Spanish Journal of Statistics
Vol. **2** No. 1 2020, Pages 41–77
doi:https://doi.org/10.37830/SJS.2020.1.04

Regular Article

Towards a modular end-to-end statistical production process with mobile network data

David Salgado[1, 2], Luis Sanguiao[1], Bogdan Oancea[3, 4], Sandra Barragán[1], Marian Necula[3]
[1]Department of Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
[2]Department of Statistics and Operations Research, Complutense University of Madrid, Spain
[3]Department of Innovative Tools in Statistics, Statistics Romania (INS), Romania
[4]Department of Business Administration, University of Bucharest, Romania

Received: November 3, 2020. Accepted: March 1, 2021.

Abstract: Mobile network data has proved to be an outstanding data source for the production of statistics in general, and for Official Statistics, in particular. Similarly to another new digital data sources, this poses the remarkable challenge of refurbishing a new statistical production process. In the context of the European Statistical System (ESS), we substantiate the so-called ESS Reference Methodological Framework for Mobile Network Data with a first modular and evolvable proposed statistical process comprising (i) the geolocation of mobile devices, (ii) the deduplication of mobile devices, (iii) the statistical filtering to identify the target population, (iv) the aggregation into territorial units, and (v) the inference to the target population. The proposal is illustrated with synthetic data generated from a network event data simulator developed for these purposes.

Keywords: Statistical production, mobile network data, end-to-end process, geolocation, Deduplication, aggregation, inference

MSC: 62-07, 62P25, 62M05, 62F15

1 Introduction

Mobile network data, i.e. digital data generated in a mobile telecommunication network by the interaction between a mobile station (mobile device such as a smartphone or a tablet) and a base transceiver station (commonly known as antenna in an imprecise way) (Miao et al., 2016), constitutes a remarkable source of information for the production of statistics in Social Science, in general, and for Official Statistics, in particular. Many one-off studies can already be found in the literature with applications in different statistical domains (González et al., 2008; Ahas et al., 2010; Phithakkitnukoon et al., 2012; Calabrese et al., 2013; Deville et al., 2014; Louail et al., 2014; Iqbal et al., 2014; Blondel et al., 2015; Douglass et al., 2015; Pappalardo et al., 2016; Raun et al., 2016; Ricciato et al., 2017; Graells-Garrido et al., 2018; Wang et al., 2018) (see Salgado et al. (2020) for a more comprehensive list).

Published by the Spanish National Statistical Institute

However, the production of official statistics in National and International Statistical Systems requests a standardized and industrialised statistical production process so that this new data source is fully integrated in the daily production framework of statistical offices. This raises remarkable challenges such as the data access conditions, new methodological and quality frameworks, a larger IT infrastructure (both in hardware and in software), a deep revision of the statistical disclosure control, and the identification of relevant aggregates (mostly included as part of legal regulations) for a diversity of stakeholders and users. Although a number of illustrative case studies dealing with official statistics can already be found in the literature (Debusschere et al., 2016; Williams, 2016; Nurmi, 2016; Izquierdo-Valverde et al., 2016; Dattilo et al., 2016; Senaeve and Demunter, 2016; Meersman et al., 2016; Reis et al., 2017; Sakarovitch et al., 2019; Galiana et al., 2018; Lestari et al., 2018), we still lack a production framework with a new statistical process.

In this line of thought, efforts in the international community (UN, 2017) and in the Europen Statistical System (ESS) (Ricciato, 2018) are under way to construct a production framework and some recent examples of an end-to-end statistical production process have been tested in a statistical office (Tennekes et al., 2020). The need for a detailed standardised and harmonised statistical process goes beyond the rise of new digital data sources, since a process-oriented production system instead of a product-oriented or even domain-oriented system is nowadays considered essential to achieve high-quality standards (UNECE, 2011). In this sense, the proliferation of one-off studies with new digital data in different statistical domains may be stressing the risk over statistical offices of reinforcing production silos, thus becoming clearly inefficient and making Official Statistics socially irrelevant (DGINS, 2018).

This article presents the fundamentals of a modular and evolvable statistical process with mobile network data to produce estimates for present population counts and origin-destination matrices as a concrete business case. This proposal constitutes the first step towards the construction of the so-called ESS Reference Methodological Framework for Mobile Network Data (see e.g. Ricciato, 2018), an initiative of the ESS embracing a set of principles to ensure consistency, reproducibilty, portability, and evolvability of data processing methods for this data source, to facilitate interworking between statistical offices and mobile network operators (MNOs) both at technical and organisational levels, and to adapt to the fast-changing technological environment of telecommunications by clearly detaching technology and statistical analysis.

We shall focus on the integral view of the process underlying its functional modularity and evolvability and on the methodological core bringing novel methods in Official Statistics with a clear goal of producing both estimates and their quality indicators (accuracy). We shall illustrate the whole process using synthetic network event data generated by a data simulator developed for these purposes. In section 2 we provide a general description of our approach setting up the general context under which this proposal is thought to be implemented. In section 3 we shall shortly describe the main functionalities of the network event data simulator as of this writing. In section 4 we provide the main contents of each of the modules comprising the statistical process, namely a generic description of the data in subsection 4.1, geolocation of mobile devices in subsection 4.2, deduplication of mobile devices carried by the same individual in subsection 4.3, statistical filtering of individuals in the target population in subsection 4.4, aggregation of device-level data into territorial units in subsection 4.5, and inference with respect to the target population in subsection 4.6. In section 5 we close with some conclusions and future prospects.

2 General description

The development, implementation, and monitoring of a statistical production process with mobile network data entail several complex and highly entangled issues. We need to solve questions regarding the access to data (including the integration with other data sources and even data from several MNOs), the development of statistical methods not traditional used in the production of official statistics, the according update and modernisation of the quality assurance framework, the deployment of the corresponding IT infrastructure, the professional and technical skills of staff necessary to execute this process, and the identification of the key target aggregates to be produced for the public good.

Official statistics play a key role in democratic socities for decision-taking and policy-making. For example, public fund allocation is usually conducted taking into account official population figures published by statistical offices. Thus, high-quality standards must be ensured and verified usually following international frameworks. In this context, in agreement with the ESS Reference Methodological Framework, an official statistical process with mobile network data must comprise the process design from the raw telecommunication data generated in the networks to the final statistical outputs. Acquiring aggregates or data at device-level from unknown preprocessing steps is not considered an option here. This is the first assumption motivating our proposal of an end-to-end process.

Mobile network data are extremely sensitive data and rightful concerns immediately arise to use them for statistical purposes. Data access is indeed an intricately complex set of legal, administrative, technical, and business issues, which we shall not dealt with here. Nonetheless, we assume three principles around which a final solution must be built:

- Privacy and confidentiality: as with any other official statistics produced from any data source by any statistical office, privacy and confidentiality of data holders and respondents must be assured. Indeed, stringent legal conditions have recently arisen to prevent privacy and confidentiality in the European context (European Parliament, 2016).
- Public good: there is an evident socioeconomic interest in extracting different insights from mobile network data valuable for the public good. This is as legitimate as the production of official statistics from traditional data sources.
- Private business interest: the production of statistical outputs and insights from mobile network data stands also as an increasing economic activity providing value and progress to the economy. Indeed, the digital data economy is targeted as a pillar in the European context.

All in all, an aligment of these three principles must be reached in practice. The proposed statistical process herein assumes that a collaborating scenario between statistical offices and MNOs through public-private partnerships, joint ventures, etc. is possible and leaves room for the design, execution, and monitoring of the different modules explained below.

In the context of the ESS, as of this writing no definitive agreement for a fully-fledged sustainable production of official statistics based on mobile network data has been reached between a national statistical office (NSO) and an MNO. Only specific short-term limited agreements for research have been reached[1]. This entails a shortage of data in NSOs to develop the statistical methodology,

[1] A remarkable exception is the compilation of international travel statistics for the balance of payments produced by the National Bank of Estonia (National Estonian Bank, 2020), not a statistical office, though.

the quality frameworks, and the software tools. Furthermore, given the extraordinarily rich and complex data ecosystem associated to a mobile telecommunication network, the identification of concrete data for statistical purposes must be undertaken (Radio network data? Core network data? Network management data? Call Detail Records?). In this sense, our strategy is to produce synthetic network event data together with a ground truth scenario so that all these aspects can be developed and investigated. In this way, more specific data requests can be formulated in agreement with the quality indicators and the ground truth computed in the simulated scenarios. Thus, our starting point will be the generation of these simulated scenarios.

A key feature of the ESS Reference Methodological Framework is the evolvability of the statistical process so that improvements and adaptations of the statistical methods to the underlying technological conditions is always possible and seamless. This justifies the approach of functional modularity (already present in modern proposals of traditional statistical processes (see e.g. Salgado et al., 2018)). By breaking the end-to-end process into modules according to the data abstraction principle we design transparent and independent production steps so that a change in one module will not affect the next module beyond the quality of the input/output interconnecting them through a standardised interface. In this proposal we do not include all necessary modules (e.g. data acquisition, substituted by the simulator) but only those core methodological stages (see Radini et al., 2020, for an architectural point of view):

- Geolocation.- This module focuses on the computation of location probabilities for each device across a reference grid used for the statistical analysis.
- Deduplication.- This module focuses on the computation of multiplicity probabilities for each device, i.e. probabilities of a given device to be carried by an individual jointly with one or several other devices. This is motivated by our interest on individuals of the target population, not on mobile devices.
- Statistical filtering.- This module focuses on the algorithmic identification of mobile devices of individuals of the target population such as domestic tourists, commuters, inbound tourists, etc.
- Aggregation.- This module focuses on the computation of probability distributions for the number of individuals detected by the network (i.e. with mobile devices) across different territorial units.
- Inference.- This module focuses on the computation of probability distributions for the number of individuals of the target population (even with no device) across different territorial units.

A cautious reader will immediately notice how the computation of probabilities is essential across the whole process. The use of probabilities, in our view, is jointly motivated by several relevant reasons. Firstly, probability distributions allow us to account for the uncertainty along the whole process, thus paving the way for the computation of quality indicators, especially those related to accuracy. Secondly, probability models provide a natural way to integrate data through priors and posteriors in a hierarchy of models. This is important because the combination of diverse data sources will not only produce statistical outputs with higher quality but it is also necessary in many cases, in particular, with mobile network data to avoid identifiability problems (see below). Thirdly, probability distributions stand as a flexible module interface between the successive production steps. In this line, we can use the total probability theorem to connect the original input data (raw telco data) with the final output data (population estimates):

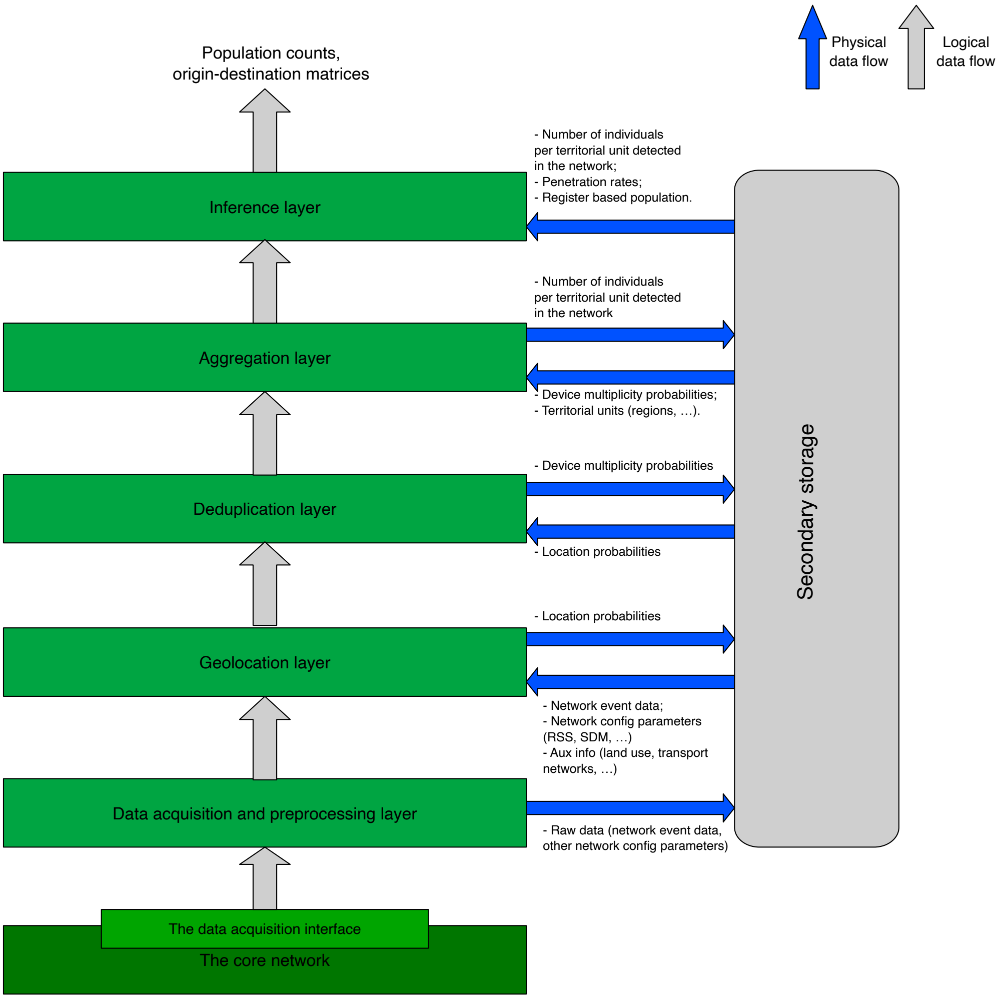
Physical data flow
Logical data flow
Population counts, origin-destination matrices
Inference layer
Aggregation layer
Deduplication layer
Geolocation layer
Data acquisition and preprocessing layer
The data acquisition interface
The core network
Secondary storage
- Number of individuals per territorial unit detected in the network;
- Penetration rates;
- Register based population.
- Number of individuals per territorial unit detected in the network
- Device multiplicity probabilities;
- Territorial units (regions, ...).
- Device multiplicity probabilities
- Location probabilities
- Location probabilities
- Network event data;
- Network config parameters (RSS, SDM, ...)
- Aux info (land use, transport networks, ...)
- Raw data (network event data, other network config parameters)

Figure 1: Modular structure of the statistical process and its software tools.

$$\mathbb{P}(z_{out}|z_{in}) = \int \mathrm{d}z_1 \int \mathrm{d}z_2 \cdots \int \mathrm{d}z_N \; \mathbb{P}(z_{out}|z_N) \cdots \mathbb{P}(z_2|z_1)\mathbb{P}(z_1|z_{in}). \tag{1}$$

The modular structure of the methodology is translated into a modular structure for the software tools. The choice of programming languages to develop these tools is motivated by multiple reasons. Firstly, software developed with the intention to be used in the future should be portable at the level of source code. Thus, portability is our first consideration. Secondly, our goal is to produce a software for statisticians, not for computer scientists. Thus, the language(s) of the implementation should be familiar for statisticians and easy to use by them. Thirdly, in the line of software development in the ESS, we planned to use only open source tools like libraries, IDEs, debuggers, profilers, etc. to maintain the software development process under a strict control regarding the associated costs. Moreover, the programming language(s) together with these tools should have a large community of programmers and users which can be seen as a free technical support. Fourthly, the programming language(s) should have support for parallel and distributed computing. Since all the algorithms involved by the our methodological approach are computational intensive, and the size of mobile network data could be very large, this is a mandatory requirement. Last but not least important, the criteria of programming efficiency and resources needed to run the software even on normal desktops/laptops are also considered.

After analysing different choices, eventually we came to the following two software ecosystems: R (R Core Team, 2020) or Python (Van Rossum and Drake, 2009). Both systems meet our criteria and have a large community of users but while Python is considered to be more computationally efficient, R is better suited for statistical purposes and it seems to gain ground among the official statistics community (Templ and Todorov, 2016; Kowarik and van der Loo, 2018). Since our target audience is the official statistics community, we decided to develop our software modules using R since it has a huge number of available packages, it has support for parallel and distributed processing, it can be easily interfaced and work together with high performance languages like C++ when the performance of plain R is not enough, it can be easily interfaced with computing ecosystems widely used in the Big Data area such as Hadoop (White, 2009) or Spark (Zaharia et al., 2016) and there are several packages allowing a neat interface between R and these systems (Oancea and Dragoescu, 2014; Venkataraman et al., 2016) which means that, if needed, all modules in our software stack can be easily integrated with such systems for a production pipeline.

Thus, to execute the process with simulated data, we have developed an R package for each module implementing the corresponding statistical methods. With a view on scalability through distributed computing and parallelization, we use secondary memory instead of main memory to pass input and output data between modules as well as execution parameters (see figure 1). In the next section we provide details about the contents of each module.

3 Network event data simulator

The simulator is a highly modular software (Oancea et al., 2019) implementing agent-based simulating scenarios with different elements configured by the user. The basic elements are:

- a geographical territory represented by a map;
- a telecommunication network configuration in terms of a radiowave propagation model;
- a population of individuals carrying 0, 1, or 2 mobile devices during their displacement;
- a displacement pattern for individuals;
- a reference grid for analysis.

The simulator works essentially by using a radiowave propagation model (Shabbir et al., 2011) to simulate the connection between the base transceiver stations (loosely, antennas) and each mobile station (device) during the displacement of each carrying individual. The connection mechanism is an extreme simplification of the real world extracting the essential features for statistical analysis. The core output data consists of a time sequence of cell IDs (loosely, antenna IDs) and network event codes (connection, disconnection, etc.) for each device along the duration of the simulation. We simulate signalling data (i.e. passive data not depending on subscribers' behaviour) instead of Call Detail Records or any other active data generated by individuals (call, SMS, Internet connections, . . .).

For the time being, since our priority is the simulator as a whole, the different elements implemented so far are kept as simple as possible. Firstly, displacement patterns of individuals are basically a sequence of stays (no movement) and random walks with/without a drift with two possible speeds (namely, walk and car speeds). The drift, the speeds, and the shares of individuals with 0, 1, and 2 devices are easily configured by the user. Only closed populations can be simulated so far, i.e. individuals cannot abandon or enter into the territory under analysis.

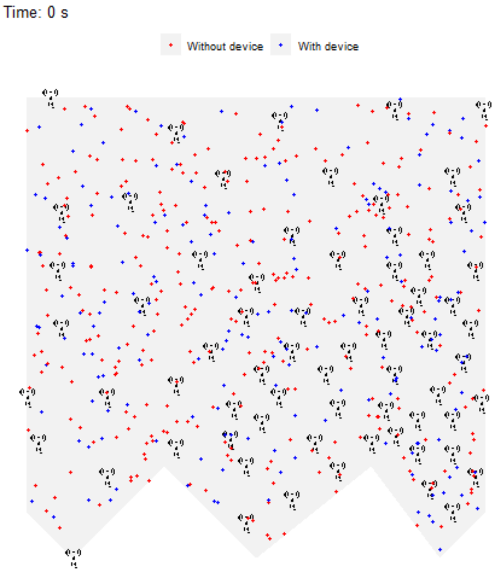

Figure 2: **Animation.** Positions of 70 antennas and drifted displacement pattern of individuals.

Secondly, an extremely simplified radiowave propagation model and a variant thereof is used in terms of the Received Signal Strength (RSS – expressed in dBm), the distance r between the BTS and the device, the emission power P, the so-called path exponent γ (quantifying the loss of signal strength) and some geometrical parameters regarding the BTS orientation (only for directional antennas (see e.g. Tennekes et al., 2020)). For omnidirectional antennas, the model is simply expressed by

$$\mathrm{RSS}(r) = 30 + 10 \cdot \log_{10}(P) - 10 \cdot \gamma \cdot \log_{10}(r). \tag{2}$$

Each device connects to the antenna producing the highest signal strength in each tile until the antenna reaches its maximum capacity. Both the emission power and the path loss are selected as input parameters by the user. A convenient variant introduced by Tennekes et al. (2020) performs a parameterised logistic transformation upon RSS producing the so-called Signal Dominance Measure:

$$\mathrm{SDM}(r) = \frac{1}{1 + \exp\left(-S_{\mathrm{steep}} \cdot (\mathrm{RSS}(r) - S_{\mathrm{mid}})\right)}, \tag{3}$$

where S_{steep} and S_{mid} are chosen according to characteristics of each radio cell. Each device connects to the antenna providing the highest signal dominance measure in each tile until the antenna reaches its maximum capacity. Both S_{steep} and S_{mid} are selected as input parameters by the user, too.

Figure 3 represents the RSS and the SDM for a given antenna in an arbitrary territory depicted as an irregular polygon with a 10 km × 10 km bounding box.

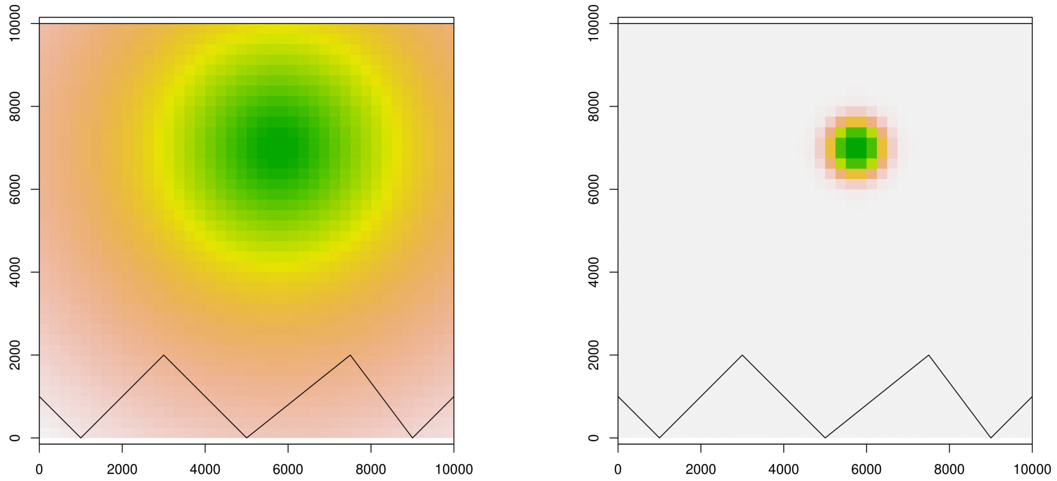

Figure 3: Received signal strength and signal dominance measure for an omnidirectional antenna according to models (2) and (3).

4 Production modules

4.1 Data

When contacting MNOs to access data, the first reaction from telecommunication engineers and data engineers in these companies is to ask "what data?" The data ecosystem of a mobile telecommunication network is extremely complex, derived from its nested cellular structure (see figure 4). Thus, a first step to use mobile network data for statistical purposes is to substantiate the meaning of these data. In this line, the use of a synthetic simulator allows us to devise an end-to-end process and to set up an empirical criterion about specific data to compile statistics accurate enough for official purposes.

Our proposed process helps us to provide a first typology of data required to reach our goal. We identify three types of data (according to organisation which generates them).

4.1.1 Mobile network data

Under this category we embrace two sorts of data related to mobile telecommunication networks. On the one hand, we need data about the configuration of the network. Basically, these are parameters entering the radiowave propagation models used in subsequent stages (see below) such as emission powers, path loss exponents, frequencies and frequency correction factors, base station heights and azimuths,... Notice that these variables do not contain information about the subscribers but they are extremely sensitive for MNOs due to the highly competitive degree of the telecommunication market. Ultimately, the variables to access will depend on the chosen model, which should be in principle chosen according to the accuracy of the final estimates and the associated acquisition costs under the public-private agreement. Access to these data does not mean whatsoever that these data should be made public or even that they have to leave MNOs'

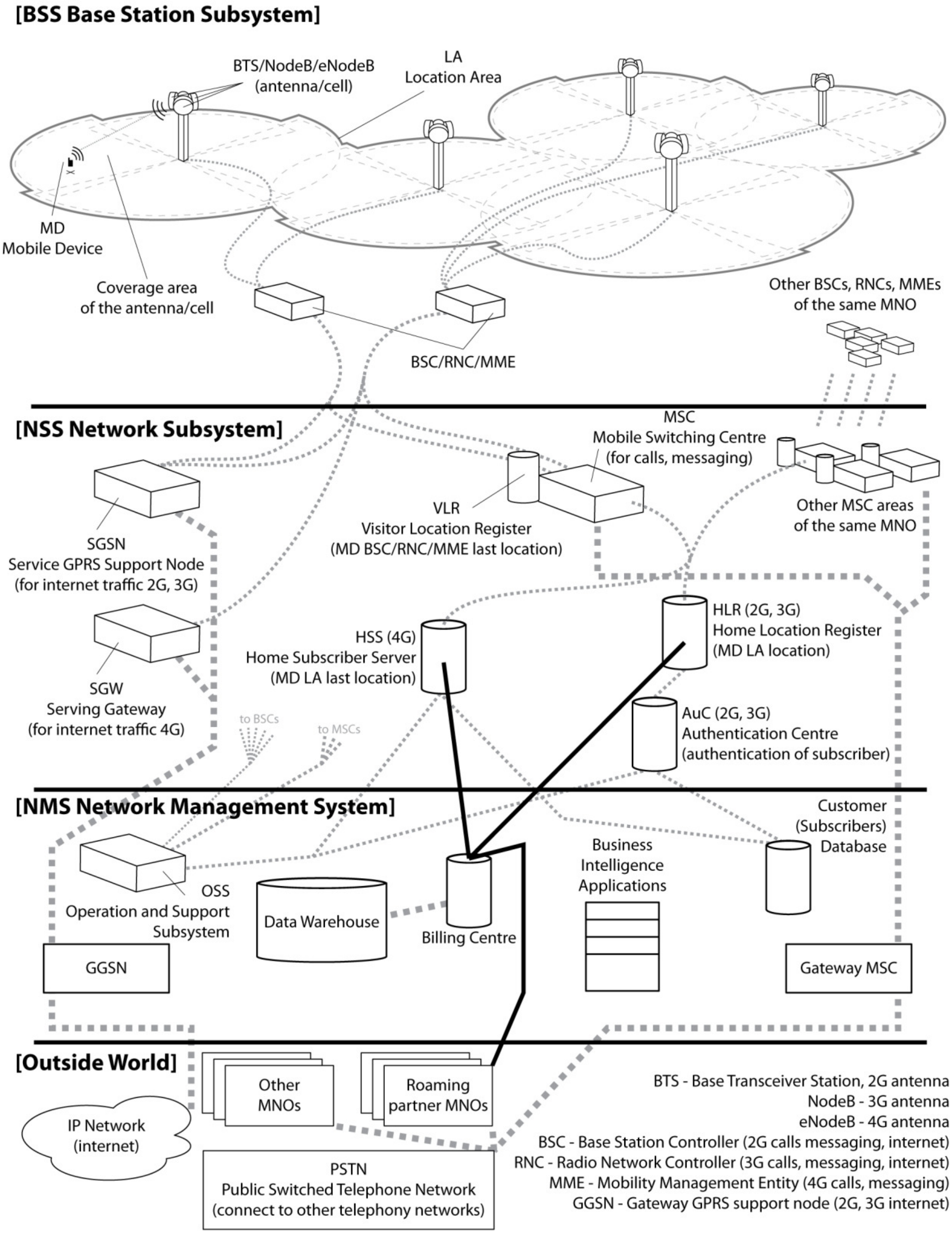

Figure 4: Nested cellular structure of a GSM-like network (taken from Positium (2016)).

information systems. This sensitive information must be kept protected also by NSOs and they are just required to be accessed to produce specific outputs in later stages. Agreements on computing these outputs and their sharing into the statistical process should be enough for our goals (see below).

On the other hand, so-called network event data generated by each mobile station (device) in the network must be accessed. These can be variables such as the cell ID (identifying the cell or sector whether the interaction between a device and a connecting antenna is established), the Time of Arrival (basically collecting the time for a signal to reach a mobile device from the connecting antenna), the Angle of Arrival (measuring the angle of the line-of-sight of a device from the connecting antenna),...These data do contain sensitive information about the subscribers. Again, not only must they be kept private but also they must be preprocessed in the MNOs' original information systems (i.e. no transmission whatsoever to NSOs). Identifying precisely what variables to use will ultimately depend on the accuracy of final estimates and associated accessing and preprocessing costs. Once more, details must be part of agreements between MNOs and NSOs.

In the illustrative example with the data simulator below we will use the emission power and path loss exponent of each base station (network configuration) together with the cell ID of each connection/signal transmission/disconnection and orientation parameters between devices and base stations every 10 seconds.

4.1.2 Auxiliary NSO information on target aggregates

This is information produced by NSOs themselves, thus providing profuse access to microdata for alternative (possibly undisclosed) aggregations in finer territorial units. They may be survey microdata, administrative data, or aggregates from any combination of sources with a relevant relationship with the target outputs of our analysis.

In the illustrative example with the data simulator below to produce present population counts and general-mobility origin-destination matrices we shall use data from the current population register or some other similar demographic operation. It is important to state that the treatment of both data sources makes a difference on their role. Whereas mobile network data will be used as the central source to produce outputs (thus gaining in both spatial and time breakdowns), the population register will enter as an auxiliary prior data source. An equal-footing integration of all data sources to produce, modify, and correct the population register is not pursued here.

4.1.3 Auxiliary (public) information on the geographic territory

As with the production of any other official statistics, the more available information to integrate, the higher expected quality for the output. In this sense, auxiliary information from (usually public) organizations such as land use or transport network configurations and schedules may be profitably integrated in the modelling exercise. For example, for the geolocation of mobile devices, prior location probabilities upon grid tiles can be fixed according to the land use features of each tile. In the wilderness this probabilities will differ a great deal from those in the city centers.

In the illustrative example below, since the geographical territory is just an arbitrary irregular polygon, we shall not use any prior information about land use or transport network. Every tile will be similar to each other.

4.1.4 Privacy-preserving data technologies

As an immediate side-effect of this complex and sensitive data ecosystem, the integration of information in stringent privacy-preserving conditions is a must. A research avenue clearly seems to be arising extending the traditional statistical disclosure control from output aggregates to also input and intermediate data.

This brings the privacy-preserving technologies (Zhao et al., 2019) into scene. However, we would like to pose the following reflection. When considering mobile network data (and probably similarly sensitive new digital data), we detect a change in society about the role of statistical officers in producing official statistics. With more traditional data sources such as survey data and administrative data, statistical officers are undisputedly endowed with the legitimacy of accessing, processing, and integrating ***personal data*** from these diverse sources. Take e.g. the construction of a business register where sensitive information from all business units in a country are compiled for further use in the statistical production process. No privacy-preserving technique is demanded in this case, in spite of which privacy and confidentiality is completely guaranteed and statistical disclosure control is fully effective. In our view, statistical offices must reclaim their traditional role as secure recepcionist of information for the public good.

However, having said this, the challenge of integrating MNOs into the statistical production process includes the management of trust and privacy-preserving techniques stand as an excellent tool in this sense.

4.2 Geolocation

The utility of mobile network data to produce statistics for the public good arises at least from three aspects. Firstly, the geospatial nature of this information makes it ideal to provide population counts and mobility-related statistics at an unprecedented spatial and time breakdown. Different social groups can be targeted (tourists, commuters, present population, etc.) provided algorithms are put into place to identify them within the datasets. Secondly, Internet traffic and the nature of donwloaded mobile apps can provide relevant insights for social analysis (see e.g. Ucar et al., 2019). Finally, and more interestingly in our view, mobile network data can provide an excellent source of network data, i.e. interactions between population units, thus paving the way for the use of network science in the production of novel statistical outputs.

Currently, the main focus of research is centered on the geolocation of mobile devices. Originally, Voronoi tessellations of the geographical territory under analysis were used to partition this territory into disjoint tiles assigning each one to a BTS. In our view, this is an oversimplification of the network, since coverage areas and sector cells of each BTS can often be intersecting (even nested) and directional. To overcome this complexity, we divide the territory into a grid of tiles and using radiowave propagation models compute the so-called event location probabilities $\mathbb{P}(\mathbf{E}_{dt} = \mathbf{e}_j \big| T_{dt} = i)$, i.e. the probability that a device d produces network event data $\mathbf{e}_j$ (e.g. the cell ID of a given BTS to which the device is connected) conditioned on being located at tile i. This conditional probability is used to compute the reverse so-called posterior location probability $\gamma_{dti} = \mathbb{P}(T_{dt} = i \big| \mathbf{E}_d)$ at each time t and each device d. The posterior joint location probabilities $\gamma_{dtij} = \mathbb{P}(T_{dt} = i, T_{dt-1} = j \big| \mathbf{E}_d)$ are also of interest for later modules. Notice that we condition upon all available network information

$\mathbf{E}_d = \{\mathbf{E}_{dt}\}_{t=0,1,\ldots}$

A first direct approach is to make use of Bayes' theorem together with the prior location probabilities $\mathbb{P}(T_{dt} = i)$ (computed according to the prior auxiliary information such as land use or transport network information): $\mathbb{P}(T_{dt} = i \big| \mathbf{E}_{dt} = \mathbf{e}_j) \propto \mathbb{P}(\mathbf{E}_{dt} = \mathbf{e}_j \big| T_{dt} = i) \cdot \mathbb{P}(T_{dt} = i)$. This is the static approach followed by Tennekes et al. (2020).

A superseding alternative is to consider the dynamical behaviour of individuals in the population and to postulate a generic transition model across the reference grid, which together with the event location probabilities computed above, enter into a hidden Markov model (HMM) as transition and emission models, respectively. Upon estimation of these model parameters, we can compute the posterior location probabilities for each device d (see figure 5). Mathematical details are provided in the appendix.

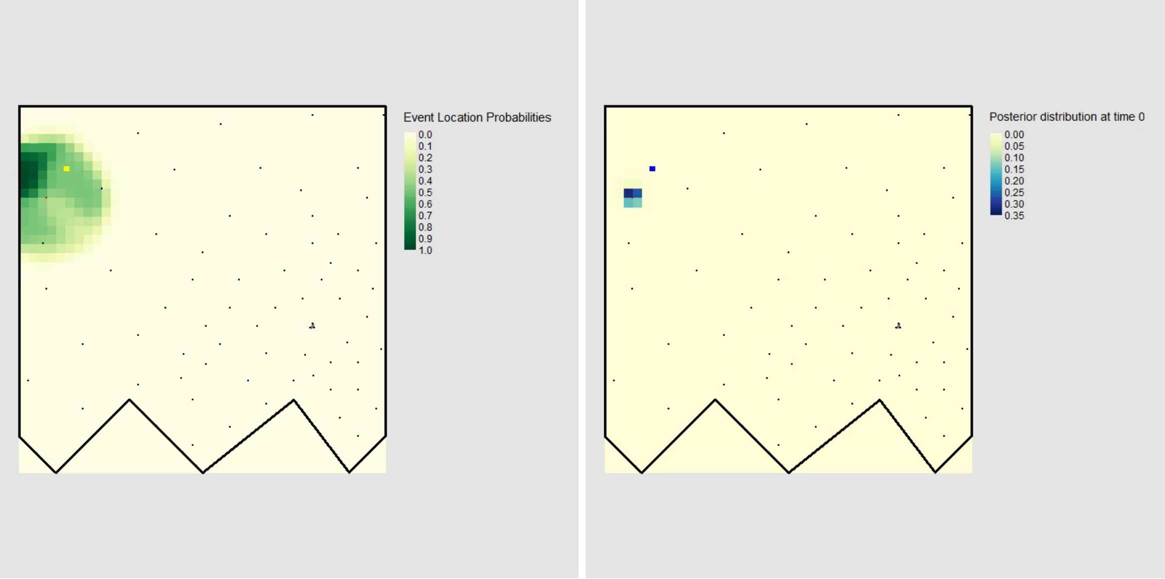

Figure 5: **Animation.** Event location probabilities [left] and posterior location probabilities [right] for a given device. True position also included.

The use of HMMs in the context of this reference grid for analysis is notably versatile and provides a generic framework to deal with multiple aspects. Firstly, at this initial stage of the project, we have defined the HMM state just as the location in the grid, but more complex states can be possibly defined taking into account the velocity, the transport mode, or a classification of anchor points (home, work, second residence, etc.). Secondly, the emission model (i.e. the event location probabilities) is built independently of the transition model, which allows MNOs to concentrate the processing of sensitive network information (antenna localizations, network parameters, etc.) on a this concrete production step. For the HMM, only the output of this step is needed, thus making it possible to undisclose and protect this sensitive information. Finally, the use of probabilities allows us to take into account the uncertainty in the estimation process from the onset. Indeed, we can define familiar accuracy indicators for the geolocation such as bias, standard deviation, and mean squared error as with traditional survey data (see appendix).

4.3 Deduplication

Since we focus on estimating population counts of individuals, not of mobile devices, we need to detect which terminals are carried over by the same individuals. We call it device multiplicity. The goal is to compute a device-multiplicity probability $p_d^{(n)}$ for each device d to be carried over by the same individual together with a total of n devices. In our simulated scenario, for computational ease, with limit the number of devices per individual to 2. Thus, we aim at computing $(p_d^{(1)}, p_d^{(2)} = 1 - p_d^{(1)})$, i.e. the probability that a device d belongs to an individual with 1 or 2 devices, respectively.

The problem of device duplicity has been often recognised as an overcoverage problem. It is usually considered ***after*** the aggregation step producing **number of devices** per territorial area and time interval. Once this aggregation step has been conducted, the challenge is really serious and may easily drive us into an identifiability problem (Lehmann and Casella, 2003) in any model estimating the number of individuals from the number of devices. The reader may easily be convinced with a simple example. Consider a population of $N^{(D)} = 10$ devices, all corresponding to a different individual, i.e. $N = 10$. Consider another population of $N^{(D)} = 10$ devices, where each individual has two devices, i.e. $N = 5$. There is no possible statistical model using only the variable $N^{(D)}$ possibly distinguishing between these two situations. In other words, we run into an identifiability problem unless more parameters are introduced, which will require the use of auxiliary information. In this simple case, we may think of a statistical model based on $(N^{(D)}, R_{dup})$ where we have introduced another parameter R_{dup} standing for the duplicity rate in the population. With these variables, the identifiability problem ameliorates, but the model complexity increases, apart from the issue about data availability (is R_{dup} really available?).

This is why we recommend to address this problem **before the aggregation step**. This has direct implications for the access agreements. According to this recommendation, the number of devices is not a target dataset in the statistical process and the device multiplicity issue must be addressed upon individual information at the device level, thus ideally in MNOs' premises (together with the geolocation step).

Another important consideration arises when considering uncertainty. It is important to remind that we target at the probability $p_d^{(n)}$ of each device d. This probability distribution will indeed be another intermediate distribution in the chain (1). We need to assess the **uncertainty** (i.e. probabilities) and not just to conduct a classification. The relevance of this will be evident in the aggregation step later on.

We have proposed two alternative approaches. On the one hand, we resort to Bayesian reasoning to test the hypothesis that two given devices d_1 and d_2 belong to the same individual. Let us denote by H_{dd} the hypothesis that device d uniquely corresponds to an individual, whereas $H_{d_1 d_2}$ stands for devices d_1 and $d_2 \neq d_1$ belonging to the same individual. Thus, we need to compute $p_d^{(1)} = \mathbb{P}\left(H_{dd} \middle| \mathbf{E}, \mathbf{I}^{\text{aux}}\right)$, where $\mathbf{E} = \{\mathbf{E}_{dt}\}_{t=0,\dots,T}^{d=1,\dots,D}$ is all network event information. We propose two procedures:

- Pair computation.- We compute $p_d^{(1)} = 1 - \max_{d' \neq d} \mathbb{P}\left(H_{dd'} \middle| \mathbf{E}_d, \mathbf{E}_{d'}, \mathbf{I}^{\text{aux}}\right)$, where

$$\mathbb{P}(H_{dd'}|\mathbf{E}_d, \mathbf{E}_{d'}, \mathbf{I}^{\text{aux}}) = \frac{\mathbb{P}(\mathbf{E}_d, \mathbf{E}_{d'}|H_{dd', \mathbf{I}^{\text{aux}}})\,\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})}{\mathbb{P}(\mathbf{E}_d|H_{dd}, \mathbf{I}^{\text{aux}})\,\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}}) + \mathbb{P}(\mathbf{E}_d, \mathbf{E}_{d'}|H_{dd'}, \mathbf{I}^{\text{aux}})\,\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})}, \tag{4}$$

with $\mathbb{P}(H_{dd'}|\mathbf{I})$, $\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})$ being prior probabilities and $\mathbb{P}(\mathbf{E}_d, \mathbf{E}_{d'}|H_{dd'}, \mathbf{I}^{\text{aux}})$, $\mathbb{P}(\mathbf{E}_d|H_{dd}, \mathbf{I}^{\text{aux}})$ standing for the likelihoods under each hypothesis, respectively.

- One-to-one computation.- Alternatively, posing $\Omega_d = \bigcup_{d'=1}^{D} H_{dd}$, we compute

$$p_d^{(1)} = \frac{\mathbb{P}\left(\mathbf{E}_d\middle|H_{dd}, \mathbf{I}^{\text{aux}}\right) \cdot \mathbb{P}\left(H_{dd}\middle|\mathbf{I}^{\text{aux}}\right)}{\mathbb{P}\left(\mathbf{E}_d\middle|H_{dd}, \mathbf{I}^{\text{aux}}\right) \cdot \mathbb{P}\left(H_{dd}\middle|\mathbf{I}^{\text{aux}}\right) + \sum_{d' \neq d} \mathbb{P}\left(\mathbf{E}_d, \mathbf{E}_{d'}\middle|H_{dd'}, \mathbf{I}^{\text{aux}}\right) \cdot \mathbb{P}\left(H_{dd'}\middle|\mathbf{I}^{\text{aux}}\right)}. \tag{5}$$

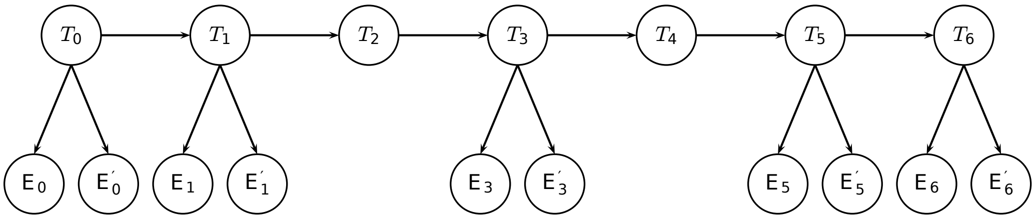

Figure 6: Extended HMM to compute $\mathbb{P}(\mathbf{E}_d, \mathbf{E}_{d'}|H_{dd'}, \mathbf{I}^{\text{aux}})$ for a given device d (subscript not included in the graphical model).

In both procedures the probabilities $\mathbb{P}(\mathbf{E}_d|H_{dd}, \mathbf{I}^{\text{aux}})$, $\mathbb{P}(\mathbf{E}_d, \mathbf{E}_{d'}|H_{dd'}, \mathbf{I}^{\text{aux}})$ are computed with the original HMM and the extended HMM represented in figure 6, respectively. Priors are computed incorporating prior information e.g. from the Customer Relationship Management Database or any other complementary information (see Salgado et al., 2020, for some details).

On the other hand, instead of focusing on the network event variables $\mathbf{E}_{dt}$, we can make use of the random location $\mathbf{R}_{dt} \in \{\mathbf{r}_i^{(c)}\}_{i=1,\ldots,N_T}$ estimated according to the posterior location probabilities γ_{dti}. Then, we can follow the same approach as the Bayesian pair computation case (4) substituting $\mathbf{E}_{dt}$ by $\mathbf{R}_{dt}$ (see Salgado et al. (2020) for details).

In figure 7 we show the results for the Bayesian one-to-one case for our illustrative example. The ROC curves show an excellent performance for the classification of devices according to their duplicity with values of the area under the curve (AUC) above 0.95. Using the simulated ground truth and a threshold of 0.50 we can also notice that very few false positive cases result (and they are due to the short period of time under analysis: basically two individuals following nearly the same sequence of coverage areas), whereas the number of false negative cases are a bit notable. This is due to devices of different individuals staying under the same coverage area during the time period: they are wrongly classified as duplicity cases of analysis. Realistic time periods of analysis will hopefully avoid these problems.

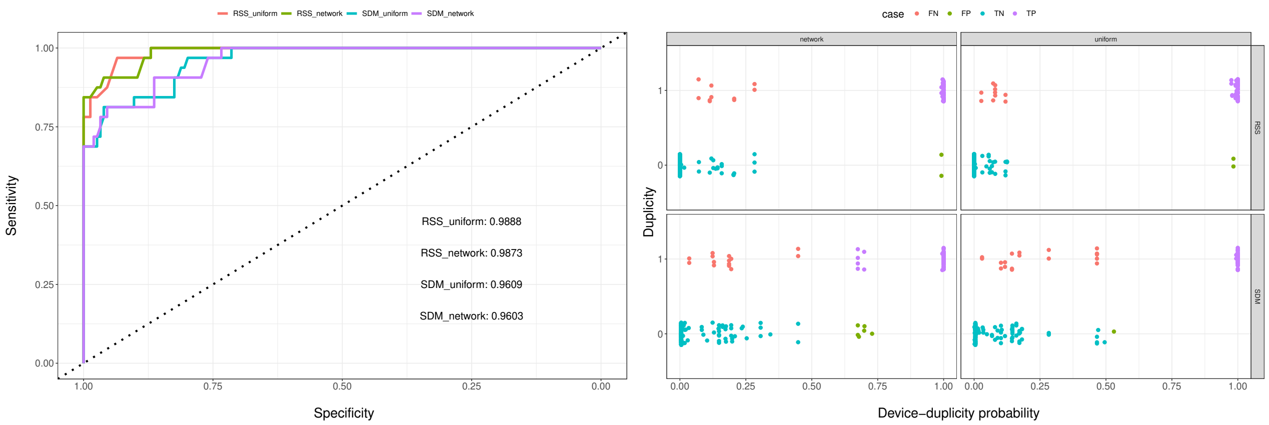

Figure 7: [Left] ROC curve for two emission models (RSS and SDM) and two HMM priors (uniform and network). [Right] Cases for two emission models (RSS and SDM) and two HMM priors (uniform and network).

4.4 Statistical filtering

As of this writing, this module is the less developed since more complex and realistic displacement patterns are needed in the simulator to study and analyse different proposals. We limit ourselves to provide a generic view. Again, we shall be focusing on analyses upon the geolocation data, i.e. upon the network event data and location probabilities derived thereof.

First of all, the target mobile network data is assumed to be basically some form of signalling data so that time frequency and spatial resolution are high enough as to allow us to analyse movement data in a meaningful way. In this sense, for example, CDR data only provides information up to a few records per user in an arbitrary day which makes virtually impossible any rigorous data-based reasoning in this line.

The use of HMMs implicitly incorporates a time interpolation which will be very valuable for this statistical filtering exercise. In this way we avoid the issues arising from noncontinuous traces approaches (see e.g. Vanhoof et al., 2018, for home location algorithms). However, a wider analysis is needed to find the optimal time scope. In turn, the spatial resolution issue is dealt with by using the reference grid. This releases the analyst from spatial techniques such as Voronoi tessellation, which introduces too much noise for our purposes. Nonetheless, the uncertainty measures computed from the underlying probabilistic approach for geolocation must be taken into account to deal with precision issues in different regions (e.g. high-density populated vs. low-density populated).

In our view, the algorithms for statistical filtering should be mainly based on quantitative measures of movement data. In particular, from the HMMs fitted to the data (especially the location probabilities) we propose to derive a probability-based coarse-grained trajectory per device which will be the basis for these algorithms. Once a trajectory is assigned to each device, different indicators and measures of movement shall be computed upon which we shall apply algorithms to determine important concepts such as usual environment, home/work location, second home location, leisure activity times and locations, etc.

A critical issue in the development of this kind of algorithms is the validation procedure. On the one hand, the use of the simulator, once more complex and realistic displacement patterns have been introduced, will offer us in the future a validation against the simulated ground truth. On the other hand, with real data two main problems need to be tackled, namely (i) the use of pseudoanonymised real data will prevent us to link mobile device records with official registers, so only indirect aggregated validation procedures can be envisaged, and (ii) the representativity of the tested sample of devices (e.g. using GPS signals) to validate the algorithm for the whole population needs to be rigorously assessed.

Thus, the starting point is the construction of a probability-based coarse-grained trajectory for each device. In our geolocation model, the state of the HMM was defined in terms of the tile where the device is positioned. Thus, the concept of trajectory follows immediately as the time sequence of states, in which we shall use the coordinates of each tile to build the so-called *path* $\{(x_{dt_0}, y_{dt_0}), (x_{dt_1}, y_{dt_1}), \ldots, (x_{dt_N}, y_{dt_N})\}$, where at each time instant t_i the spatial coordinates x_{dt_i} and y_{dt_i} for device d are specified. In more complex definitions of states, another procedure should lead us to deduce the path from the adopted concept of HMM state.

Given an HMM, it is well-known that at least two different methods can be approached to build a sequence of states, i.e. a trajectory in our case. We can compute either the most probable sequence of states or the sequence of most probable states. In mathematical terms, the former is the sequence

$$T^{*}_{dt_0:t_N} = \operatorname{argmax}_{T_{dt_0:t_N}} \mathbb{P}\left(T_{dt_0:t_N} \middle| \mathbf{E}_{dt_0:t_N}\right), \tag{6}$$

which can be computed by means of the Viterbi algorithm (see e.g. Murphy, 2012). The second method is indeed given by

$$T^{*}_{dt_0:t_N} = \left(\operatorname{argmax}_{T_{dt_0}} \gamma_{dt_0}, \operatorname{argmax}_{T_{dt_1}} \gamma_{dt_1}, \ldots, \operatorname{argmax}_{T_{dt_N}} \gamma_{dt_N}\right), \tag{7}$$

where $\gamma_{dt_j} = \mathbb{P}\left(T_{dt_j} \middle| \mathbf{E}_{dt_0:t_N}\right)$ are the posterior location (state) probabilities.

We choose the maximal posterior marginal (MPM) trajectory because it is more robust and because unimodal probabilities are expected so that differences will not be large (Murphy, 2012). Furthermore, coherence with other process modules (e.g. duplicity) using the posterior location probabilities is favoured in this way.

Once a path is assigned to each device we can compute different indicators as well as joint measures. Following Long and Nelson (2013) (see also multiple references therein) we distinguish the following groups of measures:

- Time geography.- This represents a framework for investigating constraints such as maximum travel speed on movement in both the spatial and temporal dimensions. These constraints can be capability constraints (limiting movement possibilities because of biological/physical abilities), coupling constraints (specific locations a device must visit thus limiting movement possibilities), and authority constraints (specific locations a device cannot visit thus also limiting movement possibilities).

- Path descriptors.- These represent measurements of path characteristics such as velocity, acceleration, turning angles. By and large, they can be characterised based on space, time, and space-time aspects.
- Path similarity indices.- These are routinely used to quantify the level of similarity between two paths. Diverse options exist in the literature, some already taking into account that paths are sequences of stays and displacements (see e.g. Long and Nelson, 2013).
- Pattern and cluster methods.- These seek to identify spatialâĂŞtemporal patterns from the whole set of paths. These are mainly used to focus on the territory rather than on individual patterns. They also consider diverse aspects on space, time, and space-time features.
- IndividualâĂŞgroup dynamics.- This set of measures compile methods focusing on individual device displacement within the context of a larger group of devices (e.g. a tourist within a larger group of tourists in the same trip).
- Spatial field methods.- These are based on the representation of paths as space or space-time fields. Different advanced statistical methods can be applied such as kernel density estimation or spatial statistics.
- Spatial range methods.- These are focused on measuring the area containing the device displacement, such as net displacement and other distance metrics.

Diverse indicators can be defined and used within each group (see Salgado et al. (2020) for preliminary examples on our simulated scenario). Further analysis is needed with realistic displacement patterns. With a selected set of movement indicators, we shall be able to provide a computational algorithm to substantiate the concepts of usual environment, home/work location, second home location, etc. Notice that the definition of state for the HMM could be enhanced using these concepts, thus incorporating more information into the geolocation estimation.

4.5 Aggregation

The next step is to aggregate the preceding information at the level of territorial units of analysis. These territorial units usually come from an administrative division of the geographical territory, but in general terms they will be undestood as aggregation of tiles of the reference grid. In this sense, when deciding about the choice of grid, it is highly recommended that the territorial units of analysis are taken into account from the onset. Obviously, the smaller the tiles, the higher the flexibility to define different granular levels of the territorial units.

The bottom line the aggregation step is to avoid making further modelling hypothesis as much as possible. In this line, we use probability theory to define and compute the probability distribution for the number of individuals (not devices) detected by the network using both the posterior location and device-duplicity probabilities.

It is important to make the following general remarks about our approach. Firstly, the aggregation is on the number of ***detected individuals***, not on the number of devices. This is a very important difference with virtually any other approach found in the literature (see e.g. Deville et al., 2014; Douglass et al., 2015). We take advantage of the preceding modules working at the device level to study in particular the duplicity in the number of some devices per individual. This has strong implications regarding agreements with MNOs to access and use their mobile network data for statistical purposes. The methodology devised in the preceding section to study this duplicity (or variants thereof) needs to be applied **before** any aggregation. As we can easily see, working with the number of devices instead of the number of individuals poses severe identifiability problems

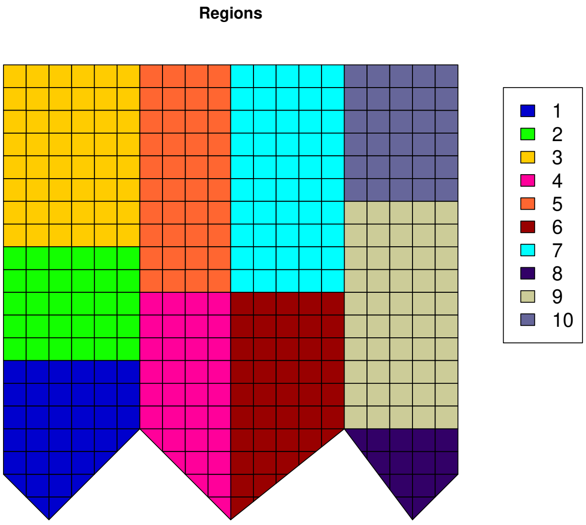

Figure 8: Territorial regions.

requiring more auxiliary information. Let us consider an extremely simplified illustrative example. Let us consider population U_1 of 5 individuals with 2 devices each one and population U_2 of 10 individuals with 1 device each one. Suppose that in order to we make our inference statement about the number N of individuals in the population we build a statistical model relating N and the number of devices $N^{(\mathrm{d})}$, that is, basically we have a probability distribution $\mathbb{P}_N(N^{(\mathrm{d})})$ for the number $N^{(\mathrm{d})}$ of devices dependent on the number of individuals, from which we shall infer N. In this situation we have $\mathbb{P}_{N^{(1)}} = \mathbb{P}_{N^{(2)}}$ even when $N^{(1)} \neq N^{(2)}$. There is no statistical model whatsoever capable of distinguishing between U_1 and U_2 (see Definition 5.2 by Lehmann and Casella, 2003, for unidentifiable parameters in a probability distribution). To cope with the duplicity of devices using an aggregated number of devices we would need further auxiliary information, which furthermore must be provided at the right territorial and time scales.

Secondly, we shall use the language of probability in order to carry forward the uncertainty already present in the preceding stages all along the end-to-end process. In another words, if the geolocation of network events is conducted with certain degree of uncertainty (due to the nature itself of the process) and if the duplicity of a given device (carried by an individual with another device) is also probabilistic in nature, then a priori it is impossible to provide a certain number of individuals[2] in a given territorial unit. For this reason, we shall focus on the probability distribution of the number of individuals detected by the network and shall avoid producing a point estimation. Notice that having a probability distribution amounts to having all statistical information about a random phenomenon and you can choose a point estimation (e.g. with the mean, the mode or the median of the distribution) together with an uncertainty measure (coefficient of variation, credible intervals, etc.).

Thirdly, the problem is essentially multivariate and we must provide information for a set of territorial units. Thus, the probability distribution which we shall provide with our proposed aggregation step must be a multivariate distribution. Notice that this is not equivalent to providing a collection of marginal distributions over each territorial unit. Obviously, there will be a correlation

[2] Notice that this same argument is valid for the number of devices.

structure, the most elementary expression of which is that individuals detected in a given territorial unit cannot be detected in another region, so that the final distribution needs to incorporate this restriction in its construction.

Finally, the process of construction of the final multivariate distribution for the number of detected individuals must make as few modelling assumptions as possible, if any. In case an assumption is made (and this should be accomplished in any use of statistical models for the production of official statistics, in our view), it should be made as explicit as possible and openly communicated and justified. In this line of thought, we shall strongly based the aggregation procedure on the results of preceding modules avoiding any extra hypothesis. Basically, our starting assumptions for the geolocation and the duplicity detection will be carried forward as far as possible without introducing new modelling assumptions of any kind.

To implement the principles outlined above, we shall slightly change the notation used in preceding chapters. Firstly we define the vectors $\mathbf{e}_i^{(1)} = \mathbf{e}_i$ and $\mathbf{e}_i^{(2)} = \frac{1}{2} \cdot \mathbf{e}_i$, where $\mathbf{e}_i$ is the canonical unit vector in $\mathbb{R}^{N_T}$ (with N_T the number of tiles in the reference grid). These definitions are set up under the working assumption of individuals carrying at most 2 devices in agreement with the deduplication step. Should we consider a more general situation, the generalization is obvious, although more computationally demanding.

Next, we define the random variable $\mathbf{T}_{dt} \in \{\mathbf{e}_i^{(1)}, \mathbf{e}_i^{(2)}\}_{i=1,\ldots,N_T}$ with probability mass function $\mathbb{P}(\mathbf{T}_{dt}|\mathbf{E})$ given by

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(1)} | \mathbf{E}_{1:D}\right) = \gamma_{dti} \times p_d^{(1)} \tag{8a}$$

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(2)} | \mathbf{E}_{1:D}\right) = \gamma_{dti} \times p_d^{(2)} \tag{8b}$$

where $p_d^{(i)}$ are the device-duplicity probabilities introduced in section 4.3. Notice that this is a categorical or multinoulli random variable. Finally, we define the multivariate random variable $\mathbf{N}_t^{\text{net}} = \left(N_{t1}^{\text{net}}, \ldots, N_{tN_T}^{\text{net}}\right)$ providing the number of individuals N_{ti}^{net} detected by the network at each tile $i = 1, \ldots, N_T$ at time instant t:

$$\mathbf{N}_t^{\text{net}} = \sum_{d=1}^{D} \mathbf{T}_{dt}. \tag{9}$$

The sum spans over the number of devices filtered as members of the target population according to section 4.4. If we are analysing, say, domestic tourism, D will amount to the number of devices in the network classified with a domestic tourism pattern according to the algorithms designed and applied in the preceding module. For illustrative examples, since we have not developed the statistical filtering module yet, we shall concentrate on present population.

The random variable $\mathbf{N}_t^{\text{net}}$ is, by construction, a Poisson multinomial random variable. The properties and software implementation of this distribution are not trivial (see e.g. Daskalakis et al., 2015) and we shall use Monte Carlo simulation methods by convolution to generate random variates according to this distribution.

The reasoning behind this proposal can be easily explained with a simplified illustrative example. Let us consider an extremely simple scenario with 5 devices and 5 individuals (thus, none of them carry two devices), and 9 tiles (a 3×3 reference grid). Let us consider that the location probabilities $\gamma_{dti} = \gamma_{ti}$ are the same for all devices d at each time instant and each tile. In these conditions $p_d^{(2)} = 0$ for all d. Let us focus on the univariate (marginal) problem of finding the distribution of the number of devices/individuals in a given tile i. If each device d has probability γ_{ti} of detection at tile i, then the number of devices/individuals at tile i will be given by a binomial variable $\text{Binomial}(5, \gamma_{ti})$. If the probabilities were not equal, then the number of devices/individuals would be given by a Poisson binomial random variable $\text{Poisson-Binomial}(5; \gamma_{1ti}, \gamma_{2ti}, \gamma_{3ti}, \gamma_{4ti}, \gamma_{5ti})$, which naturally generalizes the binomial distribution. If we focus on the whole multidimensional problem, then instead of having binomial and Poisson-binomial distributions, we must deal with multinomial and Poisson-multinomial variables. Finally, if $p_d^{(2)} \neq 0$ for all d, we must avoid double-counting, hence the factor $\frac{1}{2}$ in the definition of $\mathbf{e}_i^{(2)}$.

Notice that the only assumption made so far (apart from the trivial question of the maximum number of 2 devices carried by an individual) is the independence for two devices to be detected at any pair of tiles i and j. This independence assumption allows to claim that the number of detected individuals distributes as a Poisson-multinomial variable, understood as a sum of independent multinoulli variables with different parameters. There is no extra assumption in this derivation. The validation of this assumption is subtle, since ultimately it will depend on the correlation between the displacement patterns of individuals in the population. If the tile size is chosen small enough, we claim that the assumption holds fairly well and it is not a strong condition imposed on our derivations. On the other hand, if the tiles are too large (think of an extreme case about a reference grid being composed of whole provinces as tiles), we should expect correlations in the detection of individuals: those living in the same province will have full correlation and those living in different provinces will show near null correlation. Thus, the size of the tiles imposes some limitation to the validity of the independence assumption. Even the transport network in a territory will certainly influence these correlations. Currently, we cannot analyse quantitatively the relationship between the size of the tiles and the independence assumption with the network data simulator because we need both realistic simulated individual displacement patterns and simulated correlated trajectories (probably connected to the sharing of usual environments, home/work locations, etc.).

In figure 9 we show the high-density credible intervals with $\alpha = 0.95$ for the number of individuals N_{rt}^{net} detected by the network in each region r and each time instant t. We can compare with true values from the simulator. Deeper and wider analyses are ongoing to assess this procedure and its relationship with the geolocation and deduplication modules.

4.6 Inference

The last step comprises the estimation of the number of individuals N_{rt} in the target population in each region r at each time instant t. Notice that we aim at estimating also those individuals not detected by the network, namely, those subscribers of other MNOs and those not having a mobile device. To avoid identifiability problems, we need to make use of auxiliary information. This will basically be the register-based population figures N_r^{reg} and the penetration rates P_r^{net}. Notice, however, the different time scales of the register-based population estimates and of the network-

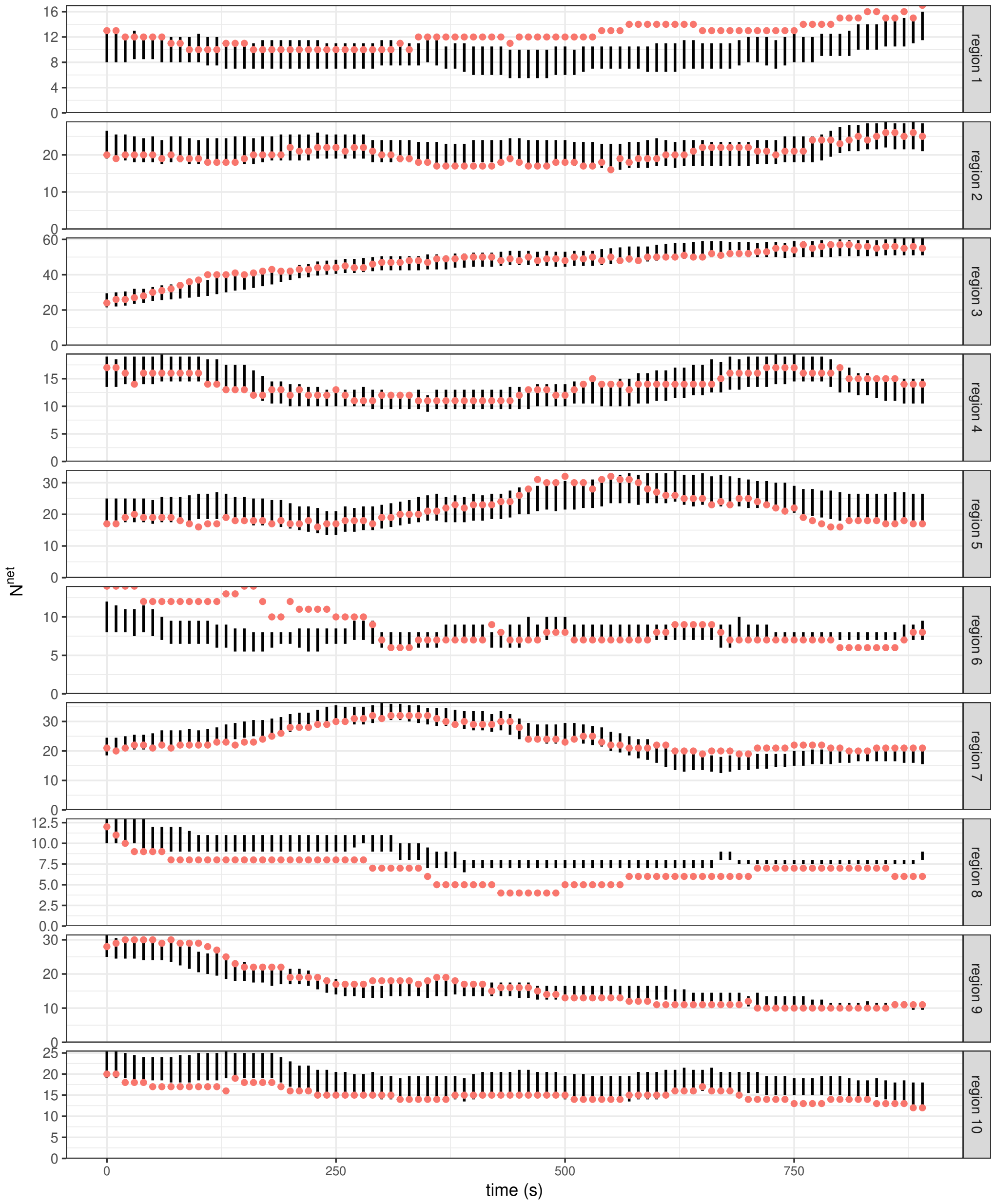

Figure 9: Credible intervals ($\alpha = 0.95$, HDI) for the number of individuals detected by the network. True values in red.

based population estimates. To avoid losing the higher time breakdown from telecommunication networks integrating at the same time the register-based population figures we shall consider a two-stage approach. Firstly, at the initial time t_0 we assume that both the register-based target population and the network-based population can be assimilated in terms of their physical location. Secondly, at later times $t > t_0$ we assume that individuals move over the geographical territory independently of the MNO, i.e. subscribers of MNO 1 will show a displacement pattern similar to those of MNO 2.

Under these general assumptions, following the approach from preceding steps, we propose to use hierarchical models (i) to produce probability distributions, (ii) to integrate all data sources, and (iii) to account for the uncertainty and the differences of concepts (present vs. residential population) and scales (time).

For the first stage the bottom line of our approach is inspired by the approach to estimate the species abundance in Ecology (Royle and Dorazio, 2009). If N_r^{net} and N_r denote the number of individuals detected by the network and in the target population, respectively, in a region r and if p_r denotes the probability of detection of an individual by the network in that region r, then we can model

$$N_r^{\text{net}} \simeq \text{Bin}\left(N_r, p_r\right), \tag{10}$$

where we have dropped out the time subscript for ease of notation. This model makes the only assumption that the probability of detection p_r for all individuals in region r is the same. This probability of detection amounts basically to the probability for an individual of being a subscriber of the given mobile telecommunication network. This assumption will be further modelled. As a first approximation, we may think of p_r as a probability related to the penetration rate P_r^{net} of the MNO in region r. We shall compute the posterior distribution $\mathbb{P}\left(N_r|N_r^{\text{net}}, \mathbf{I}^{\text{aux}}\right)$, where $\mathbf{I}^{\text{aux}}$ stands for any auxiliary information which we shall integrate into the estimation process.

As a first illustrative example of this reasoning, let us consider p_r as a fixed external parameter and try to compute the posterior probability distribution for N_r in terms of N_r^{net}, i.e.

$$\mathbb{P}\left(N_r|N_r^{\text{net}}, p_r\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}\left(N_r - N_r^{\text{net}}; 1 - p_r, N_r^{\text{net}} + 1\right) & \text{if } N_r \geq N_r^{\text{net}}, \end{cases}$$

where $\text{negbin}\left(k; p, r\right) \equiv \binom{k+r-1}{k} p^k (1-p)^r$ denotes the probability mass function of a negative binomial random variable k with parameters p and r. Once we have a distribution, we can provide point estimations, posterior variance, posterior coefficient of variation, credible intervals, and as many indicators as possible computed from the distribution. For example, if we use the MAP criterion we can provide as point estimator

$$\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{N_r^{\text{net}}}{p_r} - N_r^{\text{net}} \right\rfloor, \tag{11}$$

With the distribution we can also compute accuracy indicators such as the posterior variance, the posterior coefficient of variation, or credible intervals (see e.g. Gelman et al., 2013).

Moreover, as model assessment we can compute the posterior predictive distribution $\mathbb{P}\left(N_r^{\text{net, rep}}|N_r^{\text{net}}\right)$ and produce some indicators such as[3]

$$ppRB = \frac{\mathbb{E}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}} \middle| \widehat{N}_r^{\text{net}}\right]}{\widehat{N}_r^{\text{net}}} \qquad ppRV = \frac{\mathbb{V}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}} \middle| \widehat{N}_r^{\text{net}}\right]}{(\widehat{N}_r^{\text{net}})^2} \tag{12}$$

If we take into account the uncertainty in N_r^{net} coming from preceding modules, we can promote these indicators to random variables using the probability distribution $\mathbb{P}\left(N_r^{\text{net}}\middle|\mathbf{E}, \mathbf{I}^{\text{aux}}\right)$ and study

[3] Let us call them posterior predictive relative bias and posterior predictive relative variance.

their mean values and dispersion.

The approach described above took the detection probability p_r as an external fixed parameter built from auxiliary data sources. Furthermore, we assumed that in region r all individuals show the same probability of being a subscriber of the mobile telecommunication network. Also, the number of detected individuals N_r^{net} accumulates the uncertainty from the preceding modules, since the geolocation of mobile devices and the determination of duplicities are probabilistic. To account for this, we propose to further model these quantities, hence the hierarchical approach.

Let us firstly consider how to introduce the uncertainty in N_r^{net}. From the preceding modules we have obtained the posterior probability $\mathbb{P}\left(N_r^{\mathrm{net}}|\mathbf{E},\mathbf{I}^{\mathrm{aux}}\right)$. We still consider the detection probability p_r as an external fixed parameter. Also, we still restrict ourselves to the univariate case. Under these assumptions, the unnormalized posterior probability distribution for the number of individuals in the target population and detected by the network will be

$$\mathbb{P}\left(N_r,N_r^{\mathrm{net}}|\mathbf{E},\mathbf{I}^{\mathrm{aux}}\right)\propto \mathrm{negbin}\left(N_r-N_r^{\mathrm{net}};1-p_r,N_r^{\mathrm{net}}+1\right)\times\mathbb{P}\left(N_r^{\mathrm{net}}\middle|\mathbf{E},\mathbf{I}^{\mathrm{aux}}\right). \tag{13}$$

The normalization needs to be carried out numerically. Again, once we have the probability distribution for the random variable of interest, we can provide point estimations (MAP or posterior mean or posterior median) and accuracy indicators (posterior variance, posterior coefficient of variation, posterior IQR, credible intervals). These must be computed numerically.

The uncertainty in the detection probability p_r can be justified straightforwardly. A priori, we can think of a detection probability p_{kr} per individual k in the target population and try to device some model to estimate p_{kr} in terms of auxiliary information (e.g. sociodemographic variables, income, etc.). We would need subscription information related to these variables for the whole target population, which is unattainable. Instead, we may consider that the detection probability p_{kr} shows a common part for all individuals in region r plus some additional unknown terms, i.e. something like $p_{kr}=p_r+\mathrm{noise}$. At a first stage, we propose to implement this idea by modeling $p_r\simeq\mathrm{Beta}\left(\alpha_r,\beta_r\right)$ and choosing the hyperparameters α_r and β_r according to the penetration rates P_r^{net} and the official population data N_r^{reg}.

Let us denote by P_r^{net} the penetration rate at region r of the network, i.e. $P_r^{\mathrm{net}}=\frac{N_r^{(\mathrm{devices})}}{N_r}$. Notice that this penetration rate is also subjected to the problem of duplicities (individuals having two devices). To deduplicate, we make use of the duplicity probabilities $p_d^{(n)}$ computed in section 4.3 and of the posterior location probabilities γ_{d0r} in region r for each device d. Notice that $t=0$ according to our first generic modelling assumption. We define

$$\Omega_r^{(1)} = \frac{\sum_{d=1}^{D}\gamma_{d0r}\cdot p_d^{(1)}}{\sum_{d=1}^{D}\gamma_{d0r}}, \tag{14a}$$

$$\Omega_r^{(2)} = \frac{\sum_{d=1}^{D}\gamma_{d0r}\cdot p_d^{(2)}}{\sum_{d=1}^{D}\gamma_{d0r}}. \tag{14b}$$

The deduplicated penetration rate is defined as

$$\tilde{P}_r^{\text{net}} = \left(\Omega_r^{(1)} + \frac{\Omega_r^{(2)}}{2}\right) \cdot P_r^{\text{net}}. \tag{14c}$$

To get a feeling on this definition, let us consider a very simple situation. Let us consider $N_r^{(1)} = 10$ individuals in region r with 1 device each one, $N_r^{(2)} = 3$ individuals in region r with 2 devices each one, and $N_r^{(0)} = 2$ individuals in region r with no device. Let us assume that we can measure the penetration rate with certainty, so that $P_r^{\text{net}} = \frac{16}{15}$. The devices are assumed to be neatly detected by the HMM (i.e. $\gamma_{d0r} = 1 - O(\epsilon)$) and duplicities are also inferred correctly ($p_d^{(2)} = O(\epsilon)$ for $d^{(1)}$ and $p_d^{(2)} = 1 - O(\epsilon)$ for $d^{(2)}$). Then $\Omega_r^{(1)} = \frac{10}{16} + O(\epsilon)$ and $\Omega_r^{(2)} = \frac{6}{16} + O(\epsilon)$. The deduplicated penetration rate will then be $\tilde{P}_r^{\text{net}} = \frac{13}{15} + O(\epsilon)$, which can be straightforwardly understood as a detection probability for an individual in this network in region r. In more realistic situations, the deduplication factors $\Omega_r^{(i)}$ incorporate the uncertainty in the duplicity determination.

Now, we fix

$$\left.\begin{array}{l} \alpha_r + \beta_r = N_r^{\text{reg}} \\ \frac{\alpha_r}{\alpha_r+\beta_r} = \tilde{P}_r^{\text{net}} \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \alpha_r = \tilde{P}_r^{\text{net}} \cdot N_r^{\text{reg}} \\ \beta_r = \left(1 - \tilde{P}_r^{\text{net}}\right) \cdot N_r^{\text{reg}} \end{array}\right. \tag{15a}$$

There are several assumptions in this choice:

- On average, we assume that detection takes place with probability $\tilde{P}_r^{\text{net}}$. We find this assumption reasonable. Another alternative choice would be to use the mode of the beta distribution instead of the mean.
- Detection is undertaken over the register-based population. We assume some coherence between the official population count and the network population count. A cautious reader may object that we do not need a network-based estimate if we already have official data at the same time instant. We can make several comments in this regard:
  - A degree of coherence between official estimates by combining data sources to conduct more accurate estimates is desirable. By using register-based population counts in the hierarchy of models, we are indeed combining both data sources. In this combination notice, however, that the register-based population is taken as an external input in our model. There exist alternative procedures in which all data sources are combined at an equal footing (Bryant and Graham, 2013). We deliberately use the register-based population as an external source and do not intend to re-estimate by combination with mobile network data.
  - Register-based populations and network-based populations show clearly different time scales. The coherence we demand will be forced only at a given initial time t_0 after which the dynamical of the network will provide the time scale of the network-based population counts without further reference to the register-based population.
  - For the same model identifiability issues mentioned in the aggregation module, to estimate population counts N_r using network-based population counts N_r^{net} we need some external parameter(s). Otherwise, it is impossible. Detection probabilities are indeed these external parameters. We are modelling detection probabilities using penetration rates, which somehow already need register-based population figures. Our pragmatic approach is to identify external data sources already existing to be used in our model. These are penetration rates and register-based population counts easily collected by NSOs.

- The penetration rates P_r^{net} and the official population counts N_r^{reg} come without error. Should this not be attainable or realistic, we would need to introduce a new hierarchy level to account for this uncertainty.
- The deduplicated penetration rates are computed as a deterministic procedure (using a mean point estimation), i.e. the deduplicated penetration rates are also subjected to uncertainty, thus we should also introduce another hierarchy level to account for this uncertainty.

Then, we can readily compute the posterior distribution for N_r:

$$\mathbb{P}\left(N_r|N_r^{\text{net}},\mathbf{I}^{\text{aux}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negBetaBin}\left(N_r - N_r^{\text{net}}; N_r^{\text{net}}+1, \alpha_r - 1, \beta_r\right) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases}$$

It is a displaced negative beta binomial distribution (negBetaBin$(k;s,\alpha,\beta) \equiv \frac{\Gamma(k+s)}{k!\Gamma(s)}\frac{\mathrm{B}(\alpha+s,\beta+k)}{\mathrm{B}(\alpha,\beta)}$) with support in $N_r \geq N_r^{\text{net}}$ and parameters $s = N_r^{\text{net}}+1$, $\alpha = \alpha_r - 1$ and $\beta = \beta_r$. The mode is at

$$N^* = N_r^{\text{net}} + \left\lceil \left(\frac{\beta_r - 1}{\alpha_r}\right) \cdot N_r^{\text{net}} \right\rceil.$$

Using (15) we get as a MAP estimate:

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lceil \frac{N_r^{\text{reg}}}{\tilde{P}_r} - N_r^{\text{net}} - \frac{N_r^{\text{net}}}{N_r^{\text{reg}}}\frac{1}{\tilde{P}_r^{\text{net}}} \right\rceil, \tag{16}$$

which is very similar to (11) with the deduplicated penetration rate playing the role of a detection probability and a correction factor coming from the register-based population. The uncertainty is accounted for by computing the posterior variance, the posterior coefficient of variation, or credible intervals.

Notice that when $\alpha_r, \beta_r \gg 1$ (i.e., when $\min(\tilde{P}_r^{\text{net}}, 1 - \tilde{P}_r^{\text{net}}) \cdot N_r^{\text{reg}} \gg 1$) the negative beta binomial distribution (16) reduces to the negative binomial distribution

$$\mathbb{P}\left(N_r|N_r^{\text{net}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r+\beta_r-1}, N_r^{\text{net}}+1\right) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases}$$

Notice that $\frac{\beta_r}{\alpha_r+\beta_r-1} \approx 1 - \tilde{P}_r^{\text{net}}$ so that we do not need the register-based population (this is similar to dropping out the finite population correction factor in sampling theory for large populations). The mode is at

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{N_r^{\text{net}}}{\tilde{P}_r^{\text{net}}} - N_r^{\text{net}} \right\rfloor,$$

which is similar to (11).

We can make the model more complex by defining a new level in the hierarchy for the hyperparameters α and β (see e.g. Gelman et al., 2013) so that the relationship between these parameters and the external data sources (penetration rates and register-based population counts) is also uncertain.

For example, we can go all the way down the hierarchy, assume a cross-cutting relationship between parameters and some hyperparameters and postulate

$$
\begin{aligned}
N_r^{\text{net}} &\simeq \text{Bin}\left(N_r, p_r\right), \quad \text{for all } r = 1,\dots,R, && \text{(17a)}\\
p_r &\simeq \text{Beta}\left(\alpha_r, \beta_r\right), \quad \text{for all } r = 1,\dots,R, && \text{(17b)}\\
\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r+\beta_r}\right), \alpha_r+\beta_r\right) &\simeq \text{N}\left(\mu_{\beta r}(\beta_0, \beta_1; \tilde{P}_r^{\text{net}}), \tau_\beta^2\right) \times \text{Gamma}\left(1+\xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1,\dots,R, && \text{(17c)}\\
\left(\log\beta_0, \beta_1, \tau_\beta^2, \xi\right) &\simeq \text{f}_\beta\left(\log\beta_0, \beta_1, \tau_\beta^2\right) \times \text{f}_\xi(\xi), && \text{(17d)}
\end{aligned}
$$

where $\mu_{\beta r}(\beta_0, \beta_1; \tilde{P}_r^{\text{net}}) \equiv \log\left(\beta_0 \left[\frac{\tilde{P}_r^{\text{net}}}{1-\tilde{P}_r^{\text{net}}}\right]^{\beta_1}\right)$.

The interpretation of this hierarchy is simple. It is just a beta-binomial model in which the beta parameters α_r, β_r are correlated with the deduplicated penetration rates. This correlation is expressed through a linear regression model with common regression parameters across the regions, both the coefficients and the uncertainty degree. On average, the detection probabilities p_r will be the deduplicated penetration rates with uncertainty accounted for by hyperparameters $\beta_0, \beta_1, \tau_\beta^2$. For large population cells, the hyperparameter ξ drops out so that finally the register-based population counts N_r^{reg} play no role in the model, as above. This further hierarchy is under exploration (see Salgado et al. (2020) for some computational details). Indeed, the hierarchy can be extended also to model N_r (the so-called state process), e.g. by a Poisson distribution with parameter λ_r and keep on modelling λ_r according to some underlying process integrating more auxiliary information (see Salgado et al. (2020) for details).

For the second stage we shall focus only on closed populations, i.e. populations with individuals not allowed to enter or leave the geographical territory during the time period of estimation. This is a first step in agreement with the current status of the network event data simulator.

The basic assumption is that displacement patterns are not dependent on the subscribing MNO providing the data, i.e. individuals in a given MNO network show similar displacement patterns to those in any other network or in the target population in general. We begin by considering a balance equation. Let us denote by $N_{t,rs}$ the number of individuals moving from region s to region r in the time interval $(t-1, t)$. Then, we can write

$$
\begin{aligned}
N_{tr} &= N_{t-1r} + \sum_{\substack{r_t=1\\ r_t\neq r}}^{N_T} N_{t,rr_t} - \sum_{\substack{r_t=1\\ r_t\neq r}}^{N_r} N_{t,r_tr}\\
&= \sum_{r_t=1}^{N_T} \tau_{t,rr_t} \cdot N_{t-1r_t}, && \text{(18)}
\end{aligned}
$$

where we have defined $\tau_{t,rs} = \frac{N_{t,rs}}{N_{t-1s}}$ (0 if $N_{t-1s} = 0$). Notice that $\tau_{t,rs}$ can be understood as an aggregate transition probability from region s to region r at time interval $(t-1, t)$ in the target population. According to our general assumption we can use $\tau_{t,rs}^{\text{net}} \equiv \frac{N_{t,rs}^{\text{net}}}{N_{t-1s}^{\text{net}}}$ to model $\tau_{t,rs}$. In particular, we shall postulate $\tau_{t,rs} = \tau_{t,rs}^{\mathrm{net}}$. The probability distributions of N_{st-1}^{net} and $[\mathbf{N}_t^{\mathrm{net}}]_{sr} = N_{t,rs}^{\mathrm{net}}$ can indeed be already computed in the aggregation module.

Finally, we mention two points. On the one hand, random variables N_{rt} are defined recursively in the time index t, so that once we have computed the probability distribution at time t_0, then we can use equation (18) to compute the probability distribution at later times $t > t_0$. On the other hand, Monte Carlo techniques should be used to build these probability distributions. Once we have probability distributions, we can make point estimations and compute accuracy indicators as above (posterior variance, posterior coefficient of variation, credible intervals).

This same argument can be extended to produce origin-destination matrices. If N_{tr} and $\tau_{t,rs}$ denote, respectively, the number of individuals of the target population at time t in region r and the aggregate transition probability from region s to region r at the time interval $(t-1,t)$, then we can simply define $N_{t,rs} = N_{t-1s} \times \tau_{t,rs}$ and trivially build the origin-destination matrix for each time interval $(t-1,t)$. Under the same general assumption as before, if individuals are to move across the geographical territory independently of their mobile network operator (or even not being a subscriber or carrying two devices), we can postulate $\tau_{t,rs} = \tau_{t,rs}^{\mathrm{net}}$, as before.

One of the advantages of the simulator is that we can analyse the sensitivity of the final outputs with respect to the accuracy of the auxiliary information. In particular, we can provide perturbed values for the register-based population figures in terms of their relative bias $\mathbb{E}\left(\frac{N_r^{\mathrm{reg}} - N_r^{\mathrm{reg},0}}{N_r^{\mathrm{reg},\,0}}\right)$ and their coefficient of variation $\frac{\sqrt{\mathbb{V}(N_r^{\mathrm{reg}})}}{N_r^{\mathrm{reg},0}}$.

In figure 10 we represent the high-density credible intervals ($\alpha = 0.95$) for the target population counts in each region r at the initial time instant t_0 using the negative beta binomial model (16) and the integration formula (1). Different values of the relative bias and the coefficient of variation for the register-based population are investigated.

In figure 11 we represent the high-density credible intervals ($\alpha = 0.95$) for the origin-destination matrices of the present population using the negative beta binomial model (16) at the initial time estimation, the integration formula (1) for all time instants, and the assumption $\tau_{t,rs} = \tau_{t,rs}^{\mathrm{net}}$.

5 Conclusions and future prospects

Mobile network data is a complex data source with multiple potential uses in the production of official statistics. To achieve the daily incorporation of this data source into statistical offices many challenges must be overcome. By and large, they are both strategic and technical. To reach a successful solution, strategy and technique must have a two-way interrelation.

From the technical point of view the design and implementation of a modular end-to-end process according to the principles of the ESS Reference Methodological Framework stand as a key element. We have provided a first proposal of such a process where functional modularity and seamless evolvability arise as the main characteristics. Each module is designed to deal with different aspects of the whole complex estimation process from raw telecommunication data to final

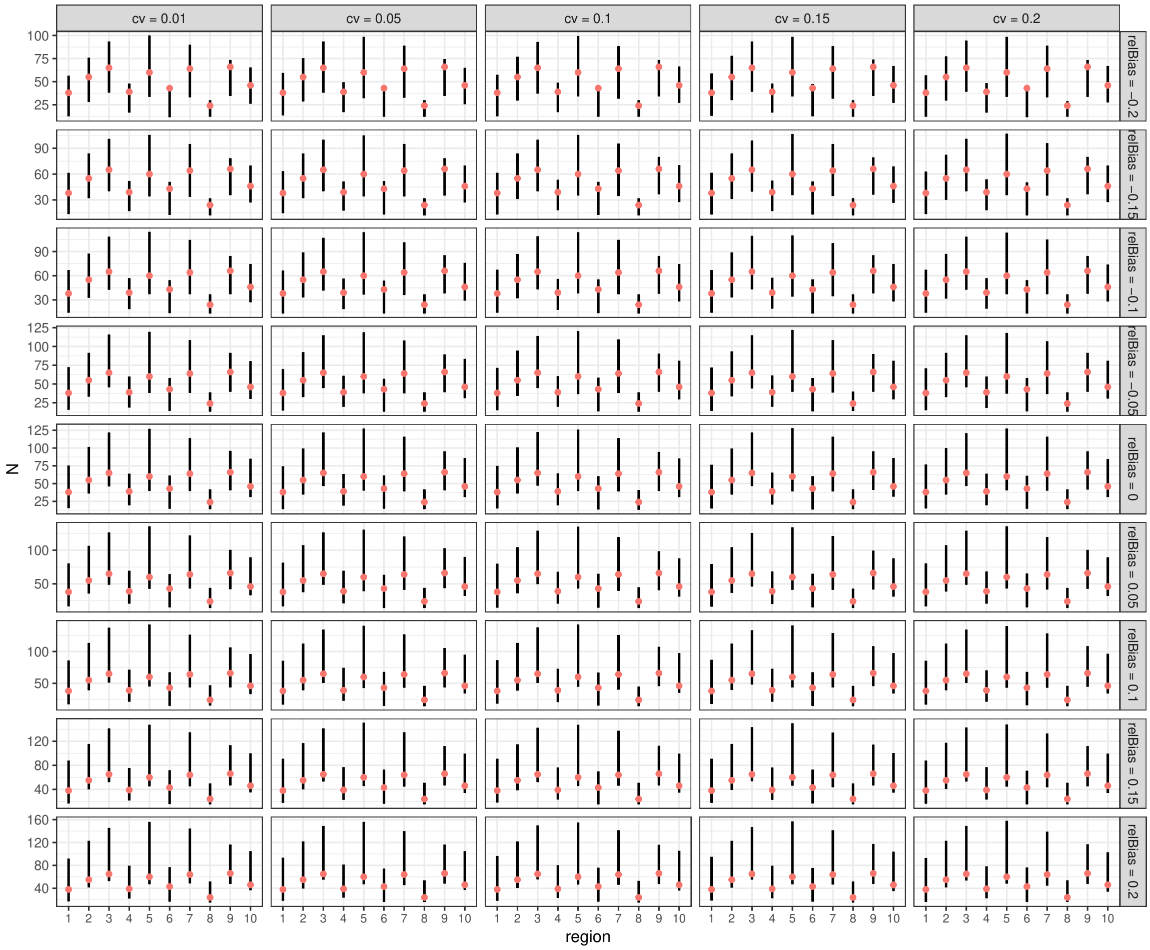

Figure 10: Credible intervals ($\alpha = 0.95$, HDI) for the number of individuals in the target population at the initial time t_0 using the negative beta binomial model and different values for the relative bias and coefficient of variation of the register-based population. True values of target population counts in red.

target population estimates together with accuracy indicators.

Since the access to mobile network data is notably limited for different reasons, we make use of a network event data simulator to provide a proof of concept. Each module needs further development. The use of HMM for geolocation of mobile devices should be further extended with more complex definitions of state and more sophisticated proposals such as continuous-time hidden Markov chains and particle-filter approaches are to be explored. Deduplication procedures and aggregation methods should be accordingly adapted. Statistical filtering algorithms for target inidividuals and anchor point identification in terms of movement analysis indicators need to be proposed and tested using both simulated and real data. Finally, inference models should be extended for open populations and be essentially multivariate including spatial autocorrelations.

Advances in the design of such a process for multiple statistical domains must be taken into account in the management of access and use of this data and agreements with MNOs.

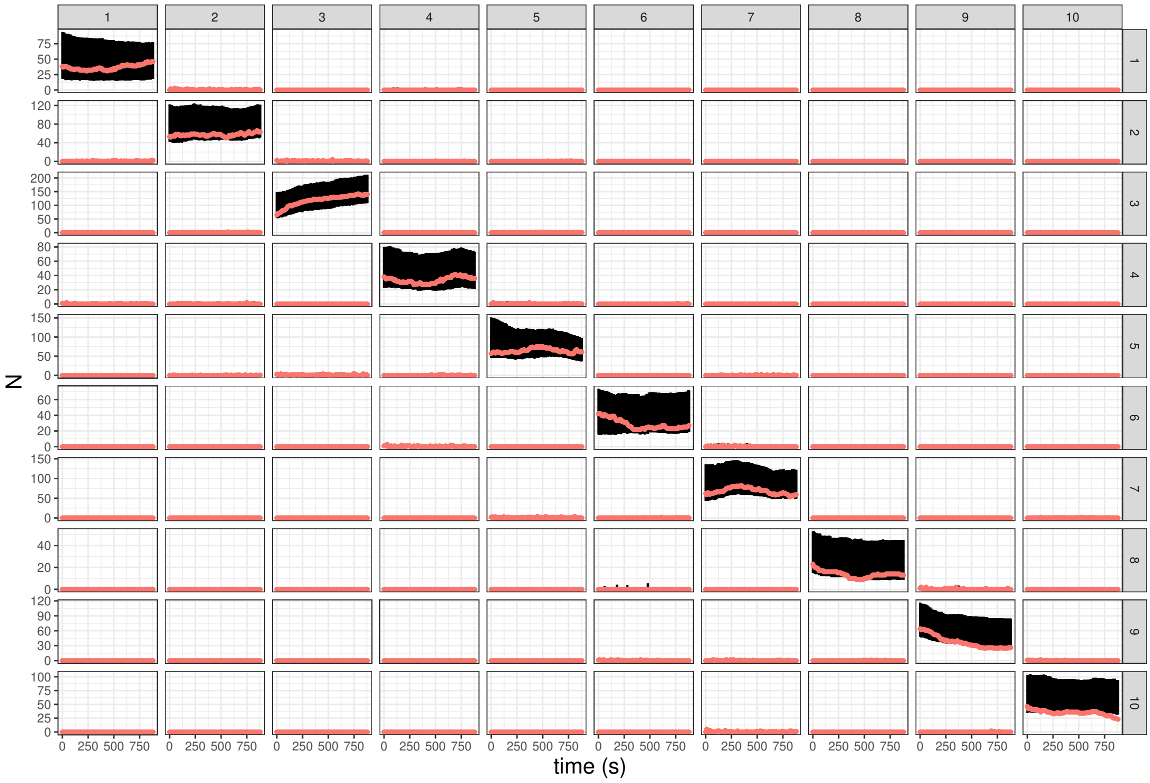

Figure 11: Credible intervals ($\alpha = 0.95$, HDI) for the O-D matrices in the target population using the negative beta binomial model and values for the relative bias 20% and coefficient of variation 20% of the register-based population. True values of target population counts in red.

A Mathematical details

A.1 Geolocation

We include some generic mathematical details to compute the posterior location probabilities from the input data. This is conducted in steps.

A.1.1 Time discretization and padding

We shall work in discrete times. To do this we need to relate three parameters, namely (i) the tile dimension l (we assume a square grid for simplicity), (ii) the time increment Δt between two consecutive time instants, and (iii) an upper bound $v_{\max}$ for the velocity of the individuals in the population. In our transition model we impose that in the time interval Δt, the device d at most can displace from one tile to an adjacent tile. Under this condition, we can trivially set $\Delta t \lesssim \frac{l}{v_{\max}}$. If in the dataset the device d is detected at longer time periods, then we artificially introduce missing values at intervals Δt between every two observed values. This artificial non-response allows us to work with parsimonious models easier to estimate instead of using more complex transition matrices.

Additionally, each observed time instance t is approximated to its closest multiple integer of Δt so that we will have as input data a sequence of time instants at multiples $t_n = \Delta t \cdot n, (n \geq 0)$ and a randomly alternate sequence of missing values and of observed event variables $\mathbf{E}_{dt_n}$.

A.1.2 Construction of the emission model

The emission model is directly built by computing the so-called emission probabilities, i.e. the event location probabilities $\mathbb{P}\left(\mathbf{E}_{t_n} = \mathbf{e}_j \middle| T_{dt_n} = i\right)$, where $\mathbf{e}_j$ is a possible value for the observed event variables $\mathbf{E}_{dt_n}$ and i denotes the tile index. We assume time homogeneity. This conditional probability is computed using the radio wave propagation model of our choice (see e.g. Salgado et al., 2020, for details).

A.1.3 Construction of the transition model

Now we specify a model for the transition between tiles (states) $\{T = i\}_{i=1,\ldots,N_T}$. For ease of explanation and notation, let us change the notation of each tile T_i to a two-dimensional index $T_{(i,j)}$. Accordingly, each tile will be specified in this section by a pair of integer coordinates. The correspondence between both enumerations is arbitrary, but fixed once it has been chosen. We again assume time homogeneity for simplicity. Thus, $\mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right)$ will denote $\mathbb{P}\left(T_{(r,s)}(t_n + \Delta t) \middle| T_{(i,j)}(t_n)\right)$ for any $t_n = 0, 1, \ldots$ We assume a square regular grid for simplicity.

The essential assumption of the model is that an individual can at most reach an adjacent tile in time Δt. Thus,

$$\mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) = 0 \qquad \max\{|r-i|, |s-j|\} \geq 2, \qquad r, s, i, j = 1, \ldots, \sqrt{N_T}. \tag{19a}$$

Now, we assume that we have no further auxiliary information to model these transitions and impose rectangular isotropic conditions:

$$\begin{aligned} \mathbb{P}\left(T_{(i\pm1,j)} \middle| T_{(i,j)}\right) = \mathbb{P}\left(T_{(i,j\pm1)} \middle| T_{(i,j)}\right) &= \theta_1 \qquad i, j = 1, \ldots, \sqrt{N_T}, & (19b)\\ \mathbb{P}\left(T_{(i\pm1,j\pm1)} \middle| T_{(i,j)}\right) &= \theta_2 \qquad i, j = 1, \ldots, \sqrt{N_T}. & (19c) \end{aligned}$$

The last set of conditions is row-stochasticity:

$$\begin{aligned} \sum_{r,s=1}^{N_T} \mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) &= 1, \qquad i, j = 1, \ldots, \sqrt{N_T}, & (19d)\\ \mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) &\geq 0, \qquad i, j, r, s = 1, \ldots, \sqrt{N_T}. \end{aligned}$$

Now back to the original notation for tiles and using the usual notation for the transition matrix $A = [a_{ij}]$, with $a_{ij} = \mathbb{P}\left(T_{jt} \middle| T_{it}\right)$ (Rabiner, 1989), conditions (19) amounts to having a highly sparse transition matrix A with up to 4 terms equal to θ_1 and θ_2 (each) per row and diagonal entries guaranteeing row-stochasticity.

In our proposed implementation, in order to seek future generalization, we will work with a generic block-tridiagonal matrix where the restrictions (19a) leading to 0 have been included, and complemented with the rest of restrictions (19b)-(19d) in matrix form. Thus, we write $C \cdot \mathrm{vec}(\tilde{A}) = \mathbf{b}$, where $\mathrm{vec}(\tilde{A})$ stands for the non-null elements of A in vector form. The rows of $[C\ \mathbf{b}]$ encode each of the restrictions (19b), (19c), and (19d). For example, $a_{12} = \theta_1$ and $a_{21} = \theta_1$ produce a row like this

$$C_i \cdot \mathrm{vec}(\tilde{A}) = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot (\cdots \cdot a_{12} \cdot \cdots \cdot a_{21} \cdot \cdots)^T = b_i = 0.$$

A.1.4 Construction of the initial state (prior) distribution

The HMM prior can be constructed according to any available information. For illustrative purposes, we consider two choices: (i) uniform prior, i.e. $\pi_i^{\mathrm{uniform}} = \frac{1}{N_T}$ and (ii) $\pi_i^{\mathrm{network}} \propto \sum_k (\mathrm{RSS}(d(\mathbf{E}_k, T_i)))$ (where RSS is expressed in watts) or $\pi_i^{\mathrm{network}} \propto \sum_k (\mathrm{SDM}(d(\mathbf{E}_k, T_i)))$, depending on the emission model. Any other choice or combination thereof is also possible (see e.g. Tennekes et al., 2020).

A.1.5 Computation of the likelihood

The likelihood is trivially computed using the numerical proviso of setting emission probabilities equal to 1 when there is a missing value in the observed variables (e.g. due to time padding). The general expression for the likelihood is

$$\begin{aligned} L(\mathbf{E}_d) &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{dt_0} = i_0\right) \prod_{n=1}^{N} \mathbb{P}\left(T_{dt_n} = i_n | T_{dt_{n-1}} = i_{n-1}\right) \mathbb{P}\left(E_{dt_n} \middle| T_{dt_n} = i_n\right) \\ &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{dt_0} = i_0\right) \prod_{n=1}^{N} a_{di_{n-1}i_n}(\boldsymbol{\theta}) \cdot \mathbb{P}\left(E_{dt_n} \middle| T_{dt_n} = i_n\right) \end{aligned} \tag{20}$$

Notice that the emission probabilities only contribute numerically providing no parameter whatsoever to be estimated.

A.1.6 Parameter estimation

The estimation of the unknown parameters $\boldsymbol{\theta}$ is conducted maximizing the likelihood. The restrictions coming from the transition model (19) makes the optimization problem not trivial. Notice that the EM algorithm is not useful. Instead, we provide a taylor-made solution seeking for future generalizations with more realistic choices of transition probabilities incorporating land use information. Formally, the optimization problem is given by:

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \\ & a_k \in [0, 1], \end{aligned} \tag{21}$$

where $\mathbf{a}$ stands for the nonnull entries of the transition probability matrix A, the objective function $h(\mathbf{a})$ is derived from the likelihood L expressed in terms of the nonnull entries of the transition matrix A, and the system $C \cdot \mathbf{a} = \mathbf{b}$ expresses the sets of restrictions from the transition model (19) not involving null rhs terms (restrictions (19b), (19c), and (19d)).

The total number of zeroes in the transition matrix A can be proven to be given by $4 \times (N_T - 4) + 4 \times (\sqrt{N_T} - 2) \times (N_T - 6) + (\sqrt{N_T} - 2)^2 \times (N_T - 9) = N_T^2 - 9 \cdot N_T + 12\sqrt{N_T} - 4$ (Salgado et al., 2020). The number of non-null components of $\mathbf{a}$ in problem (21) is $d = 9 \cdot N_T - 12\sqrt{N_T} + 4$.

The number of restrictions n_r not involving zeroes depends very sensitively on the particular transition model chosen for the displacements. In the rectangular isotropic model considered above, it can also be proven to be $n_r = 4 \cdot N_T - 4\sqrt{N_T} - 1 + 4 \times (\sqrt{N_T} - 1)^2 - 1 + N_T = 9 \cdot N_T - 12\sqrt{N_T} + 2$ (Salgado et al., 2020). Thus, the matrix C will have dimensions $n_r \times d$. Notice that $d - n_r = 2$, as expected, since we have two free parameters θ_1 and θ_2.

The abstract optimization problem is thus

$$\begin{array}{rl} \max & h(\mathbf{a}) \\ \text{s.t.} & C \cdot \mathbf{a} = \mathbf{b} \quad , \\ & \mathbf{a} \in [0,1]^d, \end{array} \tag{22}$$

where $C \in \mathbb{R}^{n_r \times d}$ and $\mathbf{b} \in \mathbb{R}^d$. The objective function $h(\mathbf{a})$ is indeed a polynomial in the non-null entries $\mathbf{a}$. This problem can be further simplified using the matrix QR decomposition. Write $C = Q \cdot R$, where Q is an orthogonal matrix of dimensions $n_r \times n_r$ and R is an upper triangular matrix of dimensions $n_r \times d$. Then we can rewrite the linear system as $R \cdot \mathbf{a} = Q^T \cdot \mathbf{b}$ and we can linearly solve variables $a_1, \ldots, a_{n_r}$ in terms of variables $a_{n_r+1}, \ldots, a_d$:

$$\begin{pmatrix} a_1 & \cdots & a_{n_r} \end{pmatrix}^T = \tilde{C}_{n_r \times (d-n_r)} \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T.$$

The system (22) then reduces to

$$\begin{array}{rl} \max & \tilde{h}(a_{n_r+1}, \ldots, a_d) \\ \text{s.t.} & 0 \leq \tilde{C} \cdot \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T \leq 1. \end{array} \tag{23}$$

In our current software implementation we resort to general-purpose optimizers. It remains for future work to find an optimised algorithm to solve (23). The solution $\mathbf{a}^*$ to problem (22) will be introduced in the transition probability matrix, which will thus be denoted by $\widehat{A}$.

A.1.7 Application of the forward-backward algorithm

Once the HMM has been fitted, we can readily apply the well-known forward-backward algorithm (see e.g. Bishop, 2006) to compute the target location probabilities γ_{dti} and γ_{tij}. No novel methodological content is introduced at this point. For our implementation, we have used the scaled version of the algorithm (see (Bishop, 2006)).

A.1.8 Model evaluation

We propose a bias-variance decomposition of the mean squared error of the estimated location as main figure of merit. We define the center of location probabilities and the root mean squared dispersion. Let us denote by $\mathbf{R}_{dt} \in \{\mathbf{r}_i^{(c)}\}_{i=1,\ldots,N_T}$ the random vector for the position of a device according to the distribution of posterior location probabilities γ_{dti}, where $\mathbf{r}_i^{(c)}$ stands for the coordinates of the center of tile c. Let us shortly denote $\bar{\mathbf{R}}_{dt} \equiv \mathbb{E}\mathbf{R}_{dt} = \sum_{i=1}^{N_T} \gamma_{dti} \mathbf{r}_i^{(c)}$. Let us also denote the true position of device d at time t by $\mathbf{r}_{dt}^*$. Then, we can decompose

$$\begin{aligned}
\mathrm{msd}_{dt} \equiv \mathbb{E}||\mathbf{R}_{dt} - \mathbf{r}^{*}_{dt}||^2 &= \mathbb{E}||(\mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}) + (\bar{\mathbf{R}}_{dt} - \mathbf{r}^{*}_{dt})||^2 \\
&= \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}\rangle\right] + \\
&\quad 2 \cdot \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \bar{\mathbf{R}}_{dt} - \mathbf{r}^{*}_{dt}\rangle\right] + \\
&\quad \mathbb{E}\left[\langle \bar{\mathbf{R}}_{dt} - \mathbf{r}^{*}_{dt}, \bar{\mathbf{R}}_{dt} - \mathbf{r}^{*}_{dt}\rangle\right] \\
&= \mathrm{rmsd}^2_{dt} + \mathrm{b}^2_{dt}. \qquad (24)
\end{aligned}$$

This decomposition motivates the definition of bias $\mathrm{b}_{dt} = ||\bar{\mathbf{R}}_{dt} - \mathbf{r}^{*}_{dt}||$ and root mean squared deviation

$$\mathrm{rmsd}_{dt} = \sqrt{\mathbb{E}\left[||\mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}||^2\right]}$$

.

Acknowledgments

The authors acknowledge M.Á. Martínez-Vidal, S. Lorenzo, M. Suárez-Castillo, R. Radini, T. Tuoto, M. Offermans, M. Tennekes, S. Hadam, and F. Ricciato for invaluable insights and debates. This work was partially supported by the European Commission through the European Statistical System [Grant Agreement Number 847375-2018-NL-BIGDATA (ESSnet on Big Data II)].

References

Ahas, Rein, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru (2010, April). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology 17*(1), 3–27.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Cambridge: Springer-Verlag New York Inc.

Blondel, Vincent D., Adeline Decuyper, and Gautier Krings (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science 4*, 10.

Bryant, John R. and Patrick J. Graham (2013, September). Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis 8*(3), 591–622.

Calabrese, Francesco, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti (2013, January). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies 26*, 301–313.

Daskalakis, C., G. Kamath, and C. Tzamos (2015). On the structure, covering, and learning of poisson multinomial distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1203–1217.

Dattilo, B., R. Radini, and M. Sabato (2016, November). How many SIM in your luggage? A strategy to make mobile phone data usable in tourism statistics. In *14th Global Forum on Tourism Statistics*.

Debusschere, Marc, Jan Sonck, and Michail Skaliotis (2016, November). Official Statistics and mobile network operator partner up in Belgium. In *OECD Statistics Newsletter*, Number 65, pp. 11–14.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem (2014, October). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

DGINS (2018). Bucharest memorandum.

Douglass, Rex W, David A Meyer, Megha Ram, David Rideout, and Dongjin Song (2015). High resolution population estimates from telecommunications data. *EPJ Data Science 4*, 4.

European Parliament (2016). *EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1*.

Galiana, Lino, Benjamin Sakarovitch, and Zbigniew Smoreda (2018, October). Understanding socio-spatial segregation in french cities with mobile phone data. DGINS18.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, and Aki Vehtari (2013). *Bayesian Data Analysis*. Taylor & Francis Ltd.

González, Marta C., César A. Hidalgo, and Albert-László Barabási (2008, June). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

Graells-Garrido, Eduardo, Diego Caro, and Denis Parra (2018, December). Inferring modes of transportation using mobile phone data. *EPJ Data Science 7*, 49.

Iqbal, Md. Shahadat, Charisma F. Choudhury, Pu Wang, and Marta C. González (2014, March). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40*, 63–74.

Izquierdo-Valverde, M., J. Prado Mascuñano, and M. Velasco-Gimeno (2016, November). Same-day visitors crossing borders a big and data approach using traffic control. In *14th Global Forum on Tourism Statistics*, Venice, Italy.

Kowarik, A. and M. van der Loo (2018). Using R in the Statistical Office: the experiences of Statistics Netherlands and Statistics Austria. *Romanian Statistical Review 2018*(1), 15–29.

Lehmann, Erich L. and George Casella (2003). *Theory of Point Estimation*. New York: Springer New York.

Lestari, Titi Kanti, Siim Esko, Sarpono, Erki Saluveer, and Rifa Rufiadi (2018, November). Indonesia's experience of using signaling mobile positioning data for official tourism statistics. In *15th World Forum on Tourism Statistics*, Cusco, Peru.

Long, Jed A. and Trisalyn A. Nelson (2013, February). A review of quantitative methods for movement data. *International Journal of Geographical Information Science 27*(2), 292–318.

Louail, Thomas, Maxime Lenormand, Oliva G. Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports 4*, 5276.

Meersman, Freddy De, Gerdy Seynaeve, Marc Debusschere, Patrick Lusyne, Pieter Dewitte, Youri Baeyens, Albrecht Wirthmann, Christophe Demunter, Fernando Reis, and Hannes I. Reuter (2016, June). Assessing the quality and of mobile and phone data as a source of statistics. In *European Conference on Quality in Official Statistics (Q2016)*, Madrid.

Miao, G., J. Zander, W. Sung, and S.B. Slimane (2016). *Fundamentals of Mobile Data Networks*. Cambridge: Cambridge University Press.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Boston, MA: MIT Press.

National Estonian Bank (2020). Methodology for the compilation of international travel statistics.

Nurmi, Ossi (2016). Improving the accuracy of outbound tourism statistics with mobile positioning data. In *15th Global Forum on Tourism Statistics*, Number from, Cusco, Peru.

Oancea, B. and R. Dragoescu (2014). Integrating R and Hadoop for Big Data analysis. *Romanian Statistical Review 2014*(2), 83–94.

Oancea, Bogdan, Marian Necula, Luis Sanguiao, David Salgado, and Sandra Barragán (2019, December). A simulator for network event data. Technical report, Statistics Romania (INS) and Statistics Spain (INE). Deliverable I.2 of Work Package I of ESSnet on Big Data II.

Pappalardo, Luca, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti (2016, June). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics 2*(1-2), 75–92.

Phithakkitnukoon, Santi, Zbigniew Smoreda, and Patrick Olivier (2012, June). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE 7*(6), e39253.

Positium (2016). Technical documentation for required raw data from mobile network operator for official statistics. ESSnet WP5 internal technical report.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Radini, Roberta, Tiziana Tuoto, Raffaella M. Acrari, and D. Salgado (2020). WPI DeliverableI.6. Quality - A proposal for a statistical production process with mobile network data.

Raun, Janika, Rein Ahas, and Margus Tiru (2016, December). Measuring tourism destinations using mobile tracking data. *Tourism Management 57*, 202–212.

Reis, Fernando, Gerdy Seynaeve, Albrecht Wirthmann, Freddy de Meersman, and Marc Debusschere (2017, March). Land use classification based on present population daily profiles from a big data source.

Ricciato, Fabio (2018). Towards a Reference Methodological Framework for processing MNO data for Official Statistics. In *15th World Forum on Tourism Statistics*, Number opera-.

Ricciato, Fabio, Peter Widhalm, Francesco Pantisano, and Massimo Craglia (2017, February). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing 35*, 65–82.

Royle, A.J. and R.M. Dorazio (2009). *Hierarchical modelling and inference in Ecology*. New York: Elsevier.

Sakarovitch, Benjamin, Marie-Pierre de Bellefon, Pauline Givord, and Maarten Vanhoof (2019). Estimating the residential population from mobile phone data, an initial exploration. *Economie et Statistique / Economics and Statistics 505-506*, 109–132.

Salgado, David, M. Elisa Esteban, María Novás, Soledad Saldaña, and Luis Sanguiao (2018, December). Data organisation and process design based on functional modularity for a standard production process. *Journal of Official Statistics 34*(4), 811–833.

Salgado, David, Luis Sanguiao, Sandra Barragán, Bogdan Oancea, and Milena Suarez-Castillo (2020). WPI DeliverableI.3. Methodology - A proposed production framework with mobile network data.

Salgado, D., L. Sanguiao, B. Oancea, S. Barragán, and M. Necula (2020). An end-to-end statistical process with mobile network data for official statistics. Submitted to EPJ Data Science.

Senaeve, Gerdy and Christophe Demunter (2016, November). When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics. In *14th Global Forum on Tourism Statistics*, Venice, Italy.

Shabbir, Noman, Muhammad T Sadiq, Hasnain Kashif, and Rizwan Ullah (2011, September). Comparison of radio propagation models for long term evolution (LTE) network. *International Journal of Next-Generation Networks 3*(3), 27–41.

Templ, M. and V. Todorov (2016, Feb). The Software Environment R for Official Statistics and Survey Methodology. *Austrian Journal of Official Statistics 45*(1), 97–124.

Tennekes, Martijn, Yvonne A.P.M. Gootzen, and Shan H. Shah (2020, May). A Bayesian approach to location estimation of mobile devices from mobile network operator data. resreport, Statistics Netherlands (CBS).

Ucar, I., M. Gramaglia, M. Fiore, Z. Smoreda, and E. Moro (2019). Netflix or youtube? regional income patterns of mobile service consumption. In *NetMob 2019*, Oxford, UK.

UN (2017). Handbook on the use of mobile phone data for official and statistics.

UNECE (2011, June). Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. UNECE (Ed.), 59th Plennay Session of Conference of European Statisticians, Item 4. High-Level Group for the Modernisation of Official Statistics.

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vanhoof, Maarten, Fernando Reis, Thomas Ploetz, and Zbigniew Smoreda (2018, December). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics 34*(4), 935–960.

Venkataraman, Shivaram, Zongheng Yang, Davies Liu, Eric Liang, Hossein Falaki, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael Franklin, Ion Stoica, and Matei Zaharia (2016). SparkR: Scaling R Programs with Spark. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD âĂŹ16, New York, NY, USA, pp. 1099âĂŞ1104. Association for Computing Machinery.

Wang, Zhenzhen, Sylvia Y. He, and Yee Leung (2018, April). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society 11*, 141–155.

White, Tom (2009). *Hadoop: The Definitive Guide* (1st ed.). OâĂŹReilly Media, Inc.

Williams, Susan (2016). Statistical uses for mobile phone data: literature review. Technical report, Office for National Statistics.

Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica (2016, oct). Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM 59*(11), 56âĂŞ65.

Zhao, C., S. Zhao, M. Zhao, Z. Chen, C.-Z. Gao, H. Li, and Y. Tang (2019). Secure multi-party computation: Theory, practice and applications. *Information Sciences 476*, 357–372.