where E_ϕ denotes expectation over the random variables $\{c_{dk}\}$. Similarly,

$$E_\phi(M_l^*) = \sum_{g=1}^{G} \phi_{lg} M_g$$

where M_g is the number of devices belonging to the users in U_g, which formally can be given as Z_g with $y_d \equiv 1$ for all $d \in D$, and M_l^* is the number of devices attributed to the contractors in U_l.

For example, suppose a device of a teenage user, say, in U_g, is contracted by her mother who belongs to U_l, where $l \neq g$. Under the model (8.5), the probability ϕ_{lg} of such an event is the same for all the teenage users in U_g, but it can differ for her father in $U_{g'}$ as long as $g' \neq g$, if his device is also contracted by the mother. Note that the model (8.5) allows the devices of a given user to be contracted by persons in different post-strata, e.g. when a teenage user has two devices contracted by either of her parents, respectively.

Let $[Z_l^*]$ be the vector of Z_l^*, and $[Z_g]$ that of Z_g. Similarly for $[M_l^*]$ and $[M_g]$. Let $[\phi_{lg}]$ be the matrix of ϕ_{lg}. Provided it is invertible, an estimator of $\bar{Z}_g = Z_g/M_g$ can be given as $\hat{\bar{Z}}_g = \hat{Z}_g/\hat{M}_g$, where

$$[\hat{Z}_g] = [\phi_{lg}]^{-1}[Z_l^*] \quad \text{and} \quad [\hat{M}_g] = [\phi_{lg}]^{-1}[M_l^*] \ .$$

Finally, the user population mean θ_g in U_g, defined in (8.2), can be written as

$$\hat{\theta}_g = \xi_g \bar{Z}_g$$

where ξ_g exists as a factor associated with U_g due to device duplication. It follows that, given the models (8.1) and (8.5) and, separately, the estimates $[\hat{\xi}_g]$ and $[\hat{\phi}_{lg}]$, an estimator of θ is given as

$$\hat{\theta}_{sud} = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\xi}_g \frac{\left[[\hat{\phi}_{lg}]^{-1}[Z_l^*]\right]_g}{\left[[\hat{\phi}_{lg}]^{-1}[M_l^*]\right]_g} \tag{8.6}$$

where $[\cdot]_g$ denotes the g-th component of the vector $[\cdot]$, and the subscript "sud" is a shorthand for 'selection error', 'user ambiguity' and 'device duplication'.

The SUD-estimator (8.6) requires the MNOs to provide $\{(Z_l^*, M_l^*) : l = 1, ..., G\}$, which are totals over devices classified according to their contractors; whereas surveys can provide the estimates $[\hat{\xi}_g]$ and $[\hat{\phi}_{lg}]$, at least in principle.

Clearly, the model (8.5) is reasonable if it can yield $[\hat{\bar{z}}_g] \approx [\bar{z}_g]$. However, we actually do not expect the assumption (8.5) to hold generally regardless the y-outcomes, because the contractor-user relationship does not occur arbitrarily in the population. For instance, it is common for parents to be the contractors of their children, but not the other way around; nor is it as common for adults to be contractors of other adult users than for child users.

Nevertheless, it is intriguing to observe that the SUD-estimator (8.6) can still be quite close to the genuine QR-estimator, provided matrices $[\hat{\phi}_{lg}]$ and $[\eta_{lg}]$ together satisfy a certain mathematical property, whether or not $[\hat{\bar{z}}_g] \approx [\bar{z}_g]$

actually. To be specific, let

$$[b_{lg}] = [\hat{\phi}_{lg}]^{-1}[\eta_{lg}]$$

where $\eta_{lg}Z_g$ is the contribution from Z_g to Z_l^*, i.e.

$$[\hat{Z}_g] = [\hat{\phi}_{lg}]^{-1}[Z_l^*] = [\hat{\phi}_{lg}]^{-1}[\eta_{lg}][Z_g] = [b_{lg}][Z_g]$$

Ideally, if the gth column of $[b_{lg}]$ can be given as $b_{gg} = 1/\sigma_{gg}$ and $b_{lg} = -\sigma_{lg}/\sigma_{gg}$ for $l \neq g$, where $\sum_l \sigma_{lg} = 1$, $\forall g$. Then, regardless $[\hat{Z}_g] \neq [Z_g]$, we would have

$$\sum_g \left(\hat{Z}_g - Z_g\right) = \sum_{g=1}^{G} b_{gg}^{-1}\Big(1 - \sum_{l=1}^{G} b_{lg}\Big) Z_g = 0 \tag{8.7}$$

In practice, as long as the columns of $[b_{lg}]$ can be approximated in this way, the SUD estimator (8.6) can be robust in the sense of nearly reproducing the genuine QR estimator, even though the model (8.5) does not hold.

Remark When non-informative contractor-user connections (8.5) are not the case *and* the condition (8.7) does not hold approximately, one would need not only to estimate $[\phi_{lg}]$ but also $[\eta_{lg}]$. This means that one needs to identify the users and the contractors of all the in-scope devices *jointly* with the different target y-outcomes. In other words, for a user with multiple devices, one needs to identify the contractor of each device, as well as whether the given device is present in each relevant situation, such as at night ($y =$ *de facto* resident place), domestic trips ($y =$ tourism trips), travel to-and-from work ($y =$ commuter), or local trips of other purposes ($y =$ trips for travel statistics). This is clearly very demanding in practice.

Part III

Application scenarios

Chapter 9

Inbound tourism

This application scenario concerns inbound tourism statistics in Spain, where the existing official statistics are based on non-MNO data from sample surveys and passenger traffic, and the experimental statistics based on MNO data.

Following Eurostat regulation and international standards, tourism refers to the activity of visitors, which may be internal, outbound or inbound, where tourism covers a subset of all trips and visitors are a subset of all travellers.

In practice, one can let the reference population of inbound tourism consist of all the inbound trips by nonresidents over a given time period, denoted by U. Let y_i be associated with each $i \in U$. For instance, let $y_i = 1$ if tourist trip i, $y_i = 0$ otherwise. Or, let $y_i = d$ be the main destination of tourist trip i, $y_i = 0$ otherwise, for $d = 0, 1, ..., D$. Or, let y_i be the number of nights spent during tourist trip i, $y_i = 0$ otherwise. The corresponding inbound tourism total is

$$Y = \sum_{i \in U} y_i$$

9.1 MNO and non-MNO data

Let each inbound trip be associated with one and only one border-crossing, such that the total number of inbound trips is the same as the total number of border-crossing (into the country). In Spain, this total can be divided into four types of border-crossing: airport, port, train, road. Figure 9.1 provides an overview of the data of Tourist Movements at Borders (FRONTUR) accordingly, which are available to the National Statistical Institute (INE). Below we highlight some aspects of the data sources.

Airport First, the Spanish Association for the Coordination and Facilitation of Time Slots (AECFA, Asociación Española para la Coordinación y Facilitación de Franjas Horarias) provides INE with the flight schedule by arrival airport. Next, the Spanish Airports and Air Navigation (AENA, Aeropuertos Españoles y Navegación Aérea) provides the list of flights actually taken place and the number of passengers on each flight. Finally, for each flight from a non-Schengen airport in the AENA list, the numbers of passengers by nationality are available, provided by the General Directorate of the National Police from the Advanced Passenger Information System (API).

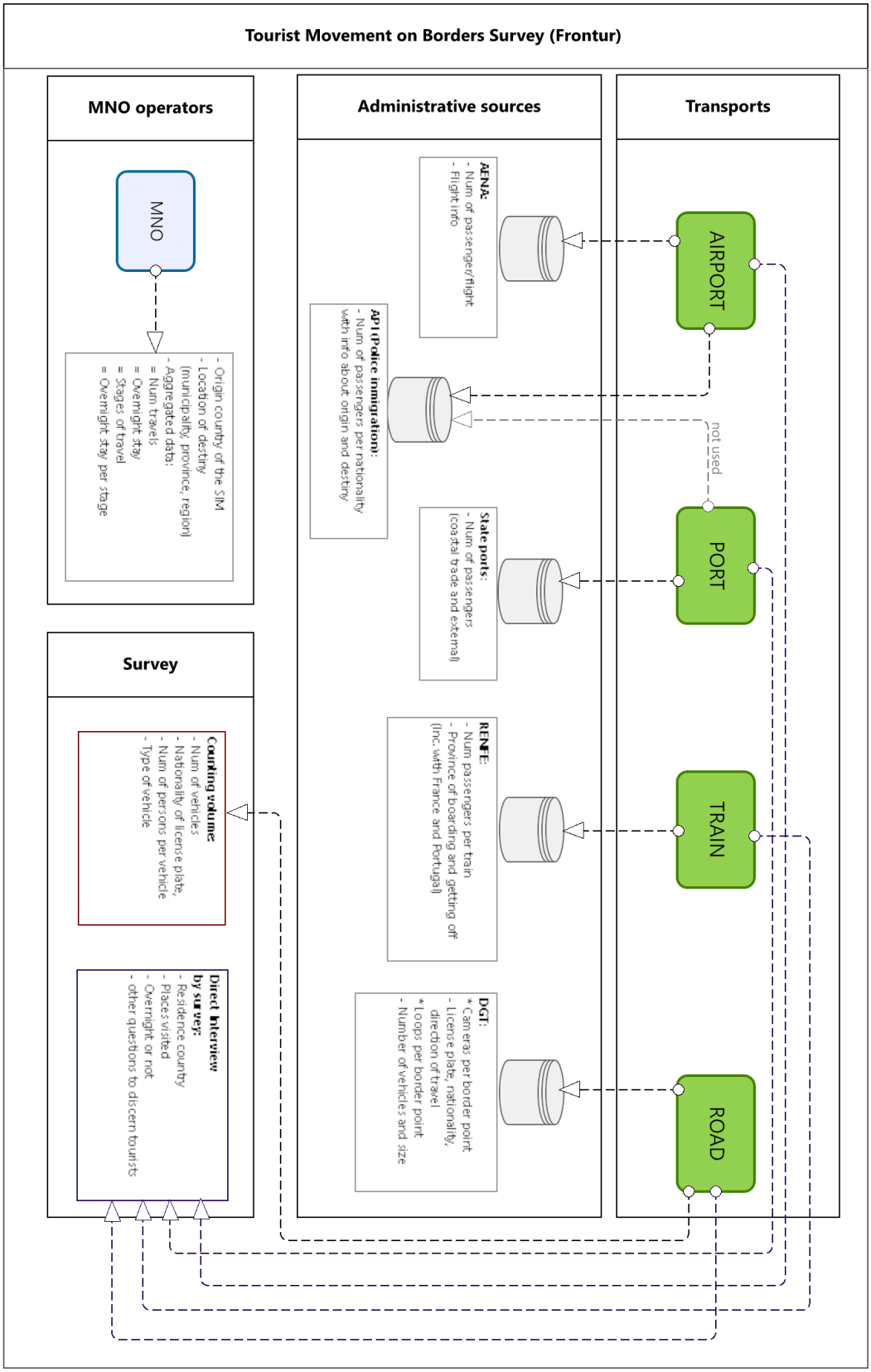

Figure 9.1: Diagram with the different sources of data in FRONTUR

Port Monthly reports of passenger totals are available from the State Ports. For each port, we obtain the number of passengers for regular disembarkation or in regular transit, for cruise disembarkation or in cruise transit.

Train The monthly total number of inbound passengers from either France or Portugal is provided by Red Nacional de los Ferrocarriles Españoles (RENFE), which is Spain's national railway company.

Road The General Directorate of Traffic (DGT) provides the sensor counts of vehicles entering and leaving at all border points, which are based on inductive loops installed at these points, where the vehicles can be classified according to their length as small, medium or large.

Additionally, counts of vehicles by video cameras installed at the main entry points to the country are available. However, the extent of missing camera counts are greater than the loop counts, such as when a camera breaks down and not repaired immediately.

Survey Sample surveys are conducted for all types of border-crossing. The total sample size is more than 450 000 each year, where the sampling design differs according to the types of border-crossing and the sampling units involve flights, ships, trains, vehicles, and individuals ultimately.

Face-to-face interviews are conducted on departure due to practical reasons as well as the information need. Tourist visitors with or without overnights are identified by the survey. Additional data are collected, such as primary destination, purpose of trip, type of accommodation, country of residence, use of travelling agency.

Capacity survey It is worth mentioning a special survey for border-crossing by road, for which the loop counts give the numbers of vehicles but not the number of passengers. An operation called *capacity counting* is carried out by surveyors managed by the DGT, during which each surveyor would count all the vehicles passing a given point, the nationality of the vehicle's license plate, the number of passengers, and the type of vehicle. Note that capacity counting is not conducted in the evenings or nights, typically from 4pm to 8am.

MNO data Since the end of 2020, the INE and each MNO have worked together to implement the FRONTUR concepts in the MNO data. The outcome includes the following monthly aggregate counts of ended trips:

- number of trips per main destination,
- number of overnights per main destination,
- number of trip stages (by different places of overnight stops),
- number of overnights by stages,
- number of same-day visits per main destination.

A series of algorithms are needed for the nano-level signals captured by mobile network antennas, the micro-level data transformation according to the definitions of tourism statistics, and the processing of the required aggregates at the macro level. The algorithms have been developed by each participating MNO, with continuous interaction from the INE.

9.2 Combining MNO and survey data

The reference population size can be obtained by passenger traffic data for the airports, ports and trains, but it needs to be estimated for the road passengers. Appendix B describes some methods that may be of interest, which however does not involve MNO data. Here we focus on combining the FRONTUR survey estimates with the MNO counts in the experimental statistics. Specifically, consider estimating the total inbound tourist trips to a given destination, using the transfer learning approach described in Section 5.1.

- Instead of the MNO counts directly, we used bias-adjusted MNO estimates for transfer learning, as will be explained.
- Transfer learning with γ by cross-validation, where $w_k = 1$ or $w_k = 1/x_k$. The results are nearly identical, so we only present the results by $w_k = 1$ below.
- Transfer learning with γ from minimising the MSE, where the variances are estimated using a state-space model.

However, it will be important to keep working with the MNOs, in order to better understand the observed systematic differences between the MNO counts and the FRONTUR survey estimates, such as when a clear seasonal bias exists in the former, and to improve the MNO counts in future.

9.2.1 FRONTUR survey vs. MNO estimates

Figure 9.2 compares the FRONTUR survey estimates to the MNO counts, for Balearic islands, Las Palmas and Santa Crus de Tenerife among the Canary islands, as well as the passenger totals (referred to as admin totals). One can observe a notable difference between the survey and MNO estimates, as well as some obvious effects due to covid-19.

The survey estimates are on average slightly lower than the passenger totals, which is reasonable since there are always passengers who are not tourists. The MNO estimates are consistently lower than the survey estimates, which suggests the former are biased. In particular, the bias for Balearic islands has a clear seasonal pattern with yearly periodicity, reaching a peak during august while falling to a minimum during winter. We would therefore like to remove such biases, before applying the transfer learning estimator.

Figure 9.3 gives the logarithm differences between the FRONTUR survey and MNO estimates. Although the beginning of the time series is unusable, the years after covid-19 display a reasonably stable yearly pattern. Indeed, by differentiating the log series with a lag of 12 months, we obtained what seems

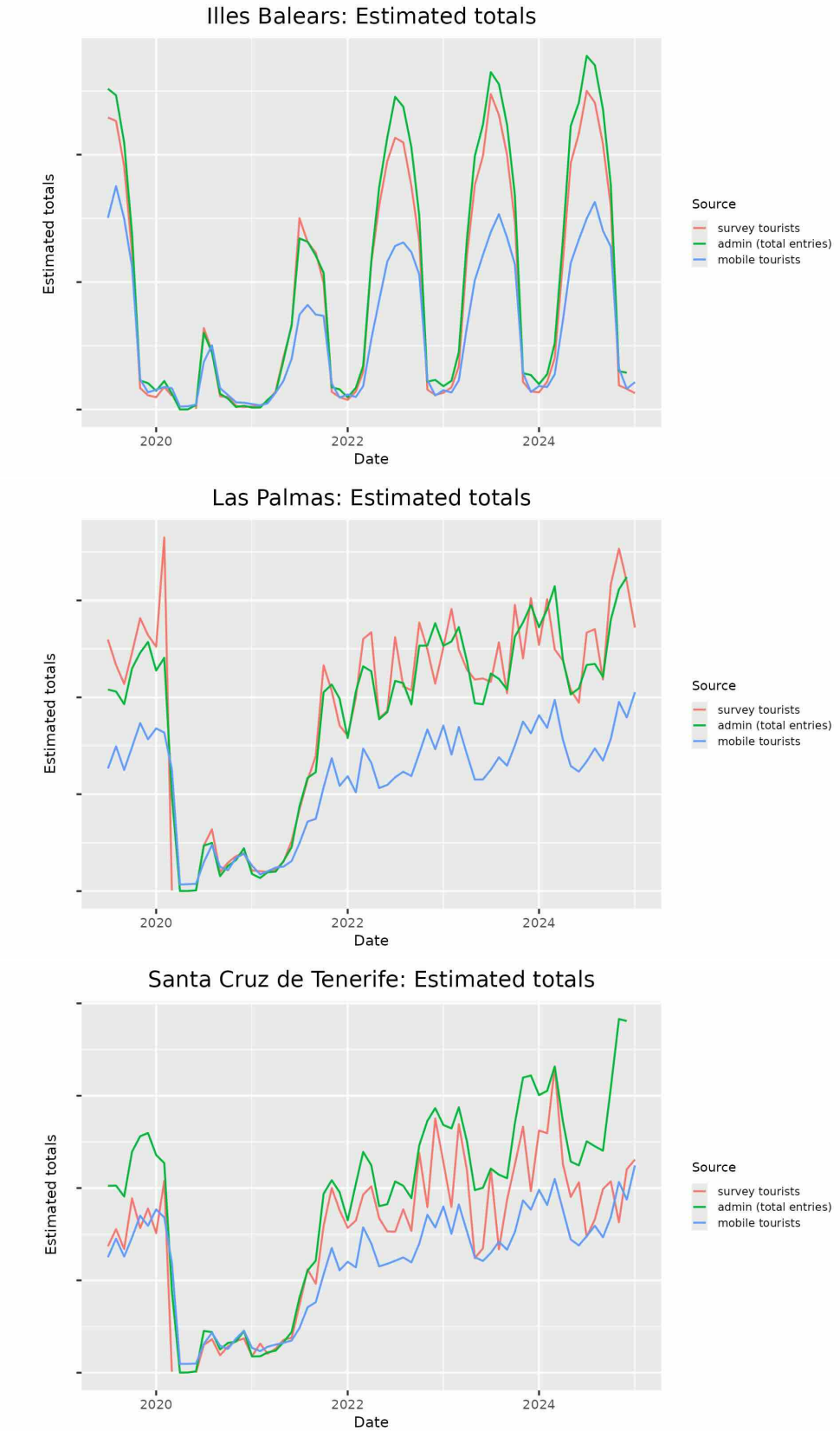

Figure 9.2: FRONTUR survey and MNO estimates given destination, passenger counts as admin total entries

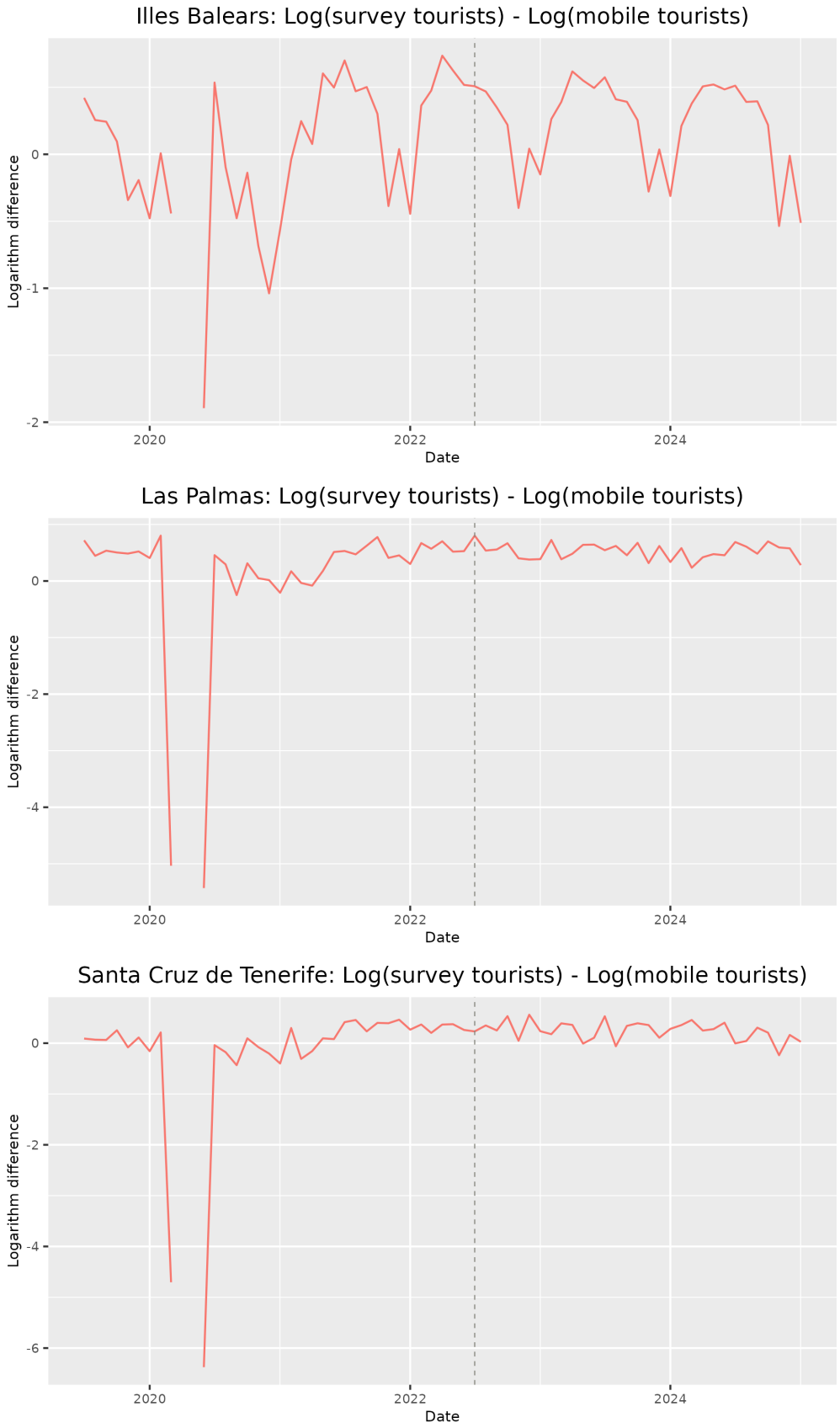

Figure 9.3: Logarithm differences between survey and MNO estimates

to be a stationary series for Balearic Islands. Note that we apply the operation only to the time points after the grey dotted line in Figure 9.3, since we deemed the data from 2020 to be unusable, even for calculating the differences for 2021. At this point, we assumed that these differences were equivalent to white noise with a variance of σ^2, which could be used to correct the bias of MNO estimator. The procedure is as follows.

Let $\hat{Y}_t$ be the FRONTUR survey estimator at time point t, and let X_t be the corresponding MNO count. Let

$$z_t = \log \hat{Y}_t - \log X_t$$

be the log difference. We assume white noise in case of seasonal MNO bias,

$$z'_t = z_t - z_{t-12} \sim N(0, \sigma^2)$$

Over a given time window $t_1, ..., t_n$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (z'_{t_i})^2$$

and the bias-adjusted MNO estimate

$$\hat{X}_t = X_t \frac{\hat{Y}_{t-12}}{X_{t-12}} e^{-\frac{\hat{\sigma}^2}{2}}$$

Meanwhile, the MNO count for neither of the Canary islands shows a clear seasonal bias, where the log difference between the survey and MNO estimates appears like white noise, except the covid-19 effects, i.e.

$$z_t = \log \hat{Y}_t - \log X_t \sim N(0, \sigma^2)$$

directly. To correct the MNO count, let

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} z_{t_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_{t_i} - \hat{\mu})^2$$

given a time window $t_1, ..., t_n < t$, such that we obtain

$$\hat{X}_t = X_t \exp\{\hat{\mu} - \frac{\hat{\sigma}^2}{2}\}$$

9.2.2 Transfer learning by cross-validation

Balearic Islands As shown in the top plot of Figure 9.4, the MNO estimates of tourist proportions behave much more erratically than the survey ones before the time point marked by the grey dotted line, which may be largely due to the effects of covid-19. The MNO estimates appear much more reasonable after the grey dotted line, where they are closer to the survey estimates, even though they can still exceed 1 from time to time.

Regarding the totals in the bottom plot of Figure 9.4, we can see that after

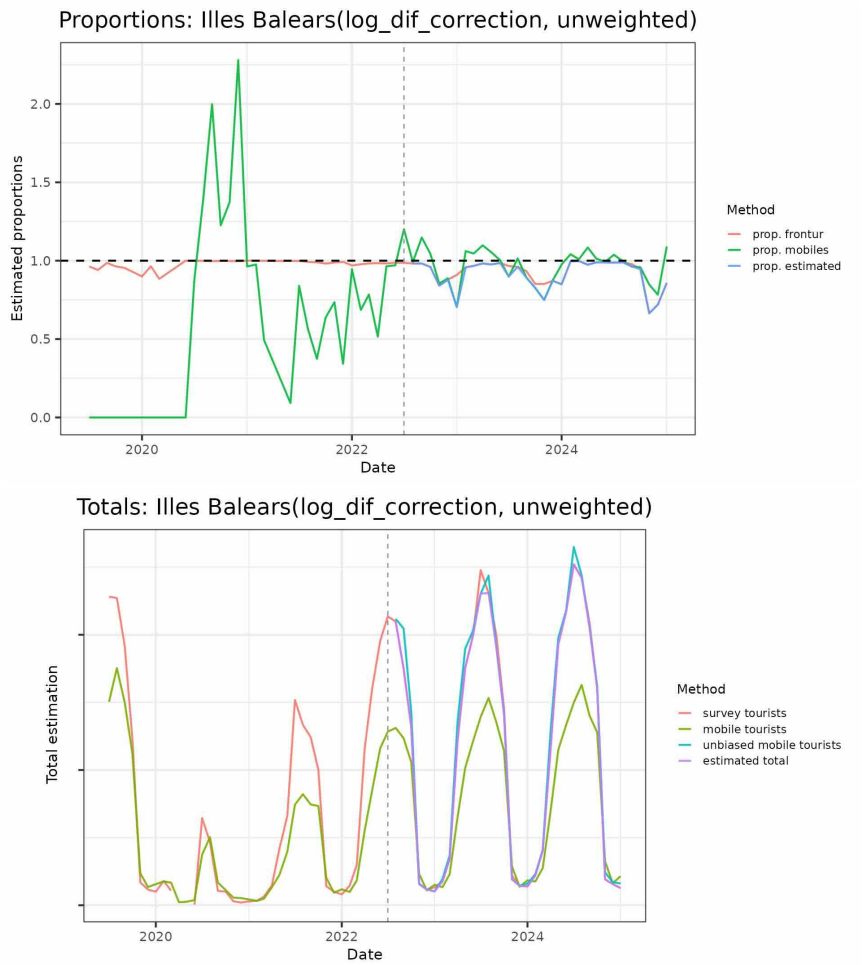

Figure 9.4: Estimates by survey, MNO and transfer learning, Balearic Islands, tourist proportion (top) or total (bottom, adjusted MNO estimates as "unbiased")

applying the bias correction to the mobile data, referred to as "unbiased" in the figure, it shows much more similar values to the survey totals. Moreover, the resulting transfer learning estimator appears to be much closer to the survey estimated totals, rarely getting closer to the adjusted mobile counts $\hat{X}_t$. Rather than choosing an intermediate value between 0 and 1, the coefficient γ for transfer learning is often either 0 or 1.

Canary Islands The Canary Islands showcase a much different behaviour, where the estimated proportion seen in the top plots of Figures 9.5 and 9.6 oscillates much more between the survey and MNO estimates, although the coefficient γ remains often either 0 or 1.

The bottom plots in Figures 9.5 and 9.6 show the estimated tourist totals. As we can see, the bias correction brings the adjusted MNO counts much closer to the survey estimates; however, given the lack of seasonality, the agreement is not as close as in the Balearic Islands.

The transfer learning estimator tends to either pick the survey estimate or the adjusted MNO count. The resulting time series seems less volatile than the 'opposite' series that would have resulted from using $\gamma' = 1 - \gamma$ instead of γ.

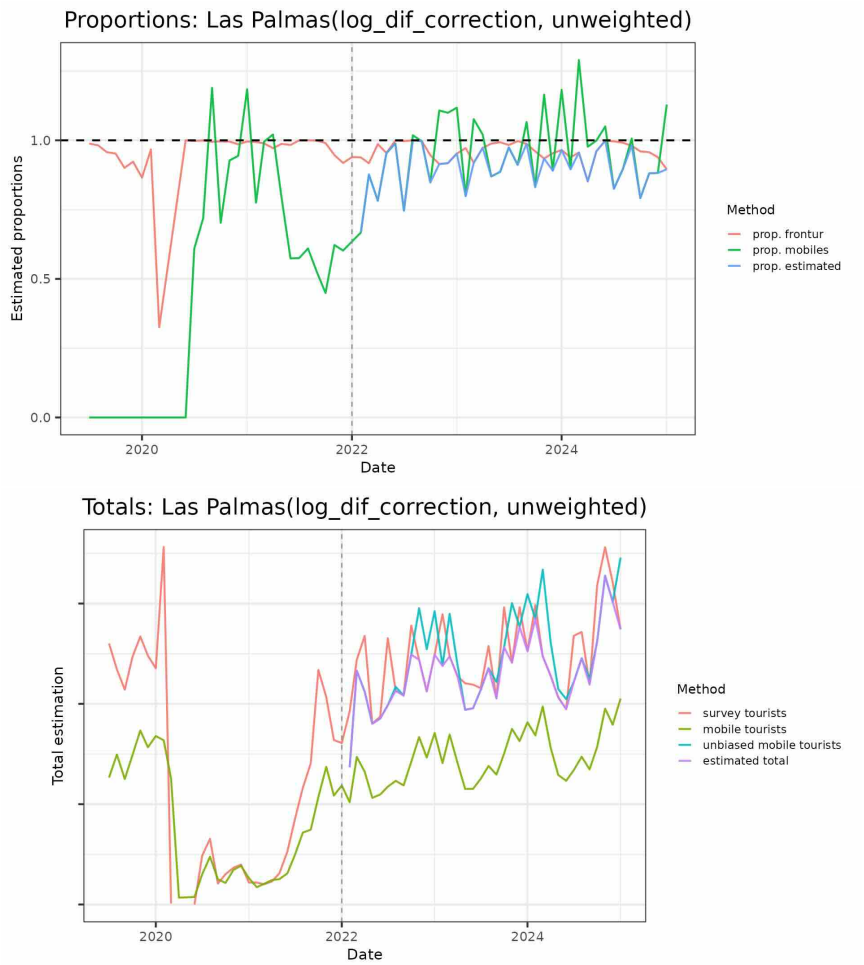

Figure 9.5: Estimates by survey, MNO and transfer learning for Las Palmas, tourist proportion (top) or total (bottom, adjusted MNO estimates as "unbiased")

9.2.3 Transfer learning using state-space models

It is actually not straightforward to estimate the sampling variances of the survey estimators, nor are the survey data available to us in this project. We therefore consider a superpopulation modelling approach to estimate the MSEs in order to determine γ for transfer learning (Section 5.1). Basically, both the survey estimator and the MNO count differ to the unknown number of tourists by their respective errors, and state-space models can be used to describe the errors over time, by which one can accommodate potential trend or seasonal effects in the bias of the MNO count as well.

Canary Islands We use the same model for both provinces in Canary Islands. Denote by x_t the (hidden) true number of tourists, assumed to follow a random walk over time t. Denote by y_t the vector of survey estimate and MNO count. We assume the survey estimator to be unbiased, while a constant bias exists

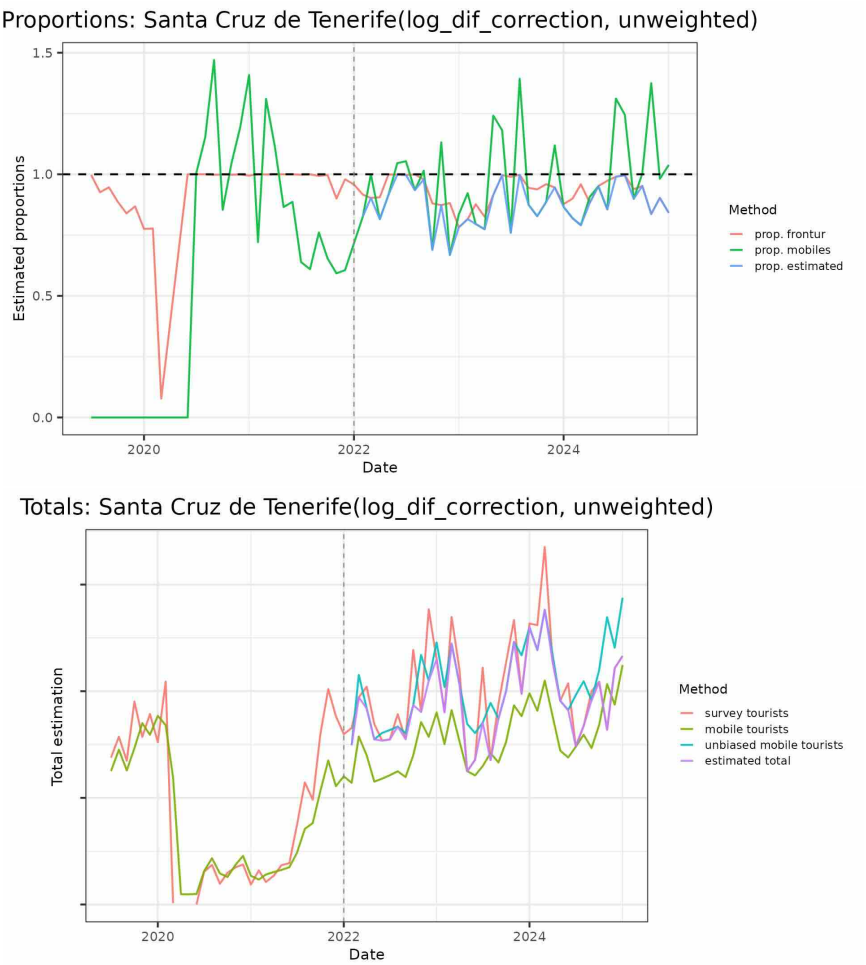

Figure 9.6: Estimates by survey, MNO and transfer learning for Santa Cruz, tourist proportion (top) or total (bottom, adjusted MNO estimates as "unbiased")

of the MNO count. Let

$$
\begin{aligned}
x_t &= x_{t-1} + \sigma e_{it} \\
y_t &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \alpha + \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \begin{pmatrix} \sigma_S & 0 \\ 0 & \sigma_M \end{pmatrix} \begin{pmatrix} e_{it}^S \\ e_{it}^M \end{pmatrix}
\end{aligned}
$$

where e_{it}'s are normally distributed with zero mean and unit variance, σ^2's are the variances, and α accounts for the bias of the MNO count.

Maximum likelihood estimation (MLE) can be achieved using the Kalman filter. It is convenient to express the variances relatively, say, to σ_S^2, as well as α^2, now that γ is invariant if the MSEs are scaled proportionally.

For Tenerife we have $\hat{\sigma}_M^2 = 0.044\sigma_S^2$ and $\hat{\alpha}^2 = 1.95\sigma_S^2$. This means that variance is much smaller for the MNO count than the survey estimator, but the large bias of the MNO count makes it worse than the survey estimator in terms of MSE. The bias-corrected MNO data estimator $y_{2t} - \hat{\alpha}$, is quite good as its variance (including that of $\hat{\alpha}$) is $0.069\sigma_S^2$, which is more than fourteen times smaller than the survey estimator. Transfer learning can still improve it a little

bit by combining it with the survey estimator:

$$\hat{x}_t^{TL} = (\gamma, 1-\gamma)\begin{pmatrix} y_{1t} \\ y_{2t} - \hat{\alpha} \end{pmatrix} \quad \text{and} \quad \gamma = \frac{0.069\sigma_S^2}{\sigma_S^2 + 0.069\sigma_S^2}$$

The variance of the TL-estimator is then $0.064\sigma_S^2$. Note that the Kalman filter estimator of x_t has a variance of $0.058\sigma_S^2$, which is even smaller, because it is optimal under the state-space model.

Similarly for Las Palmas, where we obtain $\hat{\sigma}_M^2 = 0.071\sigma_S^2$ and $\hat{\alpha}^2 = 8.61\sigma_S^2$. The variance is $0.096\sigma_S^2$ for the bias-corrected MNO count, $0.088\sigma_S^2$ for the TL-estimator, and $0.071\sigma_S^2$ for the Kalman filter estimator of x_t.

Illes Balears The seasonal behaviour of the Illes Balears series requires a slightly more complex model. Let the unknown number of tourists be

$$x_t = \mu_t + s_t + \sigma e_{it}$$

where μ_t follows the local linear trend model over time,

$$\mu_{t+1} = \mu_t + \nu_t + \sigma_\mu e_{it}^\mu$$

with ν_t as the drift or slope of the trend, and s_t the seasonal component. Let

$$y_t = \begin{pmatrix} 0 \\ 1 \end{pmatrix}\alpha + \begin{pmatrix} 1 \\ \beta \end{pmatrix} x_t + \begin{pmatrix} \sigma_S & 0 \\ 0 & \sigma_M \end{pmatrix}\begin{pmatrix} e_{it}^S \\ e_{it}^M \end{pmatrix}$$

where the bias of MNO count depends on x_t as well, as observed earlier.

MLE can be accomplished by the Kalman filter again. We obtain $\hat{\alpha}^2 = 0.46\sigma_S^2$ and $\hat{\beta} = 0.59$, which confirms that the MNO bias increases with x_t. Meanwhile, we have $\sigma_M^2 = 19\sigma_S^2$, such that the variance of MNO count is much larger than the survey estimator, completely different to the Canary Islands.

The bias-adjusted MNO count is $(y_{2t} - \hat{\alpha})/\hat{\beta}$. Its variance is $54\sigma_S^2$ when $\hat{\beta}$ is treated as a constant, which yields the transfer learning estimator

$$\hat{x}_t^{TL} = (\gamma, 1-\gamma)\begin{pmatrix} y_{1t} \\ (y_{2t} - \hat{\alpha})/\hat{\beta} \end{pmatrix} \quad \text{and} \quad \gamma = \frac{54\sigma_S^2}{\sigma_S^2 + 54\sigma_S^2}$$

The estimator would not have changed much, had one included the variance of $\hat{\beta}$ in that of the bias-adjusted MNO count.

The variance of this TL-estimator is $0.98\sigma_S^2$, which is almost identical to the survey estimator. The variance of the Kalman filter estimator of x_t is around $0.95\sigma_S^2$ towards the end of the time series and close to $0.87\sigma_S^2$ in the middle. In short, the high variance of the MNO count makes it difficult to improve the survey estimator in this case.

Chapter 10

Trips

The target of estimation related to various *trips* can be generically given as follows. Denote by $U = \{1, ..., N\}$ the population of individuals. Denote by $t = 1, ..., T$ the time duration of measurement. For instances, U may consist of all the residents of age 6 to 84 in a given country, and each t may refer to a given day over a calendar year. Let y_{kt} be the number of *in-scope* trips by individual k at time t, such that the target total is given as

$$Y = \sum_{k \in U} \sum_{t=1}^{T} y_{kt}$$

In particular, we have the following examples of y_{kt} in terms of scope.

- *All* trips regardless purpose or mode of transportation.
- *OD* trips from a given municipality (*origin*) to another (*destination*).
- *Local* trips, with origin and destination within the same municipality.
- *POI* trips to a given point-of-interest, such as Lofoton in Norway.

The trips y_{kt} may refer to a given *purpose* or *mode* of transport in addition.

Survey An estimator of Y based on the sample survey can be given as

$$\hat{Y} = \sum_{k \in U} \sum_{t=1}^{T} \delta_{kt} w_{kt} y_{kt}$$

where $\delta_{kt} = 1$ if y_{kt} is observed in the survey or $\delta_{kt} = 0$ otherwise, and w_{kt} is an estimation weight. For instance, the *design* weight is

$$w_{kt} = \pi_{kt}^{-1} \quad \text{given} \quad \pi_{kt} = \Pr(\delta_{kt} = 1)$$

according to the sampling design. In practice, however, w_{kt} usually differs to π_{kt}^{-1} due to adjustments for survey nonresponse.

MNO device counts It is possible to obtain trip counts of mobile devices based on their positions inferred from the contact signals. Denote by x_{dt} the count of device d at time t, where $d \in D$. In principle, the totals

$$X = \sum_{t=1}^{T} X_t \quad \text{and} \quad X_t = \sum_{d \in D} x_{dt}$$

may be available from each *given* mobile network operator (MNO). We shall not distinguish in notation whether X_t is obtained from a single MNO or summed over several MNOs, as long as it does not affect the estimation.

Note that the MNOs customarily apply some weighting adjustment when aggregating x_{dt} to X_t, chiefly due to varying market shares in different parts of a country. Ideally, such weighting adjustments should be avoided, since they are generally inappropriate and often unduly introduce extra noise to the counts; as discussed in Section 2.2 and Appendix A.

The quality of MNO counts X would also vary for different types of trips. For instance, take someone who leaves the office, stops at the supermarket to shop, and the kindergarten to pick up the children, before arriving home. It is standard that the journey counts 3 local trips in official statistics, but the MNO device trip count can vary due to a number of reasons,

In short, due to coverage, measurement and processing errors, the MNO count X can hardly be taken as official statistics directly.

Ensemble estimation In this chapter we consider the estimation of subtotals

$$\{Y_i : i = 1, ..., m\} \quad \text{where} \quad Y = \sum_{i=1}^{m} Y_i$$

We suppose that the survey total estimator $\hat{Y}$ is acceptable at the population level, but the variance of the survey estimator $\hat{Y}_i$ is too large for many subtotals due to the limited subsample sizes. Meanwhile, suppose the corresponding MNO trip counts are available, denoted by

$$X = \sum_{i=1}^{m} X_i$$

By treating the MNO counts as proxies to the target totals, we consider the SP modelling approach to combine the MNO and survey data.

10.1 Test of $X_i/X = Y_i/Y$

Patone and Zhang (2020) derive a test for the null hypothesis that the difference $X_i - Y_i$ is constant over i, given some big-data X_i that has negligible variance compared to the unbiased sample survey estimator $\hat{Y}_i$. This is an instance of audit sampling inference (Zhang, 2023), which uses sample surveys to make valid inference with respect to the sampling distribution of $\hat{Y}_i$. Below we adapt

the test for the null hypothesis relevant to the present context, i.e.

$$H_0 : X_i \propto Y_i \quad \Leftrightarrow \quad H_0 : X_i/X = Y_i/Y$$

Under H_0, we have $E(\hat{Y}X_i/X) = Y_i$ and $E(\hat{Y}_i/X_i) = E(\hat{Y}_j/X_j)$ over repeated sampling, where X_i is a constant of sampling. Let $P = I - \mathbf{1}\mathbf{1}^\top/m$, where I is the $m \times m$ identity matrix and $\mathbf{1}$ is the unity vector, such that $PP^\top = PP = P$. Denote by $[Z_i]$ the m-vector of $Z_i = \hat{Y}_i/X_i$. We have

$$E(P[Z_i]) = \mathbf{0} \quad \text{and} \quad V(P[Z_i]) = P\Sigma P$$

where Σ has diagonal elements $X_i^{-2}V(\hat{Y}_i)$. We can set the off-diagonal elements to 0 in practice, since each $\hat{Y}_i$ is based on a distinct subsample with negligible finite-population sampling fraction.

Now that $\sum_{i=1}^m Z_i \equiv 0$, let $[Z_i']$ be the $T-1$-vector on deleting an arbitrary component of $P[Z_i]$. Let Q be the correspond $(m-1) \times (m-1)$ submatrix of $P\Sigma P$, such that $[Z_i']$ has the $m-1$-variate normal distribution

$$[Z_i'] \sim N(\mathbf{0}, Q)$$

Let $LL^\top = Q$ be the Cholesky decomposition with lower-triangular L. We have

$$R = L^{-1}[Z_i'] \sim N(\mathbf{0}, I_{(m-1)\times(m-1)})$$

such that a test statistic for $H_0 : X_i \propto Y_i$ follows as

$$D = R^\top R \sim \chi^2_{T-1}$$

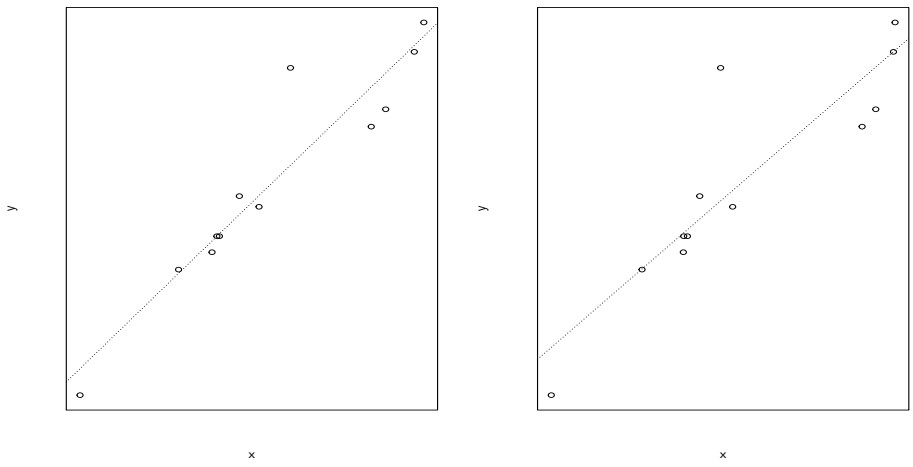

Figure 10.1: Data for test $H_0 : X_i/X = Y_i/Y$, half error variance left to right

Figure 10.1 provides an illustration using simulated data. Given a set of Y_i for $1 \leq i \leq 12$, we simulate $X_i = 1.08Y_i + e_i$ independently, where $e_i \sim N(0, (\phi Y_i)^2)$. We use $\phi = 0.05$ for the data shown to the left and $\phi = 0.07$ for the right. The regression error e_i has about half the variance to the left; the larger variance to the right is visible in the scatter plots. Given a set of stipulated standard errors of $\hat{Y}_i$, with CV values about $5\% - 6\%$, the test above yields p-value 0.609 in the left setting, whereas the p-value is 0.078 in the right setting.

10.2 Ensemble estimation

10.2.1 Linear mixed model

The model of Fay and Herriot (1979) is commonly used in small area estimation, which combines random effects and sampling errors. Treating $i = 1, ..., m$ as the 'small areas', one can let

$$\hat{Y}_i = Y_i + e_i = \beta_0 + \beta_1 X_i + v_i + e_i$$

where (β_0, β_1) are the regression coefficients, v_i is a mean-zero random effect with model variance $V(v_i) = \sigma_v^2$, and e_i the sampling error of $\hat{Y}_i$ with mean zero and sampling variance $V(\hat{Y}_i) = V(e_i)$. In particular, it is assumed that v_i and v_j are independent if $i \neq j$, e_i and e_j are independent if $i \neq j$, while v_i and e_j are independent of each other whether or not $i = j$.

Provided the variance components σ_v^2 and $V(e_i)$, the best linear unbiased predictor (BLUP) of Y_i is given as

$$\tilde{Y}_i^H = \gamma_i \hat{Y}_i + (1 - \gamma_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

where

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + V(e_i)}$$

and $(\hat{\beta}_0, \hat{\beta}_1)$ the weighted least squares estimator given $\sigma_v^2 + V(e_i)$. In practice, one would replace σ_v^2 and $V(e_i)$ by their estimates to obtain $\hat{\gamma}_i$ and the corresponding *empirical* BLUP, and constrain the final estimates of Y_i to the overall $\hat{Y}$. An induced linear-prediction (LP) estimator can be given as

$$\hat{Y}_i^{LP} = \hat{Y}\hat{p}_i^{LP} \quad \text{and} \quad \hat{p}_i^{LP} = \frac{\hat{\gamma}_i \hat{Y}_i + (1 - \hat{\gamma}_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{\sum_{j=1}^m \hat{\gamma}_j \hat{Y}_j + (1 - \hat{\gamma}_j)(\hat{\beta}_0 + \hat{\beta}_1 X_j)} \tag{10.1}$$

10.2.2 Transfer learning

Let the target proportion and the survey estimator and the MNO proxy be

$$p_i = Y_i/Y \quad \text{and} \quad \hat{p}_i = \hat{Y}_i/\hat{Y} \quad \text{and} \quad q_i = X_i/X$$

By transfer learning (TL) outlined in Section 5.2, one can combine the survey estimator and the MNO proxy to obtain the resulting TL estimator

$$\hat{p}_i^{TL} = \hat{\psi}\hat{p}_i + (1 - \hat{\psi})q_i \tag{10.2}$$

One would recover $\hat{Y}_i/\hat{Y}$ if $\hat{\psi} = 1$ or X_i/X as $\hat{\psi} \to 0$.

In the extreme case of $X_i/X = Y_i/Y$, the probability will be high for the TL estimate p_i^{TL} to be equal to X_i/X, i.e. no error at all. Meanwhile, since one still needs to estimate $(\beta_0, \beta_1) = (0, Y/X)$ for $\hat{p}_i^{LP}$ based on the m survey estimates $\hat{Y}_i$, it will be subject to the sampling errors of $\hat{Y}_i$ even in this case.

In the extreme case of uncorrelated (X_i, Y_i), the TL estimator can be reduced to the survey estimator $\hat{Y}_i/\hat{Y}$, i.e. no gain at all. Meanwhile, by virtue of the

convex combination of $\hat{Y}_i$ and $\hat{Y}/m$, the LP-estimator $\hat{p}_i^{LP}$ has a smaller MSE than $\hat{p}_i$, although $(\beta_0, \beta_1) = (Y/T, 0)$ admits nil effect of X_i.

In short, neither the LP nor the TL estimator dominates the other generally, and TL is certainly worth considering given good proxy MNO counts.

10.2.3 Bootstrap MSE estimation

Although the linear mixed model involves both the model variance $V(v_i)$ and the sampling variances $V(\hat{Y}_i)$, we propose a simple bootstrap to emulate only the sampling error, which results in a design-based MSE of any given estimator, as a common ground of uncertainty evaluation.

b1) Choose a set of plug-in values $\{Y_i^* : i = 1, ..., m\}$. Let $p_i^* = Y_i^* / \sum_{j=1}^m Y_j^*$.

b2.1) Draw $\hat{Y}_i^* \sim N\big(Y_i^*, \hat{V}(\hat{Y}_i)\big)$ independently for $i = 1, ..., m$.

b2.2) Obtain estimates $\{\hat{p}_i^*(\mu) : i = 1, ..., m\}$ by a given method signified by μ.

b3) Repeat b2.1-b2.2 to obtain $\hat{p}_{i,b}^*(\mu)$ for $b = 1, ..., B$, and the MSE estimate

$$\text{mse}_i(\mu) = \frac{1}{B} \sum_{b=1}^{B} \big(\hat{p}_{i,b}^*(\mu) - p_i^*\big)^2$$

Note that the MNO counts $\{X_i : i = 1, ..., m\}$ are held fixed as Y_i^* and p_i^* are.

The Monte Carlo error of the bootstrap estimator $\text{mse}_i(\mu)$ becomes negligible as $B \to \infty$. The sampling variance estimates $\hat{V}(\hat{Y}_i)$ are available from the survey. Regarding $\{(Y_i^*, p_i^*) : i = 1, ..., m\}$, we suggest to

- let (Y_i^*, p_i^*) be given by the best available method;
- use other plausible estimates for sensitivity analysis.

Using the design-based MSE as the common criterion allows one to discern $\hat{p}_i^{LP}$ and $\hat{p}_i^{TL}$ regardless the difference between their underlying assumptions of the source(s) of uncertainty. The additional sensitivity analysis results should hopefully agree with one's conclusion about the *relative* merits of $\hat{p}_i^{LP}$ and $\hat{p}_i^{TL}$, although the reported MSE estimates would surely have been different using a different set of plug-in values (Y_i^*, p_i^*) for the bootstrap.

10.2.4 Robust estimator

One needs to estimate $\sigma_v^2 = V(v_i)$ for the LP approach. Negative estimates arise

$$\text{if} \quad \sum_i \hat{V}(\hat{Y}_i) > \sum_i (\hat{Y}_i - \hat{\mu}_i)^2 \quad \text{given} \quad \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

The EBLUP is then simply the synthetic estimator $\hat{\mu}_i$. Similarly for $\tau_u = \sum_i u_i^2/m$ in transfer learning, if

$$\sum_i \hat{V}(\hat{p}_i) > \sum_i (\hat{p}_i - q_i)^2$$

The TL estimator is then simply the MNO proxy q_i. However, as long as $\hat{\mu}_i \neq q_i$, for $i = 1, ..., m$, it would have been imprudent to proceed as if either of them had completely negligible bias.

We notice that the situation occurs because the survey sampling variances are relatively large compared to the errors of $\hat{\mu}_i$ or q_i. In other words, $\hat{\mu}_i$ and q_i can still be quite good, because the sampling variances depend on the relevant sample sizes, whereas the errors of μ_i and q_i depend on how good the model or the MNO proxy is. The problem is that one cannot accept both $\hat{\mu}_i$ and q_i if they are not equal to each other.

It seems sensible in this situation to let a robust estimator of p_i be a convex combination of $\hat{p}_i^{LP}$ derived from $\hat{\mu}_i$ and $\hat{p}_i^{TL} = q_i$. Now that

$$\begin{aligned} E\Big((\hat{p}_i - \mu_i/Y)^2\Big) &= E\Big((\hat{p}_i - p_i)^2 - 2(\hat{p}_i - p_i)(\mu_i - Y_i)/Y + (\mu_i/Y - p_i)^2\Big) \\ &= V(\hat{p}_i) + (\mu_i/Y - p_i)^2 \\ E\Big((\hat{p}_i - q_i)^2\Big) &= E\Big((\hat{p}_i - p_i)^2 - 2(\hat{p}_i - p_i)(p_i - p_i) + (q_i - p_i)^2\Big) \\ &= V(\hat{p}_i) + (q_i - p_i)^2 \end{aligned}$$

with respect to sampling, a reasonable choice may be given as

$$\hat{p}_i^{MX} = w\hat{\mu}_i + (1 - w)q_i \tag{10.3}$$

where

$$w = \frac{\sum_i (\hat{p}_i - q_i)^2}{\sum_i (\hat{p}_i - \mu_i/Y)^2 + \sum_i (\hat{p}_i - q_i)^2}$$

For instance, $w = 0.5$ in the hypothetical case of $\mu_i/Y \equiv q_i$, as it should be.

In practice, estimates of (μ_i, Y) are needed to calculate w, for which one may as well use the OLS fit of (β_0, β_1) to compute $\hat{\mu}_i$ for extra robustness. One can still use the WLS $(\hat{\beta}_0, \hat{\beta}_1)$ for the synthetic estimator $\hat{\mu}_i$ in (10.3). Finally, one should use $\hat{p}^{MX}$ as the plug-in values for bootstrap MSE estimation.

10.3 Illustration

Let us use simulation to illustrate the methods above. Let $m = 12$, as if one were estimating the monthly total trips for the yearly trip statistics. As before, let Y_i be simulated, given which let $\hat{Y}_i \sim N(Y_i, (c_i Y_i)^2)$, where c_i is the simulated CV of the survey estimator $\hat{Y}_i$, and let $X_i - 1.08Y_i \sim N(0, (\phi Y_i)^2)$ given any chosen ϕ. The results below are obtained in two settings.

I) Let c_i vary around 0.05 and let $\phi = 0.1$.

II) Let c_i vary between 0.075 and 0.095 and let $\phi = 0.1$.

The various estimates are given in Figure 10.2 for both the settings.

In the first setting, we obtain the following key estimates:

$$(\hat{\sigma}_v^2, \hat{\tau}_u) \propto (3.277, 7.570) \quad \text{and} \quad (\hat{\bar{\gamma}}, \hat{\psi}) = (0.617, 0.800)$$

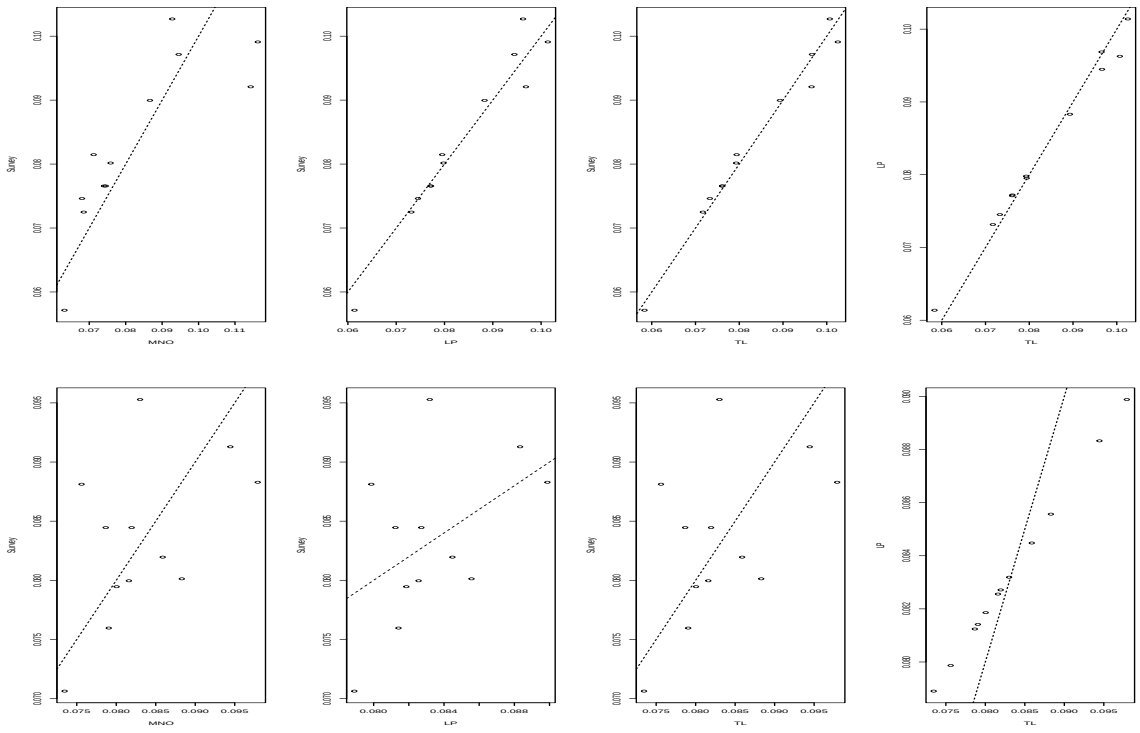

Figure 10.2: Illustration of LP, TL, MX estimators. Top, setting-I with positive estimates of (σ_v^2, τ_u); bottom, setting-II with negative estimates of (σ_v^2, τ_u)

where $\hat{\bar{\gamma}} = \sum_{i=1}^m \hat{\gamma}_i/m$ is the average shrinkage factor for the EBLUP under the linear mixed model. The TL estimator assigns on average a larger weight to the survey estimator $\hat{p}_i$ than the LP estimator, now that $\hat{\psi}$ is greater than $\hat{\bar{\gamma}}$. One can somehow see this in the top row of Figure 10.2, where the survey estimates are more tightly gathered around the TL estimates than the LP estimates, although the two estimates are highly correlated with each other as can be seen in the last plot in the top row. Moreover, both the estimates have a 'smoothing' effect on the differences between the survey estimates and MNO proxy. The robust estimator $\hat{p}_i^{MX}$ assigns a weight about $1/3$ to $\hat{p}_i^{LP}$ and $2/3$ to $\hat{p}_i^{TL}$. Let the average relative root MSE (ARRMSE) of a given estimator be

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\sqrt{\text{MSE}_i}}{p_i^*}$$

We obtain the bootstrap estimates

$$\text{arrmse}(\text{LP}, \text{TL}, \text{MX}) = (0.023,\ 0.017,\ 0.020)$$

according to which $\hat{p}_i^{TL}$ is somewhat more efficient than $\hat{p}_i^{LP}$ here.

In the second setting, we obtain the following key estimates:

$$(\hat{\sigma}_v^2, \hat{\tau}_u) = (0, 0) = (\hat{\bar{\gamma}}, \hat{\psi})$$

where the estimates of $(\hat{\sigma}_v^2, \hat{\tau}_u)$ are initially negative and thus truncated to 0. Consequently, the LP estimator of Y_i is the synthetic $\hat{\mu}_i = \hat{\beta}_0 + X_i\hat{\beta}_1$ and the TL estimator of p_i is simply q_i. As can be seen in the bottom row of Figure 10.2, the

LP estimates shrink more towards the average of p_i compared the TL estimates. The last plot there shows that the two estimates are almost perfectly correlated with each other, the one being essentially a linear transformation of the other. The robust estimator $\hat{p}_i^{MX}$ assigns a weight about 0.6 to $\hat{p}_i^{LP}$ and 0.4 to $\hat{p}_i^{TL}$ in this case. The bootstrap ARRMSE estimates are

$$\text{arrmse}(\text{LP}, \text{TL}, \text{MX}) = (0.0001,\ 0.0169,\ 0.0070)$$

according to which $\hat{p}_i^{LP}$ seems 'super-efficient' compared to $\hat{p}_i^{TL}$. The result is partly because the plug-in p_i^* is closer to $\hat{p}_i^{LP}$ than to $\hat{p}_i^{TL}$. More importantly, however, the estimator of σ_v^2 has a much higher probability of being negative during bootstrap than that of τ_u^2, such that the TL estimator varies much more during bootstrap. It would therefore be highly sensitive if one were to conclude that $\hat{p}_i^{LP}$ is as efficient as the bootstrap suggests. A more robust choice is to adopt the MX estimator in such a situation.

10.4 Application to SRVU

The Swedish Resvanor Undersökningen (SRVU) is conducted by Trafikanalys, which provides the relevant official statistics in Sweden. The target population consists of all the residents aged 6-84 in Sweden. Through cooperation with MNO-MINDS, the methods described and illustrated above have been applied to estimate the total number of local trips within each of the 21 Swedish counties, as well as the total number of trips in each calendar month. The disseminated statistics are available at

`https://www.trafa.se/transportmonster/RVU-Sverige/`

Due to data confidentiality reasons, however, we cannot disclose any details of the data or the relevant modelling and estimation results here.

Chapter 11

Commuters

In this chapter we consider an application scenario of combining MNO data with the relevant non-MNO data available at the Italian National Statistical Institute (Istat) to produce commuting statistics.

Traditionally, decennial population censuses have been the basis for official statistics of commuter flows across municipalities. However, for more timely information, Istat has shifted towards annual census surveys with reduced sample sizes. This transition presents various challenges in deriving accurate commuting statistics.

Availability of MNO data allows for the derivation of recurrent flows between home and work/study locations. These flows can potentially serve as the primary source for producing official statistics by adjusting MNO data coverage in a quasi randomisation (QR) approach, as discussed in Chapter 7.

On the other hand, in a superpopulation framework, aggregate MNO data can serve as covariates, complementing target variables obtained from census sample surveys. Spatial interaction models are widely used in transportation planning and urban studies to describe the movement of people or goods across different places. These models describe the relationship between the origin and destination of flows within a geographic area accounting for factors such as distance, attractiveness, and connectivity of the different locations, also including spatial auto-regressive components. To reduce bias that may arise due to miss-specification of the previous regression models, small area models can be applied, as discussed in Chapter 7.

In addition, a transfer learning approach can be useful to combine sample survey with MNO data, where MNO counts can serve as proxy, rather than covariates, of the target variables obtained from the survey. Transfer learning for ensemble estimation, as discussed in Section 5.2, may be considered for both in-sample and out-of-sample domains.

All the approaches mentioned above are applied here to estimate the number of commuters between the municipalities of an Italian region.

11.1 MNO data

The MNO data exploited in this work refer to the so-called Call Detail Records (CDRs), collected by a local operator in the Tuscany region during 6 weeks,

from January 2017 to February 2017. The CDR data usually include calls and text messages. The data available to this study are composed as follows: the caller ID, that is a numeric code associated to each SIM (Subscriber Identity Module) by an algorithm that guarantees anonymity; the antenna where the call originated; the date and time at which the call is originated; the duration of the call; the antenna where the call ended. For text messages, data report the date and time of the text message and the antenna from which the text message was sent. The CDRs are processed so that anonymity is ensured. During the 6 weeks of observation, the number of CDRs is just fewer than two hundred millions, divided into: more than one hundred millions of Calls and seventy millions of SMS.

The large number of observations limits missing data due to random device inactiveness or inability to transmit data due to network glitches. In addition, the anonymised SIM-level data are observed almost continuously over the time for 6 weeks, allowing us to estimate some meaningful locations, such as "home" and "work/study". In this study a very simple approach has been followed:

- "home" is the municipality where a device is more frequently located during the nighttime, defined from 8 pm to 7 am;
- "work/study" is the municipality where a device is repeatedly observed during the daytime, defined from 7 am to 8 pm.

By aggregating device data for which home and work/study have been derived, it is possible to produce home and work/study origin-destination (OD) flows, at municipal level, where only movements within the region are taken into account. One drawback of this data is that it is not possible to detect the reason of mobility, although they can indicate the frequencies of mobility.

11.2 Methods

11.2.1 Superpopulation models

Let Y denote an $m \times m$ square matrix of OD flows from each of the m origin zones to each of the m destination zones as shown in Equation (11.1). The elements on the main diagonal of the matrix represent intra-zonal flows.

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mm} \end{bmatrix} \tag{11.1}$$

Considering only the off-diagonal counts, we can produce a $M \times 1$ stacked vector $y = vec(Y)$ of these flows by origin-centric ordering with $M = m(m-1)$, where the first $m-1$ elements reflect flows from origin zone 1 ($i = 1$) to all $m-1$ destinations, and the last $m-1$ elements reflect flows from origin zone m ($i = m$) to destinations $1, \ldots, m-1$.

The classical interaction model (LeSage and Pace, 2008) for y is:

$$\mathbf{y} = \eta + \mathbf{H}\boldsymbol{\psi} + \mathbf{G}\boldsymbol{\phi} + \mathbf{X}\xi + \boldsymbol{\epsilon} \tag{11.2}$$

where $\mathbf{H} = \mathbf{H^o} \otimes I_m$ is an (M, Q) matrix of Q origin-specific variables that characterise the ability of the origin zones to produce flows, $\mathbf{G} = I_m \otimes \mathbf{G^d}$ is an (M, R) matrix of R destination-specific variables that represent the attractiveness of the destination zones, $\mathbf{X}$ is an (M, S) matrix of S factors influencing the flows, such as distance, transportation infrastructure, information between origin and destination zones including flows from MNO data, $\boldsymbol{\epsilon}$ is an $(M, 1)$ vector of errors with $\boldsymbol{\epsilon} \sim N(0, \sigma^2_\epsilon I_M)$.

A synthetic predictor of the flow y_{ij} obtained by model (11.2) depends on all the other observations via the estimated regression coefficients. To alleviate the bias due to model misspecification, one may apply small area estimators.

We assume that some direct estimators of the commuting flow from the origin i to the destination j are available, denoted by, $\hat{Y}_{\text{Dir},ij}$.

Suppose only M_s out of the M flows are observed in the sample. Let us denote by k the index for the couple (i, j) and e_k the sampling errors. Writing the covariates in the model (11.2) as a single vector, denoted by x, which depends both on origin and destination such as the MNO flows or the previous census flows, we obtain

$$\begin{aligned} \hat{Y}_{\text{Dir},k} &= y_k + e_k, \quad k = 1, \ldots, M_s \\ y_k &= \eta + \xi^\top x_k + u_k \end{aligned}$$

where x_k is the vector of covariates related to the flow, including distances (or frictions), u_k is a random effect, and e_k is the sampling error. An EBLUP estimator can be derived under this model.

A simpler estimation problem is the number of commuters originated from each municipalities $y_i = \sum_{j \neq i} y_{ij}$, when m_{os} out of m origin areas are in the sample. The small area model can then be formulated as

$$\begin{aligned} \hat{Y}_{\text{Dir},i} &= y_i + e_i, \quad i = 1, \ldots, m_{os} \\ y_i &= \eta + \beta^\top x_i + u_i \end{aligned}$$

In case of many empty sample domains, denoted by $n_i = 0$, the EBLUP would reduce to the synthetic estimator $x_i^\top \hat{\beta}$. As outlined in Chapter 5, an alternative is transfer learning given source Y_i^*, a proxy to Y_i, which may be available from the last census or other concurrent sources for all the origin areas.

This two-step approach preserves the EBLUP for the domains with $n_i > 0$ and transfer learning is only applied to the rest domains with $n_i = 0$ as an alternative to synthetic estimation.

11.2.2 Transfer learning for the in-sample domains

As in Section 5.2, we consider the sample survey yields a precise estimate for the total commuters $y = \sum_{i \in m_{os}} y_i$, but the subtotals of commuters y_i originated from the municipality i, or equivalently the sample proportion $p_i = y_i/y$ can be

improved through the known proxy breakdown q_i provided by the MNO data. In this case, transfer learning is limited to the m_{os} in-sample municipalities. The transfer learning estimator $\hat{p}_i^{TL}$ is obtained by (5.3), and the tuning parameter ψ as given in Section 5.2.1.

11.2.3 Quasi Randomisation approach

Suppose that among all the detected flows by MNO, those due to commuting by work and study m_{ij} can be identified and $m_i = \sum_j m_{ij}$. Estimates of the OD matrix or just the out-flows can then be obtained by adjusting the counts by their coverage in a QR approach, assuming non-informative selection conditionally given some socio-economic characteristics z.

We consider two QR adjustments: in the first one the pseudo probabilities are based only on the direct estimates and the MNO counts, which is referred to as an *internal* adjustment, while in the second one we rely on information derived from *external* sources, as the deduplication factor coming from an ad-hoc survey, or the proportion of subscribers of the MNO in a given group z, which is provided by the MNO itself.

Internal QR adjustment: direct estimator disaggregation

Given an accurate direct estimate of the total number of commuters at a given level l, e.g. aggregate of municipalities such as region or province, an estimate of the pseudo probability of being included in the MNO sample is

$$m_{l|z}/\hat{Y}_{l|z} \tag{11.3}$$

where $\hat{Y}_{l|z}$ are the direct estimates of commuters originating from area (province or region) l in the group z and $m_{l|z}$ the corresponding MNO count. Then a QR estimate of ij for group z is given by

$$\hat{Y}_{ij|z}^{INT} = \hat{Y}_{l|z}\frac{m_{ij|z}}{m_{l|z}} \tag{11.4}$$

and similarly

$$\hat{Y}_{i|z}^{INT} = \hat{Y}_{l|z}\frac{m_{i|z}}{m_{l|z}}. \tag{11.5}$$

The target estimate follows as

$$\hat{Y}_{ij}^{INT} = \sum_z \hat{Y}_{ij|z}^{INT} \tag{11.6}$$

or

$$\hat{Y}_{i}^{INT} = \sum_z \hat{Y}_{i|z}^{INT} \tag{11.7}$$

Unfortunately, in this study the MNO data lack socio-demo information on the users, e.g. gender or age. We therefore let z depend only on geography.

External QR adjustment: the role of the survey

Alternatively, coverage adjusting factor(s) can be estimated through a survey where the usage data of mobile phone are collected, subject to the condition that the MNOs are able to provide their aggregates broken down by the same groups considered by the adjusting factors.

The previous adjustment (11.4) breaks down the direct estimates of commuters $\hat{Y}_{i+|z}$ by the MNO ratios $m_{ii|z}/m_{i+|z}$, here we use the generic adjusting factor(s) to derive MNO-based target estimates from external source, that is a survey: for OD estimates we have

$$\hat{Y}_{ij|z}^{EXT} = \frac{m_{ij|z}}{\tau \nu_z \theta} \tag{11.8}$$

and for the number of commuters we have

$$\hat{Y}_{i|z}^{EXT} = \frac{m_{i|z}}{\tau \nu_z \theta} \tag{11.9}$$

where τ is the MNO propensity usage of the target population and it can be obtained from a survey, ν_z is the proportion of subscribers of the specific MNO in area z eventually provided by the MNO itself, and θ is the deduplication adjustment obtained from a survey.

In our application scenario, the proportion of subscribers given by the MNO referred to the general population and for this reason the actually applied formulas are:

$$\hat{Y}_{ij|z}^{EXT} = \frac{m_{ij|z}}{\nu_z^* \theta} \tag{11.10}$$

and for the number of commuters we have

$$\hat{Y}_{i|z}^{EXT} = \frac{m_{i|z}}{\nu_z^* \theta} \tag{11.11}$$

assuming ν_z^* is the proportion of subscribers among the commuters instead of the general population.

In Italy, the survey on Aspect of Daily Life is an annual sample survey which collects data on coverage and use of mobile devices (Chapter 15). The survey also collects basic demographics for all the individuals in the household, as well as data related to commuting for studying or working, working condition, income. The survey sample has not been planned to provide reliable estimates for the variables mentioned above more detailed than at the national level, hence for the purpose of this study we compute the adjustment for coverage and duplication at national level. As already said, the available MNO data do not contain any information related to the device owner, which would be ideal for computing correction factors for homogeneous groups.

11.3 Some Preliminary Results

The results shown and discussed in this section are preliminary due to the reasons described above. They serve to illustrate the aforementioned methods,

but not the quality of the estimates.

To estimate the number of commuters of the municipalities of an Italian region we have applied regression models, small area methods and the quasi randomisation approach on the following real data.

- Direct flow estimates from the Italian Population permanent census survey in 2021 and 2018. Direct estimates are used both in the super-population modeling and in the adjustment factors of the QR approach in (11.4);
- Call Details Records generated in 6 weeks (January 2017 - February 2017) from an Italian mobile operator, used as auxiliary variables in the SP models and as primary source of information in the QR approach;
- The previous census counts 2011 have been used as auxiliary in regression models as well as proxy Y^* for transfer learning.
- The survey on Aspect of Daily Life has been used as alternative source to obtain the adjustment factors in the QR approach.

Table 11.1: Summary statistics of CV (%) of direct estimates

Year	Min	Q1	Median	Mean	Q3	Max
2021	2.180	5.092	6.828	7.972	9.532	20.650
2018	2.222	5.132	6.791	8.739	10.023	47.113

Table 11.1 reports some summary statistics on the Coefficients of Variation (CVs) of direct estimates in 2021 and 2018. The municipality direct estimates of outbound commuters are reliable both in 2021 and 2018, however about 100 (or 150) municipalities are not in the sample in 2021 (or 2018) for which alternatives to direct estimates are needed.

Figures 11.1 show the relationships of the direct flow estimates and the covariates by census, MNO or administrative data. It is clear the available MNO covariates are relevant though they may not be the best. Note that whereas in 2021 the administrative data refer to the same reference year as the survey, for 2018 there is a three year mismatch. The opposite condition applies for MNO where the data reference time is closer to 2018 than 2021.

11.3.1 Results for SP modelling

Given that the outbound commuting flows range from few dozens to tens of thousands, all regression models have been applied on the log scale with bias-adjusted back-transformation for model-based prediction. Figure 11.2 shows the direct flow estimates against the Fay-Herriot (FH) model estimates with MNO and Admin covariates, both for 2021 and 2018. The plots are limited to direct estimates less than 5000 to highlight the pattern of the relationship.

Table 11.2 reports some summary statistics of the shrinkage values γ of the estimated FH models, which is the weight given to the direct vs the synthetic estimator in the EBLUP estimator, using different sets of covariates.

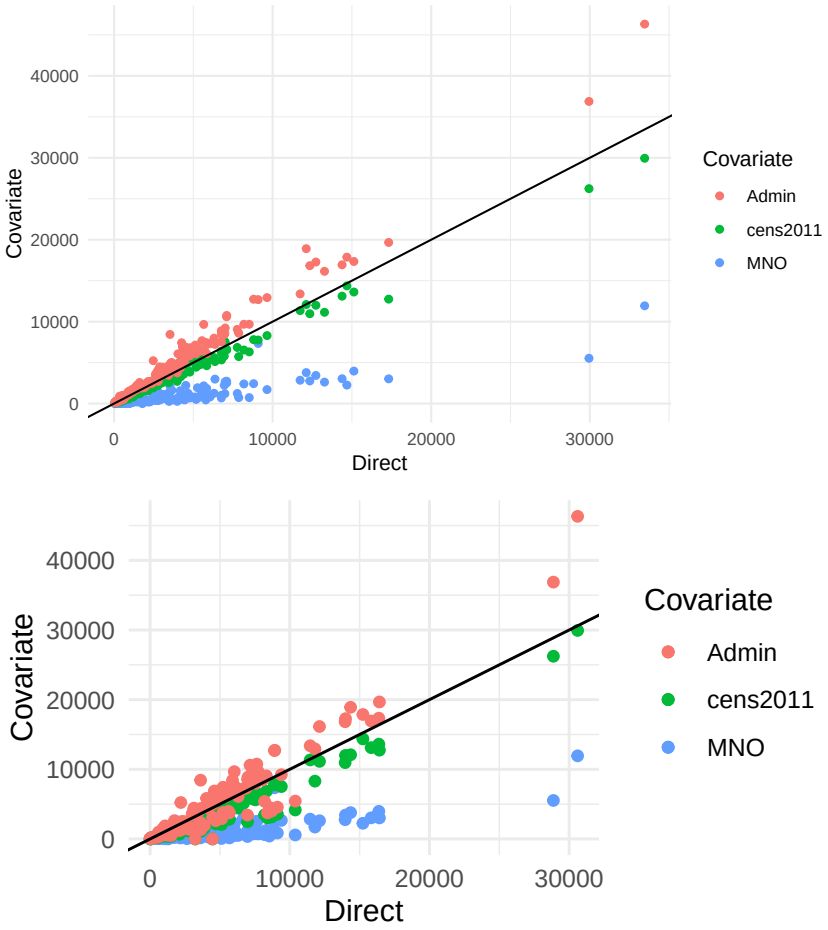

Figure 11.1: Direct flow estimates vs covariates derived from administrative data, Census 2011 and MNO data, year 2021 (top), year 2018 (bottom)

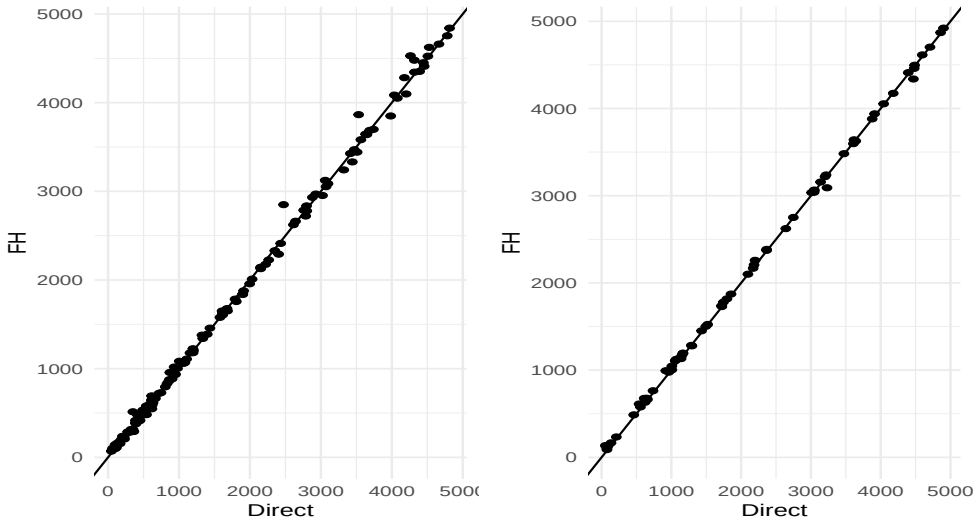

Figure 11.2: Direct and FH model estimates with MNO and Admin covariates, year 2021 on the left, year 2018 on the right

Table 11.2: Summary statistics of γ values of the FH models

Covariates	Min	Q1	Median	Mean	Q3	Max
2021						
MNO + Admin	0.35	0.71	0.83	0.78	0.90	0.98
MNO + Census11 + Admin	0.25	0.62	0.76	0.71	0.85	0.97
2018						
MNO + Admin	0.69	0.98	0.99	0.98	0.99	0.99
MNO + Census11 + Admin	0.63	0.97	0.99	0.97	0.99	0.99

Table 11.3: Relative root MSE (%) for year 2021 and 2018

2021	Min	Q1	Median	Mean	Q3	Max
MNO+Admin+GU	1.83	4.81	6.38	6.86	8.01	17.13
Census11+MNO+Admin+GU	1.98	4.60	6.17	6.46	7.62	13.64
2018						
Census11+MNO	2.32	4.89	6.18	6.59	7.98	14.89

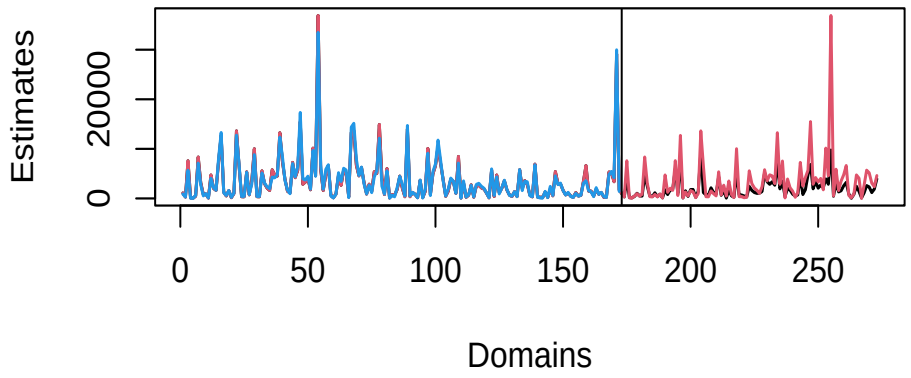

Figure 11.3: Direct Estimates, FH with covariates MNO+Admin, TL with proxy census2011, survey year 2021

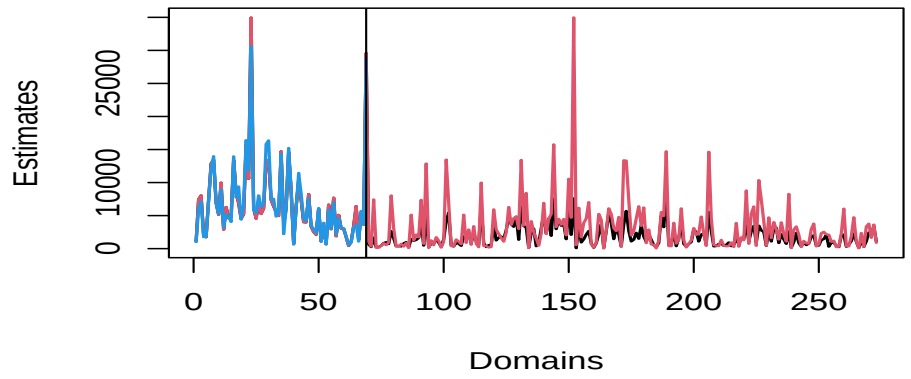

Figure 11.4: Direct Estimates, FH covariate MNO, TL with proxy census2011, survey year 2018

Table 11.3 reports some summary statistics of the relative root MSE of the EBLUP by the FH models, with best choices of covariates. They confirm the results already shown in Table 11.1, i.e. the direct estimates perform well in terms of variance and the EBLUP does not provide much improvement. Still, the direct estimates are unavailable for a large number of municipalities, for which alternative estimators are needed and the synthetic component of the EBLUP can be applied to them. It is worth noting that combining MNO data from 2018 with Census 2011 is more effective than only using one of them.

Regarding transfer learning for those out-of-sample municipalities, the tuning parameter α is 0.46 for 2021 and 0.56 for 2018. Figures 11.3 and 11.4 depicts the direct estimates, the FH model estimates and the transfer learning estimates for the in-sampled and out-of-sample municipalities. The transfer learning estimates seem more reasonable for the out-of-sample municipalities in two respects: the estimated out-of-sample outbound flows are smaller on average than the in-sample outbound flows, and there are no implausible outlying estimates compared to the EBLUP (i.e. synthetic) estimates.

11.3.2 Results for TL estimation

Figure 11.5 compares the TL ensemble estimates with the MNO proportion for the in-sample municipalities, in years 2021 and 2018. On both the occasions, the TL estimates are very close to the direct survey estimates, the estimated tuning parameter being $\hat{\psi} = 0.991$ and $\hat{\psi} = 0.993$ in 2021 and 2018, respectively. This shows that the MNO proportion here is not a good proxy of the target proportion, as already confirmed by the results in the previous subsection, where indeed the MNO counts alone are not the best covariates for the small area model, rather they contribute to improve the precision of the estimates when added as covariates along with others, especially administrative data and previous census data.

The weak relation between MNO q_i and survey $\hat{p}_i$, and the larger variability of $q_i - \hat{p}_i$ compared to the variability of the direct estimates $\hat{p}_i$ is confirmed also in the scatter plots in Figure 11.6, where the MNO proportions q_i are plotted against the sample proportions $\hat{p}_i$, and the TL estimates $\hat{p}_i^{TL}$ against $\hat{p}_i$ as well.

11.3.3 Results for QR estimation

Figure 11.7 compares the four different QR estimates with the direct estimates, for the in-sample municipalities in year 2021, depending on

- the 10 provinces (NUTS3) of the Tuscan region as l aggregation level, labelled with *prov*, vs. the whole region as a single level l, labelled with *reg*;
- adjustment factor by MNO data and direct estimates, labelled with *int*, vs. adjustment factors by ad-hoc survey data, labelled with *ext*.

It is worth noting that estimates based on adjustment at regional level *reg* are very similar using internal or external sources for the adjustment. This suggests that the operator, when providing the proportion of subscribers in the area at the regional level, is indeed estimating something very similar to

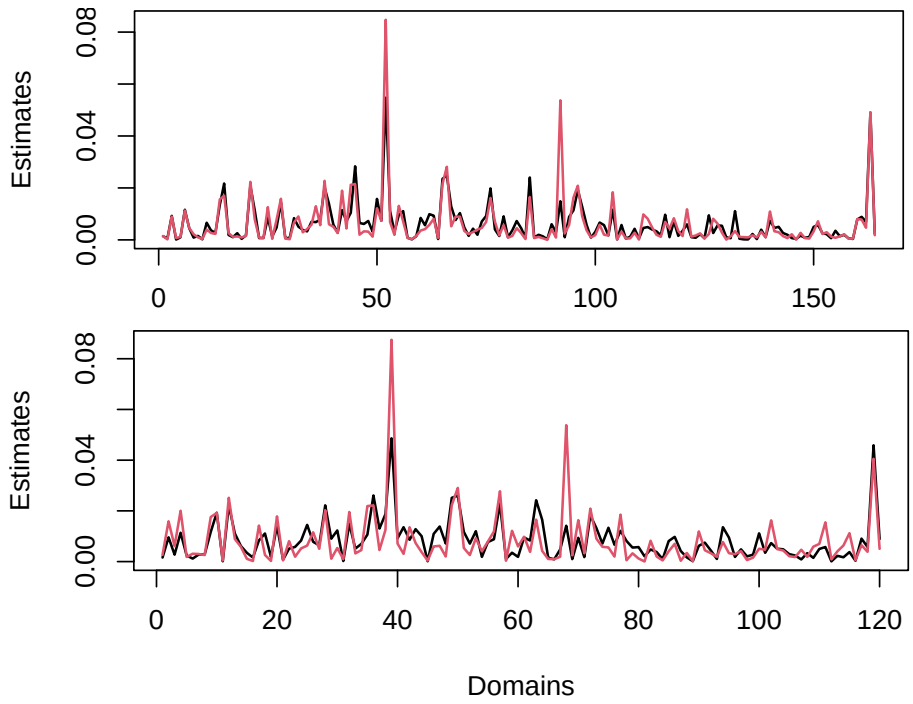

Figure 11.5: TL estimates (black) and MNO proportions (red) in year 2021 (top) and year 2018 (bottom)

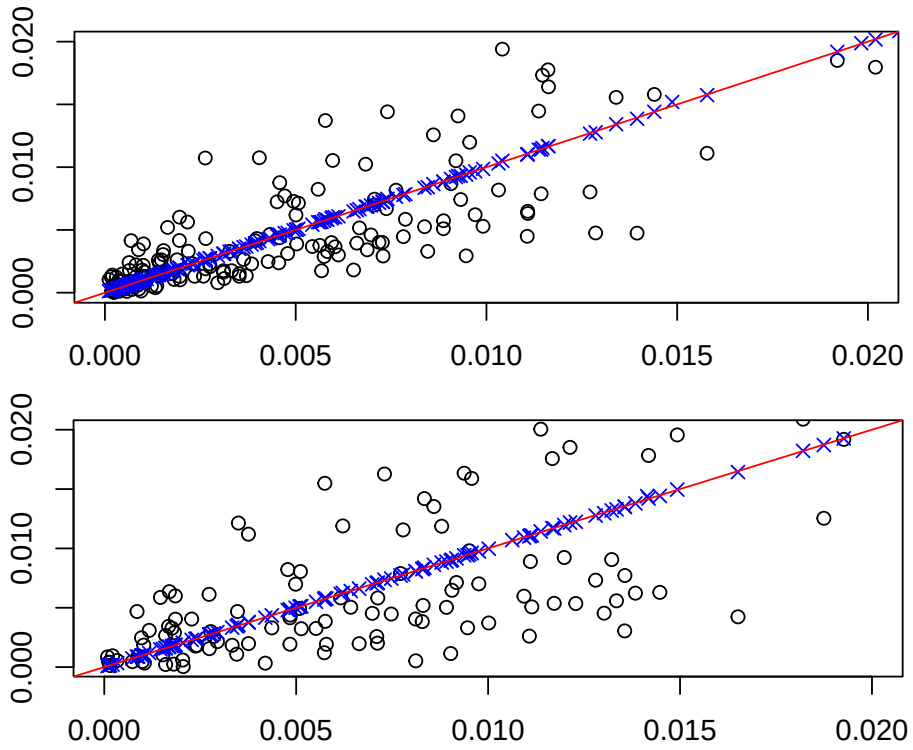

Figure 11.6: MNO proportion q_i vs. sample $\hat{p}_i$ (circles), TL estimates $\hat{p}_i^{TL}$ vs. sample $\hat{p}_i$ (crosses), year 2021 (top), year 2018 (bottom)

$m_l/\hat{Y}_l$. From conversation with the MNO we know indeed that they estimate the proportion of subscribers in the area at the regional level with n_l/N_l where n_l is the subscribers resident in the area l and the N_l is the total resident in

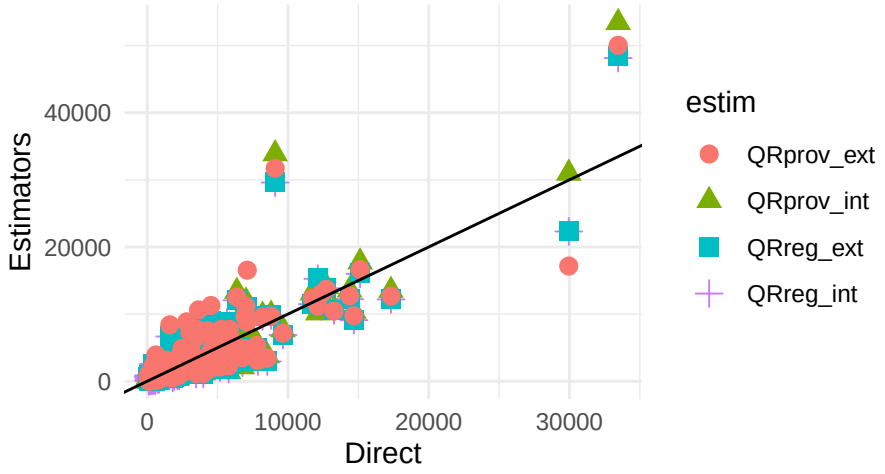

Figure 11.7: Comparison of QR estimates and direct estimates, 2021 survey.

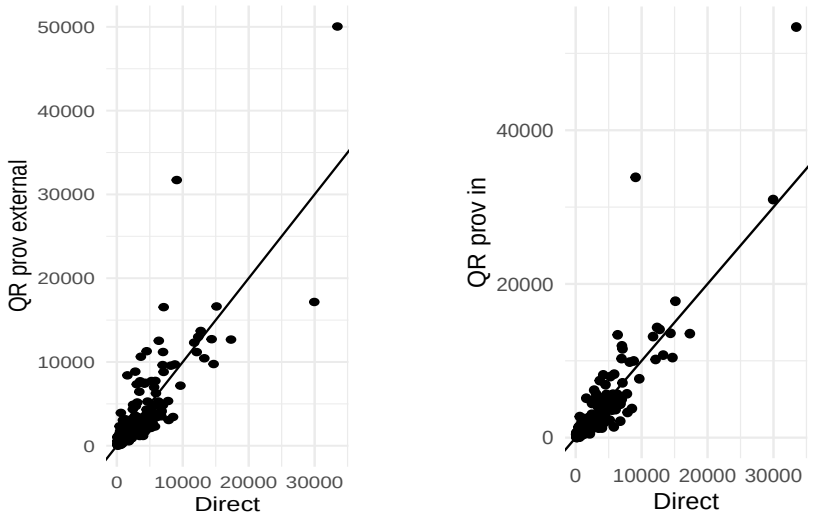

Figure 11.8: Comparison of QR estimates and direct estimates, 2021 survey, focus on external and internal QR adjustments at province level.

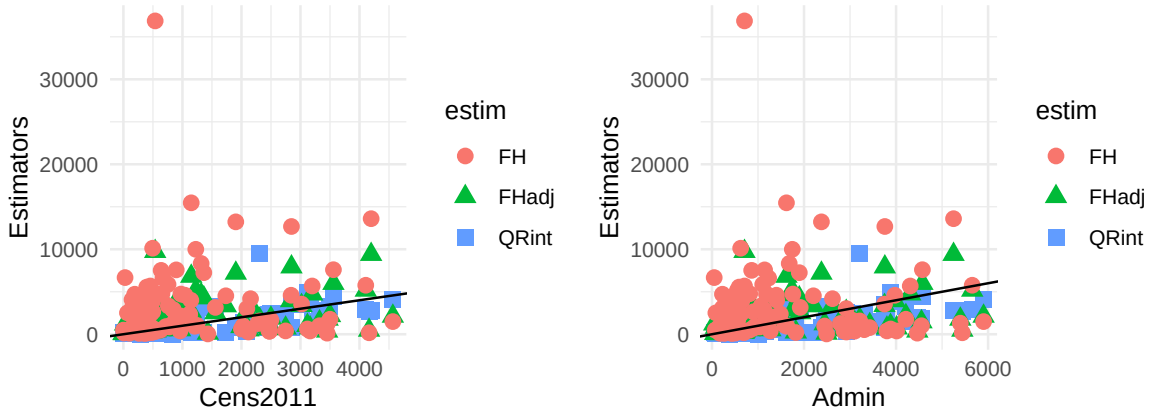

Figure 11.9: Comparison of QR and FH estimates with Census 2011 (left) and administrative data (right), year 2021 survey.

the region l according to the figures from the national statistical office. On the other side, the estimates based on adjustments at province level *prov* show larger differences, with the *ext* providing higher values for the highest commuting counts. The relationship with the direct estimates is quite good for all the estimates, including the *ext* that indeed do not use this information in the QR adjustment, despite the relationship is stronger between direct estimates and *int* QR estimates, as expected and shown in Figure 11.8.

For the out-of-sample municipalities, we can compare the QR estimates with the administrative data and the previous census data. Figure 11.9 shows the comparison between QR estimates based on province level adjustment with internal information (MNO data and census survey), Fay-Herriot estimates and adjusted Fay-Herriot estimates with transfer learning, with Census 2011 counts and administrative data, respectively. The Figure clearly shows how the FH estimates in the out-of-sample municipalities are mitigated by the TL adjustment. Both the TL-adjusted FH model and the QR estimates do not show such highest values as the basic FH model estimates do.

11.4 Concluding remarks and future works

This study has applied the methods outlined in Part II to produce preliminary results of commuter statistics by combining MNO data with other sources, namely survey, census and administrative data. Despite the available MNO data are far from what we hope to be able to exploit in the next future, it has demonstrated the potentials of MNO data for commuting statistics.

For the SP modelling approach, one would expect the MNO flow counts to be the most useful covariates, due to their high population coverage and limited measurement errors for inter-municipalities mobility, i.e. compared to past census or concurrent administrative data on home and work/study locations. However, we have only CDR data form one MNO over 6 weeks, whose coverage and measurement errors are naturally far greater than what would be admissible for real applications. Although the MNO data are outdated for both the years 2018 and 2021, for which we have other data sources, there is clear evidence that the MNO data lead to better results for 2018 than 2021. This may be taken as an indication of the potential value of MNO data, provided they can be processed and available according to the OSA's requirement.

In addition, in this experiment we could rely on the 2011 census data that still relate well to the recent surveys. But one cannot expect this to last forever, whereas one can expect to access more updated MNO data in future.

Regarding the quasi-randomisation approach, likely the adjustment that has been applied is not truly sensible, since the available MNO data do not contain any demographic characteristics on the mobile users, which vary more closely with the target population coverage by users and heterogeneous mobile device usage. See Chapters 14 and 15 for relevant details.

Several topics can be mentioned for future investigation, aided with more appropriate MNO data. First, since plausible estimates may be possible by different assumptions, without any of them clearly outperforming the others, we would like to combine the different estimates into a robust estimator, as

outlined in Chapter 7. Moreover, spatial models with or without introducing an autoregressive component may be able to improve the SP models considered here. Finally, this initial work forms a basis for the estimation of the complete OD matrix among municipalities, which is the more attractive output of the current official statistics production.

Chapter 12

Nights-spent

In many problems of forecasting, the availability of data is asynchronous, where some observations arrive earlier than others due to administrative and logistic factors. At the same time, highly informative proxy variables may be available, providing valuable signals for the missing observation. This motivates the use of forecasting techniques to produce flash estimates.

To explore this, we evaluate two methods outlined in Chapter 5, *Quasi-Transfer Learning* and *Augmented Learning*, which aim to refine predictions by exploiting structural patterns in the available data. Specifically, we assess their effectiveness in forecasting the number of nights-spent by tourists in a given municipality, using registry and MNO proxies. The approaches will be compared to some traditional forecasting methods.

12.1 Data

In the field of tourism statistics, a crucial indicator is the total number of nights spent by tourists. In Italy this indicator is provided for each municipality (local administrative unit) on a monthly basis.

The data for the total number of nights spent are provided directly by the touristic accommodations, which daily report figures on their guests to local offices, who then provide the data to Istat, typically within two months (t+2) from the reference month (t). However, a certain percentage of municipalities actually provide this information much quicker, sometimes even within the first ten days after the reference month. It is worth noting that the set of these fast response units may vary each month due to several contingency factors.

Despite this well-organised and quite quick process for the production of official indicators on the nights spent by tourist at a detailed level, the demand for faster estimates is always high, especially during some seasons, in specific regions, or in connection with some special events. The objective of our study is to investigate whether, by leveraging registry data from responding municipalities and the number of overnight stays derived from MNO data, it is possible to provide a flash estimate for municipalities that do not respond promptly. Only overnight stays are considered in this analysis, without breaking down by domestic and inbound tourism.

For this study monthly time series for official overnight stays at municipality

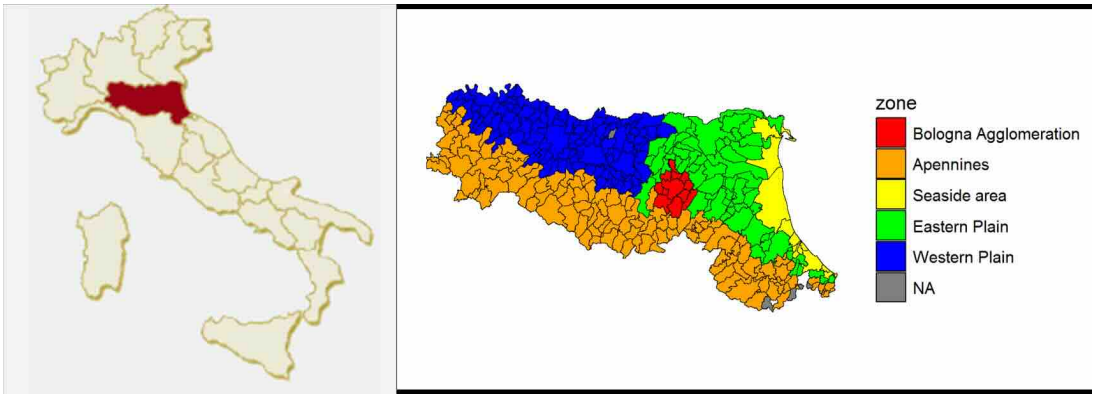

Figure 12.1: Emilia Romagna region with geographic classification zones

level are available from January 2022 to October 2023, while MNO data cover the period from August 2022 to October 2023. The data pertain to the Emilia Romagna which comprises 330 municipalities, grouped in 9 provinces, four of which are not included in our dataset due to administrative changes.

The MNO data have been elaborated according to multi-MNO algorithms, subject to some cosmetic adjustment to 'reproduce' the Official statistics. The MNO data come from a single operator, and the number of nights is counted based on a simple shared algorithm. Among others, the algorithm includes a rescaling process to account for other operators and factors beyond our control, as well as masking for values at risk of re-identification.

In addition to the MNO data, we also have access to administrative data for each municipality, such as the resident population, the surface area, the geographic classification zones and degree of urbanisation.

The Emilia Romagna regional territory can be divided in five areas as it is shown in Figure 12.1: Western Plain, Eastern Plain, Bologna Agglomeration, Apennines, Seaside area.

Table 12.1: Distribution of municipalities by geographical zone, percentage of nights-spent by geographical zone according to Register or MNO Data.

Zone	Number of Municipalities	Nights per Month (%)	
		Register Data	MNO Data
Bologna Agglomeration	12	14.1	20.8
Apennines	120	0.5	1.6
Seaside Area	14	82.6	70.4
Eastern Plain	68	1.3	3.7
Western Plain	112	1.4	3.5

From a tourism perspective, the region is primarily characterized by a high volume of seaside tourism along the coast and urban tourism in the provincial capitals, while nature-based tourism in the Apennines and the Po Delta plays a secondary role, with clear and marked seasonal differences. Table 12.1 shows the distributions of municipalities and nights-spent by geographical zones. The MNO data underestimates the seaside zone and overestimates the other zones.

12.2 Modelling

12.2.1 Simplistic approach

Given a dependent response variable y and some known features x, the most *simplistic* approach consists of training a predictor $\mu(x; s)$ on a training set s, where both y and x are available, and then applying it to a target set r, where only x is available but not y. Formally, let s_t denote the training set at time t, we obtain

$$\hat{y}_j = \mu(x_j, s_t), \quad \forall j \notin s_t \tag{12.1}$$

In our case, s_t consists of the early reporting units, the dependent variable is the number of nights spent by tourists in a given municipality according to registry data at time *t*, while the features are chosen among different possible combinations of MNO data at time *t* or earlier and registry data from some earlier time point.

12.2.2 Quasi-Transfer Learning

As outlined in Section 5.3, quasi transfer learning may target the set

$$q_t = s_t \cup r_t \tag{12.2}$$

where s_t represents the set of municipalities that have already provided the value of the response variable y_t, and r_t consists of the remaining units for which the response is not available yet. We assume that for the units r_t all the administrative and MNO features are available. Now that y_i is missing for $i \in r_t$, we introduce an augmented sample s_t^*,

$$s_t^* = s_t \cup r_{t'}^* \tag{12.3}$$

where $r_{t'}^*$ represents the units that were late one year ago, and let y_j^* be the target value of $j \in r_{t'}$ which has since become known. Notice that some adjustment may be applied to y_j^*, if it makes the relationship between y_j^* and x_j for $j \in r_{t'}$ closer to that between y_i and x_i for $i \in r_t$. In our case, to account for a trend effect primarily driven by COVID-19 recovery, we applied a +3% adjustment relative to the 2022 values.

Furthermore, we adopt the transfer scheme given in Section 5.3 and use $\mu(x, s_b^*)$ for an earlier time point b as a feature in

$$E\{\mu(x, s_t^*)\} = g(x, \mu(x, s_b^*)) \tag{12.4}$$

where $\mu(x, s_t^*)$ represents the expected value of the response variable given x, which is obtained from the set s_t^*, and $\mu(x, s_b^*)$ similarly for b instead of t.

12.2.3 Augmented Learning

As outlined in Section 5.3, augmented learning is based on an augmented sample s_t^* instead of only the sample s_t form the target domain. As mentioned

above, in this study, we augment the sample s_t of fast-response municipalities by $r_{t'}$ which are late ones for the same months in the previous year.

Note that we experimented with replicating the units s_t multiple times in s_t^* to increase their weights in the training set. However, we ultimately decided to use only s_t as no significant improvements were observed otherwise.

12.2.4 Model functions and features

We consider both linear regression and random forest models as $\mu(\cdot)$. In the linear model, the square root transformation is applied to both the registry totals and MNO proxy counts, in order to reduce the influence of extreme values and normalise the distribution of the variables. This approach improves the stability and accuracy of the model, especially considering that these variables exhibit a highly skewed distribution.

Regarding the choice of features to be included in $\mu(x, \cdot)$, we considered both spatial (cross sectional) and temporal information that could be exploited to make forecasting. After many attempts we noticed that temporal information represented with auto-regressive (lagged) features is more explanatory than the spatial one. Unfortunately we dispose of relatively short time series such that we can leverage the temporal information only to a moderate extent. We are confident that with the future availability of more data, it will be possible to make better use of the temporal information. Nevertheless, the current setup is sufficient for demonstrating the potentials of the more advanced learning approaches to the simplistic approach.

To find an adequate linear model of y_t, we started with the MNO proxy x_t of the same month and the auto-regressive (y_{t-d}, x_{t-d}) up to lag $d = 12$. We then performed a backward elimination, keeping the most significant variables. Additionally, we included a categorical variable ZONE as spatial information. The resulting model for September 2023 is

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-3} + \beta_3 y_{t-12} + \beta_4 x_t + \beta_5 x_{t-1} + \beta_6 x_{t-9} + \sum_j \gamma_j \text{ZONE}_j + \epsilon \quad (12.5)$$

and the model for October 2023 is

$$\begin{aligned} y_t = \beta_1 y_{t-1} + \beta_2 y_{t-3} + \beta_3 y_{t-6} + \beta_4 y_{t-8} + \beta_5 y_{t-12} \\ + \beta_6 x_t + \beta_7 x_{t-1} + \beta_8 x_{t-2} + \beta_9 x_{t-6} + \beta_{10} x_{t-7} \\ + \beta_{11} x_{t-8} + \beta_{12} x_{t-9} + \sum_j \gamma_j \text{ZONE}_j + \epsilon. \end{aligned} \quad (12.6)$$

We applied the same procedure for Random Forest, performing backward selection based on feature importance. Considering that Random Forest is better suited than linear regression to handle a large number of features, we also included some registry information. For both September and October, we

selected the same set of features:

$$\{y_{t-1}, y_{t-11}, y_{t-12}, x_t, x_{t-1}, x_{t-11}, x_{t-12},$$
$$\text{Resident\_Population}, \text{Coastal\_Municipality}, \text{Altitude\_center},$$
$$\text{Urbanization\_Degree}, \text{Extension\_km2}, \text{Province}\} \quad (12.7)$$

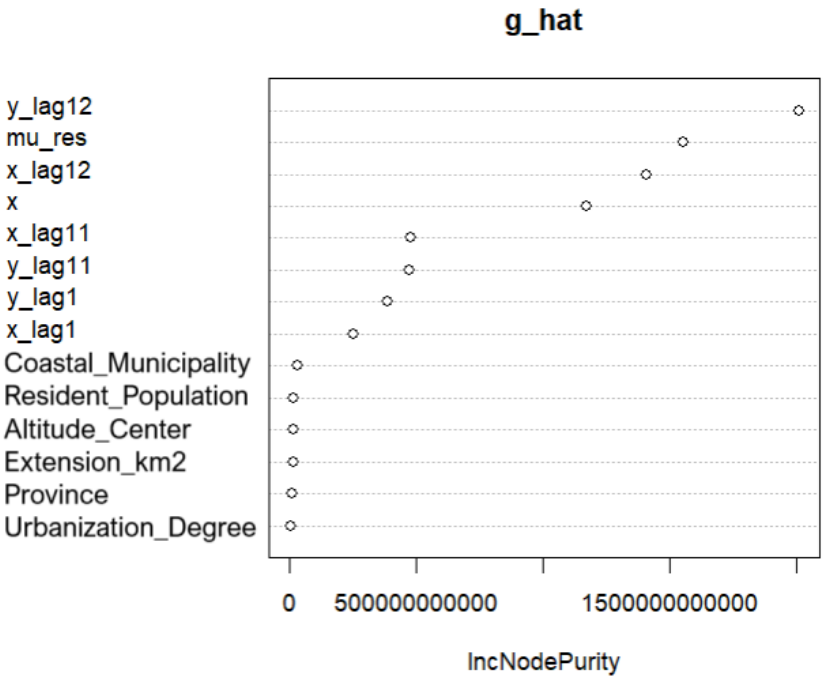

Figure 12.2: Feature Importance on $\hat{g}$

Figure 12.2 gives the feature importance for the Random Forest $\mu(x, s_t^*)$ given by (12.4), it is noteworthy that MNO proxies, as well as mu_res $\mu(x, s_b^*)$, are among the most important features. The same holds for the linear model, where some x variables and mu_res have coefficients that are significant at the 5% level in all of our experiments.

12.3 Preliminary results

We applied the three modelling approaches to September and October 2023. Note that in reality one knows whether a given municipality has a touristic vocation, and flash estimation is of interest for the municipalities with a strong touristic vocation. Since the number of overnights would otherwise be highly heterogeneous across all the municipalities, we decided to train the models only on the municipalities with a y-value greater than the median (which is 577 nights-spent in this case).

We can calculate the prediction errors of a given model against the true registry totals known to us in this study. In particular, the Mean Absolute

Error (MAE) and Mean Absolute Percentage Error (MAPE) are defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12.8}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{12.9}$$

The sum of all the predictions can be compared to the actual total as well.

Table 12.2: Prediction results for September and October 2023

	September 2023			October 2023		
Model	MAE	MAPE	Total	MAE	MAPE	Total
Linear, Simplistic	1301.94	19.93	4025369	1486.04	20.94	1766437
Linear, QTL	1206.17	19.04	4006008	1205.74	20.47	1805782
Linear, AL	1227.11	19.26	4009297	1238.69	20.35	1807020
Random Forest, Simplistic	4707.84	27.12	4144826	4625.61	35.31	2389966
Random Forest, QTL	2356.20	19.73	4104355	1900.56	21.94	1945381
Random Forest, AL	2224.88	18.98	4092401	1860.68	21.52	1941391
Total Nights-spent			4079632			1798189

Table 12.2 gives the results of MAP and MAPE, and the predicted totals are compared to the true total nights-spent. We could have simply evaluated the errors over all the units r_t of a given model obtained on the whole training set. Instead we performed 2-fold cross-validation repeatedly for 50 times. The reported errors are averages over all the 100 test sets created in this way.

The replicated test results allow us to compute a 95% Confidence Interval (CI) and a normalised Confidence Interval, defined as

$$\text{CI}_{\text{normalised},i} = \frac{\text{CI}_{\text{upper},i} - \text{CI}_{\text{lower},i}}{\hat{\bar{y}}_i} \tag{12.10}$$

where $\hat{\bar{y}}_i$ denotes the mean of the predictions for municipality i across all the repetitions. The mean of $\text{CI}_{\text{normalised}}$ over all the test units are given below.

Table 12.3: Mean Normalised CI

	Linear Model			Random Forest		
	Simplistic	QTL	AL	Simplistic	QTL	AL
September 2023	0.222	0.104	0.104	0.346	0.114	0.084
October 2023	0.301	0.097	0.098	0.384	0.087	0.055

The figures below provide more details over the different municipalities.

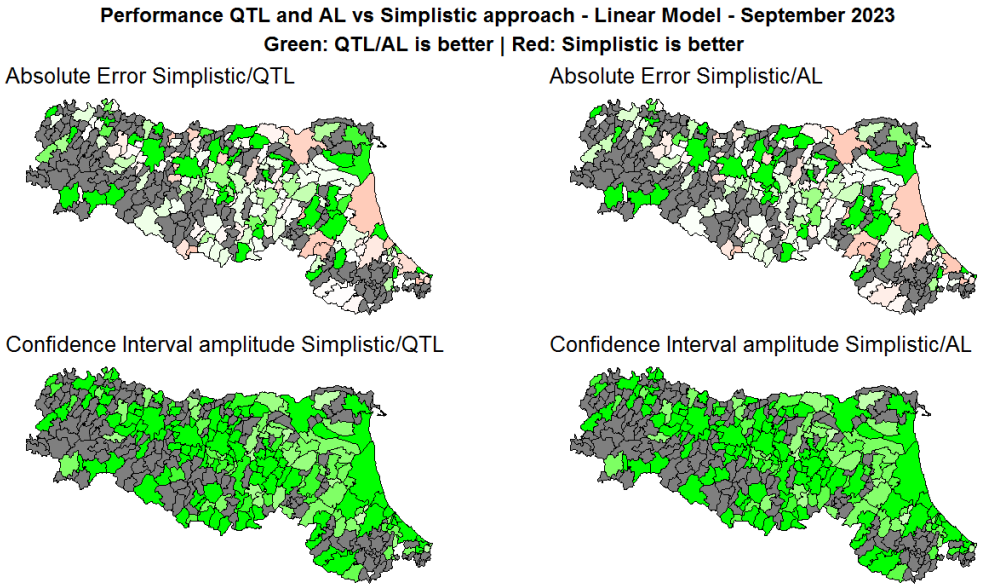

Figure 12.3: September 2023 QTL vs Simplistic - Linear Model

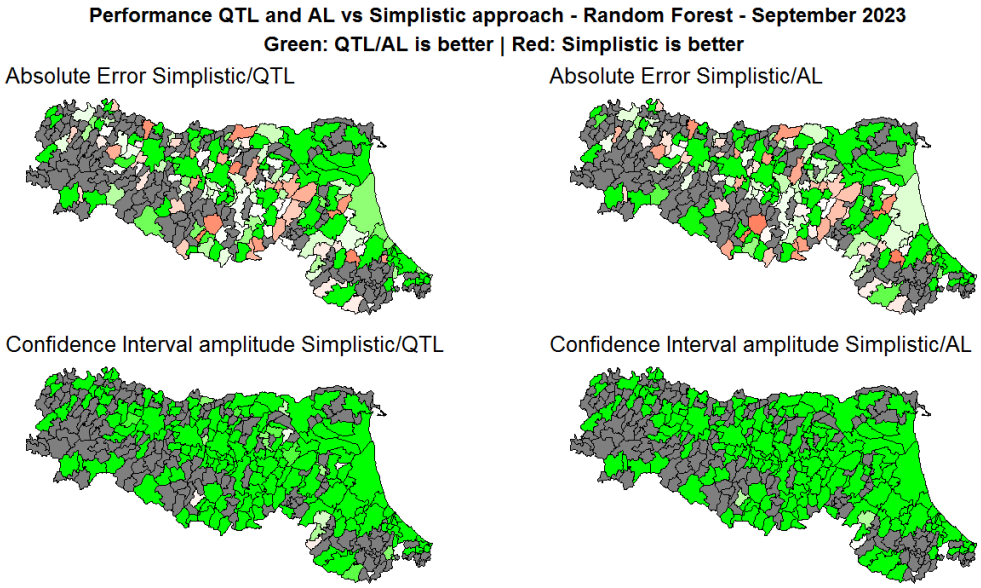

Figure 12.4: September 2023 QTL vs Simplistic - Random Forest

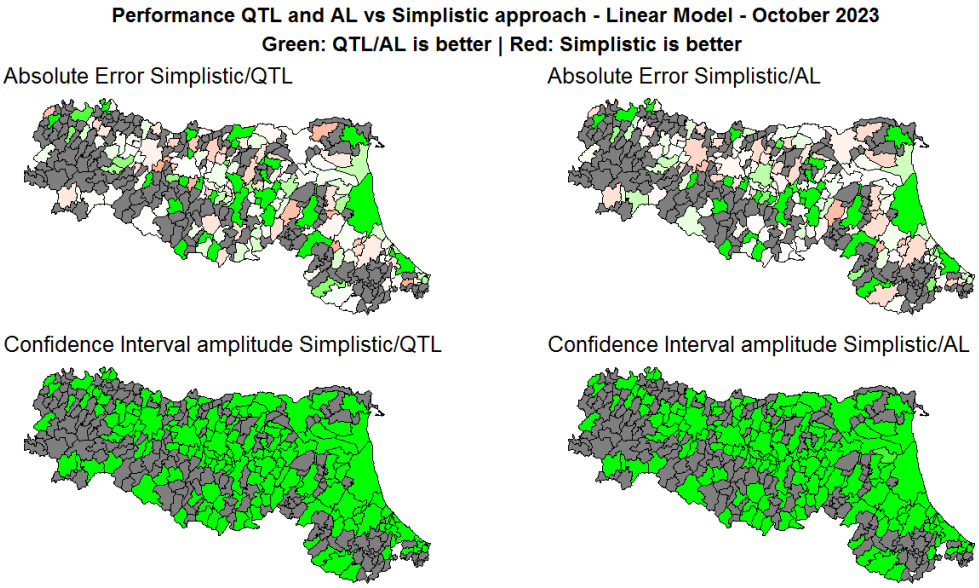

Figure 12.5: October 2023 QTL vs Simplistic - Linear Model

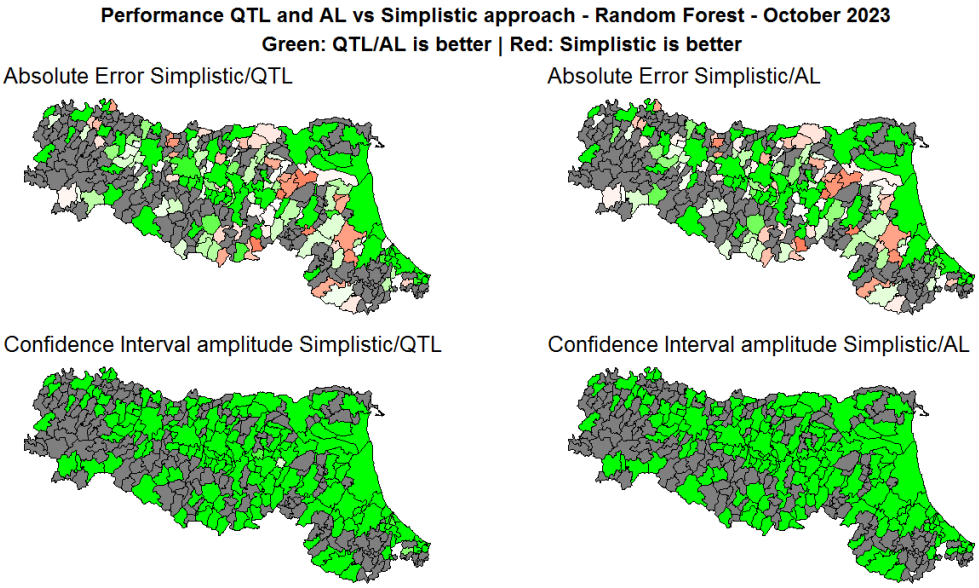

Figure 12.6: October 2023 QTL vs Simplistic - Random Forest

In the next figures, we focus on the municipalities where both QTL and AL outperform the simplistic approach in prediction, meaning that the ratios

$$\frac{\text{AbsoluteError}_{\text{Simplistic}}}{\text{AbsoluteError}_{\text{QTL}}} > 1.5 \quad \text{and} \quad \frac{\text{AbsoluteError}_{\text{Simplistic}}}{\text{AbsoluteError}_{\text{AL}}} > 1.5$$

Among these, we highlight in blue the municipalities where MAPE $\leq 10\%$, aiming to emphasize the areas where these methods achieve noteworthy results and can be strongly recommended.

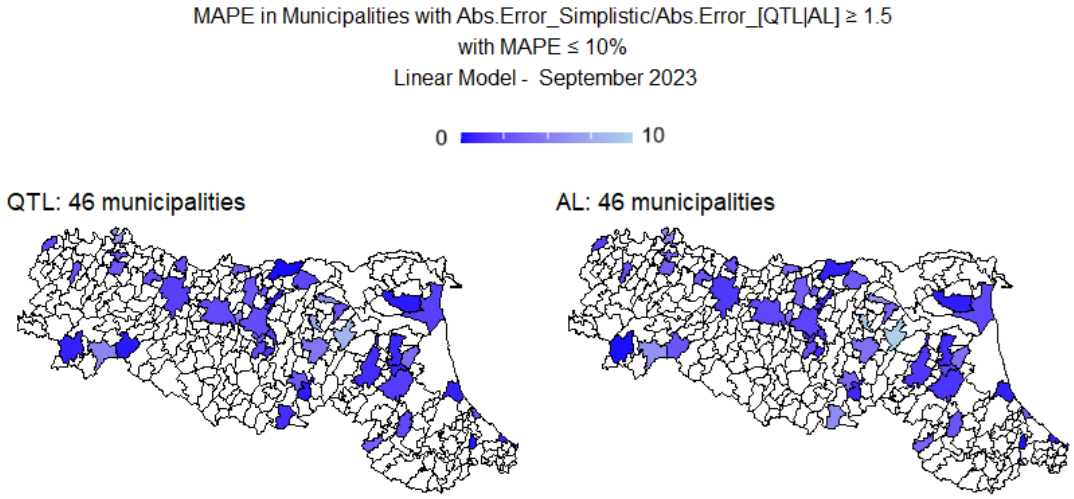

Figure 12.7: September 2023 QTL and AL MAPE - Linear Model

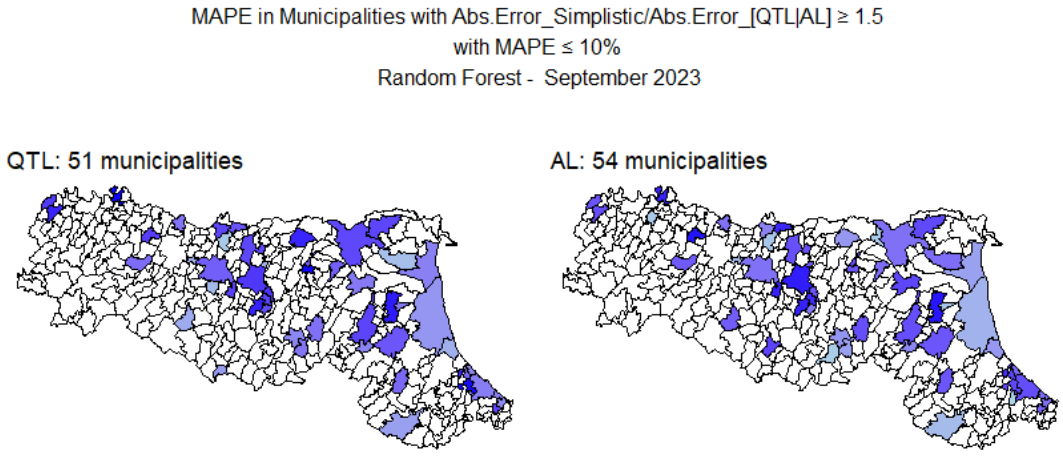

Figure 12.8: September 2023 QTL and AL MAPE - Random Forest

Observing the results, we can state that QTL and AL improve the precision of the estimates. Linear models appear to be more suitable for this prediction task, but Random Forest with QTL or AL achieves significant improvements over the Simplistic approach, reaching or even surpassing the MAPE of its

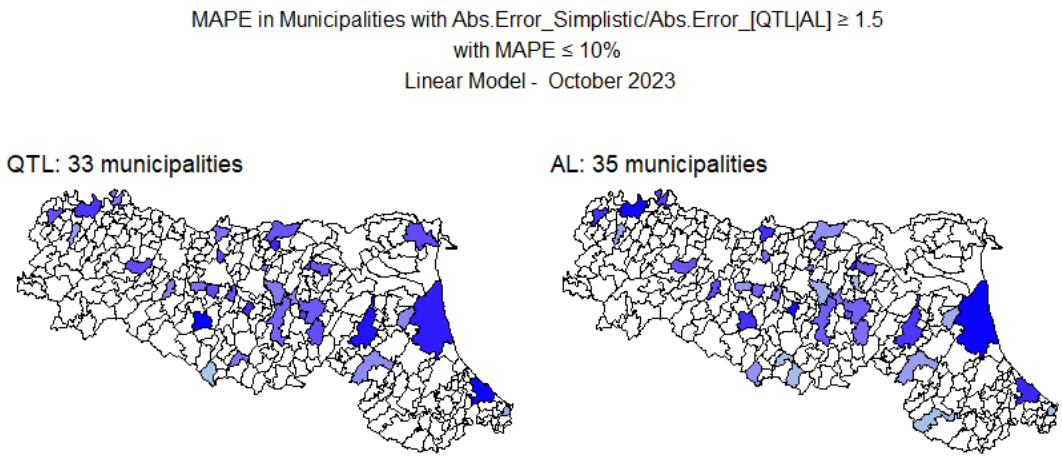

Figure 12.9: October 2023 QTL and AL MAPE - Linear Model

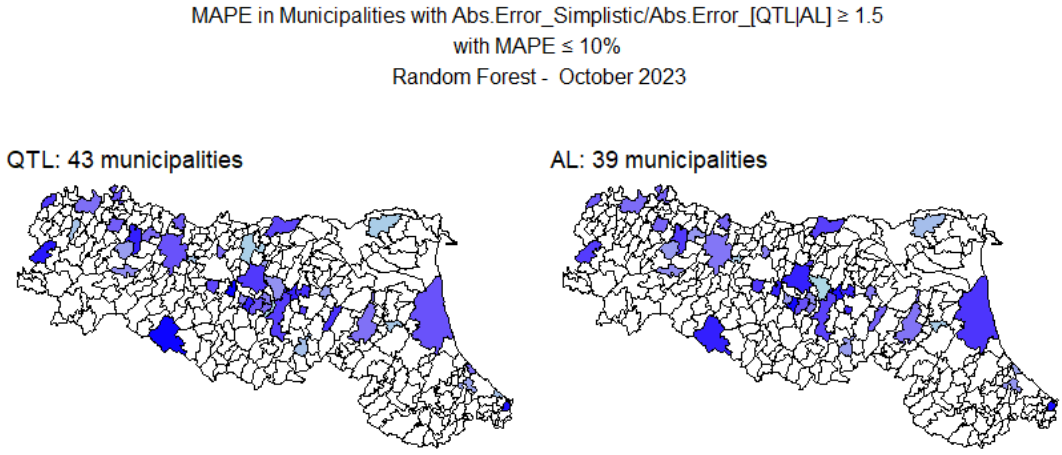

Figure 12.10: October 2023 QTL and AL MAPE - Random Forest

linear model counterpart. Finally, AL and QTL effectively improve the stability of flash estimation in terms of the confidence intervals.

12.4 Ensemble learning and future investigation

A more detailed analysis of the results above reveals that the Random Forest model performs better for municipalities with low tourism levels (under 5,000 overnight stays per month), while the linear model is more effective for the sparsely distributed units with large values. Ensemble learning that combines different models instead of relying on a single one may therefore be considered.

By a simple method of stacking linear and random forest models, we now use linear regression for municipalities where the observed target variable in the previous year exceeded a predefined threshold, whereas we use random forest for the others below this threshold. The threshold value 1300 is chosen empirically to minimise the MAPE. Approximately two-thirds of municipalities

are predicted by the linear model and the remaining third by random forest. The results are shown in Table 12.4. An improvement of about 1% MAPE can be seen against the previous results without ensemble learning.

Table 12.4: Ensemble prediction results

	September 2023			October 2023		
Model	MAE	MAPE	$TOT_{\hat{y}}$	MAE	MAPE	$TOT_{\hat{y}}$
Ensemble Simplistic	1315.20	22.78	4043084	1500.68	23.23	1778993
Ensemble QTL	1198.93	19.19	4014355	1194.13	19.15	1810638
Ensemble Augmented	1214.54	18.76	4017478	1228.30	19.35	1813292
$\mathbf{TOT_y}$			4079632			1798189

It is important to note that the APE is not uniformly distributed across the territory, which is less than 10% for about 40% of municipalities and averages at about 19%. In particular, we need to investigate the municipalities where MNO and administrative data show poor correlation, particularly those with significantly overestimated MNO values compared to administrative records. For instance, Bentivoglio municipality exhibits percentage errors of 143% and 189% in September and October, respectively. The reason may be because Bentivoglio hosts a major hospital with a high volume of patient admissions, which are misclassified as tourist presence in the MNO data.

Refining the ensemble method is another topic for future investigation. A strategy could be explored, which selects among the simplistic, quasi transfer and augmented learning approaches, depending on which performs best in a given municipality.

In this work, we tested Quasi Transfer Learning and Augmented Learning with the linear model and random forest, but many other models could be suitable for our forecasting task, such as count models (e.g., negative binomial), neural networks, and ARIMA. Further experiments could also be conducted using additional registry data, which may enhance the identification of long-term dependencies while being adjusted for short-term variations with MNO data, which are only available for recent periods.

Chapter 13

Sensor presence

Trustworthy counts of mobility or presence exist in many non-MNO sources, such as public transport passengers, museum visitors, concert audience, and so on. One may easily wish to have similar or related counts of mobility and presence, for which such trustworthy counts are lacking. For instance, for all the passengers who commute to the city by train or bus, one would like to know the number of commuters by private cars; for all the people who visit cinemas or concerts during the evening, one would like to know the number of people who dined at restaurants.

Where trustworthy counts do not exist in the non-MNO sources, prediction of them may be possible given proxy counts both where the trustworthy counts exist and where they do not exist, using the methods of statistical calibration described in Part II. For example, it is a fact that Google estimates the number of visitors to shops, museums, and so on, where trustworthy counts exist for some but not the others. For another example, Gilardi et al. (2022) combine road sensor vehicle counts with counts derived from TomTom navigation app in vehicles or mobile phones, to predict vehicle counts where there are no sensors but only the counts by TomTom app.

However, the OSA is unlikely to obtain access to Google data, or it may be troublesome to obtain GPS data from a large number of private data-holders. Can MNO engineer similar proxy counts of mobility or presence to be combined with the trustworthy counts from non-MNO sources?

13.1 Possible application scenarios

Let us illustrate by a couple of examples in the context of city-centre traffic.

Person-crossings Consider the number of *person-crossings* into (or out of) City Stockholm according to the congestion tax (Figure 13.1, top-left) during some given hours and days, where the mode of (land) travel can be

- public transport by rail, such as train, subway, tram;
- public transport by road, such as all types of bus;
- non-public motorised vehicles, such as car, lorry, motorcycle;
- other means, such as biking, walking.

We notice that the number of person-crossings refers to crossings made by persons. It is not the same as the number of (distinct) persons making the crossings, nor the number of persons present given the time and place.

Although MNOs can count the number of SIMs crossing into (or out of) Stockholm inner-city, it covers neither all the persons nor all the crossings. Non-MNO sources can only provide the target count of person-crossings by certain mode of travel (e.g. public transport), or perhaps estimates of target count (e.g. bottom-left in Figure 13.1), but they may lack the related data of movements in any case. Combing MNO data with relevant non-MNO sources can potentially generate more useful statistics.

For example, given the counts of person-crossings by public transport, one may consider compositional statistical calibration for predicting the count of person-crossings by non-public motorised vehicles. Regardless the details, this would require the MNOs to engineer proxy counts of person-crossings by the different modes of travel, which however seems infeasible given the current mobile device signalling technology and network infrastructure, in contrast to GPS data from navigation apps in vehicles or mobile phones.

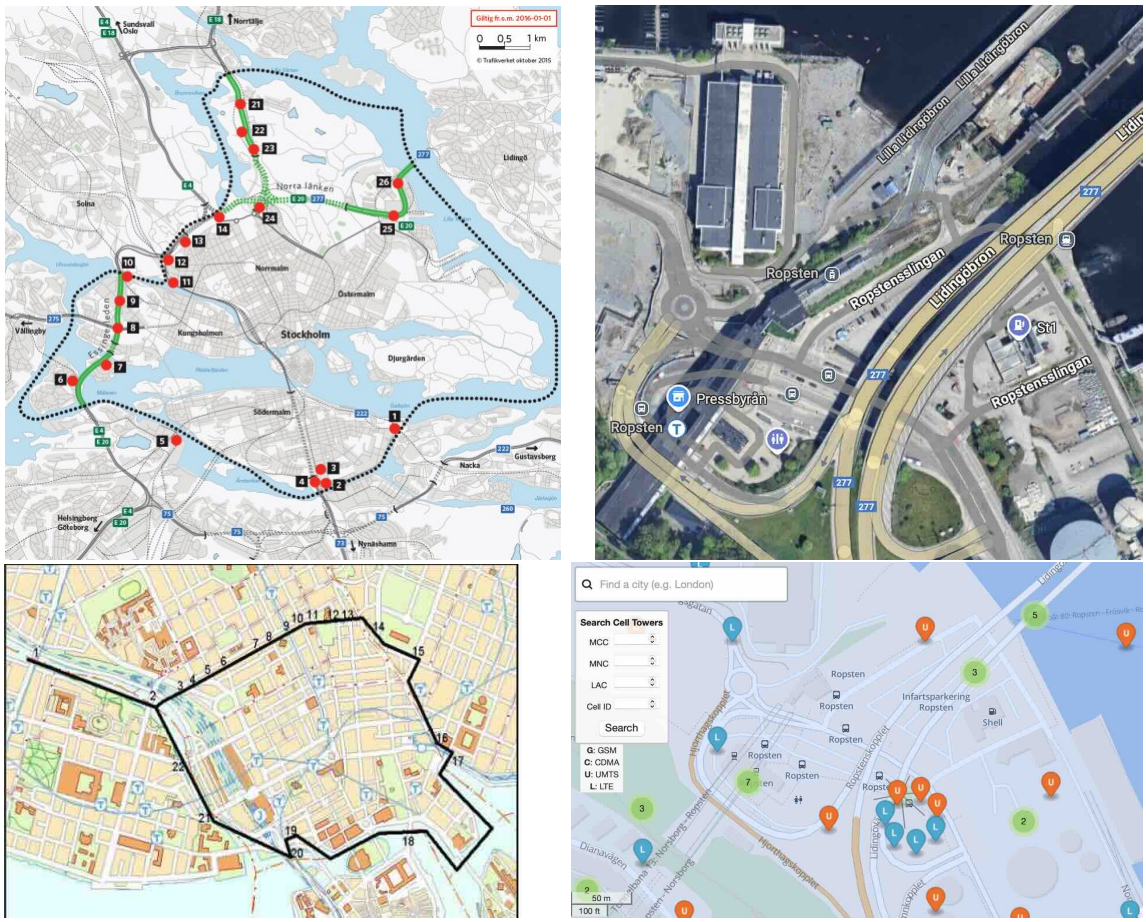

Figure 13.1: Top-left, vehicle sensors, City Stockholm (Transportstyrelsen); bottom-left, bike traffic survey locations, inner city (Trafikkontoret); top-right, aerial view surroundings vehicle sensor nr. 26, Ropsten (Google); bottom-right, antenna cells around Ropsten, numbers indicating multiple cells (OpenCellid)

Vehicle traffic The totals of the vehicles crossing into the city are recorded by the vehicle sensors (Figure 13.1, top-left) for the traffic congestion tax, denoted by y_i for $i = 1, ..., m$ generally. For traffic or environment, one may be interested in the distribution of these vehicles inside the city. For instance, how many of them passed the Central Railway Station, the National Theatre or the Zoo. Denote by $j = 1, ..., n$ these *points of interest* and y_j the target count.

Now, suppose the MNO can engineer a count of SIMs that crossed into the city at *each sensor location*, denoted by x_i for $i = 1, ..., m$. It seems then intuitive that how the x_i SIMs later spread across the city may provide useful information of the distribution of the corresponding y_i vehicles, *provided* x_i is a close proxy to y_i at each sensor location i. For instance, let x_{ij} be the number among the x_i SIMs which first passed location i and later location j. Note that this requires the MNO to keep track of all the x_i SIMs, and the fractions x_{ij} generally do not sum to 1 over $j = 1, ..., n$ and given i, i.e. $\sum_{j=1}^{n} x_{ij} \neq x_i$. A simple estimator of the number of vehicles passing POI j can now be given as

$$\hat{y}_j = \sum_{i=1}^{m} \frac{x_{ij}}{x_i} y_i$$

As another possibility, instead of the fractions x_{ij}, suppose the MNO can engineer a count x_j for each POI j, similarly to x_i at the sensor locations. One may then consider a spatial statistical calibration estimator of y_j based on $\{(x_i, y_i) : i = 1, ..., m\}$, as described in Section 6.1, provided x_i is a close proxy to y_i at each i and x_j is a close proxy to y_j at each j.

Regardless which MNO counts may be available in practice, or the details of the statistical calibration methods, a key to success is how close the MNO proxy x_k can be to the target count y_k at each specific location k.

Discussion The aerial view around the sensor location $i = 26$ at the top-right of Figure 13.1 illustrates the challenge for MNOs. The area is called Ropsten. There are several roads on more than one level, not laid out in a single direction or two opposite directions, nor are the roads necessarily parallel to each other. Moreover, looking closer, one can see at least a subway station, a tram station, a ferry landing, and several bus stops in the area. Finally, there are other facilities such as petrol station, building complex, and so on, where SIMs may be in contact with the antennas covering the sensor location — a view of the antenna cells in the area is given at bottom-right of Figure 13.1.

Thus, no matter where the exact location of the vehicle sensor nr. 26 is, or which direction it covers, it may be difficult for the MNO to 'filter out' the SIMs associated with the vehicles passing that spot and in that direction.

However, unless the MNO can engineer a sufficiently close proxy at each given location, the following problems may arise generally.

- At any given location k, complicated (or heterogeneous) topology can easily cause large systematic discrepancies between x_k and y_k, which overwhelm the sources of random error associated with x_k.

- If one treats the systematic discrepancies, say, $z_k = x_k - y_k$ and $z_l = x_l - y_l$ as 'random variables' from one location to another, then the variance of z_k over

all the locations k may be large without the additional information that can 'explain' the heterogeneous causes at the different locations.

In such cases, the individual prediction error $\hat{y}_j/y_j - 1$ may be too large to be acceptable for many points-of-interest, even though the total prediction error $\sum_{j=1}^{n} \hat{y}_j / \sum_{j=1}^{n} y_j - 1$ may be acceptable.

13.2 Exploration by simulation

The accuracy of the inferred device physical location depends on the adopted technique. Using the position of the handling antenna cell is always feasible but also the least accurate, which we shall refer to as *surrogate position* here. Triangulation or trilateration makes use of multiple cells that are in-contact with a device, where the former is based on angles (at least two cells) and the latter on distances (at least three cells). The accuracy of triangulation or trilateration depends therefore on the concentration of cell base stations, e.g. urban environments with a high cell density vs. rural areas where there do not exist multiple in-contact cells within any short time interval.

It is easier to explore the accuracy of surrogate position by simulation, since one does not need highly detailed network data or potentially resource-demanding algorithms for triangulation or trilateration.

13.2.1 Setup

The simulation tool is available at `Github_simulator`. The input data are:

- a map.wkt file defining a geographical area, which must be a closed polygon;
- an xml file configuring time, MNOs, and movement pattern (Figure 13.2);
- a file with the technical parameters and exact locations of antennas;
- a file specifying the population size, age-gender distribution, mode of move.

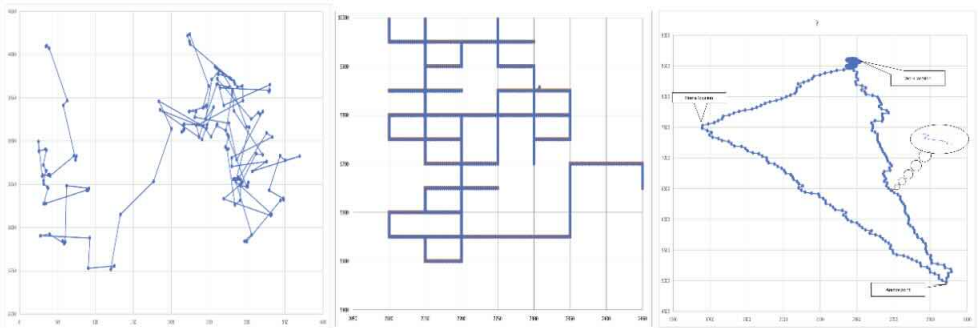

Figure 13.2: Left, Levy flights; middle, Manhattan; right, home-work

Thus, a given number of people can move randomly within the time period specified, by walking or in car, in agreement with a chosen movement pattern (illustrated in Figure 13.2). At the same time, the devices of the users will

interact with the antennas according to the technical parameters, where there may be one or two MNOs and each user may have one or two devices.

We used 400 individuals in the simulation, of which 258 have no phones. There are 70 antennas in the map, of which 50 are switched on during the simulation. The following output files are available for post-processing:

- grid.csv, which consists of 40×40 tiles each of $250m$ in length and width;
- antennas.csv, which replicates the input antenna details;
- persons.csv, including each person's location, time and mobile devices;
- AntennaInfo_MNO_MNO*.csv, which contains the signal events handled by the respective MNOs, as illustrated below;

t	Antenna ID	Event Code	Device ID	NetworkType	TA	x	y	Tile ID
0	2	0	52	4G	0	897.078045	5832.890081	221
0	4	0	90	3G	0	1510.729201	6618.276457	263
				⋮				
199	18	2	69	4G	2	2732.140804	9097.350172	365
199	3	2	38	4G	1	957.345382	4255.426410	161

Event code: 0, connect; 1, disconnect; 2, successive connect; 3, failure.
TA (time advance, length of time from device to antenna): 0, 1, 2, ...

- AntennaCells_MNO*.csv and SignalMeasure_MNO*.csv, which contain the calculated coverage area of each antenna, and signal strength at the center of each tile; the data can be used for inference of device location.

It is possible to apply the method of latent analysis (Salgado et al., 2021) to estimate the device locations. However, the computation is demanding, and one cannot necessarily expect such extra processing from the MNOs. Instead we shall focus on surrogate positions to explore the challenge for MNOs to produce proxy counts at given locations.

13.2.2 Surrogate position counts

Denote all the grids in the map by $U = \{i_1, ..., i_{1600}\}$, as well as by (h_{i1}, h_{i2}), $1 \leq h_{i1}, h_{i2} \leq 40$. Denote by (c_{k1}, c_{k2}) the coordinates of each antenna $k = 1, ..., 50$, where $0 \leq c_{k1}, c_{k2} \leq 10000$. For each antenna k, denote by $i(k)$ the grid in which the antenna is placed; denote by s_A all these *(antenna) grids*, $s_A \subset U$.

Sensor counts As illustrated in Figure 13.3, given the movements data, we can calculate the number of movements passing over each grid during the simulation, denoted by y_i for $i \in U$, as if we had a sensor to count it exactly. Note that this count includes those movements without devices or signals.

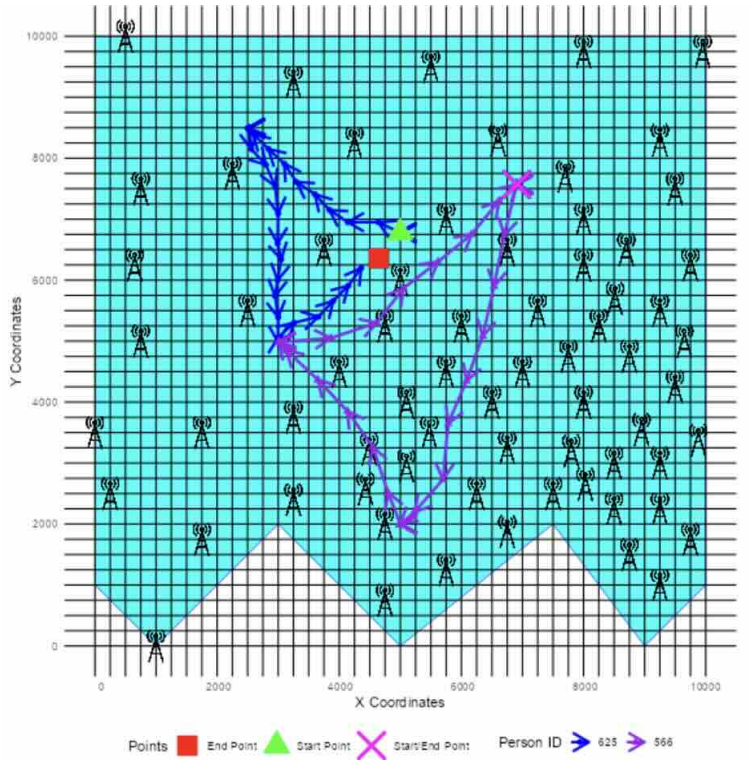

Figure 13.3: Antennas, two simulated movements

MNO counts The description applies to each MNO separately. Let $i(t;e)$ be the grid $i(k)$ of the antenna k of a signal at time point t, which is associated with an entire movement e simulated (as the two in Figure 13.3). For each movement e, let $t_1(e)$ be the first signal time point, and so on till $T(e)$ at last, yielding the $T(e)$ grids as

$$\{i(t;e) : t = t_1(e) \leq t_2(e) \leq \cdots \leq T(e)\}$$

Divide the sequence above into segments of identical grid $i(t;e)$, and extract all the distinct *signal-segment grids* of e ordered over time, denoted by

$$\{i_1(e), i_2(e), ..., i_{M_e}(e)\}$$

where $i_g(e) \neq i_{g+1}(e)$ for any $g = 1, ..., M_e - 1$ and M_e refers to the last identical signal-segment of e. For each grid $i \in U$, let the *MNO count* of movements passing over gird i in the whole simulation be

$$x_i = \sum_{e:M_e>0} \sum_{g=1}^{M_e} \mathbb{I}(i_g(e) = i)$$

13.2.3 Results

Figure 13.4 summarise the counts y_i and x_i. The two histograms are given at the top. There exists a reasonable correlation between the two, as can be seen in the bottom-left scatter plot, the linear regression of y_i on x_i has $R^2 = 0.774$ and root standard error 15.67. The 'detection' rate $100x_i/y_i\%$ is given for the grids $i \in s_A$ with $x_i > 0$ on the bottom-right, which exhibits clear spatial effects,

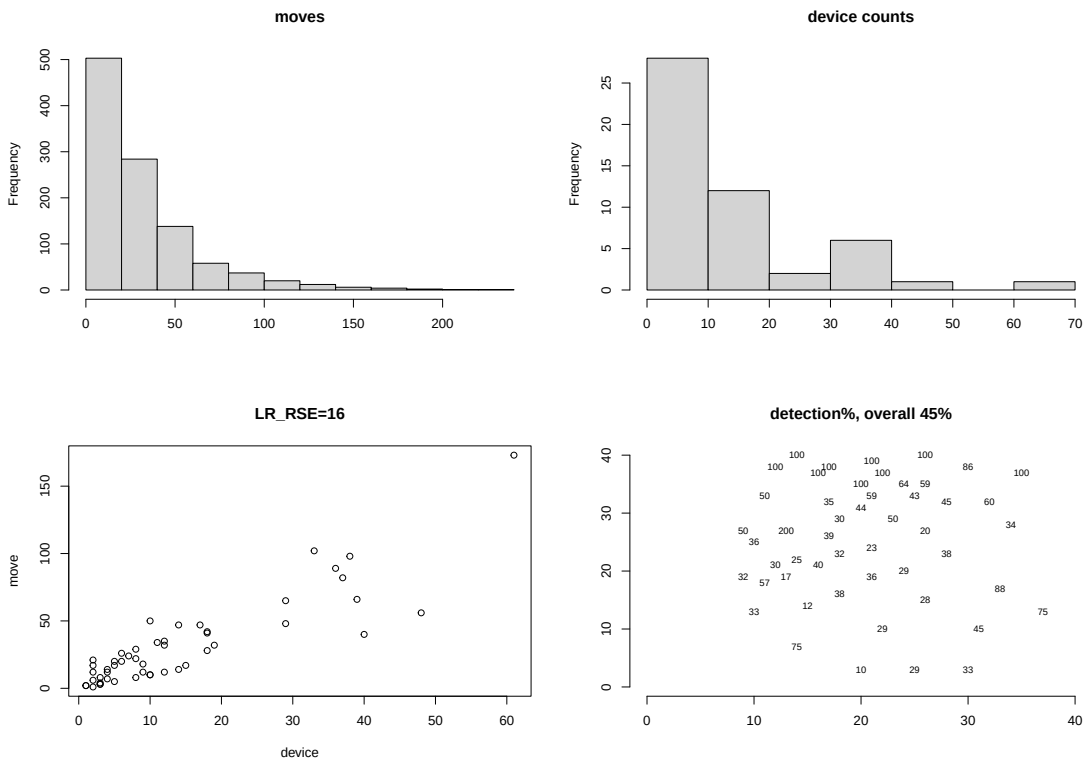

Figure 13.4: Summary of simulated grid counts of movement

e.g. about 100% close to the top of the area but decreasing to much lower rates near the bottom of the area.

Next, Figure 13.5 explores the prediction errors by out-of-bag regression. For each grid i, where $x_i > 0$, let the relative absolute error of prediction be

$$\text{rae}_i = |\hat{y}_{(i)}/y_i - 1|$$

where $\hat{y}_{(i)}$ is obtained by geographically weighted regression (GWR) on all the *other* grids except i, as explained in Section 6.1, using the weights

$$w(d_{ij}) \propto \exp\{-\alpha d_{ij}\}$$

where d_{ij} is the Euclidean distance between grids i and j. In the case $\alpha = 0$, we recover OLS regression, which yields the rae_i shown in the left plot of Figure 13.5. Those of GWR with $\alpha = 6$ are shown in the right plot.

It can be seen that spatial statistical calibration by GWR can reduce the prediction error, where the mean of rae_i is 44% compared to 55% by the OLS regression, and the rae_i no longer exhibits any clear spatial pattern (in the right plot compared to the left plot). However, an average RAE of 44% would still be too large compared to the usual accuracy thresholds in official statistics.

Therefore, the conclusion for the simulation setup here is that x_i based on surrogate cell positions is not a sufficiently close proxy to y_i for the purpose of predicting sensor counts where there are only MNO counts.

Finally, Figure 13.6 shows all the (sensor) grid counts y_i in the left plot, where the size of dots are proportional to y_i. All the MNO grid counts x_i based on surrogate positions are given in the right plot, where the dots have sizes

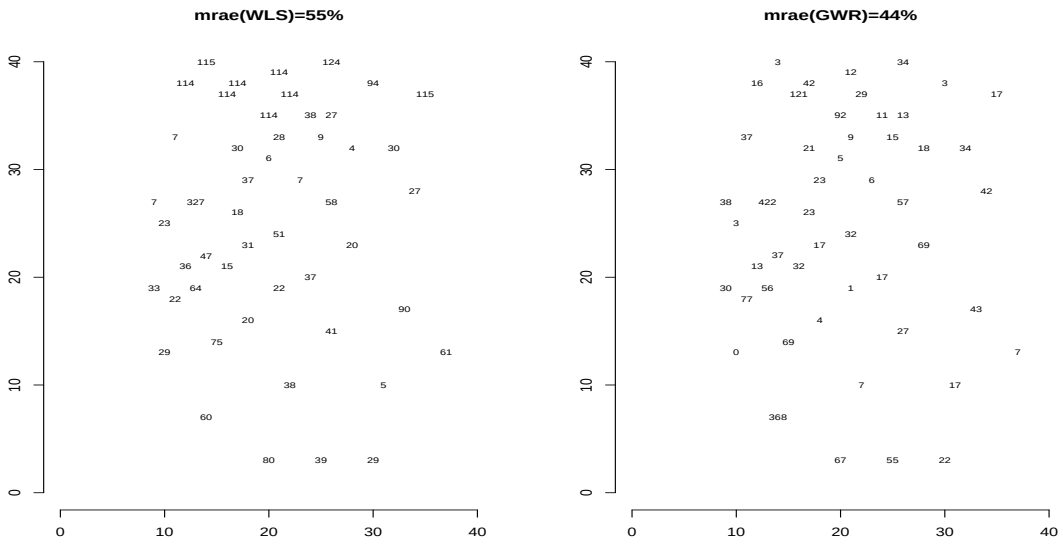

Figure 13.5: Out-of-bag relative absolute error by OLS (left) or GWR (right)

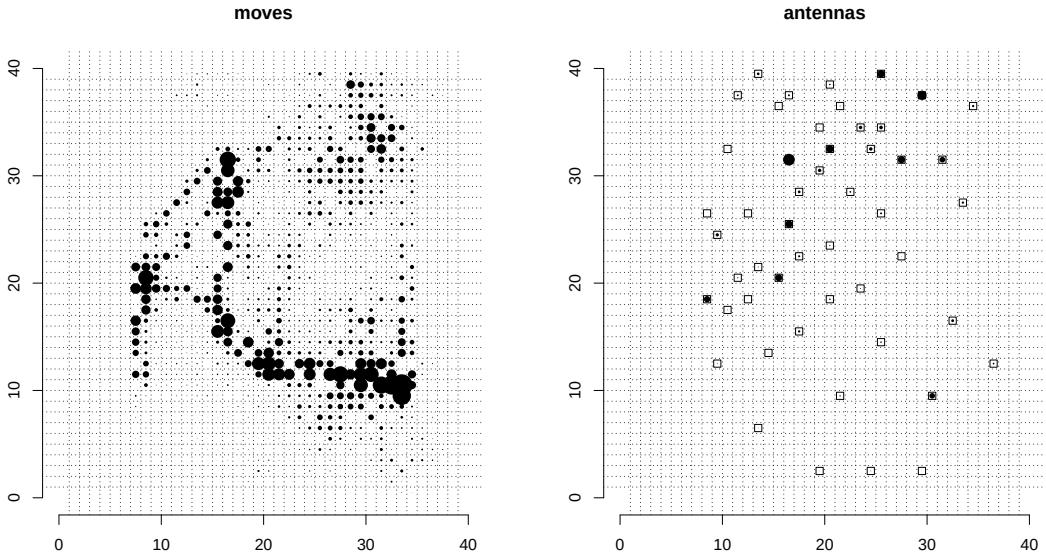

Figure 13.6: Simulated counts over grids, true (left) MNO (right)

proportional to $2x_i$ which roughly adjusts for the overall 'detection' rate

$$\sum_{i \in s_A} x_i / \sum_{i \in s_A} y_i = 0.450$$

It is evident that the spatial distribution of y_i (on the left) is very different to that of x_i (on the right). This conforms to the conjecture earlier (related to Ropsten) that the difference $y_i - x_i$ may be dominated by systematic than random errors. Moreover, it is clear that applying spatial interpolation to the observed x_i will not yield accurate prediction of y_i for $i \in U \setminus s_A$. In other words, additional information about the spatially heterogeneous relationship between x_i and y_i is needed for successful statistical calibration, as long as the MNO count x_i is not a sufficiently close proxy to the target count y_i.

Chapter 14

QR experiment

We analyse QR estimation and explore its likely effects in a proof-of-concept experiment created with register data at Statistics Norway. The mobile phone users are identified based on automatic mobile phone number search as one would have done for sample surveys in practice. (We record whether a person has a mobile phone or not, but not the phone numbers.) Several complications will be considered, such as lack of features for QR estimation, multiple devices of given mobile phone users, and ambiguity (or mis-identification) of device users. The relevant methods have been described in Chapter 8.

14.1 Setup

We draw a simple random register-household sample of persons $s \subset U$, and obtain the user sample $s_P = s \cap P$ by automatic mobile phone number search. Whatever imbalance between P and U is expected to be mirrored by s_P and s, and it is possible to explore user ambiguity since user-contractor connections within the same household are most common.

- 4×10^5 households are drawn from the Household Register (March 2024), which yields $n(s) = 662988$ persons of age 16-99. In particular, the residents of the smallest municipalities (fewer than 2000 inhabitants) have all their registered home municipalities replaced by "Annet".
- Automatic number search yields the *user sample* of $n(s_P) = 638828$ persons with Norwegian mobile phone numbers, $n(s_P)/n(s) = 96.4\%$.
- For each person in the sample s, a number of features are obtained from statistical registers with various time references, such as gender, age, native born status (March 2024), income (year 2022), activity or health status such as employed, unemployed, student (year 2022).

Notice that the population U is determined by the Household Register March 2024. Only the sample s and their associated features are made available to us for the analysis. For any post-stratification $\{U_g\}$ defined according to the features mentioned above, we have the sample proportions $\{n_g(s)/n(s)\}$ but not the population proportions $\{N_g/N\}$.

Selection error For any $k \in U$, let $\delta_k = 1$ if $k \in s_P$ in case $k \in s$, and $\delta_k = 0$ otherwise. Let $P = \{k \in U : \delta_k = 1\}$. For any given feature $\{y_k : k \in s\}$ from statistical registers, we can analyse the selection error and QR-estimation effect using the corresponding $\bar{y}(s), \bar{y}(s_P)$ and $\bar{y}_{qr}(s_P)$ as described in Section 8.3.2.

Remark Insofar as P and y-values are not obtained from the MNOs directly, this constitutes a proof-of-concept experiment rather than a real application of QR-estimation. In particular, we assume that P is suggestive of any time-specific set of users associated with devices detected by the MNOs, denoted by P_t, in terms of the selection effects. Although the selection error of $\bar{y}(s_P) - \bar{y}(s)$ is not that of a genuine MNO-count, the effect of QR-estimation is valid for any genuine MNO-count that has a similar association $Cov_N(\delta_k, y_k)$.

Lack of features Given $\{y_k : k \in s\}$, let $\bar{y}_{qr}(s_P)$ or $\bar{y}'_{qr}(s_P)$ be the QR-estimator based on post-strata $\{U_g\}$ or $\{U_{g'}\}$. While U_g depends only on features available to the MNOs such as age and registered address, $U_{g'}$ can depend on any features prepared above. Any improvements of $\bar{y}'_{qr}(s_P)$ over $\bar{y}_{qr}(s_P)$ would illustrate the complication due to lack of relevant features for QR estimation.

User ambiguity For post-stratification $\{U_g\}$, let $\mathbb{I}(k \in U_g)$ for given user k (e.g. of age 16-18) instead be determined by someone else (e.g. the eldest person) in the same household, where the latter is the stipulated service contractor. For example, instead of (age 16, student) of user k, the contractor in the same household may have (age 45, employed), in which case the user k will cause an error for any post-stratification involving age and employment status.

14.2 User sample balance

Any given feature is *balanced* in the user sample s_P compared to s (or U) if its distribution in s_P is the same as that in s (or U). If the user sample is balanced (or nearly so) in terms of all the features known for both s_P and s (or U), then one might hope it to be reasonably balanced in terms of other features that are only known in s_P but not in s (or U). This is the basic intuition for examining subset balance for indications of selection error, where the subset may refer to mobile phone users, respondents in sample surveys, and so on. Reversely, given any feature that is imbalanced in s_P compared to s (or U), it seems intuitive that adjusting the imbalance might help to reduce the selection error of s_P.

From `ssb.no` we obtain the population statistics that are deemed the closest to those of Household Register March 2024, where the latter is the sampling frame of s, in terms of age, gender, registered home municipality. Figure 14.1 compares the distribution of age in s_P, s and U (of the reference population statistics). While the distributions of age are close to each other in s and U, the eldest persons (say, 70+) are clearly under-represented in the mobile phone user sample s_P, and the persons of age up to 60 are over-represented — without the youngest persons having a notably more biased representation though.

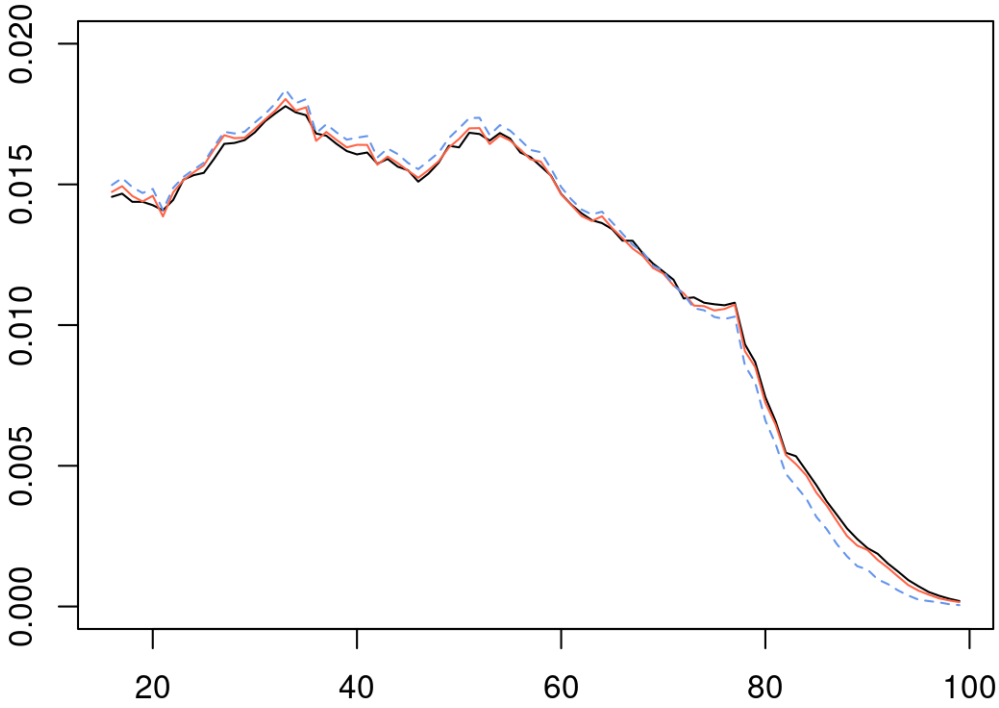

Figure 14.1: Age in s_P (dashed), s (red), U (black)

Figure 14.2 compares the distributions of (registered home) municipality, which are visually indistinguishable from each other here. The 357 Norwegian municipalities can be divided into 6 groups of Centrality according to standard of Statistics Norway, and Table 14.1 gives the distributions of Centrality. The additional group Annet in s and s_P has unknown Centrality due to the reason explained before, whereas the problem does not exist in the reference population statistics. There is very little difference between s_P and s, suggesting that the mobile phone user proportion hardly varies by Centrality.

Table 14.1 shows the user sample balance in terms of several other features: Fem for female, Emp for employed, Nat for native born, Stud for student, NAV for receiving a particular health-care-related benefit, Hsh-1 as an indicator for single-person household, and Pov an indicator if a person's income is below 60% of the median among all the persons in s. Population reference is given for Fem; whereas the other features are only compared between s_P and s.

Relatively speaking, Hsh-1 is the feature that differs most (by about 4%) from the sample to the user sample, followed by Emp and Stud with about 3% difference between s and s_P; whereas Fem differs the least by about 0.4%.

Clearly, the mobile phone user sample is not balanced in every respect, and age is the feature available to the MNOs which may be most relevant for QR adjustment (compared to gender or home municipality).

14.3 Selection error, QR adjustment

Various post-stratification can be created for QR estimation. In particular, the following features are used for the results below.

- Age, 6 groups: 16-20, 21-35, 36-50, 51-67, 68-77, 78-99.

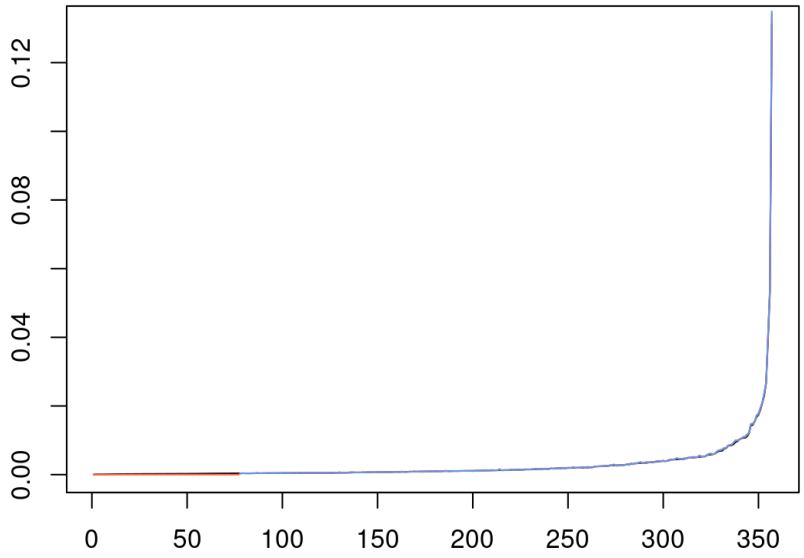

Figure 14.2: Municipality in s_P (dashed), s (red), U (black)

Table 14.1: Proportions in user sample s_P, sample s, or reference population U. Fem, female; Emp, employed; Nat, native born; Stud, student; NAV, a type of benefit; Hsh-1, single-person household; Pov: below 60% of median income.

	Centrality							
	1	2	3	4	5	6	Annet	Fem
s_P	0.185	0.273	0.258	0.145	0.097	0.024	0.017	0.496
s	0.184	0.271	0.258	0.146	0.098	0.025	0.018	0.498
U	0.183	0.272	0.257	0.146	0.102	0.039	–	0.498
	Emp	Nat	Stud	NAV	Hsh-1	Pov		
s_P	0.641	0.800	0.062	0.140	0.252	0.191		
s	0.621	0.793	0.060	0.138	0.262	0.193		

- Cent for Centrality, 7 groups including Annet.
- Emp for employed or not, 2 groups.
- Nat for native born or not, 2 groups.
- Edu for highest level of education, 5 groups including unknowns.

Table 14.2 gives $\bar{y}(s), \bar{y}(s_P)$ and $\bar{y}_{qr}(s_P)$ by various post-stratification, where the y-outcome is income, commuter or not, student or not, and female or not. In particular, commuter is the case if a person has home and workplace in different municipalities according to the statistical registers. The respective estimated standard errors are given in parentheses.

The difference $\bar{y}(s_P) - \bar{y}(s)$ is an unbiased estimator of the mobile phone user selection error, and the difference $\bar{y}_{qr}(s_P) - \bar{y}(s)$ is an unbiased estimator of the remaining error of the given QR-estimator. These are given in Table 14.3, where the QR-estimator uses post-stratification (Age, Centrality) specifically.

For the two outcomes with the largest relative selection errors, commuter and student, QR adjustment according to Age and Centrality removes 50% or

Table 14.2: Estimates (QR if unstated), standard errors (SEs) in parentheses

Method	Income		Commute (%)		Student (%)		Female (%)	
$\bar{y}(s)$	399011	(553)	34.5	(0.069)	5.99	(0.028)	49.8	(0.049)
$\bar{y}(s_P)$	404186	(578)	35.5	(0.072)	6.16	(0.029)	49.6	(0.052)
(I) Age	403407	(575)	35.0	(0.071)	6.06	(0.029)	49.6	(0.052)
(II) Cent	404063	(576)	35.5	(0.072)	6.15	(0.029)	49.6	(0.052)
(III) Emp	400526	(571)	34.5	(0.070)	6.49	(0.031)	49.7	(0.053)
(IV) Nat	403994	(577)	35.4	(0.071)	6.13	(0.029)	49.6	(0.052)
(V) Edu	401950	(570)	35.3	(0.071)	6.34	(0.030)	49.5	(0.053)
(I)×(II)	403292	(574)	35.0	(0.071)	6.06	(0.029)	49.6	(0.052)
(I)×(II)×(III)	401667	(571)	34.5	(0.070)	6.23	(0.030)	49.7	(0.052)
(I)×(II)×(IV)	402590	(569)	34.8	(0.070)	6.00	(0.029)	49.7	(0.053)
(I)×(II)×(V)	401910	(568)	34.8	(0.070)	6.05	(0.029)	49.7	(0.052)

Table 14.3: Estimated errors, $\bar{y}_{qr}(s_P)$ by (Age, Cent), SEs in parentheses

	Income		Commute (%)		Student (%)		Female (%)	
$\bar{y}(s_P) - \bar{y}(s)$	5174	(67)	1.06	(0.006)	0.17	(0.003)	-0.19	(0.017)
$\bar{y}_{qr}(s_P) - \bar{y}(s_P)$	-893	(34)	-0.51	(0.002)	-0.10	(0.001)	0.05	(0.005)
$\bar{y}_{qr}(s_P) - \bar{y}(s)$	4281	(73)	0.55	(0.006)	0.07	(0.003)	-0.14	(0.017)

more of the initial selection error. Since s_P is essentially balanced for Centrality compared to s (Table 14.1), the adjustment can almost entirely be attributed to Age in this experiment. The adjustment is smaller for the other two selection errors of mobile phone users, which are smaller relatively speaking (i.e. 1.3% for income and -0.4% for female proportion).

Selection error bound Given the sign of $\hat{\theta} - \theta$, one can derive a bound for the magnitude (i.e. absolute value) of selection error.

First, if $\hat{\theta} > \theta$, as the case with commuter, student and income here, then the maximum possible selection error is achieved if $y_k = 0$ for any $k \in U \setminus P$, such that

$$|\hat{\theta} - \theta| \;\leq\; \hat{\theta} - \frac{N(1-f)\hat{\theta}}{N} = f\hat{\theta} \tag{14.1}$$

where $N(1-f)\hat{\theta}$ is the total of y among the mobile phone users, given f as the proportion of non-users in the population.

To illustrate, let $f = 3.6\%$ be as observed in s, likewise for $\hat{\theta}$, such that

$$\text{student, } \hat{\theta} = 5.99\% \quad \Rightarrow \quad f\hat{\theta} = 0.2\% \approx 6.16\% - 5.99\% = \bar{y}(s_P) - \bar{y}(s)$$
$$\text{commuter, } \hat{\theta} = 34.5\% \quad \Rightarrow \quad f\hat{\theta} = 1.2\% \approx 35.5\% - 34.5\% = \bar{y}(s_P) - \bar{y}(s)$$

In both the cases, the error bound is quite close to $\bar{y}(s_P) - \bar{y}(s)$, which suggests that the proportion among non-users is much closer to 0 than that observed among the mobile phone users. They illustrate the possible magnitude of large MNO selection errors given similar non-user proportion f, where $\hat{\theta}$ refers to any

genuine y-value that can be obtained from the MNOs.

Next, if $\hat{\theta} < \theta$, then the selection error is most extreme if $y_k = y_{max} = \max_{k \in U} y_k$ for any $k \in U \setminus P$, such that

$$|\theta - \hat{\theta}| \ \leq \ \frac{1}{N}\Big(N(1-f)\hat{\theta} + N f y_{max}\Big) - \hat{\theta} = f(y_{max} - \hat{\theta}) \tag{14.2}$$

To illustrate, let $\hat{\theta} = 0.496$ for the female proportion, where $y_{max} = 1$ by definition and $f = 3.6\%$ as observed in s. It follows that

$$f(y_{max} - \hat{\theta}) = 1.8\% \gg 49.8\% - 49.6\% = 0.2\%$$

where the bound considerably exceeds the actual $\theta - \hat{\theta}$ of about 0.2% here. Note that the female proportion illustrates the possible magnitude of small MNO selection errors of any y-value obtained from the MNOs directly.

QR error bound Let $W_g = N_g/N$. Let $f = \sum_g W_g f_g$, where f_g is the non-user proportion in U_g. If $\hat{\theta}_g \geq \theta_g$ for any $g = 1, ..., G$, then

$$|\hat{\theta}_{qr} - \theta| = \sum_g W_g(\hat{\theta}_g - \theta_g) \ \leq \ \sum_g W_g f_g \hat{\theta}_g$$

similarly as the bound $f\theta$ given by (14.1). If $\hat{\theta}_g \leq \theta_g$ for any $g = 1, ..., G$, then

$$|\theta - \hat{\theta}_{qr}| = \sum_g W_g(\theta_g - \hat{\theta}_g) \ \leq \ \sum_g W_g f_g (y_{g,max} - \hat{\theta}_g)$$

where $y_{g,max}$ is the maximum y-value in U_g.

QR effect One can define a *relative effect (REff)* of the QR-estimator $\hat{\theta}_{qr}$ against the unadjusted $\hat{\theta}$ as the ratio of their error bounds. Let

$$\gamma = \begin{cases} \frac{\sum_g W_g f_g \hat{\theta}_g}{f\hat{\theta}} & \text{if } \hat{\theta}_g \geq \theta_g, \forall g \\ \frac{\sum_g W_g f_g (y_{g,max} - \hat{\theta}_g)}{f(y_{max} - \hat{\theta}_g)} & \text{if } \hat{\theta}_g \leq \theta_g, \forall g \end{cases} \tag{14.3}$$

We obtain $\hat{\gamma}$ on replacing $\{(W_g, f_g, \hat{\theta}_g)\}$ by their sample estimates.

Table 14.4 illustrates the use of (14.3), in terms of the commuter, student and female proportions. The error of user sample $\bar{y}(s_P)$ and its bound are given first. Note that in case of negative error (such as Female), the bound is given for its absolute value. The error of QR-estimator $\bar{y}_{qr}(s_P)$ and its bound are given next, which may vary with the post-strata, either by Age (6 groups) or Emp in addition (12 groups). Finally, the relative error

$$\frac{\bar{y}_{qr}(s_P) - \bar{y}(s)}{\bar{y}(s_P) - \bar{y}(s)}$$

and the REff by (14.3) are given, and whether the condition in (14.3) regarding $\hat{\theta}_g$ vs. θ_g is satisfied. Notice that, in practice, the condition needs to be assumed

Table 14.4: Error bound and REff, all in %. Condition satisfied if $\hat{y}_g(s_P) \geq \bar{y}_g(s)$ for Commute, Student, or if $\hat{y}_g(s_P) \leq \bar{y}_g(s)$ for Female, for $g = 1, ..., G$.

Outcome	Commute		Student		Female	
Post-strata	Age	(Age, Emp)	Age	(Age, Emp)	Age	(Age, Emp)
$\bar{y}(s_P) - \bar{y}(s)$	1.06		0.17		-0.02	
Bound of $\bar{y}(s_P)$	1.29		0.22		1.81	
$\bar{y}_{qr}(s_P) - \bar{y}(s)$	0.55	-0.01	0.07	0.24	-0.01	-0.00
Bound of $\bar{y}_{qr}(s_P)$	0.78	0.23	0.12	0.30	1.87	1.95
Relative error	51.8	-0.65	41.1	>100	68.6	22.2
REff	60.6	17.8	55.4	>100	>100	>100
Condition	Yes	No	Yes	Yes	No	No

when one applies (14.3) since its truth is unknown.

The results can be summarised as follows. First, REff >100% is a reliable indicator that the QR-estimator may not be an improvement by the given post-stratification, such as when using (Age, Emp) for student. Next, the REff can be a good indication of the actual relative error of $\hat{\theta}_{qr}$ compared to $\hat{\theta}$ when the condition of (14.3) is satisfied, such as REff 60.6% for relative error 51.8% for commuter given the post-strata by Age. However, even when the condition is not exactly satisfied, REff can still indicate whether the QR-estimator can improve the unadjusted estimator, e.g. REff 17.8% indicates correctly that the QR-estimator by (Age, Emp) is a large improvement for commuter.

14.4 Lack of features

As long as one cannot utilise secure multiparty computation, lack of features is likely to be the case in many applications, since QR-estimation must be limited to the features available to the MNOs even though better adjustment features are available at the OSA. It is therefore of interest to examine the loss of accuracy due to lack of features as a necessary cost caused by the absence of secure multiparty computation facility.

The post-stratified QR-estimator by (Age, Cent) is applicable if these features known to the MNOs in the absence of user ambiguity. To see the loss of gains, which could be achieved if it is possible to utilise additional features that are unknown to the MNOs, one may consider the other QR-estimates in Table 14.2 which depend on Emp, Nat or Edu as well. For each of the first 3 outcomes, there exists clearly such QR-estimates that are either approximately unbiased or have a much small bias, such as those marked in orange; the error for Female is small anyway. This demonstrates the gains of confidential computing settings, which enable flexible QR-estimation that can use any relevant features available in sources external to the MNOs.

However, despite the bias, a QR-estimator may be fit-for-purpose if standard survey sampling is too costly to achieve the same accuracy. Take the unbiased Horvitz-Thompson (HT) estimator based on survey sampling, denoted by $\hat{\theta}_{HT}$. To achieve the same mean squared error (MSE) as the QR-estimator, the SE of

$\hat{\theta}_{HT}$ would need to be equal to the error of $\hat{\theta}_{qr}$, assuming that the variance of the latter is negligible based on the MNO big data. In other words, the break-even point occurs at the sample size yielding

$$V(\hat{\theta}_{HT}) = (\hat{\theta}_{qr} - \theta)^2$$

where $V(\hat{\theta}_{HT})$ is the variance of $\hat{\theta}_{HT}$. This provides a tangible measure of the cost saved by using the QR-estimator instead of survey sampling.

To illustrate with the commuter proportion, suppose the bias of the QR-estimator by (Age, Cent) is 0.55% (as in Table 14.3). One can calculate the SE of the HT-estimator as the size of a simple random household sample varies. The sample size that achieves the same MSE is found to be close to 10^5. In Norway, this is about the same sample size as the Labour Force Survey, which is the largest continuous sample survey in Norway.

Auditing One can use audit sampling (Zhang, 2021a) to assess the error of a given big-data estimate. For instance, in practice, one needs to estimate the MSE of $\hat{\theta}_{qr}$,

$$E\{(\hat{\theta}_{HT} - \hat{\theta}_{qr})^2\} = (\theta - \hat{\theta}_{qr})^2 + V(\hat{\theta}_{HT})$$

where $V(\hat{\theta}_{HT})$ is the sampling variance of $\hat{\theta}_{HT}$. For the unbiased estimator of MSE($\hat{\theta}_{qr}$) to be equal to the unbiased estimator $\hat{V}(\hat{\theta}_{HT})$ of $V(\hat{\theta}_{HT})$, we need

$$(\hat{\theta}_{HT} - \hat{\theta}_{qr})^2 - \hat{V}(\hat{\theta}_{HT}) = \hat{V}(\hat{\theta}_{HT}) \quad \Leftrightarrow \quad (\hat{\theta}_{HT} - \hat{\theta}_{qr})^2 = 2\hat{V}(\hat{\theta}_{HT})$$

Other forms of valid design-based auditing inference are possible as well. For instance, for a significance test of $H_0 : \hat{\theta} = \theta$, where $\hat{\theta}$ is the direct mobile phone user estimator, the p-value of 0.05 is achieved if $\hat{\theta}_{HT}$ deviates from $\hat{\theta}$ by about 2 times SE($\hat{\theta}_{HT}$). Since the selection error of $\hat{\theta}$ for commuter proportion (Table 14.3) is about twice that of the QR-estimator above, the same household sample size mentioned above is needed for this purpose.

MNO misclassification In case a desirable feature for post-stratification is unknown to the MNOs, suppose it is still possible for the MNOs to generate a proxy feature, which however would cause a misclassification error whenever the proxy feature differs to the true feature for a given person.

For a simple exploration of the matter, take the commuter proportion, for which QR estimation by Emp is desirable (Table 14.2) but the feature is unknown to the MNOs. Suppose now the MNOs generates Emp* based on the signal data, such that the proportion of Emp* is equal to the official statistics on Emp, where the event Emp* $\neq$ Emp occurs independently of Emp, such that the extent of misclassification can be controlled by the overall misclassification rate alone. Let the resulting QR-estimator be

$$\bar{y}^*_{qr}(s_P) = \sum_{g=1}^{G} \frac{n_g(s)}{n(s)} \bar{y}^*_g(s_P)$$

where $\bar{y}^*_g(s_P)$ is the user sample means in the post-strata by Emp* instead of

Emp, while $n_g(s)/n(s)$ refers to the post-strata by Emp as it should be.

Table 14.5: QR-estimator $\bar{y}^*_{qr}(s_P)$ by misclassification rate r, all in %

			$\bar{y}^*_{qr}(s_P)$ by Emp*			
$\bar{y}(s)$	$\bar{y}(s_P)$	$\bar{y}_{qr}(s_P)$	$r = 0.5$	$r = 1$	$r = 5$	$r = 10$
34.46	35.52	34.46	34.64	35.27	35.55	35.91

Table 14.5 shows the QR-estimates given the overall misclassification rate r, all of which are calculated by simulation of Emp* and the simulation errors are negligible compared to any of the differences seen in the table. The QR-estimator is initially affected by about as much as the misclassification rate, e.g. $34.64/34.46 - 1 = 0.5\%$ given $r = 0.5\%$, and gradually to a lesser extend, e.g. $35.91/34.46 - 1 = 4.2\% < r = 10\%$. However, the resulting QR-estimate becomes nearly equal to the unadjusted user sample mean $\bar{y}(s_P)$ given a modest 5% error of Emp* randomly. In conclusion, using MNO proxy feature is easily a haphazard alternative to enabling secure multiparty computation.

14.5 User ambiguity

Although we cannot study the potential effects of device duplication here since we do not have the number of devices associated with the user sample, user ambiguity is a more fundamental problem that can be explored by stipulating various user-contractor connections. The matter is of critical importance to effective uses of MNO data, because it can potentially remove the need for additional data collection about the use-contractor connections (e.g. by ad hoc sample surveys) and the potential errors induced by any adopted adjustment method in addition.

Table 14.6 shows the results if QR-estimation treats each user aged 16-18 in the same way as the eldest person in the household (as the stipulated service contractor). An error may occur whenever a user aged 16-18 is not the eldest in the household. The estimators $\bar{y}(s)$ and $\bar{y}(s_P)$ are the same as in Table 14.2, which are reproduced here for easy comparison.

Clearly, user ambiguity can result in erratic effects of QR adjustment across different y-outcomes. The selection error of $\bar{y}(s_P)$ may be greatly increased or sometimes reduced by a given QR estimator, depending on how the user misclassification error is confounded with the selection error in the different post-strata. Moreover, user misclassification tends to increase the range of the various QR estimates for each given outcome.

Additional adjustment

We can examine the SUD estimator (8.6). In the above experiment of user ambiguity due to persons aged 16-18, there are no duplicated devices such that $\xi_g \equiv 1$ for $1 \leq g \leq G$. The probability matrix $[\phi_{lg}]$ has only multiple non-zero values in any column g if it involves persons aged 16-18; there is only a diagonal element 1 in the other columns.

Table 14.6: Estimates (QR if unstated), standard errors in parentheses

	Income		Commute (%)		Student (%)		Female (%)	
$\bar{y}(s)$	399011	(553)	34.5	(0.069)	5.99	(0.028)	49.8	(0.049)
$\bar{y}(s_P)$	404186	(578)	35.5	(0.072)	6.16	(0.029)	49.6	(0.052)
(I) Age	388405	(545)	34.7	(0.073)	7.08	(0.036)	49.6	(0.057)
(II) Cent	404063	(576)	35.5	(0.072)	6.15	(0.029)	49.6	(0.052)
(III) Emp	396522	(565)	33.2	(0.067)	6.49	(0.031)	49.9	(0.053)
(IV) Nat	404317	(579)	35.5	(0.072)	6.16	(0.029)	49.6	(0.052)
(V) Edu	397075	(551)	35.0	(0.071)	6.13	(0.029)	49.4	(0.053)
(I)×(II)	388243	(544)	34.7	(0.073)	7.08	(0.036)	49.6	(0.057)
(I)×(II)×(III)	385089	(538)	33.2	(0.068)	8.35	(0.047)	49.7	(0.059)
(I)×(II)×(IV)	385651	(536)	33.6	(0.077)	8.55	(0.126)	49.8	(0.111)
(I)×(II)×(V)	385746	(538)	34.1	(0.072)	6.96	(0.042)	49.6	(0.065)

Table 14.7: Adjustment by models (8.1) and (8.5) given post-strata by Emp or Age. Post-stratum means $[\bar{z}_l^*]$ by contractors, $[\bar{z}_g]$ by users, $[\hat{\bar{z}}_g]$ estimated.

	Commute			Student		
$[\phi_{lg}]$ by Emp	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.928 0.002	0.002	0.000	0.000	0.112	0.171	0.112
0.072 0.998	0.533	0.555	0.555	0.036	0.000	0.033
Overall	0.332	0.345	0.345	0.065	0.065	0.063
$[\phi_{lg}]$ by Age	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.446 0 0 0 0 0	0.338	0.216	0.338	0.307	0.520	0.307
0.004 1 0 0 0 0	0.422	0.422	0.422	0.082	0.081	0.081
0.300 0 1 0 0 0	0.434	0.463	0.443	0.066	0.009	0.044
0.238 0 0 1 0 0	0.395	0.413	0.398	0.046	0.001	0.028
0.008 0 0 0 1 0	0.096	0.096	0.095	0.004	0.000	0.003
0.003 0 0 0 0 1	0.024	0.024	0.023	0.003	0.000	0.002
Overall	0.347	0.350	0.350	0.071	0.061	0.061

Table 14.7 illustrates the adjustments for commuter or student. The post-strata are either given by Emp or Age. The matrix $[\phi_{lg}]$ is directly calculated from $\{\nu_{kj} : j \in s_P, k \in s\}$, which is the same whether the y-outcome is commuter or student. The means $\bar{z}_l^*$ refer to the post-strata by contractors, which is someone else for about 55.6% of the persons aged 16-18; see ϕ_{11} by Age in Table 14.7. The means $\bar{z}_g$ are calculated for the true user post-strata, where $\bar{z}_g \equiv \bar{y}_g$ here since $\xi_g \equiv 1$ in this experiment. The means $\hat{\bar{z}}_g$ are the estimates of $\bar{z}_g$ by the model (8.5), specified as the last fraction in (8.6). The 'overall' mean of $\bar{z}_g$ is the genuine QR-estimator, that of $\bar{z}_l^*$ directly is the naïve QR-estimator, and the SUD-estimator (8.6) is based on $\hat{\bar{z}}_g$.

Clearly, the model (8.5) can yield $[\hat{\bar{z}}_g] \approx [\bar{z}_g]$ in some situations, such as using Emp for commuter. In all the other cases, however, $[\hat{\bar{z}}_g]$ is often closer to the observed $[\bar{z}_g^*]$ than the actual $[\bar{z}_g]$, which means that the model (8.5) does not hold generally. Nevertheless, it is intriguing to observe that the SUD-estimator

(8.6) is quite close to the genuine QR-estimator in all the cases, whether or not $[\hat{\bar{z}}_g] \approx [\bar{z}_g]$ actually. This shows that the condition (8.7) for robust QR-estimation holds approximately here.

Additional user ambiguity scenarios

The robustness of the SUD-estimator (8.6) noted above is of value in practice. That is, even as user ambiguity causes generally bias for group-specific MNO mean estimator of θ_g for population groups $g = 1, ..., G$, the SUD-estimator of the population mean θ may still be close to the genuine QR-estimator in the absence of user ambiguity, provided the condition (8.7) holds approximately.

Table 14.8: Scenario Father. SUD-estimator (8.6) given post-strata by Emp or Age. Post-stratum means $[\bar{z}_l^*]$ by contractors, $[\bar{z}_g]$ by users, $[\hat{\bar{z}}_g]$ estimated.

	Commute			Student		
$[\phi_{lg}]$ by Emp	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.938 0.002	0.002	0.000	0.000	0.122	0.171	0.122
0.062 0.998	0.536	0.555	0.555	0.031	0.000	0.027
Overall	0.334	0.345	0.345	0.065	0.065	0.063
$[\phi_{lg}]$ by Age	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.505 0.00 0.0 0 0 0	0.263	0.216	0.262	0.444	0.520	0.444
0.000 0.98 0.0 0 0 0	0.422	0.422	0.422	0.081	0.081	0.081
0.262 0.01 1.0 0 0 0	0.441	0.463	0.456	0.052	0.009	0.021
0.228 0.01 0.0 1 0 0	0.399	0.413	0.407	0.038	0.001	0.011
0.004 0.00 0.0 0 1 0	0.096	0.096	0.096	0.002	0.000	0.001
0.000 0.00 0.0 0 0 1	0.024	0.024	0.024	0.000	0.000	0.000
Overall	0.345	0.350	0.350	0.075	0.061	0.061

Table 14.9: Scenario Mother. SUD-estimator (8.6) given post-strata by Emp or Age. Post-stratum means $[\bar{z}_l^*]$ by contractors, $[\bar{z}_g]$ by users, $[\hat{\bar{z}}_g]$ estimated.

	Commute			Student		
$[\phi_{lg}]$ by Emp	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.928 0.004	0.004	0.000	0.000	0.114	0.171	0.115
0.072 0.996	0.533	0.555	0.555	0.035	0.000	0.032
Overall	0.333	0.345	0.345	0.065	0.065	0.063
$[\phi_{lg}]$ by Age	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.389 0.00 0.0 0 0 0	0.293	0.216	0.292	0.386	0.520	0.386
0.006 0.97 0.0 0 0 0	0.421	0.422	0.422	0.082	0.081	0.081
0.444 0.02 1.0 0 0 0	0.428	0.463	0.447	0.079	0.009	0.037
0.161 0.01 0.0 1 0 0	0.402	0.413	0.407	0.029	0.001	0.012
0.000 0.00 0.0 0 1 0	0.096	0.096	0.096	0.000	0.000	0.000
0.000 0.00 0.0 0 0 1	0.024	0.024	0.024	0.000	0.000	0.000
Overall	0.344	0.350	0.350	0.075	0.061	0.061

To explore the matter further, we consider now some additional scenarios of contractor-user connections among members of the same household.

- *Father* Let a father with at least one child of age 18 or less in the household be the contractor of all his children regardless their age.
- *Mother* Let a mother with at least one child of age 18 or less in the household be the contractor of all her children regardless their age.
- *Household-I* Let the eldest person in the household be the contractor of all the household members regardless their age or relationship otherwise.
- *Household-II* Let the person with the highest education level in the household be the contractor of all the household members.

While the first two scenarios may be similar to the scenario explored previously, the last two scenarios can be regarded as extreme cases where the number of contractors is minimised as long as contractor-user connections are restricted to exist only within the households.

Tables 14.8 and 14.9 give the results under the first two scenarios above, which are arranged in the same way as Table 14.7. We note that the SUD-estimate by (8.6) remains close to the genuine QR-estimate (unaffected by user ambiguity) in both the cases, regardless the target y and the post-strata. Next, on closer inspection of the flow probabilities $[\hat{\phi}_{lg}]$ introduced by the model (8.5), we can see that these additional scenarios caused only very small changes to these probabilities given post-stratification by Emp, such that the SUD-estimates must be close to those in Table 14.7. Meanwhile, although the changes are not negligible in the case of post-stratification by Age (e.g. $\hat{\phi}_{11}$ is 0.446 in Table 14.7, 0.505 in Table 14.8 and 0.389 in Table 14.9), it has transpired that the condition (8.7) holds approximately in all these scenarios, such that the SUD-estimator remain close to the genuine QR-estimator.

Table 14.10 gives the results under the scenario Household-I, which is more striking than the other scenarios earlier. The flow probabilities of model (8.5) deviate now considerably from the identity matrix whether by Emp or Age, where all the lower-triangle elements are non-zero for QR-adjustment by Age and the diagonal elements are much less dominant than before. Likewise, Table 14.11 gives the results under the scenario Household-II, where the flow probability matrix of model (8.5) is no longer lower-triangle and none of the diagonal elements is 1 any more.

For QR-estimation by Emp, the model (8.5) works well for commuter since $[\hat{\bar{z}}_g] \approx [\bar{z}_g]$ in both the scenarios, but it clearly does not hold for student since $[\hat{\bar{z}}_g]$ is closer to $[\bar{z}_l^*]$ than to the genuine $[\bar{z}_g]$ in both Table 14.10 and Table 14.11. This demonstrates that the assumption (8.5) of non-informative contractor-user connections cannot hold regardless the target outcome. Nonetheless, the overall SUD-estimates remain close to the corresponding genuine QR-estimates in both the scenarios.

For QR-adjustment by Age, the model (8.5) holds neither for commuter nor student. However, the overall SUD-estimates remain close to the genuine QR-estimates in all the cases, exhibiting remarkable robustness in these extreme

Table 14.10: Scenario Household-I. SUD-estimator (8.6) by post-strata Emp or Age. Post-stratum means $[\bar{z}_l^*]$ by contractors, $[\bar{z}_g]$ by users, $[\hat{\bar{z}}_g]$ estimated.

	Commute			Student		
$[\phi_{lg}]$ by Emp	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.802 0.073	0.078	0.000	0.001	0.086	0.171	0.092
0.198 0.927	0.495	0.555	0.554	0.050	0.000	0.044
Overall	0.337	0.345	0.345	0.063	0.065	0.063
$[\phi_{lg}]$ by Age	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.16 0.00 0.00 0.00 0.00 0	0.280	0.216	0.280	0.450	0.520	0.450
0.02 0.76 0.00 0.00 0.00 0	0.420	0.422	0.421	0.091	0.081	0.089
0.40 0.11 0.87 0.00 0.00 0	0.432	0.463	0.455	0.074	0.009	0.019
0.40 0.12 0.11 0.92 0.00 0	0.407	0.413	0.416	0.055	0.001	0.005
0.01 0.01 0.02 0.07 0.92 0	0.143	0.096	0.073	0.007	0.000	0.000
0.00 0.00 0.01 0.01 0.08 1	0.051	0.024	0.013	0.004	0.000	0.000
Overall	0.342	0.350	0.350	0.088	0.061	0.061

Table 14.11: Scenario Household-II. SUD-estimator (8.6) by post-strata Emp or Age. Post-stratum means $[\bar{z}_l^*]$ by contractors, $[\bar{z}_g]$ by users, $[\hat{\bar{z}}_g]$ estimated.

	Commute			Student		
$[\phi_{lg}]$ by Emp	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.775 0.057	0.066	0.000	0.001	0.095	0.171	0.102
0.225 0.943	0.489	0.555	0.554	0.046	0.000	0.039
Overall	0.329	0.345	0.345	0.065	0.065	0.063
$[\phi_{lg}]$ by Age	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$	$[\bar{z}_l^*]$	$[\bar{z}_g]$	$[\hat{\bar{z}}_g]$
0.21 0.00 0.01 0.01 0.00 0.00	0.275	0.216	0.228	0.351	0.520	0.446
0.04 0.86 0.06 0.05 0.01 0.01	0.421	0.422	0.424	0.079	0.081	0.083
0.45 0.06 0.87 0.05 0.02 0.01	0.428	0.463	0.466	0.076	0.009	0.024
0.29 0.07 0.05 0.85 0.07 0.03	0.392	0.413	0.416	0.046	0.001	0.007
0.01 0.01 0.01 0.03 0.85 0.07	0.113	0.096	0.076	0.004	0.000	0.000
0.00 0.00 0.00 0.00 0.04 0.89	0.035	0.024	0.014	0.002	0.000	0.000
Overall	0.343	0.350	0.350	0.076	0.061	0.061

scenarios of contractor-user connections, where the face-value QR-estimate is considerably biased for student, i.e. 0.088 or 0.076 vs. 0.061.

In summary, the experiment here has demonstrated that the model (8.5) does not hold generally, but it is still possible for the condition (8.7) to hold approximately in many scenarios of user ambiguity (including those examined above), provided which the SUD-estimator (8.6) can remain close to the genuine QR-estimator (unaffected by user ambiguity).

14.6 Conclusions

We have reported a proof-of-concept experiment on QR-estimation using a household sample and features obtained from various register sources, where the mobile phone users are identified via automatic mobile phone number search (as one would have done for sample surveys in practice). The relevant methods and analyses are explained and illustrated, so that the experiment can be replicated or adapted by others.

The experiment results have provided (i) clear evidence that MNO data are often subject to selection errors, (ii) relevant indications of the maximum MNO selection errors that can be encountered in reality, as well as (iii) the likely gains of QR adjustment by features known for the MNO users, such as age, registered home municipality. The selection error bounds and REffs will be helpful in applications in this respect.

Moreover, the experiment has highlighted the gains that can be achieved by secure multiparty computation in situations where, despite of QR-estimation, the selection error is deemed unacceptably large for the users or stakeholders of the relevant statistics, which however can be removed by other features that readily exist in sources external to the MNOs.

User ambiguity can still present a major challenge in countries where the user information is not required or compulsory for mobile phone services. Although the SUD-estimator (8.6) is remarkably robust in all the scenarios explored in our experiment, where the scenario Household is quite extreme, such robustness cannot be taken for granted for all conceivable device-user-contractor connections, because we have shown that the simplifying model (8.5) of non-informative connections does not hold generally.

A possibility that may help to reduce the error caused by device duplication and user ambiguity is to limit the MNO counts to the devices that are almost always active. This allows one to assume that these devices are ever-present regardless the target outcome variable, although it can increase the selection effect since fewer device users would be included in the MNO counts. This option is however not possible to be explored in this experiment.

Where user ambiguity is not a systemic issue, QR-estimation is the most cost-effective and broadly applicable approach for producing official statistics based on MNO macro data, even without the possibility of secure multiparty computation that can further improve the selection error adjustment. Audit sampling (Zhang, 2021a) can be applied to inaugurate and periodically renew the official statistics status. The QR approach may be studied and, if viable, applied to other big data sources such as digital payment transactions.

Chapter 15

QR pseudo experiment

15.1 Introduction

The Aspect of Daily Live (ADL) is an annual sample survey that aims to investigate the habits and problems that people in Italy face on a daily basis.

The survey is based on a two-stage probabilistic stratified sample; the primary sampling units (PSUs) are municipalities, while the secondary sampling units (SSUs) are households (HHs). The PSUs are stratified by region (NUTS2; 21 regions) and by type of municipality (6 categories, also taking into account the number of inhabitants). Within each region, larger municipalities are separated into a *take-all* stratum (sampled with certainty), while the remaining municipalities are further stratified according to their size (*take-some* strata) and within each stratum a sample of PSUs is selected with probability proportional to their size. At the second stage, a random sample of HHs is selected (at least 24 HHs per PSU) within each sampled municipality (PSU).

A sample of about $24\,000$ Italian HHs was selected for year 2023, but the responding ones were $18\,565$ (about 75%). The responding HHs correspond to $41\,772$ individuals (1610 with age<=5 and $40\,162$ with age>5).

To compensate for unit nonresponse the adopted strategy consists in a correction of the HH *base* weight (obtained as the inverse of the inclusion probability) with response rate in the homogeneous response group (with respect to the response propensity) to which the HH belongs. The corrected weights are then calibrated to reproduce known population totals at regional level (totals of individuals by type of municipality, gender crossed with eight age groups and citizenship crossed with gender; 24 known totals). The calibration uses a truncated logarithmic distance function and is constrained to assign an equal calibrated weight to all the individuals belonging to each HH.

All persons belonging to the sampled HHs are interviewed, except those aged less than or equal to 5 years. Data relating to the whole HH are provided by one of its adult members (age > 17). The main data on the HH and its members are collected through a CAPI interview, while the remaining individual data are collected through a PAPI self-administered questionnaire (but for those aged <14, a proxy response is provided by an adult member of the HH).

The ADL survey collects a high number of variables at both HH and personal level; the variables of interest for investigating coverage and use of mobile phone (MP) devices are:

- HH level: does the HH own a mobile phone? (No/Yes)
- HH level: if the HH owns a mobile phone device, how many MPs? (1-9)
- Person level: if a HH owns a mobile phone device, then each HH member (age>5) is asked about the "frequency" of usage of the mobile phone device (6 categories: 1="every day", 2="few times a week", 3="once a week", 4="a few times a month", 5="a few times a year", 6="never")

Other variables of interest at person's level are: gender, age, education level, professional status, main source of income, commuting for work or study, etc.

15.2 Households covered by mobile phones

There are 545 HHs (2.9%) which declared that they do not own a mobile phone; these are mainly HHs with an elderly member. The number of HHs in Italy without a MP is estimated to be 733 928 HHs (2.8%).

Table 15.1: Distribution of Households by ownership of mobile phone devices

Own MP	Resp. HHs	Estimated No. of HHs	Rel. freq. HH
No	545	733 928	0.03
Yes	18 020	25 467 504	0.97
Total	18 565	26 201 432	1.00

The information on MPs ownership permits to assess whether the subset of HHs with a MP device has characteristics that differ from the whole target population of Italian HHs.

The comparison by geographical areas does not highlight differences.

Table 15.2: Distribution of HHs with and without MP devices by geographical area (%)

Area	HHs with MP	SE	All HHs	Sampl. err.	Difference
N-W	28.02	*0.44*	27.94	*0.39*	0.09
N-E	19.79	*0.36*	19.77	*0.33*	0.02
Center	20.48	*0.37*	20.31	*0.33*	0.17
South	21.07	*0.38*	21.29	*0.34*	-0.22
Islands	10.63	*0.26*	10.69	*0.25*	-0.06
Total	100.00		100.00		0.00

Also the type of municipality does not show marked differences.

Table 15.3: Distribution of HHs with and without MP devices by type/size of municipality (%)

Type	HH with MP	SE	All HH	SE	Difference
Metrop	16.42	*0.33*	16.26	*0.30*	0.16
Surr. Metr.	13.93	*0.30*	13.84	*0.28*	0.09
<=2000	5.26	*0.18*	5.35	*0.18*	-0.09
2001-10000	21.80	*0.38*	22.04	*0.35*	-0.24
10001-50000	25.29	*0.42*	25.25	*0.37*	0.04
>50000	17.29	*0.34*	17.25	*0.31*	0.04
Total	100.00		100.00		0.00

The comparison by number of members of the HH shows a slight difference for HHs consisting of only one person, which are under-represented (-1.4%) among those owning a MP device. This is due to the fact that one-person HHs are often made up of an elderly person who has difficulty in using technological devices such as the MPs.

Table 15.4: Distribution of HHs with and without MP devices by HH size (%)

HH size	HH with MP	SE	All HH	SE	Difference
1	34.26	*0.50*	35.69	*0.43*	-1.43
2	29.04	*0.45*	28.60	*0.39*	0.44
3	18.56	*0.35*	18.06	*0.32*	0.50
4	13.95	*0.30*	13.57	*0.28*	0.38
>4	4.20	*0.16*	4.08	*0.16*	0.12
Total	100.00		100.00		0.00

15.3 People in households with mobile phones

Unfortunately, the survey does not collect information on MP ownership at the individual level, but only asks about frequency of use; this is a question in the self-administered PAPI version of the individual questionnaire (only for individuals aged >5; but for those aged <14 it is a proxy response). It is a categorical variable with six categories ranging from "use every day" to "never use" (1='every day', 2='a few times a week', 3='once a week', 4='a few times a month', 5='a few times a year', 6='never')[1].

Focusing on the subset of surveyed HHs that reported owning an MP device, it is possible to compare the distribution of frequency of use by some socio-demographic characteristics. "Daily" users (category (1='every day') are almost the 85% of individuals in HHs with one or more MP devices. A slightly higher frequency of use (+1%) is observed for men.

[1]Note that about 900 people in HHs with an MP refused to answer this question, so we decided to impute it using a two-stage procedure based on random forest and then random hotdeck imputation.

Table 15.5: Distributions of frequency of usage of MP devices by gender (%)

Frequency of usage	M	F
1	86.7	85.7
2	6.2	6.4
3	0.8	0.7
4	1.0	1.0
5	0.8	1.0
6	4.4	5.1
Total	100.0	100.0

An analysis of usage by age shows that 95% of people in 15-54 age group use a MP device daily. The proportion of daily users falls below 60% for the age groups at the extremes of the distribution.

Table 15.6: Distributions of MP device usage frequency by age groups (%)

	Frequency of usage						
Age group	1	2	3	4	5	6	Total
(5,10]	36.2	31.7	6.3	5.6	3.6	16.6	100
(10,14]	88.3	6.0	0.8	0.8	0.4	3.6	100
(14,17]	97.0	1.6	0.3	0.0	0.3	0.8	100
(17,24]	98.0	0.7	0.3	0.1	0.3	0.5	100
(24,34]	97.1	1.5	0.1	0.3	0.2	0.7	100
(34,54]	96.2	2.2	0.3	0.2	0.4	0.8	100
(54,65]	91.8	4.8	0.4	0.6	0.3	2.1	100
(65,74]	81.6	9.0	0.5	1.2	1.2	6.4	100
(74,105]	55.2	15.9	1.7	3.0	3.4	20.7	100

Let's now focus the attention to people with age greater than 14 years, as it corresponds to the target population of many surveys in official statistics (36 844 individuals in the sub-sample of responding HHs). The subset of $n_{sp} = 31\,748$ (86.2%) persons that: (a) live in a HH that owns one or more MP devices; and, (b) declare to use the MP device every day; can be considered as a sample representative of the users of MP devices (with age > 14) and, more generally, as the sample that could be obtained by drawing units from a frame of subscribers to MP services, in the absence of two major issues:

- *device duplication*: more MP devices used by the same person; and
- *user ambiguity*: a MP device used by a person other than the person who signed the SIM card contract.

Similarly, the statistical outputs provided by this sub-sample could be considered as those that would be obtained by analyzing data from a mobile network operator (MNO), always in the absence of issues related to user ambiguity and device duplication.

The comparison by geographical area does not show marked differences.

Table 15.7: Distributions of daily users of MP devices by geographical area compared to the corresponding marginal distribution estimated from the whole sample (%; only people with age>14)

Area	MP daily users	SE	All	SE	Difference
N-W	27.29	*0.30*	26.90	*0.24*	0.38
N-E	19.54	*0.26*	19.59	*0.21*	-0.04
Center	19.90	*0.26*	19.89	*0.21*	0.01
South	22.46	*0.28*	22.77	*0.22*	-0.31
Islands	10.80	*0.19*	10.85	*0.16*	-0.04
Total	100.00		100.00		0.00

Table 15.8: Distributions of daily users of MP devices by type/size of municipality compared to the corresponding marginal distributions of all individuals (only people with age>14; %)

Type/size municip.	MP daily users	SE	All	SE	Difference
Metrop	15.44	*0.23*	15.14	*0.19*	0.30
Surr. Metr.	14.45	*0.22*	14.26	*0.18*	0.20
<=2000	4.79	*0.14*	5.14	*0.12*	-0.35
2001-10000	22.09	*0.28*	22.55	*0.22*	-0.45
10001-50000	26.01	*0.30*	25.98	*0.23*	0.03
>50000	17.22	*0.24*	16.94	*0.19*	0.27
Total	100.00		100.00		0.00

There are no major differences when comparing by gender and by citizenship. While there are some differences in the estimates of educational attainment, the proportion of people with a secondary or tertiary education is actually higher among the users of MP devices compared to the total target population (this implies an under-coverage of around 5% of people with no education or primary education). There are also differences in occupational status, with MP users over-representing those in employment and under-representing those in retirement, while there are no differences for students. MP users over-represent those whose main income is from dependent employment, while there are no differences for those whose main income is from self-employment. Finally, MP users over-represent commuters for work, while the difference is negligible for commuters for study.

Table 15.9: Estimates (%) based on MP users or all individuals age>14

	MP daily users	SE	All	SE	Difference
Gend=F	50.86	*0.40*	51.55	*0.31*	-0.69
Gend=F, Age 18-64	38.16	*0.34*	34.74	*0.26*	3.41
IT citizen	91.97	*0.59*	92.40	*0.40*	-0.43
University degree	19.79	*0.24*	17.85	*0.20*	1.94
Secondary school	42.36	*0.36*	39.20	*0.28*	3.16

Parameters of interest	MP daily users	SE	All	SE	Difference
Employed	49.89	*0.40*	44.99	*0.29*	4.90
Student	9.55	*0.17*	8.48	*0.15*	1.07
Retired	16.84	*0.22*	21.40	*0.21*	-4.56
Employee	40.07	*0.35*	36.03	*0.27*	4.04
Self-employed	9.71	*0.17*	8.86	*0.15*	0.86
Commuter	31.00	*0.30*	27.78	*0.24*	3.22
- Study	5.35	*0.13*	4.75	*0.11*	0.60
- Work	25.65	*0.27*	23.03	*0.22*	2.62

15.4 Adjustment by post-stratification

As shown in the previous table, the sub-sample of users of MP devices introduces a *selection error* in the estimation of educational attainment, the number of persons with a job, the number of persons commuting to work and the number of persons whose main income is from form-dependent work. To compensate for this error, a *quasi-randomisation* (QR) estimation approach can be applied. The starting point is simply a *post-stratification* (PS). We tested different post-strata settings obtained according to different combinations of few variables:

a) geographical area (RIP, 5 categories);

b) type and size of municipality (DOM, 6 categories)

c) gender (2 categories)

d) age in classes (15-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, >74);

e) gender crossed with classes of age (16 cells)

f) RIP*DOM (30 cells);

g) RIP*DOM*GENDER (60 cells);

h) RIP*DOM*AGE.CL (240 cells)

The variables chosen for the PS are those that can be expected to "accompany" the data that could be provided by a MNO (assuming there are no errors due to user ambiguity and device duplication). In our case the totals used in PS are estimated directly from the whole sample and in most of the cases correspond to the true population totals, as many of the variables used in post-stratification are also used in survey weights calibration.

The following table compares the estimated proportions provided by (i) the full sample of adults (reference), (ii) the sub-sample of people living in HHs with an MP device and using it every day, and (iii) the PS estimator applied to the latter sub-sample with different PS settings.

Table 15.10: Estimates (%) based on all, MP device users, PS (a) - (h). Gender=F* for female and age 18-64.

	All	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Gend=F	51.5	50.9	50.9	50.9	51.5	51.3	51.5	50.9	51.5	51.2
Gend=F*	34.7	38.2	38.6	38.6	39.1	34.9	34.8	38.6	39.1	34.9
IT citizen	92.4	92.0	92.0	92.0	92.0	92.6	92.6	92.0	92.0	92.6
University	17.8	19.8	20.0	19.9	20.0	19.1	19.1	19.9	19.9	19.0
Secondary	39.2	42.4	42.8	42.9	42.8	40.9	40.9	42.9	42.8	40.8
Employed	45.0	49.9	50.4	50.5	50.4	45.9	45.9	50.4	50.3	45.8
Student	8.5	9.6	8.5	8.5	8.5	8.5	8.5	8.5	8.5	8.5
Retired	21.4	16.8	17.0	17.0	17.0	22.0	21.8	17.0	17.0	21.9
Employee	36.0	40.1	40.5	40.5	40.5	36.7	36.8	40.5	40.4	36.7
Self-empl.	8.9	9.7	9.8	9.8	9.8	9.1	9.1	9.8	9.8	9.0
Commuter	27.8	31.0	30.8	31.0	30.8	28.3	28.3	31.0	30.9	28.3
- Study	4.7	5.3	4.9	4.9	4.9	4.8	4.8	4.9	4.9	4.8
- Work	23.0	25.7	25.9	26.1	25.9	23.5	23.5	26.1	26.0	23.5

The following table shows the differences between the various estimates based on MP users, with or without PS, and the corresponding ones obtained from the all the individuals (given in the first column of the previous table).

Table 15.11: Difference of estimates (%) by MP device users with or without PS (a) - (h) to that by all individuals.

	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Gend=F	-0.7	-0.7	-0.7	0.0	-0.3	0.0	-0.7	0.0	-0.3
Gend=F, age 18-64	3.4	3.9	3.9	4.4	0.2	0.1	3.9	4.4	0.2
IT citizen	-0.4	-0.4	-0.4	-0.4	0.2	0.2	-0.4	-0.4	0.2
University degree	1.9	2.2	2.1	2.2	1.2	1.2	2.1	2.1	1.1
Secondary school	3.2	3.6	3.7	3.6	1.7	1.7	3.7	3.6	1.6
Employed	4.9	5.4	5.5	5.4	0.9	0.9	5.4	5.3	0.8
Student	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Retired	-4.6	-4.4	-4.4	-4.4	0.6	0.4	-4.4	-4.4	0.5
Employee	4.0	4.5	4.5	4.4	0.7	0.7	4.4	4.4	0.7
Self-employed	0.9	1.0	1.0	0.9	0.2	0.2	1.0	0.9	0.2
Commuter	3.2	3.0	3.2	3.0	0.5	0.5	3.2	3.1	0.6
- Study	0.6	0.1	0.2	0.1	0.0	0.0	0.2	0.2	0.1
- Work	2.6	2.9	3.1	2.8	0.5	0.5	3.0	2.9	0.5

Selection error in estimating relative frequencies of some of the categories of the considered variables (education level, with occupation, main source of income, and commuting for work) is attenuated (but not removed) by the by PS (d) (by age classes), PS (e) (crossing gender and age classes) and PS (h) (crossing area, type/size of municipality and age classes). This latter PS setting seems to produce slightly better results but is also the one having many cells (240) and potentially the risk of empty or very small cells. Should this be the case,

however, one can easily modify the post-strata accordingly without losing the essential benefits from using all these post-stratification variables.

15.5 Assessing the effect of user ambiguity

To HHs declaring to own a MP device, the ADL survey asks about the number of MPs owned (from 1 to 9) and this information permits to get an estimate of the number of MPs owned by the Italian HHs (as shown later). The data on the number of owned MPs permits also to perform some experiments to asses the effect of user ambiguity, i.e. what would happen if instead of a sample of individuals (using or not the MP) we have a sample of MPs with the associated info about the subscriber of the SIM card within the device. This situation resembles the case of data that are expected to be provided by a MNO.

We assume that the subscriber of a SIM card contract must be over 17 years old (reflecting the main practice in Italy, although there may be some exceptions) and carry out three simulation experiments:

1. RND: select at random a number of adult HH members equal to the number of MP owned by the HH. Selection is done with replacement to tackle the case of more MPs than adult HH members. In addition, the selection, when possible, is limited to the HH members that declare to use MP more frequently (basically every day).

2. EDU: all the MPs owned by the HHs are assigned to one adult member chosen at random within those with highest education level and being a frequent user of MP (use it every day).

3. OCC: all MPs owned by the HHs are assigned to an adult member chosen at random among those who use the MP every day and are employed as managers or middle managers or employed with main income coming from self-employment; in the absence of an adult with this characteristic, we consider the educational level as in the EDU scenario.

The samples generated by the three simulations scenario return an estimate of the population size corresponding to the estimated number of "used" MPs provided by the ADL survey, i.e. $\hat{M}_U = 51\,710\,424$ (we discard MPs owned but never used by any HH member, $\hat{M}_{NU} = 749\,815$, as will be explained later).

In RND simulation experiment only PS involving the age in classes, i.e. (d), (e) and (h), makes it possible to reduce the effect of the selection error, as shown in the following table, which shows the differences between the achieved estimates and the reference ones obtained from the entire ADL survey.

Table 15.12: Difference of estimates (%) by MP device users with or without PS (a) - (h) to that by all individuals age>17.

	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Gend=F	-0.6	-0.6	-0.6	0.0	-0.2	0.0	-0.6	0.0	-0.2
gend=F, age in 18-64	3.4	3.5	3.4	3.9	0.1	0.0	3.4	3.9	0.1

Table 15.12: Difference of estimates (%) by MP device users with or without PS (a) - (h) to that by all individuals age>17.

	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
IT citizen	-0.5	-0.5	-0.5	-0.6	0.1	0.1	-0.5	-0.5	0.2
University degree	2.0	2.0	1.9	2.0	1.1	1.0	1.9	1.9	1.0
Secondary school	2.8	2.8	2.8	2.8	1.3	1.3	2.8	2.7	1.2
Employed	5.6	5.5	5.6	5.5	1.0	1.0	5.5	5.4	0.9
Student	0.3	0.3	0.3	0.3	0.1	0.0	0.3	0.3	0.1
Retired	-4.6	-4.6	-4.6	-4.6	0.5	0.4	-4.6	-4.7	0.4
Employee	4.5	4.4	4.5	4.5	0.7	0.7	4.4	4.3	0.6
Self-employed	1.1	1.0	1.1	1.0	0.2	0.2	1.0	1.0	0.2
Commuter	3.0	2.9	3.2	2.9	0.4	0.4	3.0	3.0	0.4
- Study	0.2	0.3	0.3	0.3	0.1	0.1	0.3	0.3	0.1
- Work	2.8	2.7	2.9	2.7	0.3	0.3	2.8	2.7	0.3

In EDU simulation experiment, the PS based on age groups still reduces the selection error but less than in the previous experiment; obviously, PS is effective to reduce error in estimating females in the age group 18-64 years and also the error in estimating retired people as well as commuters; on the contrary, it has no effect on the selection error in estimating the education level that is used to generate the experiment data (all the SIM cards in the HH are assigned to the adult HH member with the highest education level and being a frequent user of the MP device).

Table 15.13: Difference of estimates (%) by MP device users with or without PS (a) - (h) to that by all individuals age>17. Gender=F* for female and age 18-64.

	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Gend=F	0.7	0.7	0.7	0.0	0.4	0.0	0.7	0.0	0.4
Gend=F*	6.5	6.5	6.5	6.0	1.7	0.0	6.5	6.0	1.7
IT citizen	-0.5	-0.5	-0.5	-0.5	0.5	0.5	-0.5	-0.5	0.5
University degree	12.4	12.3	12.3	12.3	10.5	10.2	12.2	12.2	10.2
Secondary school	5.8	5.8	5.9	5.8	3.6	3.7	5.8	5.9	3.5
With occupation	9.1	9.0	9.1	9.2	3.4	3.6	9.0	9.0	3.2
Student	0.1	0.2	0.1	0.1	-0.9	-0.9	0.2	0.1	-0.8
Retired	-5.9	-6.0	-5.9	-5.9	1.1	0.8	-6.0	-6.0	1.0
Employeed	7.6	7.5	7.6	7.7	2.6	2.8	7.5	7.6	2.4
Self-employed	1.4	1.4	1.4	1.4	0.7	0.7	1.4	1.4	0.7
Commuter	5.4	5.3	5.6	5.5	1.4	1.5	5.5	5.5	1.4
- Study	0.3	0.3	0.3	0.3	-0.3	-0.3	0.4	0.4	-0.3
- Work	5.1	5.0	5.2	5.2	1.7	1.9	5.1	5.2	1.8

In the OCC simulation scenario we observe results similar to those in EDU with some differences; here PS involving the age in groups seems effective in reducing the selection error when estimating occupied people, retired ones, and other variables related to occupation, as well as commuters. Selection error in estimating education level seems only slightly reduced.

Table 15.14: Difference of estimates (%) by MP device users with or without PS (a) - (h) to that by all individuals age>17. Gender=F* for female and age 18-64.

	User	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Gend=F	-0.3	-0.3	-0.3	0.0	-0.1	0.0	-0.3	0.0	-0.2
Gend=F*	4.5	4.6	4.5	4.7	0.4	0.0	4.5	4.7	0.4
IT citizen	-0.4	-0.3	-0.3	-0.4	0.5	0.5	-0.3	-0.3	0.5
University degree	6.5	6.4	6.4	6.5	5.0	4.9	6.4	6.3	4.9
Secondary school	5.5	5.5	5.6	5.5	3.6	3.6	5.6	5.5	3.5
Employeed	7.5	7.4	7.5	7.5	2.4	2.4	7.4	7.4	2.3
Student	0.4	0.5	0.4	0.4	-0.3	-0.3	0.5	0.5	-0.3
Retired	-5.4	-5.4	-5.4	-5.4	0.7	0.5	-5.4	-5.4	0.6
Employee	6.1	6.0	6.1	6.1	1.7	1.8	6.0	6.0	1.7
Self-employed	1.4	1.4	1.4	1.4	0.6	0.6	1.4	1.3	0.6
Commuter	4.6	4.5	4.7	4.5	1.3	1.3	4.6	4.6	1.3
- Study	0.4	0.4	0.4	0.4	0.0	-0.1	0.4	0.4	-0.1
- Work	4.2	4.1	4.3	4.2	1.3	1.4	4.2	4.2	1.4

The results obtained in these simulation experiments, and in particular in artificial samples generated in EDU and OCC scenarios, where all the SIM cards allocated in the MP devices of an HH are subscribed to by a single member of the HH with specific characteristics, show that post-stratification by age can reduce the bias due to both section error and user ambiguity in the estimation of some characteristics, such as the percentage of commuters, which is one of the areas where data from MNOs, together with other "traditional" data sources, are expected to improve the relevance, the timeliness and accuracy of statistical outputs.

The main lesson to be learnt from these experiments is that it would be really important to have data from the MNO on the age and gender of the SIM card subscriber, which should in principle be available at least for "residential" contracts (SIM card not owned by a company and not used for "machine-to-machine" data exchange; see the next section for more details).

Note that we just apply the PS estimator and do not the SUD estimator (8.6) because we are aware that it would return more or less the same results obtained with PS in the absence of user ambiguity and device duplication.

15.6 Some indications about device duplication

In the ADL survey, HHs owning an MP device are asked to report the number of devices owned, with a maximum of 9. This latter variable, although affected by some measurement errors, can be considered a good proxy of the number of SIM cards owned by the HH. It is a proxy because there may be MP devices with two distinct physical SIM cards, or with one/two physical SIM cards and one or more "electronic" SIMs (eSIM).

The ADL survey returns an estimate of the number of MPs owned by the Italian HHs equal tp $\hat{M} = 52\,460\,236$; as mentioned before, the information about usage of MPs allow us to distinguish between "used" and "non-used" MPs:

$$\hat{M} = \hat{M}_U + \hat{M}_{NU} = 51\,710\,424 + 749\,815$$

In our approach, unused MPs are those belonging to a HH where all members declare to "never use" the MP. These unused MPs, if switched off most of the time, are not expected to be counted in the data that could be provided by an MNO. On the contrary, if they are switched on but not used to make calls, send messages, etc., the SIM(s) will still interact with the network antenna and therefore be counted, unless the MNO discards from its counts those associated to SIM cards that do not make calls, etc. in a given time interval. Obviously, if the MP is switched on but not taken when commuting to work, this will introduce a negative bias in the estimation of the number of work commuters based on the data from MNO.

Given all the MP devices counted in the survey, to estimate the device duplication rate we need to know the denominator, i.e. the owners. The survey does not provide this information, but for HH consisting of only one person, we can attribute all MPs to that person.

There are 430 HHs with a single person but with more than one MP, i.e. 7.4% of the HHs with a single person, so the average number of MPs per person in HHs with a single member is 1.10.

Table 15.15: Rough estimate of average number of mobile phone devices per person (age>17)

HH Size	No. of MPs	No. Adults	Average
HH size = 1	9 562 257	8 724 522	1.10
HH size > 1	42 897 983	39 900 681	1.08
Total	52 460 239	48 625 203	1.08

Unfortunately, for HHs with more than one member, we cannot consider all members and we should somehow limit the subset of people. Although in the ADL survey data there are many non-adult daily users of MPs, it is common practice for a SIM card to be subscribed by an adult member (aged >17) of the HH, so we decided to consider two alternative hypotheses:

a) consider only people with age>17 who declare to use the MP, even if occasionally (we exclude those who declare to never use the MP);

b) consider only people with age>17 who declare to use the MP every day.

With these opposite hypotheses the estimated number of MP devices per person ranges from 1.12 to 1.18.

Table 15.16: Possible estimates of average number mobile phone devices per person (age>17)

HH size	No. of MPs	No. Persons (a)	Avg (a)	No. Persons (b)	Avg (b)
HH size=1	9 562 257	8 724 522	1.10	8 724 522	1.10

HH size	No. of MPs	No. Persons (a)	Avg (a)	No. Persons (b)	Avg (b)
HH size>1	42 897 983	38 224 342	1.12	35 600 988	1.20
Total	52 460 239	46 948 864	1.12	44 325 510	1.18

It is worth noting that the number of MPs estimated by the survey, $\hat{M} = 52\,460\,236$, is significantly lower than the number of active "human" SIM cards[2] in Italy in 2023, which is estimated to be around 78 450 000 (annual average), according to the Italian Authority for the Regulation of Communications (AGCOM). The latter figure can be split in "residential" SIM cards (belonging to a person), estimated to be 68 016 050 (86.7%), and "business" cards (belonging to a company), estimated to be equal to 10 443 950 (13.3%) (these figures do not include the "machine-to-machine" SIM cards[3] . In practice, the survey estimate (52 460 239) would represent 66.9% of the total number of SIM cards, and 77.2% of privately owned SIM cards.

Unfortunately, the ADL survey refers to MP devices and not to SIM cards, and the estimated number of MPs is expected to be inferior to the number of SIM cards because a MP can store one or more distinct SIMs (both physical and electronic cards); in addition the survey collects data on HH and therefore does not count MPs owned by people not being part of HHs. Finally, the survey, counting the MP devices, does not distinguish whether the SIM card in a MP is a business or a residential one.

At this stage, it is not clear what kind of data could be provided by an MNO for "business" SIM cards; the absence of data on the person using the SIM card (i.e. data related only to the company) may represent an additional source of error in the estimation process. Let's look at some examples: if the business SIM cards are placed in MP devices assigned to managers or also to employees traveling for work in a company, the absence of data on the real user would not represent a problem if the user also has a personal MP device with a "residential" SIM card; on the contrary, it could represent a problem if the user has only one MP device with a "business" SIM card, as it would become a source of undercoverage when making statistics based on data related to "residential" SIM cards. In Italy, this latter situation is not uncommon and occurs for self-employed persons.

Finally, it should be noted that these figures on SIM cards provided by AGCOM refer to declarations of active SIM cards made by MNOs, which have an interest in showing high market shares and therefore include in their active SIM card count those that connect to the network only a few times. In this sense, the data provided by AGCOM is affected by an overestimation. In any case, dividing the number of "residential" SIM cards owned by individuals (68 016 050) by the number of adults in Italy in 2023 (49 473 395) gives an average of 1.37 cards per adult, an estimate higher than that provided by the survey; the average is expected to be higher if we include in the denominator the real owners

[2]"Human" SIM cards are those where all exchanges of data or commands originating from an MP are due to human intervention, as opposed to "machine-to-machine" M2M cards.

[3]"machine-to-machine" (M2M) SIM cards are those where there is an exchange of data, commands, etc. between electronic devices without human intervention (or with a very limited human intervention). M2M cards are around 29 400 000

(unknown), but on the other hand it is expected to decrease due to the overestimation of the numerator.

15.7 Some final considerations

The analyses shown in the previous Sections are based on data collected in the annual ADL sample survey that investigates habits of Italian HHs including those related to ownership and use of MP devices; although not at the required level of detail, these data permits to assess coverage of MPs and potential errors that may arise in deriving statistics form a non-random sub-sample of MP users, in particular the selection error that is non-negligible for some characteristics (e.g. commuters for work).

The survey data show a good coverage of the Italian population by MPs devices; about 96% of people aged between 14 and 54 years are users of MPs devices, while the coverage drops to about 92% for people in the age group 55-64 and to 82% for people in the age group 65-74. In other words, it is difficult to reach older people with MP, who tend to live alone (HHs with only one member). Age is a powerful auxiliary variable that, when used in a simple post-stratification estimation approach, makes it possible to compensate for errors related to under-coverage of the whole population (selection error). It is worth noting that the post-stratification setting is rather limited to relatively few variables related to SIM card contractors that are expected to be provided by MNOs, presumably gender, age in classes and a few others (perhaps citizenship and place/region of birth).

In order to assess the impact of user ambiguity, i.e. the fact that an MP device is used by a person other than the subscriber of the contract of the associated SIM card, we carried out some simulation experiments which showed that, in the presence of an extreme situation where all SIM cards in an HH are assigned to a single adult member with certain characteristics (higher education level, with an occupation, etc.), an additional bias is introduced and its direction is not easy to predict (it depends on the assumed mechanism related to the assignment of all SIM cards to one member). This error sometimes is reduced by a simple post-stratification based on age groups although not for all the target parameters to be estimated. One could consider more complex estimators such as the SUD estimator (8.6), which tackle simultaneously the various errors.

Correcting for the user ambiguity error via the SUD estimator requires knowledge of the distribution of subscribers with respect to the chosen post-stratification setting and an estimate of ϕ_{lg}, the probability that a SIM used by a person in the g-th post-stratum has a contractor in the l-th post-stratum. Estimating these probabilities requires an ad hoc data source (e.g. a small sample) where both user and subscriber data are available. In our study, we could only estimate the ϕ_{lg} in a simulation setting, whereas in a "standard" setting, we are only expected to know the marginal distributions; in particular, the marginal distribution of users with respect to the PS variable, ϕ_{+g}, can be estimated using the survey data, while the marginal distribution of contractors, ϕ_{l+}, should be derived from data provided by the MNO. Thus we are in an *uncertain* sit-

uation, where it is only possible to determine some bounds on the individual probabilities using the Fréchet-Bonferroni property:

$$\max\{\phi_{l+} + \phi_{+g} - 1; 0\} \leq \phi_{lg} \leq \min\{\phi_{l+}; \phi_{+g}\}$$

In some cases it may be known that some $\phi_{lg} \backsimeq 0$, as for example in post-stratification by age groups when considering SIM cards used by a non-adult (age<18) and whose contractor is still a non-adult (age<18), a situation quite rare in Italy and that may not happen at all (structural zeros) in presence of rules that impose that the contractor must be and adult. It is worth noting that the matrix of $\hat{\phi}_{lg}$ should be invertible to be included in the SUD estimator. Finally, it should be remembered that the SUD estimator also requires an estimate of device duplication per post-stratum, which is not straightforward.

Appendix A

Why we ask MNO to count but not weight: A toy example

Suppose there are only two cities. Let (N_1, N_2) be the *de jure* resident totals. Let (p_1, p_2) be the proportions of the residents living in the same cities, such that the *de facto* numbers of people living in the two cities are

$$Y_1 = N_1 p_1 + N_2(1 - p_2) \qquad \text{and} \qquad Y_2 = N_1(1 - p_1) + N_2 p_2 \tag{A.1}$$

Consider estimating (Y_1, Y_2) based on MNO users and their *de facto* residence.

Suppose the MNO weights its users by *MNO-home city* with respect to the official *de jure resident* total in that city, where m_i is the MNO count of users with *de facto* MNO-home city i in this case, which yields

$$m_1 \frac{N_1}{m_1} \equiv N_1 \quad \text{and} \quad m_2 \frac{N_2}{m_2} \equiv N_2$$

This is clearly futile, since it simply reproduces the *de jure* populations totals, which is biased as long as

$$(N_1, N_2) \neq (Y_1, Y_2)$$

Adjustment Let (α_1, α_2) be the proportion of MNO users by resident city. Let m_{ij} be the number of MNO users with *de facto* MNO-home city i and *de jure* resident city j, where $i, j = 1, 2$, such that

$$m_1 = m_{11} + m_{12} = N_1 \alpha_1 p_1 + N_2 \alpha_2 (1 - p_2) \tag{A.2}$$

$$m_2 = m_{21} + m_{22} = N_1 \alpha_1 (1 - p_1) + N_2 \alpha_2 p_2 \tag{A.3}$$

In the special case $\alpha_1 = \alpha_2 = \alpha$, the user proportion is the same everywhere, such that

$$\frac{N_1 \alpha_1 + N_2 \alpha_2}{N_1 + N_2} = \alpha = \frac{m_1 + m_2}{N_1 + N_2}$$

and

$$\begin{cases} m_1 \overset{\text{(A.2)}}{=} m_{11} + m_{12} = N_1 \alpha p_1 + N_2 \alpha (1 - p_2) \overset{\text{(A.1)}}{=} \alpha Y_1 \\ m_2 \overset{\text{(A.3)}}{=} m_{21} + m_{22} = N_1 \alpha (1 - p_1) + N_2 \alpha p_2 \overset{\text{(A.1)}}{=} \alpha Y_2 \end{cases}$$

Weighting the MNO user counts $\{m_i\}$ with respect to the overall population

total $N_1 + N_2$ *without* regard to MNO-home city suffices, since

$$m_1 \frac{N_1 + N_2}{m_1 + m_2} = \frac{m_1}{\alpha} = Y_1 \quad \text{and} \quad m_2 \frac{N_1 + N_2}{m_1 + m_2} = \frac{m_2}{\alpha} = Y_2$$

However, generally, the MNO user proportions differ in the two cities, i.e.

$$\frac{m_{11} + m_{21}}{N_1} = \alpha_1 \quad \neq \quad \frac{m_{12} + m_{22}}{N_2} = \alpha_2$$

By collecting the terms in (A.2) and (A.3) according to either *de jure* resident city, we have

$$\frac{m_{11}}{m_{11} + m_{21}} = \frac{N_1 \alpha_1 p_1}{N_1 \alpha_1} = p_1 \quad \text{and} \quad \frac{m_{22}}{m_{12} + m_{22}} = \frac{N_2 \alpha_2 p_2}{N_2 \alpha_2} = p_2$$

Weighting the MNO counts $\{m_{ij}\}$ by *de jure* resident city with respect to *de jure* resident totals $\{N_j\}$ suffices, since

$$\begin{aligned}
\frac{m_{11}}{\alpha_1} + \frac{m_{12}}{\alpha_2} &= N_1 \frac{m_{11}}{m_{11} + m_{21}} + N_2 \frac{m_{12}}{m_{12} + m_{22}} = N_1 p_1 + N_2 (1 - p_2) = Y_1 \\
\frac{m_{21}}{\alpha_1} + \frac{m_{22}}{\alpha_2} &= N_1 \frac{m_{21}}{m_{11} + m_{21}} + N_2 \frac{m_{22}}{m_{12} + m_{22}} = N_1 (1 - p_1) + N_2 p_2 = Y_2
\end{aligned}$$

Notice that ignoring the different user proportions (α_1, α_2) and weighting MNO user counts $\{m_i\}$ by the overall user proportion would now yield

$$\begin{aligned}
m_1 \frac{N_1 + N_2}{m_1 + m_2} - Y_1 &= \Big(N_1 \frac{m_1}{m_1 + m_2} + N_2 \frac{m_1}{m_1 + m_2} \Big) - \Big(N_1 p_1 + N_2 (1 - p_2) \Big) \\
&= N_1 \Big(\frac{m_1}{m_1 + m_2} - p_1 \Big) + N_2 \Big(p_2 - \frac{m_2}{m_1 + m_2} \Big)
\end{aligned}$$

which is generally not zero if $\alpha_1 \neq \alpha_2$, and

$$m_2 \frac{N_1 + N_2}{m_1 + m_2} - Y_2 = -\Big(m_1 \frac{N_1 + N_2}{m_1 + m_2} - Y_1 \Big)$$

Summary Weighting MNO users by *de facto* MNO-home city with respect to *de jure* resident totals is misconceived, because it confuses the two resident concepts. Weighting MNO users by *de jure* resident city with respect to *de jure* resident totals is viable, which requires the MNO to provide the counts $\{m_{ij}\}$ instead of $\{m_i\}$ or the weights $\{N_i/m_i\}$.

Now, it is clear to the discerning readers that any weighting adjustment must correspond to some particular assumption of the selection mechanism of MNO users from all the residents. Indeed, as explained in Section 7.3,

- weighting by the overall user proportion α is based on the assumption that all the MNO users form a random sample from all the residents, regardless of their *de jure* or *de facto* residence;
- weighting by *de jure* resident city is based on the assumption that MNO users are random samples by *de jure* resident city, respectively, among the

N_1 or N_2 residents that are known;

- weighting by *de facto* resident city is based on the assumption that MNO users are random samples by *de facto* resident city, respectively, among the unknown number of Y_1 or Y_2 residents.

We notice that it is possible to adjust with respect to the unknown (Y_1, Y_2) as well, see e.g. Section 7.3, but weighting by N_i/m_i would be mistaken.

Finally, since in reality there exist multiple MNOs as well as other statistics of interest than *de facto* resident totals, the only viable approach is for the MNOs to provide their user counts as specified by the OSA, so that the OSA can undertake the appropriate adjustment for the MNO user selection effect. This is why we ask MNOs to count but not weight.

Appendix B

Road passenger numbers

The number of inbound road passengers defines the corresponding reference population for inbound tourism, which needs to be estimated. This can be achieved in two steps: first, impute the missing loop counts of vehicles; second, predict the number of passengers given the number of vehicles. Below are the results so far, which are to be improved by further investigation.

B.1 Imputation of loop counts

Loop counts of vehicles are missing during the periods of time when a loop stops working. We first impute the total count of vehicles, then disaggregate the imputed total by size (small, medium, large).

- The vehicle counts are given every fifteen minutes, which are aggregated to hours. Partially complete hours will be disregarded and treated as missing.
- An hourly model is trained and applied to the partially complete days, after which we train a daily model to complete the imputation.

We train deep learning models for both hourly or daily imputation.

- Two LSTM models to process the loop count time series: past data gets processed forwards and future data gets processed backwards, with respect to the hour to impute. These layers are aimed at extracting time features.
- Three layers to process the contemporaneous counts from all the other loops and cameras. These layers are aimed at extracting spatial features.
- Both time series and contemporaneous features provide input to three dense layers in order to generate the output.

Figures B.1 and B.2 illustrate the imputation results with or without the camera counts in addition. It is clear that the patterns obtained are much more reliable with the camera counts.

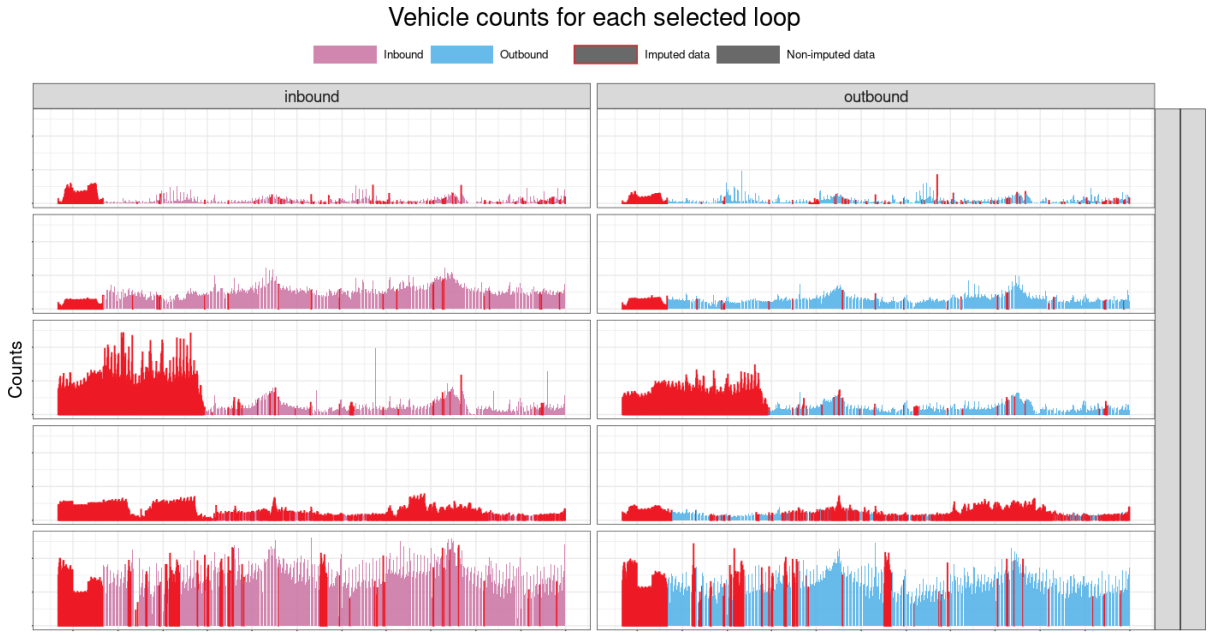

Figure B.1: Imputation of vehicle counts without camera counts

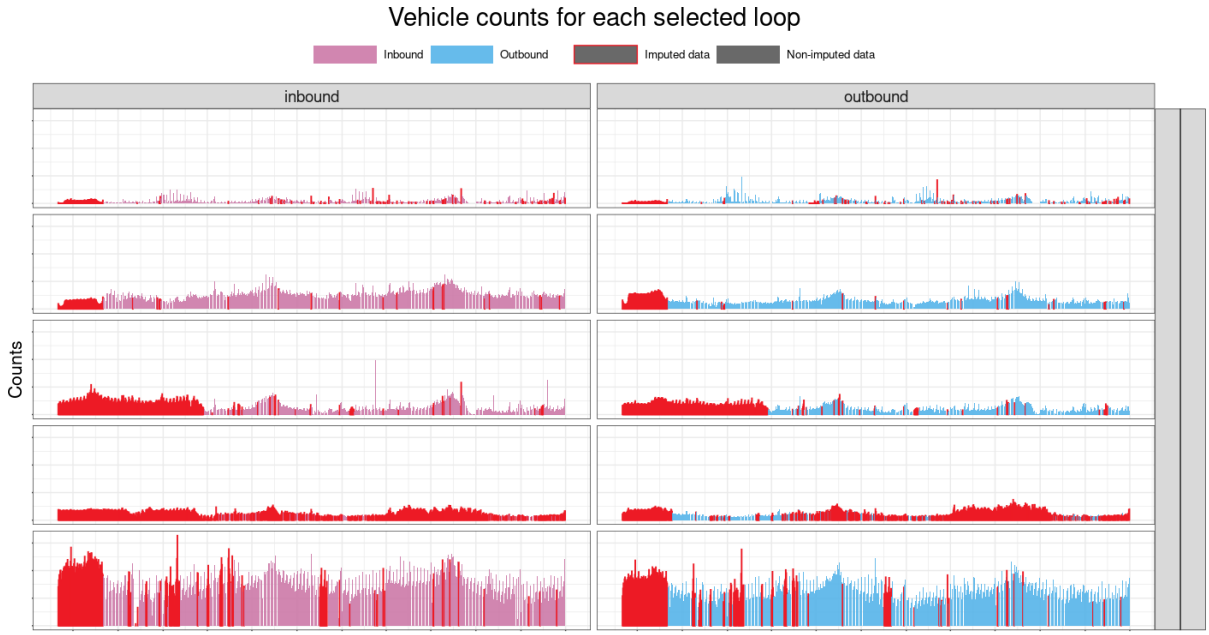

Figure B.2: Imputation of vehicle counts with camera counts

B.2 Prediction of passenger numbers

There are a couple of complications to modelling the passenger number per vehicle. First, capacity counting is only carried out from 8am to 4pm. Next, it is difficult to count large vehicles such as buses. In addition, to apply the model to vehicle counts, one cannot directly distinguish the different types of vehicles of the same size, such as buses vs. trucks, and there are two entry points that do not have loops at all. Nevertheless, random forest models by

vehicle size have been considered using the following features:

year, month, day, hour, day of week, region, province, road type, bordering country, Spanish holiday (Yes, No), foreign holiday (Yes, No), no. vehicles.

Data from years 2021, 2022 and 2023 have been used for training, where cross-validation is grouped by hour. Year 2024 was used as the test data. We used either the R package ranger or mice. In the first case, we considered three possibilities of weighting the training set observations: no weights, weights as the inverse of hourly vehicle total, weights as the inverse of hourly passenger total. In the second case, an oversampling procedure is applied to the training data due to the unbalanced observations regarding the hours.

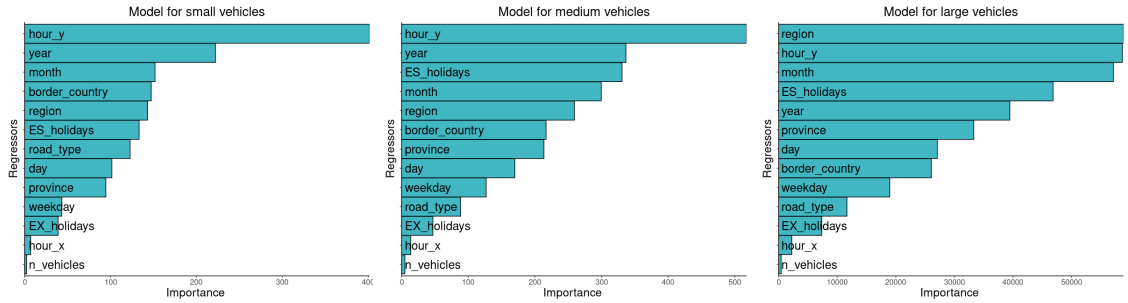

Figure B.3: Feature importance for random forest models by vehicle size

Figure B.3 gives the feature importance for the random forest models by vehicle size. Hour of the day is seen to be important regardless the vehicle size. This suggests that imputation of vehicle counts by the hourly model should be considered even for the days with completely missing loop counts.

Bibliography

[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. and 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] Deville, J.-C., and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87:376-382.

[15] 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.

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

[17] ESSNet MNO-MINDS (2024) Deliverable 3.1 Preliminary Report on Methodology, Work Package 3 *Methodologies and open source tools for integrating MNO and non-MNO data sources*.

[18] Fay, R.E. and Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 85:398-409.

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

[20] 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

[21] 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

[22] Gu, T., Han, Y. and Duan, R. (2024). Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society Series B*, https://doi.org/10.1093/jrsssb/qkae111.

[23] 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.

[24] 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

[25] James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1:361-379.

[26] 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.

[27] 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.*

[28] 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

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

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

[31] 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.

[32] LeSage, J. and Kelley Pace. R (2008). Spatial econometric methods for modeling origin-destination flows. In *Journal of regional science*, 48:941-967

[33] 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*.

[34] 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.

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

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

[37] 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

[38] 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

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

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

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

[42] Pan, S.J. and Yang, Q. (2019). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345-1359.

[43] Pannekoek, J. and Zhang, L.-C. (2015). Optimal adjustments for inconsistency in the presence missing data. *Survey Methodology*, 41:127-144.

[44] Patone, M. and Zhang, L.-C. (2020). On two existing approaches to statistical analysis of social media data. *International Statistical Review*, 89:54-71.

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

[46] 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.

[47] 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

[48] 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

[49] 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.

[50] 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.

[51] Salgado, D., Sanguiao-Sande, L., Oancea, B., Barragán, S. and Necula, M. (2021). An end-to-end statistical process with mobile network data for official statistics. *EPJ Data Science*, 10, 20. https://doi.org/10.1140/epjds/s13688-021-00275-w

[52] 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)*.

[53] 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

[54] 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

[55] 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

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

[57] 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

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

[59] 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

[60] 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.

[61] 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

[62] 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

[63] 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.

[64] 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

[65] 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.

[66] 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.

[67] 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

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

[69] Zhang, L.-C., Sanguiao-Sande, L. and Lee, D. (2025). Design-based predictive inference. *Journal of Official Statistics*, 41:404-432.

[70] Zhang, L.-C. and Haug, J.K. (2025) Turnover Flash Estimation by Purposive Sampling and Debit Card Transactions. *Journal of Official Statistics*, 41:382-403.

[71] Zhang L.-C. and 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