ESSnet Big Data II

Grant Agreement Number: 847375-2018-NL-BIGDATA

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

Workpackage I
Mobile Network Data

Deliverable I.3 (Methodology)
A proposed production framework with mobile network data

Final version, 26 November, 2020

Prepared by:

David Salgado (INE, Spain, coord.)

- Luis Sanguiao (INE, Spain)
- Sandra Barragán (INE, Spain)
- Bogdan Oancea (INS, Romania)
- Milena Suarez-Castillo (INSEE, France)

Workpackage Leader:

David Salgado (INE, Spain)
david.salgado.fernandez@ine.es
telephone : +34 91 5813151
mobile phone : N/A

Contents

1

General Introduction

This document contains the methodological content of the work package on mobile network data for the project ESSnet on Big Data II. Hereafter we draw a modular design to incorporate mobile network data into the production of official statistics. The main goal is not to provide concrete methodological solutions for concrete cases but to pave the way for a generic production framework which can be adapted to the different statistical needs and statistical domains. Nonetheless, even being generic, our proposal already suggests particular statistical methods and techniques to approach different aspects of the production (geolocation, inference, accuracy, etc.).

This document is divided into the following chapters. In this general introduction we set the scene for our proposal for a production framework with mobile network data. In chapter 2 we draw a generic picture of the whole end-to-end process describing in very general terms the content of each module constituting the process. In chapter 3 we describe the statistical methods for the geolocation of mobile devices. In chapter 4 we propose a classification algorithm to detect devices carried by the same individual (device duplicity classification). In chapter 5 we deal with the filtering and identification of statistical units in the data sets. In chapter 6 we show how to estimate the number of individuals detected by the network. In chapter 7 we set up an inferential framework connecting the mobile network datasets with the target population under study. Finally, in chapter 8 we comment on future prospects and several open issues.

1.1. Big Data methodology

It is a well known fact that Big Data sources and new digital data sources in general bring a strong methodological challenge for Official Statistics due to the impossibility of using probability sampling, which is at the core of survey methodology and the traditional statistical production process. New methods must be provided. Thus, it is tempting to state that there does not exist any Big Data methodology. In our view, this statement requires qualification.

In the production environment in a statistical office, survey methodology provides a fully-fledged framework providing many techniques for the different process steps leading to the dissemination of statistical products. The official statistician can easily find different solutions for a variety of practical situations for frame construction, sample selection, data collection, data integration, data error detection and treatment, estimation, statistical disclosure control, etc. (see e.g. Pfeffermann and Rao, 2009a,b). Every consideration regarding a change or extension of the production methodology should take this production environment as a frame of reference.

In this line of thought, for all new digital data sources we truly lack a production framework with similar characteristics to survey methodology. But it is not true that statistical methods and techniques do not exist to produce statistics with these data sources. There does exist a myriad of methods (see e.g. Bishop, 2006; Hastie et al., 2009; Murphy, 2012) (see also WPK.10, 2020). Now the challenging goal is to construct a production framework offering official statisticians different rigorous approaches for each aspect in the statistical process, now enlarged with new data sources. Indeed, this is the compelling leit motiv behind the present ESSnet (ESS, 2019) and the past edition (ESS, 2018) within the European Statistical System.

1.2. Modularity

Survey methodology and official statistical production already underwent strong efforts for the modernization of the statistical process with survey data. Production models such as the GSBPM and GSIM were developed with this aim. They already show one of the main characteristics of industrialised, standardised, and modern production systems: functional modularity. Functional modularity can be seen as the application of four principles, namely, modularity, abstraction, layering, and hierarchy (see Saltzer and Kaashoek (2009) for an application to computer system design; also Salgado et al. (2018) for a proposal for statistical production).

Modularity is just the division of the production process into separate parts, i.e. into modules. Abstraction is the design and division of these modules so that their interaction takes place only through the interfaces making the internals of each module as much independent of the internals of the rest of modules. Layering and hierarchy are two principles to organize all modules by minimizing their interaction, thus coping with complexity. In traditional statistical production, the aforementioned methodological techniques (frame construction, sample selection, data editing, imputation, etc.) are indeed modules naturally derived from the underlying statistical methodology.

With new data sources, especially for those highly technology-dependent, functional modularity in the statistical process should be central. Regarding mobile network data there already exist specific proposals in this line by Eurostat (see e.g. Ricciato, 2018). Indeed, we shall completely adhere to this approach and embed our methodological proposals in this framework. We shall refer to it as the ESS Reference Methodological Framework (RMF) for Mobile Network Data.

The defining feature of the RMF is the detachment of the strongly technological stratum of this data source from the statistical analysis producing official statistics. This detachment is absolutely necessary for mobile network data since these are produced in an ever-changing technological environment in a highly competitive economic sector (telecommunications). The goal is double: to reach a sustainable production process not subjected to rapid and abrupt changes in the technology but simultaneously incorporating all the potentiality this data brings to the production of official statistics.

The basic ideas behind the RMF can be represented in figure 1.1. We divide the whole process into three big modules, namely, the data (D-) layer, the convergence (C-) layer, and the statistical (S-) layer. The D-layer is dedicated to the retrieval of raw telco data and the preliminary processing needed to arrange them for statistical purposes. This layer produces data comprising information about the network events (calls, SMS/MMS, Internet connection, antenna connection, etc.). There exist different technologies associated to this layer (GSM, UMTS, LTE, . . .) and different approaches to the geolocation of network events (cell ID, RSS, ToA, geographical mapping, tracking with filters, . . .). The C-layer embraces the processing

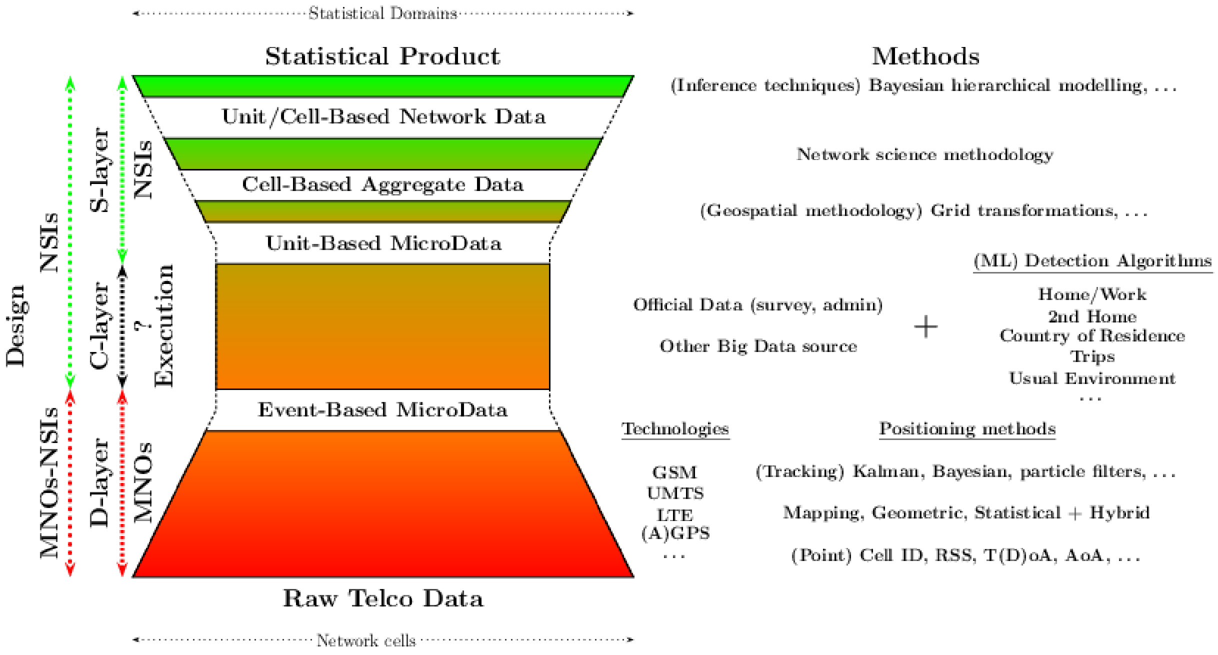
Statistical Domains
Statistical Product
Methods
(Inference techniques) Bayesian hierarchical modelling, . . .
Unit/Cell-Based Network Data
Network science methodology
Cell-Based Aggregate Data
(Geospatial methodology) Grid transformations, . . .
Unit-Based MicroData
(ML) Detection Algorithms
Home/Work
2nd Home
Country of Residence
Trips
Usual Environment
. . .
Official Data (survey, admin)
Other Big Data source
+
Event-Based MicroData
Technologies
Positioning methods
GSM
UMTS
LTE
(A)GPS
. . .
(Tracking) Kalman, Bayesian, particle filters, . . .
Mapping, Geometric, Statistical + Hybrid
(Point) Cell ID, RSS, T(D)oA, AoA, . . .
Raw Telco Data
Network cells
Design
NSIs
MNOs-NSIs
S-layer
C-layer
D-layer
NSIs
?
Execution
MNOs

Figure 1.1: Representation of the Reference Methodological Framework. See also (Ricciato, 2018).

of this event-based microdata to produce unit-based microdata. Data may also be enriched with auxiliary variables, possibly even from official data. Algorithms for detecting home/work locations, second home, usual environment, trips, etc. are to be developed upon this layer. Special regards need the combination of data from several Mobile Network Operators (MNOs). Up to details, this is indeed considered in the RMF proposed by Eurostat, however we shall not consider this scenario in this document, since we are still facing important problems regarding the access and we prioritise the single-MNO case so far. This remains for future work. Finally, in the S-layer all the statistical processing and analysis needs to be carried out for the different statistical domains under study. Questions regarding different time and spatial breakdown and especially the inference techniques must be considered here.

It is important to make a distinction between the design and the execution of the production process. Statistical offices must be part of the design of the whole process, i.e. of the three layers, even the D-layer. For the execution, the S-layer should be executed by statistical offices, but the other two layers should be subjected to a close analysis together with MNOs reaching an optimal agreement. This entails direct consequences regarding the access conditions discussed in a separate deliverable. So far, it seems clear that the D-layer shall be executed by the MNOs themselves. The execution details are definitively entangled to the data access conditions and privacy-preserving analyses must be a guiding principle in their design. As a working hypothesis all throughout this work package, we assume that all sensitive data never leave their original premises at MNOs' information systems, so that preprocessing takes place in situ.

1.3. What data?

Mobile network data for statistical purposes does not exist. Indeed this term makes reference to an abstraction which should be given a concrete substantiation within the extraordinarily complex data ecosystem associated to cellular telecommunication networks. By mobile network data different meanings can be understood depending both on the underlying technology (GMS, UMTS, LTE, . . .) and the statistical topic under stake. For statistical purposes, it is indeed the different statistical topics which should provide scope and context to the telco data to be retrieved, accessed and used. Ultimately, the combined scopes for all statistical domains should provide an answer for the question about what data to use, especially in the long term under sustainable production conditions.

The key challenge is to identify not only traditional statistical needs which can be satisfied probably with higher degrees of spatial and time breakdown but especially novel statistical insights about society thus broadening the role of statistical offices in providing useful information for the social good. In this line, for the time being we have identified three cores for data provision of statistical information, namely (i) geospatial information, (ii) Internet use, and (iii) social interactions.

By geospatial information we mean those telco data related to subscriber mobility. A cellular telecommunication network is characterised, among other things, for its physical implementation over the whole range of a geographical territory thus offering the possibility to geolocate people. This is a first core for statistical information and analysis for Official Statistics in diverse domains (demography, tourism, transport, . . .).

However, MNOs can also provide information about the use of Internet connections (volume, apps, etc.). This information may be used for different sociodemographic analyses. For example, correlations between sorts of apps and income can be analysed (Ucar et al., 2019).

These two topics, nonetheless, may be understood as a refinement of more or less traditional topics in Official Statistics. The richness of this new data source, in our opinion, opens the door to further types of sociodemographic and economic insights. In particular, mobile network data can contain information about the interactions of subscribers paving the way for network science studies (see e.g. Moro et al., 2019). This is a new field in Official Statistics, which traditionally has focused on estimates of totals and similar quantities.

In this document we propose the initial steps for a production framework focused only on the geolocation. Notice however that geolocation is also to be used in the more novel insights. Thus, we need firstly to provide a standardised approach to make use of this type of information provided by mobile network data. The reader can find many studies in the literature focusing on the geolocation information (see e.g. Ahas et al., 2007; Deville et al., 2014; Douglass et al., 2015; Dobra et al., 2015), but all of them are one-off studies, i.e. not providing a framework for a continuous statistical production as in a statistical office. The challenge is indeed to build this framework.

2

The modular structure

This chapter is devoted to provide a generic description of the modular structure of the proposed end-to-end production process with mobile network data. Modularity is a key element in any production process, thus also in statistical production (see e.g. Salgado et al., 2018). Hereafter, we identify the main modules as well as their inputs and outputs. This set of modules has been designed in full alignment with the Reference Methodological Framework for the production of official statistics with mobile network data (Ricciato, 2018) (see figure 1.1 in chapter 1).

The process is proposed up to a certain level of granularity. Details such as, e.g., the visualization techniques for the statistical output are not described in this work. We focus on the main steps driving us from raw telco data to statistical products. We propose five big modules, namely, (i) the geolocation of network events, (ii) device duplicity classification, (iii) statistical filtering, (iv) aggregation, and (v) inference. We must remind the extreme difficulty in getting access to this data, especially for a production sustainable in time, which entails consequent issues in building and testing methodological proposals. This absence of real data, up to some extent, is covered by the use of synthetic data from the network event data simulator developed in a parallel research track in this work package (WPI.2, 2019). Needless to say, in real production some adaptations will need to be undertaken, but the main characteristics of the production framework are intended to be described in this document.

In section 2.1 we provide the generic description of the geolocation of mobile devices, together with different submodules and their inputs and outputs. In section 2.2 we describe in general terms the classification algorithm for mobile devices carried by the same individual. In section 2.3 we describe intermediate steps to filter the geolocated data for the target population under analysis. In section 2.4 we provide a generic description of the aggregation procedure. In section 2.5 we include the principles to provide inference for the traditional problems (target population counts) tackled by Official Statistics. We include a further discussion in section 2.6.

2.1. Geolocation of mobile devices: the event and transition models

The first feature which mobile network data provides for statistical purposes is the geolocation of mobile devices through the constant interaction between a mobile device and a telecommunication network, whose result will be called **network events**. These comprise the set of variables generated, collected, and stored in the different computing systems associated with the interaction between a mobile device and a telecommunication network. There exists a high level of technological complexity in how these network events are generated and how the communication between mobile devices and antennas is produced (see e.g. Miao et al., 2016). Moreover, this technology is constantly evolving, thus the modular view detaching the technological aspects from the statistical process is necessary (Ricciato, 2018). We aim at condensing the geolocation information comprised in the network. This will be undertaken by using (i) probability theory and (ii) a reference grid over the network service area (the geographical territory). These two elements will be explained in further detail in chapter 3. For the time being, we just define a **reference grid** over the territory under analysis. Each pixel in the grid will be referred to as a **tile**. A natural suggestion for this grid in the European context is the geographical infrastructure from the INSPIRE Directive (INSPIRE, 2007), but other options are possible. We shall condense the geolocation information in the so-called **location probability** or **location distribution** or, simply, **location**, which we define as follows.

Firstly we need to introduce the available information (from the network and any other information) to compute these location distributions. Let us denote by $\mathbf{E}_d(t)$ the set of variables associated to an event of mobile device d at time $t \in [t_0, t_f]$. These variables can be the cell ID where the device is detected, the timing advance of the signal (TA), the angle of arrival of the signal (AoA), or any other telco variable related to the location of the device (see WP5.2 (2017) for a description of the variables in a mobile network dataset for statistical purposes). The generality is introduced on purpose to cover as many different situations as possible. When the time variable is dropped, we mean that all time instants are considered, i.e. $\mathbf{E}_d \equiv (\mathbf{E}_d(t_0), \mathbf{E}_d(t_1), \ldots, \mathbf{E}_d(t_f))$.

Let us denote by $\mathbf{I}(t)$ the set of variables comprising any sort of auxiliary information such as land use, geographic information, etc. This information shall be incorporated into the estimation of the location probabilities. We assume the same convention when dropping the time variable.

We define by $T_d(t)$ the random variable for the tile where a mobile device d is located at a time instant t. A set of tile indices will be denoted with calligraphic letters $\mathcal{T}$ so that a territorial region r can be denoted as a set of tiles $r = \bigcup_{i \in \mathcal{T}_r} T_i$.

We shall denote by $\gamma_{di}(t) \equiv \mathbb{P}\left(T_{di}(t) \middle| \mathbf{E}_d, \mathbf{I}\right)$ the probability for a device d to be located at tile T_i at time t conditioned upon all observed and auxiliary variables for the whole time interval $[t_0, t_f]$. Also, we shall denote by $\gamma_{dij}(t, t') \equiv \mathbb{P}\left(T_{di}(t), T_{dj}(t') \middle| \mathbf{E}_d, \mathbf{I}\right)$ the probability for a device d to be located at tile T_i at time t and tile T_j at time t' conditioned upon all observed and auxiliary variables for the whole time interval $[t_0, t_f]$. These will be referred to as **location probabilities** and **joint location probabilities**, respectively.

The target mathematical objects for the module on the geolocation of network events will be these location and joint location probabilities $\gamma_{di}(t) \equiv \mathbb{P}\left(T_{di}(t) \middle| \mathbf{E}_d, \mathbf{I}\right)$ and $\gamma_{dij}(t, t') \equiv \mathbb{P}\left(T_{di}(t), T_{dj}(t') \middle| \mathbf{E}_d, \mathbf{I}\right)$. It is important to point out that by focusing on $\gamma_{di}(t)$ and $\gamma_{dij}(t, t')$, being probabilities, we can integrate the uncertainty in the locations and movements from the onset. This will be relevant for quality issues later on. Also, as a particular case, if conditions are met so that we can pinpoint where every mobile device d is located, this will be reduced to trivial cases with zero-one probabilities. The joint probabilities will be important for the inference module (section 2.5). Furthermore, the language of probability is very adequate to treat complexity in data (Murphy, 2012). We shall distinguish between two different situations: (i) static analysis and (ii) dynamical analysis.

2.1.1. Static approach

In this situation, we shall compute the location probabilities at each time instant t independently, i.e. not taking account the evolution of the mobile devices in trajectories. As we shall see, this is conducted with a direct use of Bayes' theorem:

$$\mathbb{P}\left(T_d(t)\middle|\mathbf{E}_d(t),\mathbf{I}(t)\right) \propto \mathbb{P}\left(T_d(t)\middle|\mathbf{I}(t)\right)\mathbb{P}\left(\mathbf{E}_d(t)\middle|T_d(t),\mathbf{I}(t)\right), \tag{2.1}$$

for each time instant t independently. We shall refer to $\mathbb{P}\left(T_d(t)\middle|\mathbf{I}(t)\right)$ as the **prior location probability** or, simply, **prior location**. Also, we shall refer to $\mathbb{P}\left(\mathbf{E}_d(t)\middle|T_d(t),\mathbf{I}(t)\right)$ as the **location likelihood**, **event location probability** or, simply, **event location** depending on the context (notice that very often the difference between a likelihood and a probability reduces to a matter of focus on variables and parameters). In chapter 3 we provide details about how to compute these probabilities.

2.1.2. Dynamical approach

When we consider the dynamical behaviour of mobile devices, we can model the dynamics in the grid and interrelate these probabilities to extract more information from the data.

We propose to use hidden Markov models (see e.g. Bishop, 2006) to model this behaviour (see figure 2.1). Firstly, this means that the time-continuous behaviour of events $\mathbf{E}_d(t)$ and of locations $T_d(t)$ are coarse-grained into finite-time sequences of averaged events $\mathbf{E}_{dt}$ and discrete random variables T_{dt} (notice that time is also a subscript now). Secondly, we define an unobserved Markov chain for the time sequence $t = 0, 1, 2, \ldots, T$ of locations T_{dt} in the grid. Remember that the real locations of each mobile device are never actually observed. Thirdly, what we observe at every time instant t is the set of variables $\mathbf{E}_{dt}$ defining each event[1].

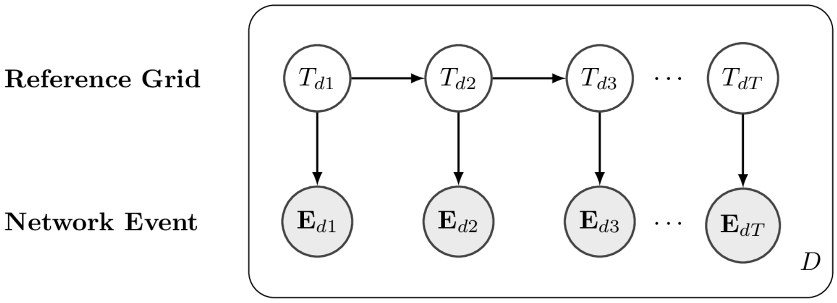

Figure 2.1: Hidden Markov model for the dynamical behaviour of a mobile device d.

Our goal with this statistical model shall be to infer about the locations and movements using the information from the events, i.e. to estimate the location probabilities $\gamma_{dti} \equiv \mathbb{P}\left(T_{dti}|\mathbf{E}_d, \mathbf{I}\right)$ and joint probabilities $\gamma_{dt,ij} \equiv \mathbb{P}\left(T_{dti}, T_{dt-1j}|\mathbf{E}_d, \mathbf{I}\right)$ in this dynamical context taking into account the information from all events.

In chapter 3 we provide methods to conduct this estimation with an adaptation of the so-called forward-backward algorithm (see e.g. Bishop, 2006). This algorithm amounts to

[1] We may relax this to having observations at a subsequence of time instants $\{t_{i_1}, t_{i_2}, \ldots\} \subset \{0, 1, \ldots, T\}$, depending on the sort of raw telco data we have (e.g. Call Detail Records). The proposed methodology can indeed deal with missing variables (see below).

incorporating information from the past and the future in the estimation procedure of the locations. The two core elements in this HMM are (i) the so-called **transition probabilities** $\mathbb{P}\left(T_{dt}|T_{dt-1},\mathbf{I}\right)$ to model the dynamical behaviour and (ii) again the so-called **emission probabilities** $\mathbb{P}\left(\mathbf{E}_{dt}\middle|T_{dt},\mathbf{I}\right)$ to model the observation process (i.e. the location likelihoods or event locations in the static context). We shall refer to them as the transition model and the event model, respectively.

Indeed, the data provenance for these two models are so different and play such a different role within the RMF that they must be considered as independent (sub)modules. For the event model, the input data will be the raw telco data such as cell ID, timing advance (TA), and similar available and accessible information in the network. For this reason, this must be executed in the D-layer. For the transition model, the input data will mainly concern land use and model assumptions about the individual movement patterns. No telco data is needed. For this reason, this must be executed in the S-layer (hence, the separation of these two models).

Both the emission and transition models are put together with each device data as input data into the HMM[2] so that the location γ_{dti} and joint location probabilities γ_{dtij} are obtained in the C-layer as the first core intermediate dataset for the S-layer. Metrics to assess the performance of this production step are put off to the deliverable on quality.

It is important to underline that we do not claim to provide a single-shot specific solution for any statistical purpose but an inferential and production step versatile enough to be adapted to different statistical needs and statistical domains. In this line, the concept of state and the Markov property provide great flexibility to model very diverse situations (see e.g. Bishop, 2006).

2.2. Device duplicity classification

An important module comprises the task devoted to estimate device-individual duplicity probabilities, i.e. probabilities of individuals carrying more than one mobile device. To do this we shall use the HMM fitted for each mobile device. We formulate this task as a classification problem to decide whether each given device d corresponds to an individual carrying[3] only one device (1:1 correspondence) or two devices (2:1 correspondence). The key element is to produce a probability of such an event for each device.

Details will be provided in chapter 4. The bottom line of our approach is simple. We shall make a comparison of each device with all other devices. We shall make two proposals. First, we shall conduct an analysis based upon the spatial distributions $\{\gamma_{dti}\}$ and compute the probability distribution for the distance between the centers of location probabilities[4] of a pair of devices. Then we shall establish a threshold. With simulated data we can perform a ROC analysis to assess the selection of this threshold.

For the second proposal we shall follow a Bayesian approach and compute the probabilities of duplicities conditional on the network event information $\mathbf{E}_d$ for all devices d and any auxiliary information $\mathbf{I}$. These probabilities are computed using Bayes' theorem with some weakly informative priors and the likelihoods coming from the HMMs models and a new HMM model

[2] Notice that the static context is somewhat contained in the dynamical model just by using trivial transition probabilities.

[3] For simplicity we shall assume that an individual carries at most two devices. The generalization is evident.

[4] See section 3.4 for this concept.

under the assumption of an individual producing pairs of network event data variables $\mathbf{E}_{d_1}$ and $\mathbf{E}_{d_2}$.

We build a probabilistic classifier upon this setting producing for each mobile device d a Bernoulli variable $\omega_d \simeq \text{Ber}\,(p_d)$ providing the probability p_d for device d to correspond to an individual carrying two devices.

The input data for this module shall be basically the same input data for the HMM. In practice, both modules are recommended to be executed side by side. The output data will be the duplicity probabilities p_d for each device. The throughput amounts to the concrete algorithm to be applied, which also determines the input parameters.

2.3. Statistical filtering

A central concept in the estimation exercises of population density is the definition of **target population**. This definition drives us to select devices in our set of location probabilities. For example, should the target population be the population of domestic tourists in a country, then we must filter those location and joint probabilities corresponding to devices with a dynamical behaviour of a domestic tourist.

The design and implementation of algorithms to filter target devices/units in the mobile network data will require intermediate computation and algorithms such as the identification of home/work locations, second home locations, commuting patterns, usual environments, trips and so on (WP5.3, 2018). Given the current difficulties to access real data and the higher level of complexity in simulating these concepts in a realistic way with the network event data simulator, we shall focus on estimating *present population* providing thus population counts at different levels of territorial and time disaggregation. The set of target devices will be denoted by $\{d_1, \dots, d_{D^{\text{target}}}\} \subset \{1, \dots, D\}$.

For this module on filtering the input data is the set of location and joint location probabilities $\{\gamma_{dt\cdot}, \gamma_{dt\cdot\cdot}\}_{t=0,\dots,T}^{d=1,\dots,D}$ and the duplicity probabilities $\{p_d\}_{d=1,\dots,D}$ together with auxiliary information such as land use. The output data is the filtered set of location and joint probabilities for the target population at stake $\{\gamma_{d_it,\dots}\}_{t=1,\dots,T}^{i=1,\dots,D^{\text{target}}}$ and the corresponding duplicity probabilities $\{p_{d_i}\}_{i=1,\dots,D^{\text{target}}}$. The throughput will depend on details about the target population. Input parameters will in turn depend on the algorithms and the target population under analysis.

2.4. Aggregation

A key step in the estimation of population counts is the aggregation step from data at the device level to territorial cells $r = \bigcup_{i \in \mathcal{T}_r} T_i$. Our reasoning to approach this step can be easily summarised as follows. The location of each device is provided by a discrete probability distribution over the set of tiles, thus **the number of individuals detected by the network in a given tile will also be a discrete random variable**.

It is important to underline that we focus on the number of individuals detected by the network, *not on the number of devices*. The number of devices is never used in our proposed statistical process. This has strong implications for the access agreements with MNOs. Currently, we claim that it is at the microlevel (data per device) where we should deal with the device duplicity issue.

The construction of a random variable for the number of individuals detected by the network can be formalised using probability theory under very mild assumptions. We shall define $\mathbf{N}^{\text{net}}$ as a multivariate discrete random variable providing the number of individuals detected by the network at each region r of interest, so that $\mathbf{N}^{\text{net}} = \left(N_1^{\text{net}}, \ldots, N_R^{\text{net}}\right)^T \in \mathbb{N}^{\times R}$. We have dropped time subscripts for ease of notation. If we define a multivariate random variable $\mathbf{T}_d$ for each device d with values on the unit vectors $\mathbf{e}_{dr}$ or $\frac{1}{2} \cdot \mathbf{e}_{dr}$ with probabilities $\gamma_{dr} \cdot (1 - p_d)$ and $\gamma_{dr} \cdot p_{dr}$, respectively, then

$$\mathbf{N}^{\text{net}} = \sum_{d=1}^{D^{\text{target}}} \mathbf{T}_d \simeq \text{Poisson-multinomial},$$

which follows a Poisson-multinomial distribution under the mild assumption of independence among all devices. The aggregate probabilities γ_{dr} are trivially computed as the sum of the original location probabilities γ_{di}.

Notice that a similar argument is also valid to compute the probability distribution of the number of individuals detected by the network $N_{t(r|s)}^{\text{net}}$ going from region s at time $t-1$ to region r at time t. Just define the conditional probability $\gamma_{dt(r|s)} \equiv \frac{\gamma_{dtsr}}{\gamma_{dt-1s}}$ and follow the same reasoning as before.

For the module on aggregation the input data is the set of target location and joint location probabilities $\{\gamma_{d_i t,\ldots}\}_{t=0,\ldots,T}^{i=1,\ldots,D^{\text{target}}}$ and the duplicity probabilities $\{p_{d_i}\}_{i=1,\ldots,D^{\text{target}}}$. The output data is the distribution of the number of individuals detected in each territorial cell of interest and of the number of movements between pair of tiles (thus, also the origin-destination matrices). Again, performance metrics are also put off to the deliverable on quality.

2.5. Inference

This module focuses on inferring the number of individuals in a population using the number of individuals detected in the network and auxiliary information. Again, the goal is not to provide a definitive specific solution for all statistical domains but to identify and to introduce an inferential framework versatile enough to be adapted to the different statistical needs in statistical production. We also proceed in steps: for the time being we shall consider only closed populations. Open populations remain for future work.

Our proposal is to mimick the ecologists' approach to the species abundance problem (Royle and Dorazio, 2014) and to use hierarchical Bayesian techniques (Gelman et al., 2013) inside a probabilistic graphical model (Koller and Friedman, 2009) to compute the posterior distribution for the number of individuals $\mathbb{P}\left(\mathbf{N}_{\cdot t} \middle| \mathbf{E}_{1:D}, \mathbf{N}_{\cdot 0}^{\text{REG}}, \mathbf{P}_{\cdot 0}^{\text{net}}, \mathbf{I}\right)$ at each time t conditioned on the detected events $\mathbf{E}_{1:D}$, the register-based (resident) population $\mathbf{N}_{\cdot 0}^{\text{REG}}$, the MNO penetration rates $\mathbf{P}_{\cdot 0}^{\text{net}}$, and any auxiliary information $\mathbf{I}$. We comment on this choice:

- Bayesian inference is not new to Official Statistics. It is one of the options for small area estimation (Rao and Molina, 2015). More recently, it has been proposed to estimate demographic accounts integrating different administrative registers (Bryant and Graham, 2013). Also, Bayesian proposals to overcome the inferential schizophrenia with design-based inference were proposed some years ago aiming at providing a homogeneous inferential framework at all levels of population disaggregation (Little, 2012).
- It provides a natural statistical framework to integrate the information coming from different sources (as e.g. in Bryant and Graham (2013)).

- Having a probability distribution as output we may provide point and interval estimates as well as accuracy measures. Furthermore, in the Bayesian paradigm we also have the posterior predictive distribution so that model checking indicators can be straightforwardly computed (Gelman et al., 2013).

For the present population estimation problem, we propose two assumptions:

- There exists an initial time instant t_0 in which the registed-based (resident) population and the present population can be assimilated (i.e. individuals are at their home location at 6:00am).
- The dynamics of subscribers in an MNO can be assimilated to the dynamics of individuals in the general population (i.e. individuals move around the territory independently of the MNO they are subscribed to).

The estimation is proposed as an exercise with a probabilistic graphical model such as in figure 2.2. The outputs from the first modules provide the conditional probability distributions $\mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot t}\middle|\mathbf{E}_{1:D},\mathbf{I}\right)$ and $\mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot\cdot t}\middle|\mathbf{E}_{1:D},\mathbf{I}\right)$, where we have abstracted away in the notation the intermediate variables and parameters from the hidden Markov model (i.e. $\mathbf{E}_{1:D}$ stands for the plate containing the hidden Markov model, $\mathbf{N}^{\text{net}}_{\cdot t}$ denotes the vector of numbers of individuals detected by the network at stake in each territorial unit at time t, $\mathbf{N}^{\text{net}}_{\cdot\cdot t}$ denotes the matrix of numbers of individuals detected by the network at stake at consecutive times $t-1$ and t, and $\mathbf{I}$ stands for any auxiliary information).

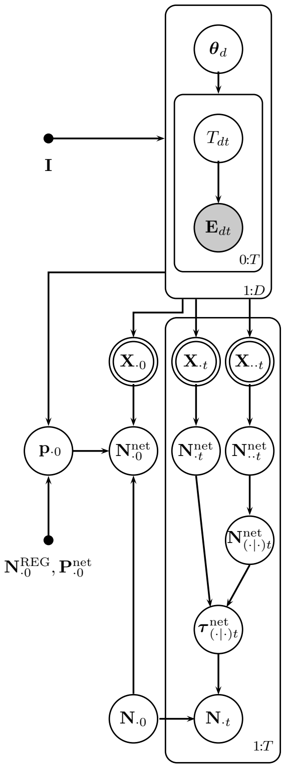
$\boldsymbol{\theta}_d$
T_{dt}
$\mathbf{E}_{dt}$
0:T
1:D
$\mathbf{I}$
$\mathbf{X}_{\cdot 0}$
$\mathbf{X}_{\cdot t}$
$\mathbf{X}_{\cdot\cdot t}$
$\mathbf{p}_{\cdot 0}$
$\mathbf{N}^{\text{net}}_{\cdot 0}$
$\mathbf{N}^{\text{net}}_{\cdot t}$
$\mathbf{N}^{\text{net}}_{\cdot\cdot t}$
$\mathbf{N}^{\text{net}}_{(\cdot|\cdot)t}$
$\mathbf{N}^{\text{REG}}_{\cdot 0}, \mathbf{P}^{\text{net}}_{\cdot 0}$
$\boldsymbol{\tau}^{\text{net}}_{(\cdot|\cdot)t}$
$\mathbf{N}_{\cdot 0}$
$\mathbf{N}_{\cdot t}$
1:T

Figure 2.2: A proposed PGM for the present population estimation.

The estimation procedure is thus devised in two steps. At the initial time (first step), we propose a hierarchical model for the number of individuals $\mathbf{N}_{\cdot 0}$ with different layers in the hierarchy depending on the available information. The model can be specified using the graph in figure 2.2 to unravel the conditional dependencies in the probability distributions:

$$\begin{aligned}\mathbb{P}\left(\mathbf{N}_{\cdot 0}\middle|\mathbf{E}_{1:D},\mathbf{N}^{\text{REG}}_{\cdot,0},\mathbf{P}^{\text{net}}_{\cdot 0},\mathbf{I}\right) &= \sum_{\mathbf{N}^{\text{net}}_{\cdot 0}}\mathbb{P}\left(\mathbf{N}_{\cdot 0},\mathbf{N}^{\text{net}}_{\cdot 0}\middle|\mathbf{E}_{1:D},\mathbf{N}^{\text{REG}}_{\cdot,0},\mathbf{P}^{\text{net}}_{\cdot 0},\mathbf{I}\right)\\ &= \sum_{\mathbf{N}^{\text{net}}_{\cdot 0}}\mathbb{P}\left(\mathbf{N}_{\cdot 0}\middle|\mathbf{N}^{\text{net}}_{\cdot 0},\mathbf{P}^{\text{net}}_{\cdot 0}\right)\cdot\mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot 0}\middle|\mathbf{E}_{1:D},\mathbf{I}\right)\end{aligned}\tag{2.2}$$

As stated above, the probability $\mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot 0}\middle|\mathbf{E}_{1:D},\mathbf{I}\right)$ comes from the preceding modules. Now, we need to compute the probability $\mathbb{P}\left(\mathbf{N}_{\cdot 0}\middle|\mathbf{N}^{\text{net}}_{\cdot 0},\mathbf{P}^{\text{net}}_{\cdot 0}\right)$. We used hierarchical modelling techniques for this:

$$\begin{aligned}\mathbb{P}\left(\mathbf{N}_{\cdot 0}\middle|\mathbf{N}^{\text{net}}_{\cdot 0},\mathbf{P}^{\text{net}}_{\cdot 0}\right) &\propto \mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot 0},\mathbf{N}_{\cdot 0}\middle|\mathbf{P}^{\text{net}}_{\cdot 0}\right)\\ &\propto \sum_{\mathbf{p}_{\cdot 0}}\mathbb{P}\left(\mathbf{N}^{\text{net}}_{\cdot 0}\middle|\mathbf{N}_{\cdot 0},\mathbf{p}_{\cdot 0}\right)\cdot\mathbb{P}\left(\mathbf{N}_{\cdot 0}\right)\cdot\mathbb{P}\left(\mathbf{p}_{\cdot 0}\middle|\mathbf{P}^{\text{net}}_{\cdot 0}\right),\end{aligned}\tag{2.3}$$

where $\mathbf{p}_{\cdot 0}$ stands for detection probabilities of individuals by the telecommunication network.

The conditional probability distribution $\mathbb{P}\left(\mathbf{N}_{\cdot 0}^{\text{net}} \middle| \mathbf{N}_{\cdot 0}, \mathbf{p}_{\cdot 0}, \mathbf{P}_{\cdot 0}^{\text{net}}\right)$ represents the first layer in the model:

$$\mathbf{N}_{\cdot 0}^{\text{net}} \simeq \prod_{r=1}^{R} \text{Binomial}\left(N_{r0}, p_{r0}\right) \tag{2.4}$$

The random variables N_{r0} and p_{r0} are further specified in the next layers. In our example above, these are made dependent on deterministic quantities. Notice that we should carry out a modelling exercise, on the one hand, for the system (the population size $\mathbf{N}_{\cdot 0}$) and, on the other hand, for the observation process (the detection probabilities $\mathbf{p}_{\cdot 0}$). Once we normalized the distribution (2.3), we introduce it in (2.2) so that we have the target conditional distribution for the initial population size at each region r.

Now, in the second step, we model the dynamical behaviour (at the aggregated region level) based on the behaviour at the device level. To this end, we define a parameter $\tau_{(r|s)t}^{\text{net}}$ as an aggregate transition probability or proportion estimation of devices going from region s to region r from time $t-1$ to time t by

$$\tau_{(r|s)t}^{\text{net}} = \frac{N_{(r|s)t}^{\text{net}}}{N_{t-1s}^{\text{net}}}, \qquad t \geq 1.$$

Then we model the number of individuals N_{rt} at region r at time t as follows. We set by definition:

$$\begin{aligned}
N_{rt} &= \sum_{r_t=1}^{R} N_{(r|r_t)t} - \sum_{\substack{r_t=1\\ r_t\neq r}}^{R} N_{(r_t|r)t} && \text{(2.5a)}\\
&= \sum_{r_t=1}^{R} \frac{N_{(r|r_t)t}}{N_{r_t t-1}} \times N_{r_t t-1} && \text{(2.5b)}\\
&\equiv \sum_{r_t=1}^{R} \tau_{(r|r_t)t} \times N_{r_t t-1} && \text{(2.5c)}\\
&= \sum_{r_1=1}^{R} \cdots \sum_{r_t=1}^{R} \prod_{i=1}^{t} \tau_{(r|r_i)i} N_{r_1 0}, \qquad r = 1, \ldots, R, && \text{(2.5d)}
\end{aligned}$$

and model recursively $\tau_{(r|s)t}$ in terms of $\tau_{(r|s)t}^{\text{net}}$ for all $r, s = 1, \ldots, R$ and $t \geq 1$ (e.g. by simply equating $\tau_{(r|s)t} = \tau_{(r|s)t}^{\text{net}}$).

The input data for this module is the set of distributions for N_{rt}^{net} and $N_{(r|s)t}^{\text{net}}$. The output data is the set of posterior distributions for N_{rt}. Notice that we can further provide point and interval estimates using these distributions. The throughput is provided by the above modelling exercise. The set of input parameters is provided also in this modelling exercise.

2.6. Discussion

We underline the basic working assumptions for our proposed end-to-end process.

- **Modularity**.- It is absolutely fundamental to design and execute the statistical process in independent modules. This functional modularity entails far-reaching consequences in three respects:

  1. The independence between modules is achieved by designing input and output data sets hiding the throughput of each module from the rest. This is why in the process any building block should follow the generic structure of a statistical production step (see figure 2.3).

  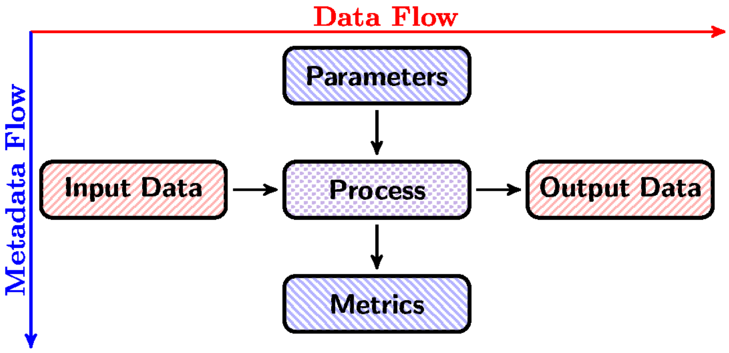

  Figure 2.3: Building block for the statistical process (adapted from Loo (2019)).

  2. The modularity allows us to update, modify, and adapt each module according to the statistical domain, statistical needs, and data ecosystem we face in each case. In this document we do not aim at providing definitive solutions for each different statistical need, but to pave the way for a general production framework. More specific solutions will certainly be developed in the future. With a functionally modular approach, the evolution of this framework is feasible.
  3. The modularity also allows us to approach the issue about data access with a higher degree of flexibility. The design and execution of the process are intimately related to data access (e.g. should NSIs have access to raw telco data?). Upon an agreement about the design of the process, if modularity is followed, the execution can be undertaken in steps whose inputs and outputs can be thought of as intermediate products where the access issue is more easily solved (e.g. having access only to some intermediate and partially aggregated dataset).

  In this chapter we are following a top-down approach so that the level of granularity of the process is still low. In later chapters, when providing specific proposals, we shall descend to finer tasks with more details.

- **Probabilistic Graphical Models**.- One of the most paralysing issues in the incorporation of new data sources to the production of official statistics is the impossibility to use sampling designs and design-based inference (Smith, 1976). Thus, we are driven to the use of model-based inference by fitting models to the concrete input data in each case. Statistical learning algorithms (Hastie et al., 2009) can indeed be understood as an exercise of model-based inference where models are continuously fitted according to new input data.

Official Statistics is facing the challenge to provide a homogeneous inferential framework as versatile and robust as survey methodology. Notice that in design-based inference the sampling designs and the multipurpose property of sampling weights thereof are used to provide a systematic solution to the inference problem in a finite population for *all* variables and *all* population domains[5]. The estimation procedure is thus sustainable for an arbitrary number of target variables. The introduction of models will require to fit a model for every target variable in every population domain. If data volume and data variety grows exponentially, a systematic approach will certainly be needed.

In the case of mobile network data, which we approach in this work package, we resort to probabilistic graphical models for the following reasons:

1. Model-based inference fundamentally amounts to (i) understanding data as realizations of random variables and (ii) assigning probability distributions to these random variables to conduct the inference or estimation. Thus, we are basically diving into probability theory.
2. New data sources, in concrete mobile network data, are characterised by the volume and velocity of generation (hence the famous Vs in the definition of Big Data), i.e. by the complexity derived from the amount of data (apart from their lack of statistical metadata forcing us to a higher degree of preprocessing in comparison to survey data or, even, administrative data). To deal with this complexity, *graph theory* has proven to be a versatile tool (Barabási, 2018).
3. The combination of probability theory and graph theory leads naturally to probabilistic graphical models (Koller and Friedman, 2009), which allows us to cope with the complexity while facing inference in a rigourous setting.
4. The use of probability theory also allows us to propose and compute quality indicators, in particular, to assess accuracy in a natural way. New ingredients in Quality Assurance Frameworks in Official Statistics will be needed (such as model assessment), but there already exists plenty of statistical methods to do that.

By and large, with the new data sources in Official Statistics one may be tempted to let data speak for themselves so that the natural stepwise process in three stages (preprocessing, to prepare data; inferential model, to connect the dataset with the target population; visualization, to visualize the results) is often reduced to introducing preprocessed data into complex visualization software frameworks. Inference and quality need to remain as key in official statistical production. Probabilistic graphical models arise as potential tool to deal with both inference and data complexity in a rigourous way.

In figure 2.4 we summarise the modular structure in terms of the collection of production steps (in rectangular shape) with input and output data (in circular shape) to be executed in each layer of the RMF. In next chapters, we provide concrete methodological proposals for these modules. Notice that we do not deal with the roles of each actor in the process, i.e. we limit ourselves to provide methodological proposals not taking into account management aspects about responsibilities and resources of both MNOs and NSIs in this process. For the time being, as indicated in figure 2.4, we shall assume that basically all modules except the last one is executed by MNOs. This is to be tackled in the access agreements between NSIs and MNOs.

[5] Up to problem with small area estimation (see Little, 2012).

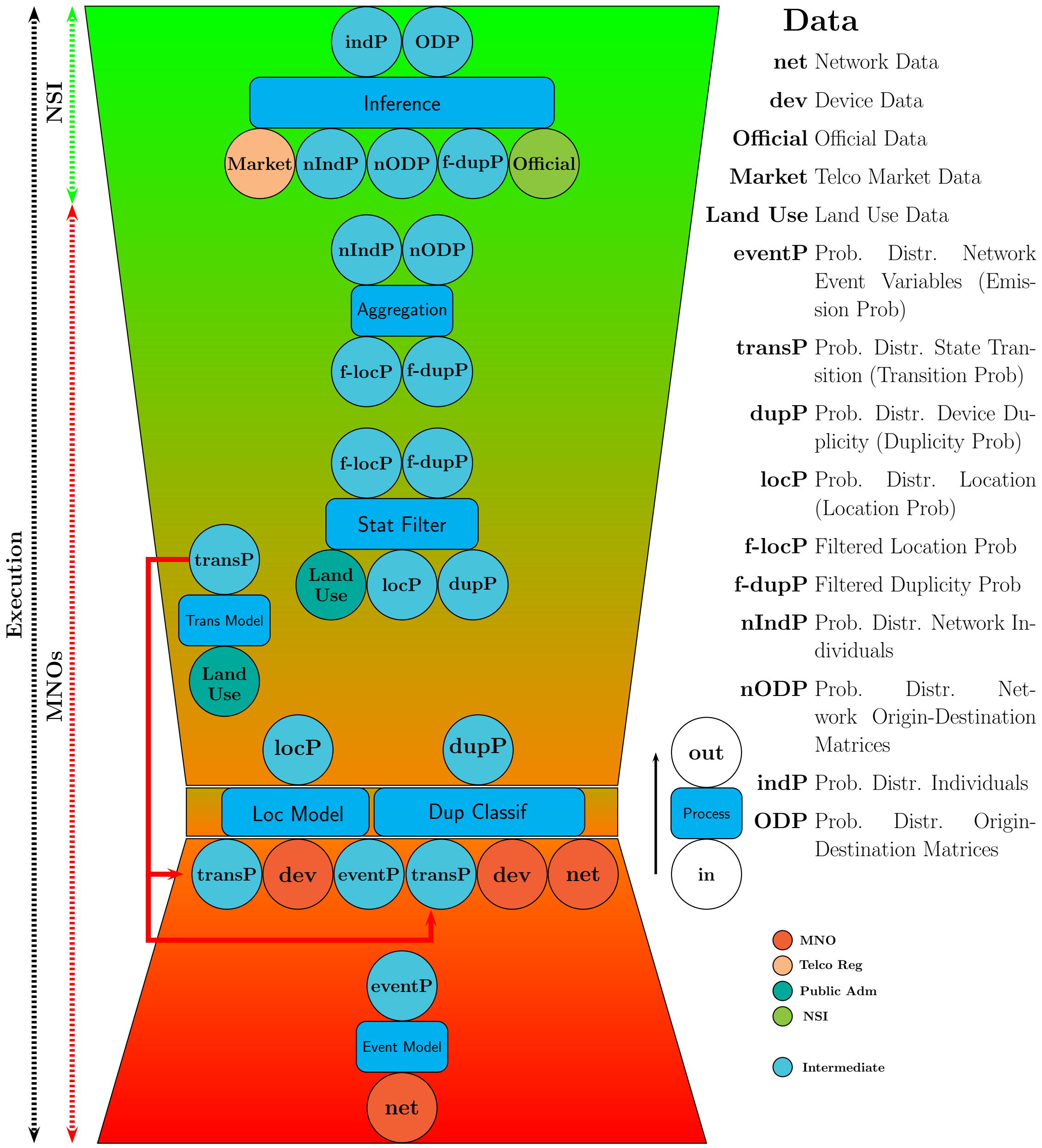
Execution
NSI
MNOs
indP
ODP
Inference
Market
nIndP
nODP
f-dupP
Official
nIndP
nODP
Aggregation
f-locP
f-dupP
f-locP
f-dupP
Stat Filter
transP
Trans Model
Land Use
Land Use
locP
dupP
locP
dupP
Loc Model
Dup Classif
transP
dev
eventP
transP
dev
net
eventP
Event Model
net
out
Process
in
Data
net Network Data
dev Device Data
Official Official Data
Market Telco Market Data
Land Use Land Use Data
eventP Prob. Distr. Network Event Variables (Emission Prob)
transP Prob. Distr. State Transition (Transition Prob)
dupP Prob. Distr. Device Duplicity (Duplicity Prob)
locP Prob. Distr. Location (Location Prob)
f-locP Filtered Location Prob
f-dupP Filtered Duplicity Prob
nIndP Prob. Distr. Network Individuals
nODP Prob. Distr. Network Origin-Destination Matrices
indP Prob. Distr. Individuals
ODP Prob. Distr. Origin-Destination Matrices
MNO
Telco Reg
Public Adm
NSI
Intermediate

Figure 2.4: A proposed statistical process structure.

Finally, all modules are chained together in the statistical process as an application of the total probability theorem. If we denote in an abstract way by z_k the different intermediate datasets in the process and by z_{in} and z_{out} the initial input and final output data, then we can write:

$$\mathbb{P}\left(z_{out}|z_{in}\right) = \int \mathrm{d}z_1 \int \mathrm{d}z_2 \cdots \int \mathrm{d}z_N \mathbb{P}(z_{out}|z_N) \cdots \mathbb{P}(z_2|z_1)\mathbb{P}(z_1|z_{in}). \tag{2.6}$$

Every module amounts to building the intermediate probability distribution $\mathbb{P}(z_k|z_{k-1})$ with a statistical model. Notice how the uncertainty propagates along the process so that our final probability distribution $\mathbb{P}\left(z_{out}\right)$ already takes into account the different sources of uncertainty. Once we have this final probability distribution, we can provide final point estimates and uncertainty measures (e.g. with the variance and credible intervals).

3

Geolocation of mobile devices

This chapter provides the foundations to incorporate the geolocation information of a cellular telecommunication network into the production of official statistics. This geolocation information is generated by the electromagnetic interaction between each mobile device and the antennas of the network spread over the whole geographical territory covered by the network. The technology underlying the device-antenna connections is becoming more and more complex (see e.g. Miao et al., 2016). Following the approach outlined in preceding chapters, in agreement with the RMF, we aim at detaching the technological substrate from the statistical analysis. This will be undertaken by using (i) probability theory and (ii) a reference grid over the network service area (the geographical territory). These two elements will allow us to model the geolocation of network events independently of the underlying technology and any other changing conditions thus paving the way for stable production conditions.

This chapter is organised as follows. In section 3.1 we introduce and motivate the choice of the two aforementioned basic elements of our approach. In section 3.2 we present the foundations for a static analysis of the geolocation of network events as a natural continuation of the work already developed by WP5.3 (2018). In section 3.3 we extend this analysis to incorporate the dynamical behaviour of subscribers in the model. In section 3.4 we present some results on synthetic data. We close this chapter with some future prospects in section 3.5.

3.1. The basic elements

We have briefly outlined the challenges about the statistical methodology to use in the production of official statistics when new digital data sources are incorporated into the statistical process. A systematic approach like survey methodology needs to be found to provide statistical methods and solutions not only for the different statistical domains but also for new sociodemographic and economic insights arising from the new data sources. In this line, we defend the idea that heuristic methods for every different situation must be avoided at all costs. If in the traditional production of official statistics based upon survey methodology there was (still is) an urgent need to provide production standards and good practice, now the urgency for Official Statistics is lethal (DGINS, 2018). It will be impossible to afford the production of official statistics using new data digital sources with heuristic non-standardised statistical methods. The trend provided by survey methodology with survey data must be followed now with the absence of sampling designs.

In this line, we claim that probability theory should be considered the spinal cord of the methods used to process new digital data and to provide rigorous inference from the new huge data sets to the target populations under analysis. Probability theory is versatile enough to provide the practitioner with diverse methods to conduct estimations and to deal with uncertainty and accuracy on a firm basis. All inferential problems can be tackled using probability theory.

Furthermore, as we all know by the meaning of the term Big Data, new data sources (and mobile network, in particular) bring unprecedented volume and velocity of generation. This means we need to deal with a huge number of random variables. This leads naturally to the use of probabilistic graphical models (see e.g. Koller and Friedman, 2009), which combine probability theory and graph theory to deal with this information complexity. We shall follow this path in our model for the geolocation of network events.

The second element is intimately related to the geospatial dimension of the information under analysis. Ideally, spatiotemporal coordinates could be assigned to each network event thus paving the way for geostatistical methods. However, this degree of precision is usually both neither attainable in practice nor necessary in theory. The cost-benefit tradeoff does not seem favourable nowadays (perhaps in the future wireless communication technology will be able to provide coordinates at a very low cost). For this reason, a grid of reference over the service area of the network is enough for statistical purposes. The dimensions of the grid tiles will be discussed and analysed with the data simulator later on to understand their implications on the final estimates. These tiles will be simply labelled by indices $1, 2, 3, \ldots, N_T$. A set of tile indices will be denoted with calligraphic letters $\mathcal{T}$ so that a territorial region r can be denoted as a set of tiles $r = \bigcup_{i \in \mathcal{T}_r} T_i$.

The two basic elements in our model are mathematically represented by defining a random variable $T_d(t)$ for each mobile device d at time t whose support is the set of tile labels. Notice that this is a discrete random variable for each device d and time instant t. The target quantities will be the probability mass functions associated to each random variable $T_d(t)$ and the joint probability functions for two consecutive time instants t and t'.

To construct the target probability functions we shall make use of all available information, namely, of data coming from the telecommunication network and of data comprising the auxiliary information. We adhere to the notation introduced in chapter 2. We shall denote by $\mathbf{E}_d(t)$ the set of variables associated to an event of mobile device d at time $t \in [t_0, t_f]$. Upon the interaction or control of the interaction between each mobile device d and an antenna, some variable values will be generated in the information systems of the network. Notice the generality of this statement. These variables can be the antenna cell ID where the device is detected, the timing advance of the signal (TA), the angle of arrival of the signal (AoA), or any other telco variable related to the location of the device (see WP5.2 (2017) for a generic description of the variables in a mobile network dataset for statistical purposes). The values of each component of $\mathbf{E}_d(t)$ will depend on the concrete variables taken from the information systems of the network.

As before, we shall adopt the convention that when the time variable is dropped, we mean that all time instants are considered, i.e. $\mathbf{E}_d \equiv (\mathbf{E}_d(t_0), \mathbf{E}_d(t_1), \ldots, \mathbf{E}_d(t_f))$.

Let us also introduce the auxiliary information in our model. By auxiliary information we mean any source of data other than the telecommunication network used to assist the estimation process. It could be official statistical data (e.g. residential population density), national telecommunication regulator data (e.g. penetration rates), or geographical data (e.g. land use). We shall denote this auxiliary information at time t by $\mathbf{I}(t)$, which comprises the set of variables encompassing any sort of auxiliary data source. We shall adopt the same convention when dropping the time variable.

Thus, in mathematical terms, we shall focus on:

$$\gamma_{di}(t) \equiv \mathbb{P}\left(T_{di}(t)|\mathbf{E}_d, \mathbf{I}\right) = \mathbb{P}\left(T_d(t) = i|\mathbf{E}_d, \mathbf{I}\right) \tag{3.1a}$$

$$\gamma_{dij}(t, t') \equiv \mathbb{P}\left(T_{di}(t), T_{dj}(t')|\mathbf{E}_d, \mathbf{I}\right) = \mathbb{P}\left(T_d(t) = i, T_d(t') = j|\mathbf{E}_d, \mathbf{I}\right). \tag{3.1b}$$

We convene in calling (3.1a) **location probability** for mobile device d to be in tile i at time t and calling (3.1b) **joint location probability** for mobile device d to be in tiles i and j in consecutive time instants t and t'. Notice that these probabilities are conditioned upon all the available information from the past and the future at each time instant t.

It is important to point out that by focusing on probabilities, we can integrate the uncertainty in the locations and movements from the onset. This will be relevant for quality issues later on, since it paves the way to assess accuracy. Additionally, probabilities are versatile enough to adapt to many diverse situations. For example, in the particular case in which we can pinpoint where every mobile device d is located, this will be reduced to trivial cases with zero-one probabilities (degenerate random variables).

3.2. Static analysis

This first approach was already introduced in the past edition of this project (ESS, 2018) (see WP5.3 (2018)). This approach has been recently extended to include both cell ID and timing advance variables in the computation of the location probabilities by Tennekes et al. (2019).

The key idea of this static analysis is to make use of Bayes' rule:

$$\gamma_{di}(t) = \mathbb{P}\left(T_{di}(t)|\mathbf{E}_d(t), \mathbf{I}(t)\right) \propto \mathbb{P}\left(\mathbf{E}_d(t)|T_{di}(t), \mathbf{I}(t)\right) \times \mathbb{P}\left(T_{di}(t)|\mathbf{I}(t)\right), \tag{3.2}$$

and to compute separately the location likelihood (or event location probability) and the prior location probability. Both computations are undertaken depending very sensitively on the available information.

As a first example, if no data is available about the coverage area of each antenna cell or about the technical parameters allowing us to use radio propagation models (see below), we can resort to purely geographical considerations (Avouac et al., 2019). We can model the network coverage using a Voronoi tessellation with the antenna sites as reference points. Notice that this modelling is not incorporating important features of the telecommunication network as the directionality of the antennas and the overlapping character of the coverage cells. In our mathematical notation, the event variables $\mathbf{E}_d(t)$ will thus stand for the Voronoi cell (i.e. the coverage cell) where the device d has been detected at time t. Indeed, the rigorous reasoning is: a device d is detected at time t by the antenna j, which is associated to Voronoi cell v_j. Thus, in our notation, we have $E_d(t) = v_j$.

A second example comes from the use of radio wave propagation models. Following Tennekes et al. (2019), which extends our previous work (WP5.3, 2018), we can set $\mathbf{E}_d(t) = (c_d(t), \tau_d(t))$ to comprise the antenna cell ID $c_d(t)$ and the time advance $\tau_d(t)$ detected by the network. The antenna cell ID makes reference to the identification of the sector cell of a Base Transceiver Station (BTS) associated to an antenna site (see figure 3.1). The time advance variable may be interpreted as a measure of distance between the mobile device and the antenna site. More precisely, for illustrative purposes, if an event contains a value $\tau \in \{0, 1, \ldots, 1282\}$ (like in 4G networks) for the time advance variable, the associated mobile device will be located approximately in an annulus centered around the antenna site of a given width Δ ($78m$ for 4G networks) and an inner circle of radius $\tau \cdot \Delta$ (see Tennekes et al., 2019). This establishes the maximal distance between the device and the antenna site in $99996m \approx 100km$, which can be used for modelling values later on.

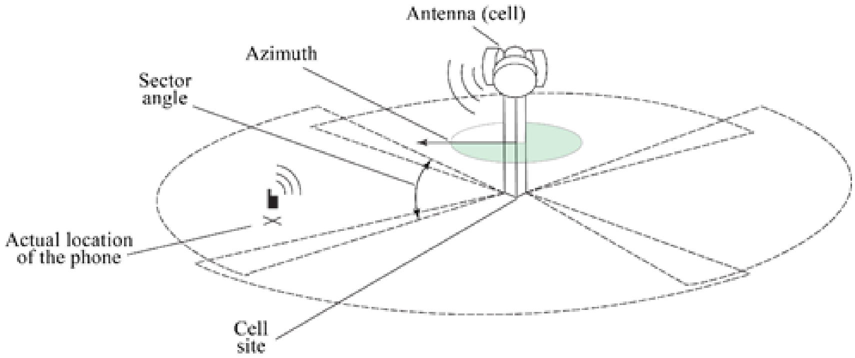

Figure 3.1: Representation of the sector cells of a BTS.

The concrete contents of the event variables $\mathbf{E}_d$ will thus depend on the access agreement and the actual available information provided by the MNO. Different computations will be possibly undertaken depending on this information.

3.2.1. Computation of event location probabilities

The computation of the event location probabilities $\mathbb{P}\left(\mathbf{E}_d(t)|T_{di}(t), \mathbf{I}(t)\right)$ is essential not only for the static analysis but also for the dynamical analysis presented later on in section 3.3. Since the time variable play no role in the static analysis, for ease of notation we shall drop the variable t.

For our first example using only geographical considerations with a Voronoi tessellation, Avouac et al. (2019) thus propose the following direct computation:

$$\mathbf{P}(E_d = v_j|T_{di}, \mathbf{I}) = \frac{\text{area}(v_j \cap T_i)}{\text{area}(T_i)}. \tag{3.3}$$

Notice that basically no information about the network is used (except the antenna sites), thus we are missing features as the directionality of the antennas and the overlapping character of the coverage cells.

The second example allows us to provide a more complex computation. The event location probability can be modelled as

$$\mathbb{P}\left(c_d, \tau_d|T_{di}, \mathbf{I}\right) = \mathbb{P}\left(\tau_d|c_d, T_{di}, \mathbf{I}\right) \times \mathbb{P}\left(c_d|T_{di}, \mathbf{I}\right), \tag{3.4}$$

so that we need to model the time advance and the antenna cell ID in turn. For the time advance modelling, notice that we just need to make geometrical considerations involving the TA annulus, the sector cell $c_d(t)$ and the tile dimensions (Tennekes et al., 2019). For the cell ID modelling, an underlying signal propagation model can be used together with some parameters for the BTS. This approach, basically using a simplified radio propagation model, was developed in the past edition of this ESSnet (ESS, 2018) (see also Tennekes et al., 2019). We provide an executive summary for completeness.

Firstly, the received signal strength (RSS) of a device d connected to an omnidirectional antenna cell $c_d(t)$ at time t separated by a distance r is modelled as

$$\text{RSS}(t) = \text{RSS}_0 - 10 \cdot \gamma \cdot \log_{10}(r), \tag{3.5a}$$

where RSS_0 (in dBm) is the reference received signal at a distance $r = 1m$ and γ stands for the *path loss exponent*, which expresses the reduction of propagation due to reflection, diffraction, and scattering of the signal. Usual values of γ are $\gamma = 2$ for free space and $\gamma = 4$ for urban environments. The reference received signal is expressed in terms of the antenna power P (in W) as

$$\text{RSS}_0 = 30 + 10 \cdot \log_{10}(P/1W). \tag{3.5b}$$

If the antenna is directional (e.g. divided into 3 sectors of 120°), we need to modify equation (3.5a) to account for the directionality:

$$\text{RSS}(r) = \text{RSS}_0 - 10 \cdot \gamma \cdot \log_{10}(r) - S_{\text{axi}}(\delta_{T-a}, \alpha_a, \psi_a) - S_{\text{elev}}(\epsilon_{T-a}, \beta_a, \theta_a), \tag{3.5c}$$

where

- ψ_a stands for the azimuth angle, that is the angle between the north direction and the direction of the cell sector (seen from above). It can take any value between $0°$ and $360°$;
- θ_a stands for the angle of the elevation plane (angle between the horizon plane and the tilt of the antenna). Usually this angle is around $4°$ as much;
- α_a stands for the horizontal beam width, that is the angular amplitude in the elevation plane where the signal loss is $3dB$ (halved power). The exact angles in the elevation plane where the signal loss is $3dB$ are $\psi_a \pm \alpha_a/2$. Usual values are around $65°$;
- β_a stands for the vertical beam width, that is the angular amplitude in the vertical plane orthogonal to the elevation plane where the signal loss is $3dB$. The exact angles in this orthogonal plane where the signal loss is $3dB$ are $\theta_a \pm \beta_a/2$. Usual values are around $9°$.

With this model, we can compute the RSS from any antenna for a device d in any tile T. The distance to be used in equation (3.5a) is the distance between the antenna site and the centroid of the tile. Details about the angular components of the RSS need to be worked out in collaboration with the MNOs, since each antenna has each own pattern.

In our model, thus, a first step is the computation of the received signal strength $\text{RSS}(T, c_d)$ for each device d when located at each tile T and connected to each cell c_d. Now, to compute $\mathbb{P}(c_d | T_{di}, \mathbf{I})$ for each cell c_d and each i we need to consider the physical connection between the device and the antenna. This is complex and depends on a diversity of factors (load balancing, handover, ...). For this reason, a logistic transformation is applied on the RSS computed above returning a simulated signal quality in terms of which the connection is established. This transformation is given by:

$$s(c_d, T_{di}) \equiv \frac{1}{1 + \exp\left(-S_{\text{steep}}\left(\text{RSS}(c_d, T_{di}) - S_{\text{mid}}\right)\right)} \tag{3.6}$$

In this way, the connections are modelled not according to absolute values of the RSS but of relative comparisons in terms of the RSS and other factors gathered in the heuristic parameters S_{steep} and S_{mid}. Then, we set

$$\mathbb{P}\left(c_d \middle| T_{di}, \mathbf{I}\right) \propto s(c_d, T_{di}), \tag{3.7}$$

that is, the device has larger probability of connection to the antenna (cell c_d) with highest value of s. Indeed, this resembles the physical connection. Notice that the normalization factor in (3.7) is provided by the sum $\sum_c s(c, T_{di})$ (hence the need to have all values).

It is very important to notice that this computation must be carried out in conjunction with MNO experts and thus it must be considered in the D-layer of the RMF to be naturally executed in MNOs' premises. This or similar models substantiate the notion of mobile network data. Notice that for the application of this proposal we need antenna parameters (orientation angles, power, path loss exponents) and load balancing simulation (S_{steep} and S_{mid}). Access to this data does not necessarily entail that this information needs to leave MNOs' information systems. Access agreements should reach collaboration proposals for statistical offices to get these event location probabilities whereas sensitive information for MNOs is simultaneously protected. Indeed, the event location probabilities themselves do not need either to leave MNOs' information systems, since this is only a first step in the whole computation. Should we access richer raw telco variables, more complex radio propagation models could be used, potentially impinging on the computation of these event location probabilities.

Under this approach, all these considerations strongly suggests to define a submodule within the geolocation module focused on the computation of event location probabilities. The input data and input parameters for this submodule constitute the raw telco data needed for the computation. The output data will be the event location probability matrix (say, one row per tile and one column per cell). The throughput will amount to applying the above model or a possibly more sophisticated model incorporating more information from the network. Quality issues will not be dealt with in this document.

3.2.2. Prior location probabilities

The next step is the computation of the prior location probabilities $\mathbb{P}\left(T_{di} \middle| \mathbf{I}\right)$ (we also drop out the time variable). The role of the auxiliary information is now clearly seen. Tennekes et al. (2019) already point out several options:

- Uniform prior.- No auxiliary information is used and equal-probability distribution is assigned to each tile:

$$\mathbb{P}_{\text{Unif}}\left(T_{di} \middle| \mathbf{I}\right) = \frac{1}{N_T}, \tag{3.8a}$$

  where N_T is the number of tiles.

- Land use prior.- Administrative sources about land use can be used so that one expect more number of devices in an urban area that in a lake or a forest. We can specify this prior as

$$\mathbb{P}_{\text{LandUse}}\left(T_{di} \middle| \mathbf{I}\right) \propto \bar{N}_i^{(D)}, \tag{3.8b}$$

  where $\bar{N}_i^{(D)}$ is the expected number of devices in tile i coming from the administrative sources. A variant of this proposal can be considered by using some proportional measure. For example, let us consider K categories for land use (e.g. urban, main road, other land, water) and consider their relative amounts of devices $u_k \in [0, 1]$ (e.g. $u_{\text{urban}} = 0.9$, $u_{\text{road}} = 0.5$, $u_{\text{other}} = 0.1$, $u_{\text{water}} = 0.01$). Let us denote by w_{ik} the proportion of tile i dedicated to land use k. Then, we can model $\bar{N}_i^{(D)} \propto \sum_{k=1}^{K} u_k \cdot w_{ik}$. The drawback using administrative sources is their radically different time scale and adaptation to transient phenomena (e.g. festivals, concerts, sport events, ...).

- Network prior.- Network measures can also be used by setting

$$\mathbb{P}_{\text{Network}}\left(T_{di}(t)\middle|\mathbf{I}(t)\right) \propto \sum_{c} s(c, T_i), \tag{3.8c}$$

  where s is computed according to (3.6).

- Composite priors.- Whenever we have several options for choosing priors, we can also combine them in a convex combination:

$$\begin{aligned}\mathbb{P}_{\text{Comp}}\left(T_{di}\middle|\mathbf{I}\right) &= \pi_{\text{Unif}} \cdot \mathbb{P}_{\text{Unif}}\left(T_{di}\middle|\mathbf{I}\right) + \\ &\quad \pi_{\text{LandUse}} \cdot \mathbb{P}_{\text{LandUse}}\left(T_{di}\middle|\mathbf{I}\right) + \\ &\quad \pi_{\text{Network}} \cdot \mathbb{P}_{\text{Network}}\left(T_{di}\middle|\mathbf{I}\right),\end{aligned} \tag{3.8d}$$

  where $\pi_{(\cdot)} \in [0, 1]$ represents the contribution of each type of prior to the final composite prior. These coefficients must be chosen according to some preliminary auxiliary information.

Again, we see a natural submodule for the computation of prior locations. Depending on the choice of prior, this production step can be executed by the statistical office independently of mobile network data (not the case for the network prior). It seems also natural to us to consider this production step a submodule of the geolocation globally embedded in the D-layer of the RMF. In this case, the information systems of the MNOs' should also receive the data coming from administrative sources or, at least, the minimal amount for these computations to be carried out.

As input data and input parameters we need the information at tile level to be used in the choice of prior. The output data will be the prior location probabilities for each device and each tile. Quality metrics will not be dealt with in this document.

3.2.3. An illustrative example

We can illustrate the application of this model using the data simulator in a toy example. We begin by configuring a set of 18 omnidirectional antennas over a service area (a given irregular polygon). The dimensions of this irregular service area are $10km \times 10km$ and the tiles are $500m \times 500m$. We specify for each antenna the coordinates, the power, the path loss factor, and the parameters S$_{\text{steep}}$ and S$_{\text{mid}}$ (see figure 3.2).

We have set up an urban area in the bottom right corner of the polygon (you see a higher number of antennas) whereas a rural area has been configured in the top left corner. This is observed also in the choice of the power and path loss exponent parameters (not shown in the visualization). The result of the computation of the event location probabilities for some antennas is depicted in figure 3.3. Notice how the farther away from an antenna site, the lower

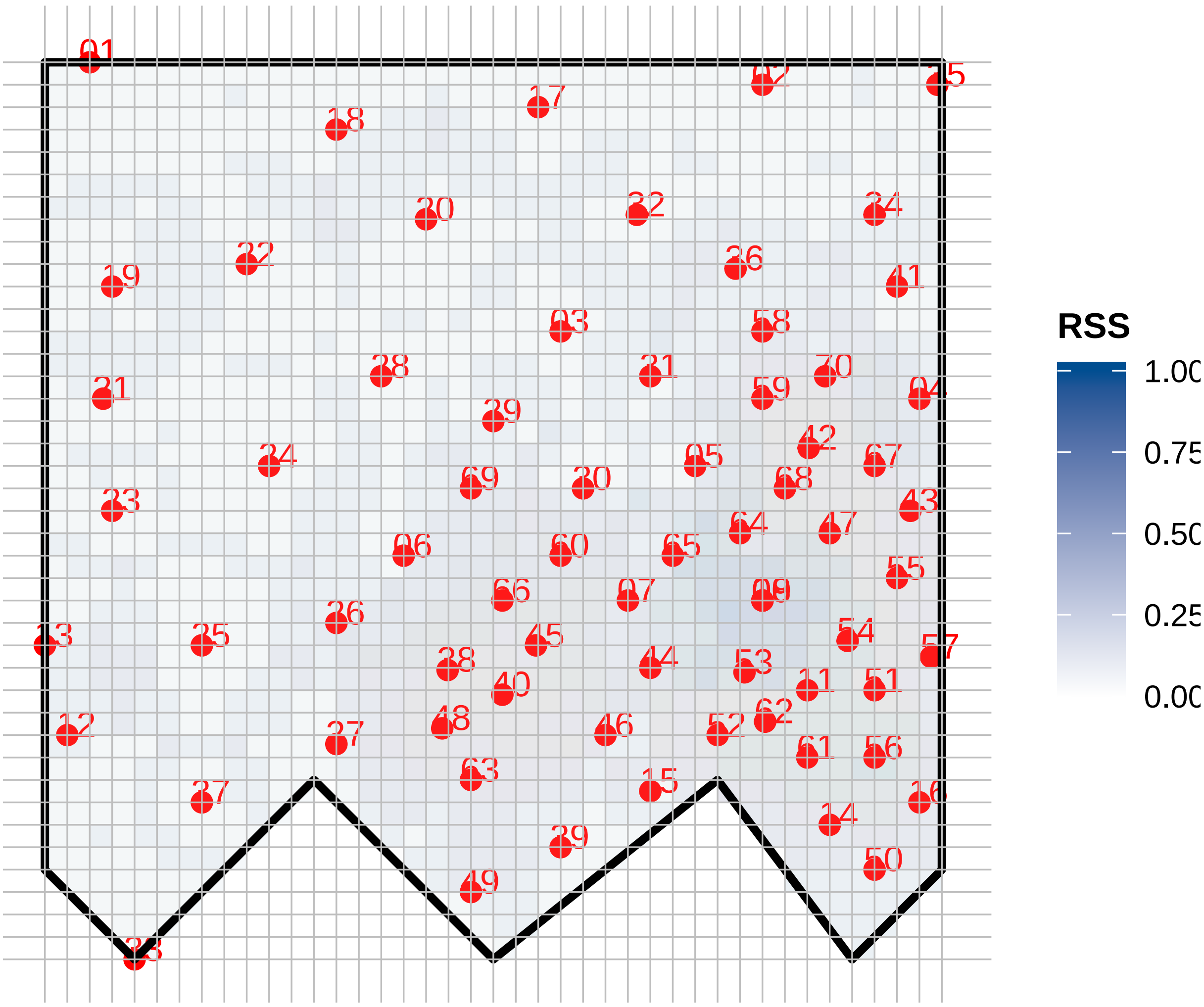

Figure 3.2: Network configuration.

the probability to be connected to that antenna, as expected.

To compute the posterior location probabilities (3.1a) we need firstly to choose the prior (network in our example) and just implement the corresponding formulas above. We show in the animation below how a mobile device is traced in terms of posterior location probabilities. As the device moves around the polygon, each antenna with the running connection is shown together with the posterior location probabilities for each tile.

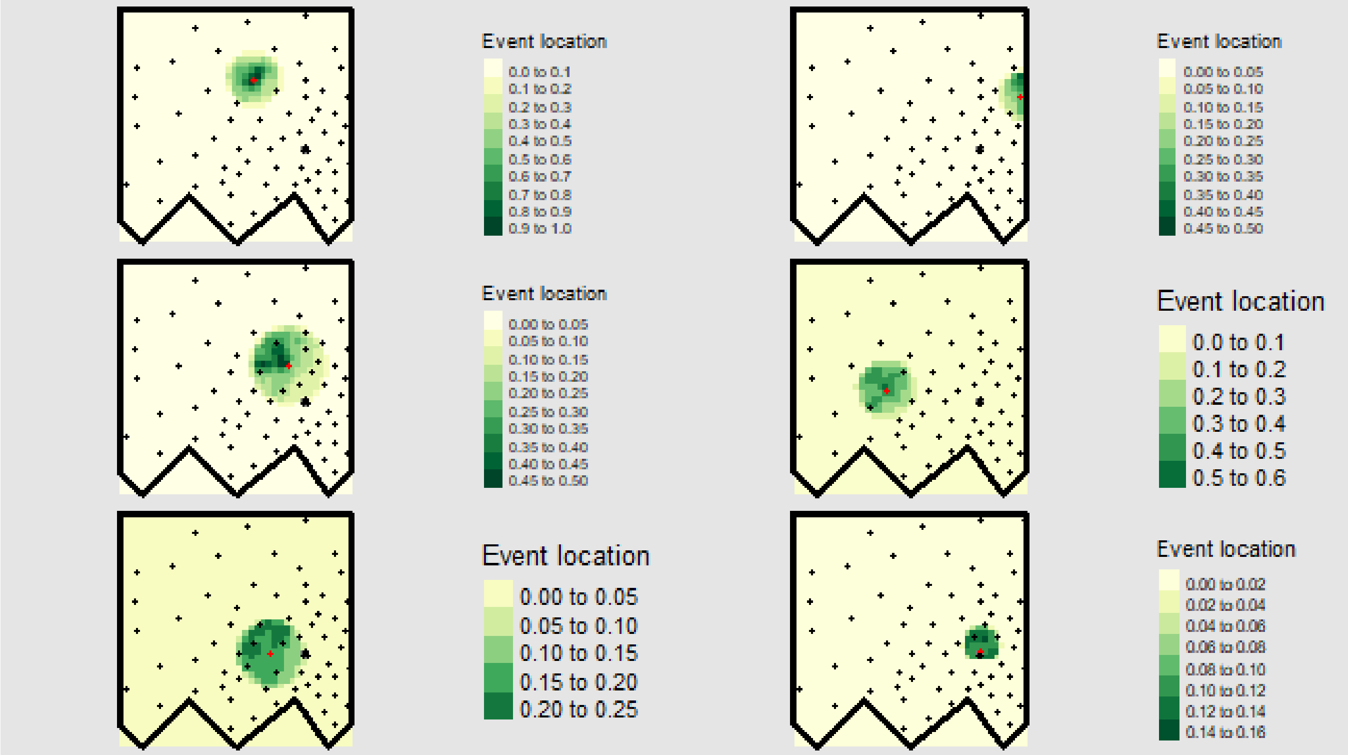

Figure 3.3: Examples of event location probabilities.

3.3. Dynamical analysis

In the preceding approach no use is made of the dynamical characteristics of the underlying phenomenon, i.e. of the movements of devices around the geographical territory. Now, we incorporate this consideration still under the general working assumptions of using probabilities and a reference grid. Thus, the dynamical behaviour will enter naturally into the model by considering transition probabilities between the tiles. A closer reflection, indeed, strongly suggests the kind of statistical tool to model this situation. On the one hand, according to the static analysis above, we have the event location probabilities $\mathbb{P}\left(\mathbf{E}_d(t)\middle|T_{di}(t),\mathbf{I}\right)$ relating observed data (network event data $\mathbf{E}_d$) with unobserved variables (the location tile $T_d(t)$). On the other hand, now we introduce a transition probability $\mathbb{P}\left(T_{di_1}(t_1)\middle|T_{di_2}(t_2)\right)$ to move from tile i_1 at time t_1 to tile i_2 at tile t_2. These two mathematical elements can be easily viewed as the so-called emission and transition probabilities of a hidden Markov model (Rabiner, 1989; Bishop, 2006; Murphy, 2012).

In appendix A we provide technical details about the construction of an HMM to estimate the geolocation of devices in this proposed framework. Here we focus on the generic approach underlining those elements making this proposal a generic framework potentially adapted to diverse situations depending on data availability and reinforcing the idea of modularity defined in the RMF (see chapter 2).

Figure 3.4: Device movement traced by connected antenna and posterior location probabilities.

3.3.1. Discrete vs. continuous HMMs

Our setting focuses on a reference grid to provide information about the geolocation of mobile devices. This amounts to working with probability distributions over the tiles of this reference grid, but this is a choice. It could be possibly to focus instead on providing probability distributions on the position coordinates of the mobile devices, thus driving us to continuous HMMs.

Our choice can be justified both for computational simplicity and due to data availability. A priori, in most current practical cases available data can only provide geolocation information in a reliable way in a discrete regime. For example, raw telco data only about the connecting antenna is available, which makes very hard to consider position coordinates. Notice that there exist methods to pinpoint devices in a territory covered by a mobile telecommunication network (see e.g. Gezici, 2007), but this data access and data preprocessing scenario for statistical purposes needs much further joint work between MNOs and NSIs.

As a potential drawback of the discrete choice is that it may drive us to the so-called change of support problem in geostatistics (Gelfand, 2010), because sooner or later in some statistical domains or to fulfill a concrete unplanned statistical need, new territorial cells will need to be investigated. In the continuous version this problem will not arise.

To reduce the impact of this potential drawback, we recommend to use a reference grid with tile size as small as possible searching for a trade-off between computational cost and efficiency. Having probabilities γ_{dti} and $\gamma_{dij,t}$ over a grid with small tiles T_i we should be able to provide estimates for territorial regions in many statistical domains.

3.3.2. The concept of state

In the current proposal we focus on two sorts of random variables. On the one hand, we aim at providing the probability distribution of the geolocation $T_d(t)$ of each mobile device d. On the other hand, we observe the raw telco variables $E_d(t)$ about events of each device d detected in the network.

The first set of variables comprises the so-called *state* of the Markov model. This is a rich and versatile concept. Our first natural step has been to specify the state of each device with the geolocation in terms of grid tiles T_{di}, but again this is a modelling choice.

We can also model the state of a device both in terms of discrete position (tile) and discrete velocity (e.g. direction and speed in terms of number of tiles). Even we can make it more complex using also transport mode (depending on velocity regime, e.g.). Many possibilities arise depending on the statistical needs.

In the illustrative examples considered so far herein, we will only use the original proposal of specifying the state in terms of the tile location.

3.3.3. Emission probabilities and event location probabilities

The connection between the observed raw telco data and the rest of variables for the estimation process takes place in the computation of the emission probabilities, which in our preceding parlance were identified as the event location probabilities.

In section 3.2.1 we show two examples of computation of these probabilities in terms first of simple geometrical considerations and second of a simple radio propagation model. At this point we want to underline the following points:

- The concrete computation exercise is not important for the HMM framework. This will depend on data availability, which ultimately will be derived from the MNO-NSI agreement. Should MNOs and NSIs agree to work on a rich data environment and sophisticated radio propagation models, emission probabilities will be more accurate and will pay off in the final estimates. Possibilities will directly depend on the partnership agreements between MNOs and NSIs.
- The computation of the emission probabilities enter into the HMM as an input and is not expressed in terms of model parameters to be estimated later on. This departs from usual practice in HMMs (see e.g. Bishop, 2006), but it is intended so in order to pursue modularity and gain efficiency in the overall process. Under this prescription, once emission probabilities are computed, no further access or use of raw telco data is necessary, thus minimising all risks and costs associated to this sensitive data. Ultimately, even we can reduce the data request and data preprocessing to MNOs to the computation of these emission probabilities. However, our recommendation is to pursue joint work as much as possible even beyond the computation of these emission probabilities.

3.3.4. Transition probabilities and auxiliary information

Transition probabilities constitute the mathematical element introducing the dynamical behaviour in the modelling exercise. They express in probabilistic terms the movement patterns between tiles in the reference grid. Notice that these probabilities must be expressed in terms of the state chosen for the devices. In our original choice, since the state is specified by the location tile T_{di}, these transition probabilities are computed as usual as conditional probabilities $\mathbf{P}\left(T_{dit}|T_{djt-1},\mathbf{I}\right)$, where $\mathbf{I}$ stands for the available auxiliary information (not usually explicitly included in these expressions).

In appendix A we provide technical details about the construction of this transition matrix and show a concrete illustrative example with rectangular isotropy across the grid (see figure A.1). Here we want to underline the many possibilities to model the dynamical behaviour:

- The transition probabilities in the rectangular isotropic model are reduced to two possible values depending on the direction of the movement (either up-down or in diagonal). We assume space homogeneity across the grid. Notice that this is an extreme choice. The other extreme choice would be to set a different values for each entry of the transition matrix (with evident problems for the estimation if we do not have massive amounts of data). The wisest procedure seems to be the use of land use information, making a classification of tiles depending on this information, which can be homogeneous across the grid (e.g. different sets of probabilities for tiles with roads, with pedestrian zones, with residential buildings, etc.).
- The procedure devised in appendix A is generic enough to consider this wealth of choices for the transition matrix, so that the effort can be concentrated on the modelling choices.

3.3.5. Time homogeneity

As stated in the appendix, we have assumed time homogeneity throughout the whole modelling exercise. Again, this is not an essential assumption and it has been adopted for clarity's sake. It is an extreme choice whereas in the opposite extreme you find a different transition matrix at each time instant t. This would drive us into estimation problems in case of not having massive amounts of data.

The relaxation of this assumption should be made in agreement with other modelling assumptions, especially regarding the incorporation of auxiliary information through the transition probabilities and possibly if the radio propagation model also depends on time for the generation of network events (e.g. due to climatic conditions).

3.3.6. Parsimony

As explained in the appendix, we have deliberately built a parsimonious model by choosing the time increment Δt so that at most one tile can be reached in this time interval. In this way, the restrictions (A.1a) are automatically satisfied and the transition matrix is sparse. To do this, we even pad the sequence of connecting antennas with missing values and let the forward-backward algorithm to impute these values with predictions.

Again, this is a modelling choice and many more possibilities can be considered. For example, we can even assume a distribution decreasing with distance depending on a single parameter (to keep parsimony) and then repeat the whole process with another likelihood and another optimization problem.

In any case, we recommend to keep parsimony as an important feature in the modelling exercise in order to arrive at manageable likelihood functions and practically solvable optimization problems.

3.3.7. Scalability

All our proposal is thought to set up generic statistical principles and so far does not embrace implementation issues. In this sense, the geolocation estimation based on HMM is thought to be applied to each device so that the task is embarrassingly parallelizable. This is a computational challenge which, in our view, should be tackled in a separate stage of the process. Nonetheless, the computational treatment should not exhaust the methodological possibilities to improve the efficiency in the geolocation estimation. Once the end-to-end process from raw telco data to the final estimation of the number of individuals in each territorial region is devised, even sampling procedures over the whole dataset may be possibly explored to reduce computational load.

3.3.8. Static analysis as a particular case of an HMM

Intuitively, the static models proposed in section 3.2 should be recovered as a concrete case of an HMM. This is, indeed, the case. Let us consider the posterior location probabilities γ_{dit_k} for an arbitrary device d (see eq. (3.1a)). These can be written as

$$\gamma_{dit_k} = \frac{\alpha_{dit_k}\beta_{dit_k}}{\mathbb{P}\left(\mathbf{E}\right)}, \tag{3.9a}$$

where we have followed the usual notation (see e.g. Bishop, 2006) to denote

$$\alpha_{dit_k} = \mathbb{P}\left(E_{dt_1}, \dots, E_{dt_k}, T_{dit_k}\right) \tag{3.9b}$$

$$\beta_{dit_k} = \mathbb{P}\left(E_{dt_{k+1}}, \dots, E_{dT}|T_{t_k}\right) \tag{3.9c}$$

$$\mathbb{P}\left(\mathbf{E}\right) = \sum_{i_1=1}^{N_T}\cdots\sum_{i_T=1}^{N_T}\mathbb{P}\left(T_1 = i_1\right)\left(\prod_{k=2}^{N_T}\mathbb{P}\left(T_{t_k} = i_k|T_{t_k-1} = i_{k-1}\right)\right)\prod_{k=1}^{T}\mathbb{P}\left(E_k|T_{t_k} = i_k\right). \tag{3.9d}$$

If the transition matrix is chosen to be the identity matrix, i.e. $\mathbb{P}\left(T_{t_k} = i_k|T_{t_{k-1}} = i_{k-1}\right) = \delta_{i_k i_{k-1}}$, where δ_{ij} stands for the Kronecker delta function, then it is straightforward to show that

$$\gamma_{dit_k} = \frac{\mathbb{P}\left(T_{t_k} = i\right)\prod_{k=1}^{T}\mathbb{P}\left(E_{t_k}|T_{t_k} = i\right)}{\sum_{j=1}^{N_T}\mathbb{P}\left(T_{t_k} = j\right)\prod_{k=1}^{T}\mathbb{P}\left(E_{t_k}|T_{t_k} = j\right)}, \tag{3.9e}$$

which is exactly the core expression for the static analysis (see eq. (3.2)).

3.4. A view on results

To provide a view on the kind of results obtained from the application of the HMM framework we have produced a synthetic dataset with 70 omnidirectional antennas over a territory with bounding perimeter of 10km × 10km as represented in figure 3.3. There are 500 people with 218 devices (186 subscribers) moving according to a random walk with drift (see (WPI.2, 2019)). For the transition probabilities we use the rectangular isotropic model devised in appendix A. The software tools implementing this model will be explained elsewhere. We shall concentrate

Figure 3.5: Device movement traced by connecting antenna and an HMM. The red point denotes the connecting antenna; the blue point denotes the device.

on the outputs and comment on these results to get acquainted with the features.

The main result is the set of posterior location probabilities γ_{dti}, $i = 1, \ldots, N_T$, $t = 0, 1, 2, \ldots, T$ for each mobile device d. This evolving set of posterior location probabilities can be represented as in animations 3.4 (for the static analysis) and 3.5 (for the dynamical analysis).

.

To assess the output let us define two summarization quantities related to the posterior location probabilities:

1. Let us define the *center of location probability* $\mathbf{cp}_{dt}$ of device d at time t as

$$\mathbf{cp}_{dt} = \sum_{i=1}^{N_T} \gamma_{dti} \left(x_i^{(c)} \; y_i^{(c)} \right)^T, \tag{3.10}$$

   where $x_i^{(c)}, y_i^{(c)}$ stand for the x and y coordinates of the centroid of tile i. This can be understood as an estimation of the position of the device according to the posterior mean. Notice that this quantity plays a similar role to a first-order spatial moment for the distribution γ_{dti}.

2. Let us define the *radius of location probability dispersion* rd_{dt} of device d at time t with respect to position $\mathbf{r}^*_{dt} = (x^*_{dt}\ y^*_{dt})^T$ as

$$rd_{dt}(\mathbf{r}^*_{dt}) = \sqrt{\sum_{i=1}^{N_T} \gamma_{dti} \left[(x_i^{(c)} - x^*_{p,dt})^2 + (y_i^{(c)} - y^*_{dt})^2 \right]}. \tag{3.11}$$

where (x^*_{dt}, y^*_{dt}) stands for the true x and y coordinates of the device d at time t. This can be understood as a root mean square distance with respect to a reference position. Notice that this quantity plays a similar role to a standard spatial deviation for the distribution γ_{dti}. We can also generalize this definition by using an alternative distance function:

$$rd_{dt}(\mathbf{r}^*_{dt}) = \mathrm{d}\left(\mathbf{r}^{(c)}_{dt}, \mathbf{r}^*_{dt}\right). \tag{3.12}$$

For the time being, we shall concentrate on the Euclidean distance, but the Manhattan distance is also meaningful.

Now, we shall compute these quantities for each device in the simulation and produce the following figures of merit:

1. **Distance to the true position of the device**.- We want to focus on the Euclidean distance between the center of location probabilities and the true position of the device at each time instant t. We take advantage of our knowledge of this true position in the simulation exercise. To this end, we compute $\mathbf{c}_{p,dt}$ for each device d and each time t and the Euclidean distance $\|\mathbf{c}_{p,dt} - \mathbf{r}^*_{dt}\|$, where $\mathbf{r}^*_{dt}$ denotes the true position vector of device d at time t. Notice that this figure of merit is intended to measure the bias in terms of distance in the geolocation estimation.

2. **Root mean square dispersion**.- We want to focus on the dimensions of the spatial probability distribution γ_{it} in a similar way as standard deviation measures the dispersion of a random variable. We define

$$\mathrm{rmsd}_t = rd_{dt}(\mathbf{c}_{p,dt}). \tag{3.13}$$

Notice that this figure of merit is intended to measure the dispersion with respect to the center of location probabilities, thus playing a similar role to the standard deviation in random variables.

Obviously, these figures of merit are not exhaustive and we can propose more (e.g. to measure the kurtosis, concentration, etc.). Having the set of probability distributions γ_{it} and the true values many choices arise.

We compute these figures in our simulation exercise for three geolocation models, namely the static approach with uniform and network priors and the HMM approach with a uniform initial prior. In figures 3.6 you can see two illustrative examples: the first one corresponds to a device with non-identity estimated transition matrix (the device moves around) whereas the second corresponds to a device with an identity estimated transition matrix (the device is mainly motionless. You can see how when motion is present the HMM captures in a better way the location of the device. When motion is not present, the three models are more or less equivalent (remind that the initial prior for the HMM is the uniform prior).

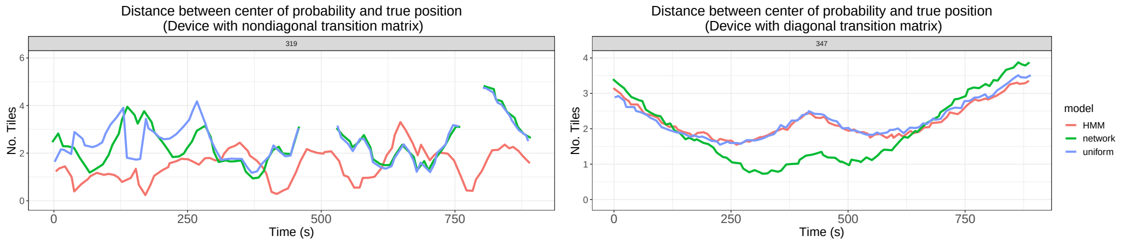

Figure 3.6: Distance between the center of location probabilities and true position of the device under three different models.

We can investigate these differences for all the devices. In figure 3.7 we represent the distribution of distances between the center of location probabilities and true positions for those devices with a non-identity transition matrix (motion). In figure 3.8 we represent the same figure of merit for devices with identity transition matrix (no motion).

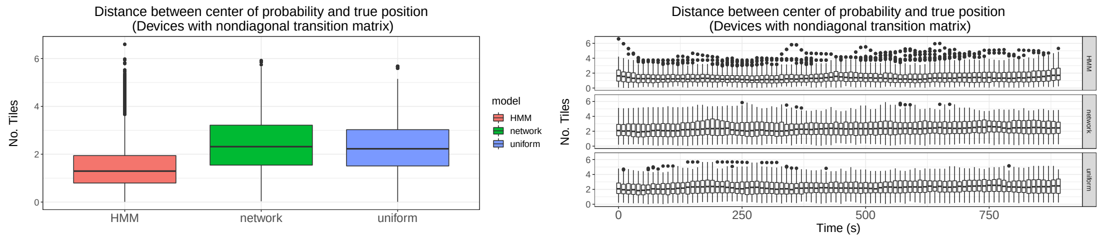

Figure 3.7: Distance between the center of location probabilities and true position under three different models. Devices with motion.

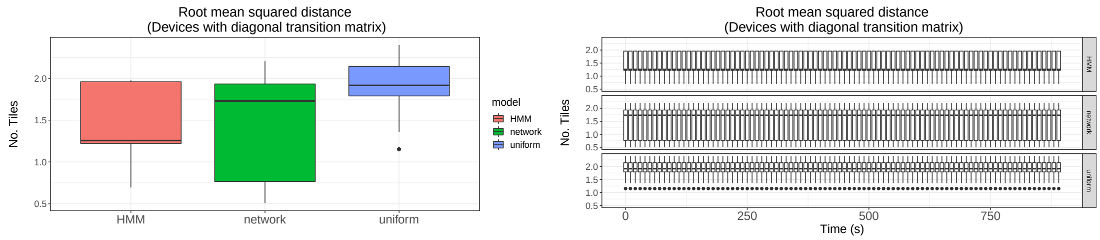

Figure 3.8: Distance between the center of location probabilities and true position under three different models. Devices with no motion.

For the root mean square dispersion, similar results are represented in figures 3.9, 3.10, and 3.11.

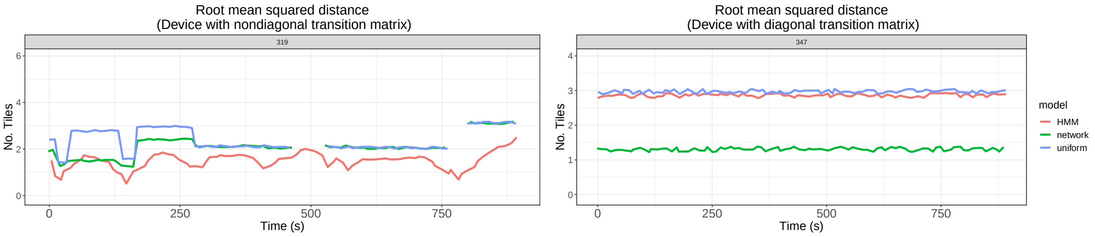

Figure 3.9: Root mean square dispersion of the device under three different models.

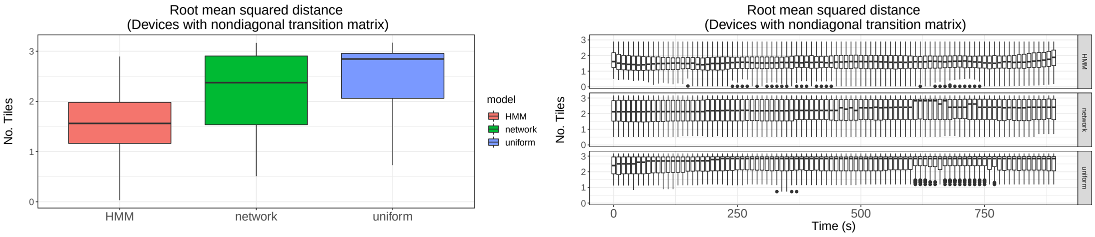

Figure 3.10: Root mean square dispersion under three different models. Devices with motion.

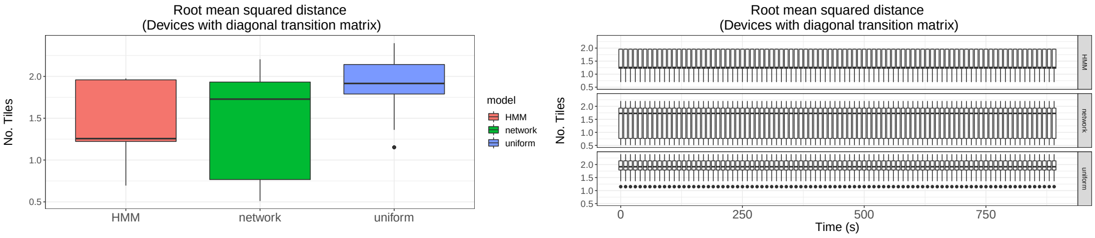

Figure 3.11: Root mean square dispersion under three different models. Devices with no motion.

Observe that these two figures of merit may be understood as playing the same roles as bias and variance in traditional accuracy assessment. We can make the following comments:

- In general terms, the dynamical approach performs better than the static analysis both in terms of distance to true positions and of the precision (size) of the spatial probability distribution when the device is moving. When the device is motionless, the models perform similarly, even slightly better for a network prior, which incorporates more prior information (see figure 3.6). This suggests that more priors should also be incorporated in the HMMs.

- It is relevant to mention that due to grid boundary effects and the non-dynamical character of the emission model, when the device is located close to the boundary of the map, where only one antenna can be possibly connected, the static analysis may produced a more accurate estimation. Since the emission model amounts to randomly choosing an antenna for the connection (like throwing a dice with probabilities b_{ai}) and in this time interval only the geographically external antenna is connected, the likelihood immediately drives the model to conclude that the most external tiles are more probable for the location. These effects should be taken into account in our HMM. This strongly suggests that territorial zones close to the coast should be carefully modelled.

- There exist time instants where the device is not connected to an antenna (see figure 3.12). In the static analysis, we cannot estimate the geolocation whereas with the dynamical approach we can.

- There exist time instants where an antenna oscillation phenomenon is detected because the mobile device moves in the frontier of two neighboring coverage areas (see figure 3.13). In both static analyses, this phenomenon in the handover mechanism is observed in the distance between the center of probabilities and the true position (notice the oscillations). However, in the HMM it is exactly the opposite. It is in these time instants in which we can estimate the geolocation of the device with higher accuracy. In our view, this is due to having more information (from two antennas) than otherwise, thus the dynamical model gains in accuracy.

- There exist some times instants where the dynamical information does not seem to improve the performance over the static models. In our view, this can be explained by our choice of initial priori and/or the choice of state, where only the geolocation is used. In physical dynamical systems you need both position and velocity to account for the inertia in order to compute (thus predict) the trajectory. In our case, when the device undertakes abrupt changes of direction, the model needs some time to adapt. More extensive modelling exercises with more complex states should be conducted to assess the differences.

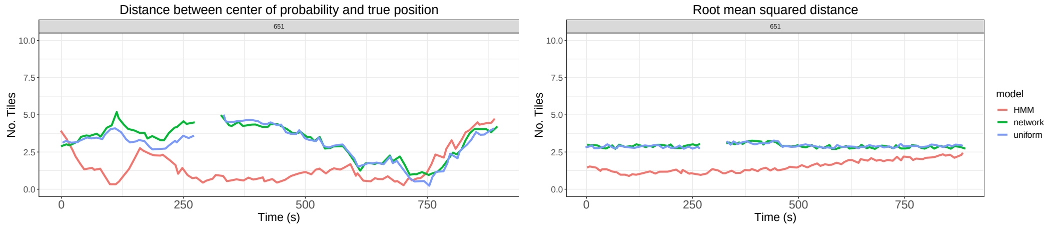

Figure 3.12: Comparison for device 651 under three different models.

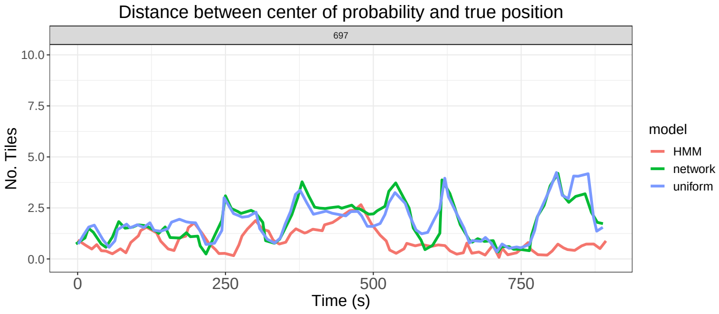

Figure 3.13: Distance between the center of location probabilities and the true position for device 697 under three different models.

To sum up, we have a modelling framework possibly integrating many diverse pieces of information to estimate the geolocation of each mobile device. The framework is fully thought in a modular way according to the RMF so that each module can be independent executed thus adapting to many different situations, especially regarding data access and data availability. Furthermore, the resulting posterior location probabilities γ_{dti} allow us not only to provide the geolocation estimates but also to account for uncertainty in the estimation process. This will have beneficial effects in later stages (uncertainty is transmitting from one stage to the following).

3.5. Future prospects

In our view, our proposal is a very first step to construct a production framework which needs to be substantiated in different aspects and further explored:

- The choice of the time increment Δt in the model (as suggested in the appendix) can be further elaborated. Apart from an educated guess for an a priori value of the maximum velocity for a device, more sophisticated estimates for an upper bound in terms of the connecting antennas and their geometrical arrangement should be explored.
- As many different emission models as possible should be tested first with simulated data and later on, if possible, with real data. The emission models are deeply rooted in data access conditions and data availability, which ultimately depend on the partnership models between MNOs and NSIs. Assessing the different possibilities will allow us to take better decisions about the optimal conditions to access and use mobile network data.
- A lot of work needs to be conducted to incorporate land use information and possibly more auxiliary information in the transition model. Obviously, in real territories we will find tiles corresponding to roads, parks, residential areas, work areas, leisure areas, etc. All this information could be possibly introduced in the transition model in terms of different parameters.
- One-tile movement models are not compulsory and more complex choices should also be explored.
- Different grid sizes should also be explored to assess their effects on the geolocation estimates.
- All potential models should also be explored with more complex movement behaviours in the simulator (also to be undertaken in the future).
- More figure of merits for the output posterior location probabilities should also be explored (e.g. the distribution entropy) to select the most appropriate ones.
- The scalability issue needs to be approached for this proposal to be applicable in real-life conditions. In this sense, both the pure computational approach and extra methodological elements should be developed.

For the current project, we have prioritised the construction of a proposal for an end-to-end framework providing a skeleton which can be further substantiated with joint work in the future.

4

Device duplicity

The statistical unit of analysis with mobile network data is the individual of a target population, not a mobile device. This clearly introduces the problem of device multiplicity in our statistical process, i.e. a given individual can carry more than one device with him/her, thus introducing the problem of multiple counting. For simplicity, throughout all our proposed process we shall make the simplifying assumption that each individual can carry at most 2 devices. The generalization to 3 or more devices is straightforward and, in any case, this circumstance is so rare that practically we are not making a strong assumption.

The bottom line of our approach amounts to classify every single device d in our dataset as corresponding to an individual with only one device (1:1 correspondence between devices and individuals) or as corresponding to an individual with two devices (2:1 correspondence between devices and individuals). This classification will be probabilistic, thus assigning a probability p_d of duplicity to each device d.

This chapter is organized as follows. In section 4.1 we begin by making general considerations about this module. In section 4.2, we propose a Bayesian approach directly connecting with the underlying HMM for geolocation. In section 4.3 we adopt the Bayesian approach to use the estimated distance between pair of devices. Finally, in section 4.4 we make considerations about the computational issues.

4.1. General considerations

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, 1998) in any model estimating the number of individuals in the target population 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 duplicity 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 of each device d to have a 2:1 device-individual correspondence. This probability distribution will indeed be the intermediate distribution in the chain (2.6). We need it to assess the **uncertainty** in this classification and not just to conduct the classification. The relevance of this exercise will be evident in the aggregation step in chapter 6.

4.2. Bayesian approach with network event data

4.2.1. Computation of the duplicity probability

To follow naturally the language of probability adopted with the use of HMMs for the geolocation, let us denote by D_{d1} the (Boolean) event in which the device d has a 1:1 device-individual correspondence. Equivalently, D_{d2} will denote the (Boolean) event in which the device d has a 2:1 device-individual correspondence. We shall compute probabilities making use of the network event information and auxiliary information already used for the construction of the underlying HMMs. With this notation we have $p_d = \mathbb{P}\left(D_{d2}|\mathbf{E},\mathbf{I}\right)$ and $q_d = 1 - p_d$ or, equivalently, $q_d = \mathbb{P}\left(D_{d1}|\mathbf{E},\mathbf{I}\right)$, where $\mathbf{E}$ stands for the network event variables $\mathbf{E}_d = \{E_{dt_0}, E_{dt_1}, \ldots, E_{dt_N}\}$ of all devices d and $\mathbf{I}$ stands for any available auxiliary information.

To compute p_d for each device d we proceed in two steps:

a) Let us denote by $D_{d_1d_2}$ the event "devices d_1 and d_2 are carried by the same individual" (duplicity event), so that the complement $D^c_{d_1d_2}$ stands for "devices d_1 and d_2 are not carried by the same individual" (no duplicity event). We argue as follows. The probability p_d for a device d is the probability of pair-duplicity $p_{dd'} \equiv \mathbb{P}\left(D_{dd'}|\mathbf{E},\mathbf{I}\right)$ corresponding to the device d' more similar to d. Thus, we can write

$$p_d = \max_{d' \neq d} \mathbb{P}\left(D_{dd'}|\mathbf{E},\mathbf{I}\right). \tag{4.1}$$

This will be the equation connecting the duplicity probability of each device d with the pair-duplicity $p_{dd'} \equiv \mathbb{P}\left(D_{dd'}|\mathbf{E},\mathbf{I}\right)$. Notice trivially that $p_{dd} = 0$ for all d.

b) We must compute the pair-duplicity probabilities $p_{d_1d_2}$. In particular, we shall use the set of network event variables $\mathbf{E}_{d_1}$ and $\mathbf{E}_{d_2}$ for both devices d_1 and d_2, so that $\mathbf{E}$ is reduced to $\mathbf{E} = \{\mathbf{E}_{d_1}, \mathbf{E}_{d_2}\}$. Then, we use Bayes' theorem to write

$$\begin{aligned}\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{E},\mathbf{I}\right) &= \frac{\mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2}\right)\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)}{\mathbb{P}\left(\mathbf{E}|D_{d_1d_2}\right)\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right) + \mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2},\mathbf{I}\right)\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)} \\ &= \frac{1}{1 + \frac{\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)}{\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)} \times \frac{\mathbb{P}\left(\mathbf{E}|D_{d_1d_2},\mathbf{I}\right)}{\mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2},\mathbf{I}\right)}},\end{aligned} \tag{4.2}$$

where $\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)$ and $\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)$ are prior probabilities for the duplicity and non-duplicity events and $\mathbb{P}\left(\mathbf{E}|D_{d_1d_2},\mathbf{I}\right)$ and $\mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2},\mathbf{I}\right)$ stand for the likelihoods under each hypothesis $D_{d_1d_2}$ and $D^c_{d_1d_2}$, respectively. All these probabilities are computed below.

The prior probabilities $\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)$ and $\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)$ should be computed making an optimal use of the auxiliary information $\mathbf{I}$. If no auxiliary information at the device level is available, we could for instance reason as follows. Let N^{D} denote the total number of devices and N^{ext} denote the estimated total number of individuals according to an external source. Then we can assume

$$\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right) = \frac{2\times\left(N_D - N^{\mathrm{ext}}\right)}{\binom{N_D}{2}}.$$

If some auxiliary information at the device level is available (for instance from the Customer Relationship Management database), if two devices d_1 and d_2 reside in different provinces, then naturally $\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right) = 1$ and $\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right) = 0$, hence $\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{E},\mathbf{I}\right) = 1$ irrespectively of the underlying HMMs for geolocation.

We shall denote the prior odds ratio of devices d_1 and d_2 by

$$\mathrm{OR}_{12} \equiv \frac{\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)}{\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)}$$

and the log odds ratio of devices d_1 and d_2 as $\lambda_{12} \equiv \log\left(\mathrm{OR}_{12}\right)$.

The likelihoods are computed using the HMMs presented in chapter 3. Let us start by considering $\mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2},\mathbf{I}\right)$. Under this assumption, indeed we have two independent models so that we can write

$$\mathbb{P}\left(\mathbf{E}|D^c_{d_1d_2},\mathbf{I}\right) = \mathbb{P}\left(\mathbf{E}_{d_1}|\mathbf{I}\right)\times\mathbb{P}\left(\mathbf{E}_{d_2}|\mathbf{I}\right), \tag{4.3}$$

where the likelihoods $\mathbb{P}\left(\mathbf{E}_{d_i}|\mathbf{I}\right)$, $i = 1, 2$, are just taken from the HMMs adjusted from the network event data of each device i, respectively.

Now the novelty arises from the computation of $\mathbb{P}\left(\mathbf{E}|D_{d_1d_2},\mathbf{I}\right)$. A brief reflection will immediately suggest the reader that we are again in an HMM but with a set of paired observed variables, as represented in figure 4.1.

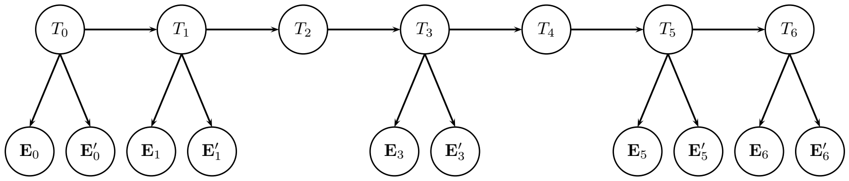

Figure 4.1: HMM for an individual carrying two devices, thus adapted to a double event.

The emission probabilities now read $\mathbb{P}\left(\mathbf{E}_t|T_t,\mathbf{I}\right) = \mathbb{P}\left(\mathbf{E}_{t1}|T_t,\mathbf{I}\right)\times\mathbb{P}\left(\mathbf{E}_{t2}|T_t,\mathbf{I}\right)$ and the whole formalism can be applied *mutatis mutandi* to this situation.

Thus, all we need are the likelihoods L_1 and L_2 for device d_1 and d_2 independently, and the likelihood L_{12} for the paired-observed model. If we use the log-likelihoods ℓ_1, ℓ_2, and ℓ_{12}, then our pair-duplicity probability will be given by

$$p_{d_1d_2} \equiv \mathbb{P}\left(D_{d_1d_2}|\mathbf{E},\mathbf{I}\right) = 1 - \frac{1}{1+\exp\left(\lambda_{12}+\ell_{12}-\ell_1-\ell_2\right)}. \tag{4.4}$$

Once the pair-duplicity probabilities have been computed, the duplicity probability p_{d_n} for each device d_n can be readily computed:

$$p_{d_n} = \max_{m\neq n}\left(1 - \frac{1}{(1+\exp\left(\lambda_{nm}+\ell_{nm}-\ell_n-\ell_m\right))}\right). \tag{4.5}$$

This duplicity probability for each device d_n will be the main output of the duplicity classification module, which will allow us to tackle the overcoverage problem before moving to the aggregate level. It is advisable to have a feeling about how the pair-duplicity probabilities $p_{d_1d_2}$ relate to the duplicity probability p_d. Let us consider different situations:

- Mobile device d_n with a 1:1 correspondence.
  Under this assumption, we have $p_{d_nd_m} \approx \epsilon \ll 1$ for all $m \neq n$. In these conditions we have

  $$p_{d_n} = \epsilon. \tag{4.6}$$

  Thus, we have the expected result. No duplicity is detected for d_n.

- Mobile device d_n with a 1:1 correspondence and a false positive.
  Under this assumption there exists another device m^* such that $p_{d_nd_{m^*}} = \xi_{m^*} \not\ll 1$ (but d_n and d_{m^*} does not pertain to the same individual) and $p_{d_nd_m} \approx \epsilon \ll 1$ for all $m \neq m^*, n$. In these conditions we have

  $$p_{d_n} = \xi_{m^*}. \tag{4.7}$$

  Thus, the false positive error is propagated to the duplicity probability p_{d_n}. In theory, this is the expected behaviour since the duplicity probability reflects what is happening with device m^* according to $p_{d_nd_{m^*}}$, but in practice this may pose a severe problem forcing us to compute the pair-duplicity probabilities $p_{d_nd_{m^*}}$ as accurately as possible. In particular, the use of auxiliary information I in the computation of prior probabilities $\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)$ is critical. Again, this impinges on the data availability and access agreements with MNOs.

- Mobile device d_n has a 2:1 correspondence with device m^*.
  Under this assumption, $p_{d_nd_{m^*}} = 1-\epsilon^*$, with $\epsilon^* \ll 1$, and $p_{d_nd_m} = \epsilon \ll 1$ for all $m \neq n, m^*$. In these conditions, we have

  $$p_{d_n} = 1-\epsilon^*. \tag{4.8}$$

  Thus, we have the expected result. Duplicity with device d_{m^*} is detected.

- Mobile device d_n has a 2:1 correspondence with device m^* and a false positive.
  Under this assumption (i) $p_{d_nd_{m^*}} = 1-\epsilon^*$, with $\epsilon^* \ll 1$, (ii) there exists another device $\bar{m}$ such that $p_{d_nd_{\bar{m}}} = 1-\bar{\epsilon}$, with $\bar{\epsilon} \ll 1$ (but they pertain to different individuals), and (iii) $p_{d_nd_m} = \epsilon \ll 1$ for all $m \neq n, m^*, \bar{m}$. In these conditions, we have

  $$p_{d_n} = \max\{1-\epsilon^*, 1-\bar{\epsilon}\}. \tag{4.9}$$

  Thus, we still have the expected result. Indeed, we see that the 2:1 correspondence case is robust against the misclassification of other pairs, in contraposition to the 1:1 correspondence case, which is very sensitive.

- Mobile device d_n has a 2:1 correspondence with device m^* and all false negatives. Under this assumption (i) $p_{d_n d_m} = \epsilon \ll 1$ for all $m \neq n$ (even for device d_{m^*}). In these conditions we have

$$p_{d_n} = \epsilon. \tag{4.10}$$

Thus, we get a wrong result. Notice that in this missclassification case all other devices m must yield wrong pair-duplicity pairs $p_{d_n d_m}$. In this line, it is important to scrutinize all the available information to detect possible duplicities according to this data. If data leads us to the impossibility of detect duplicities, we run into a serious identifiability problem. These considerations must be taken into account in the access agreements with MNOs to use the optimal information.

4.2.2. Illustrative example

We make use of the network event data simulator built in this ESSnet (WPI.2, 2019) to also illustrate the above proposal. To compute the pair-duplicity probabilities $p_{d_n d_m}$ we need to select priors $\mathbb{P}\left(D_{d_n d_m}|\mathbf{I}\right)$, which we choose as argued above, so that

$$\mathbb{P}\left(D_{d_n d_m}|\mathbf{I}\right) = \frac{N_D - N^{\text{ext}}}{\binom{N_D}{2}} = \frac{2 \cdot \left(N_D - N^{\text{ext}}\right)}{N_D \cdot (N_D - 1)}. \tag{4.11}$$

Thus, the odds ratio OR_{nm} will be readily computed. The likelihoods ℓ_n and ℓ_{nm} are readily computed using the HMM presented in chapter 3. In figure 4.2 we represent the ROC curve and corresponding AUC for the set of pair-duplicity probabilities $p_{d_1 d_2}$ under this approach. This curve clearly shows our ability to discern whether a given pair corresponds to the same individual or not.

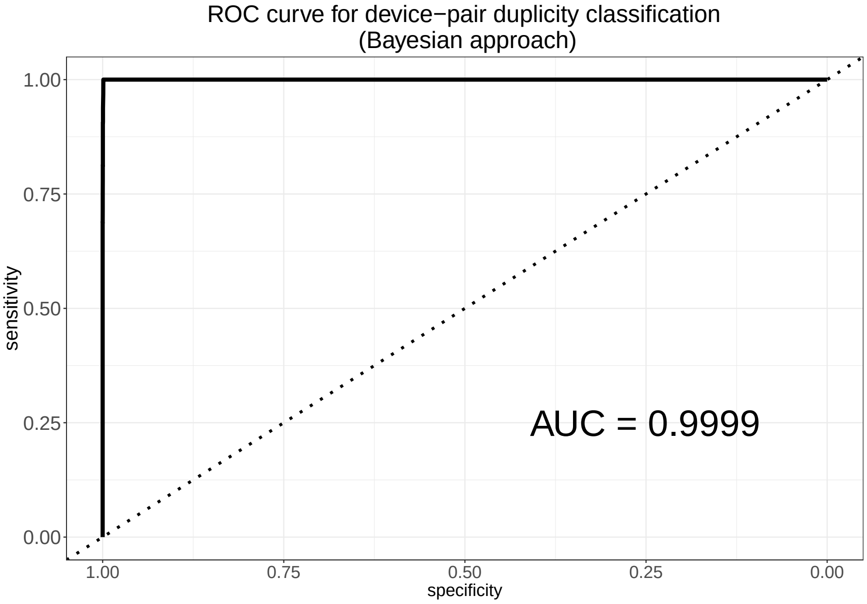

Figure 4.2: ROC curve and AUC for the Bayesian approach to device-pair duplicity classification.

Once pair-duplicity probabilities $p_{d_1 d_2}$ are computed, the next step is to find the device-duplicity probabilities p_d. Again, we use ROC analysis to show our capacity to discern whether a given device belongs to the 1:1 or 2:1 class. This is represented in figure 4.3.

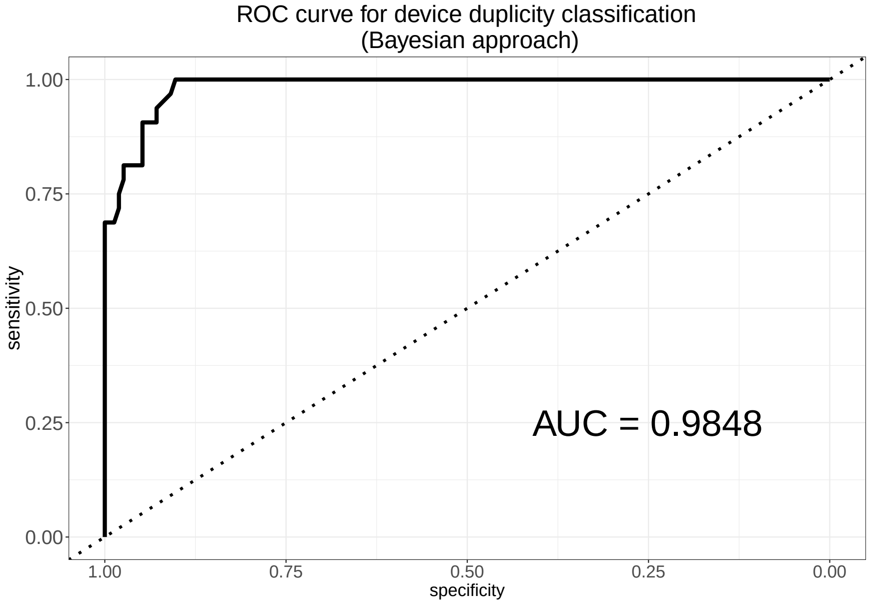

Figure 4.3: ROC curves and AUC for the Bayesian approach to device duplicity classification.

4.3. Approach based on the distance between centers of location probabilities

4.3.1. Computation of the duplicity probability

Our second proposal follows a similar Bayesian approach, but now instead of using network event variables $\mathbf{E}$ we will focus on properties of the trajectories derived from the HMM and in particular from the location probability distributions $\{\gamma_{dti}\}$ of all devices d.

In this line of thought, we again write

$$\begin{aligned}\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{X},\mathbf{I}\right) &= \frac{\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2}\right)\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)}{\mathbb{P}\left(\mathbf{X}|D_{d_1d_2}\right)\mathbb{P}\left(D_{d_1d_2},\mathbf{I}\right)+\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2},\mathbf{I}\right)\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)}\\ &= \frac{1}{1+\frac{\mathbb{P}\left(D_{d_1d_2}|\mathbf{I}\right)}{\mathbb{P}\left(D^c_{d_1d_2}|\mathbf{I}\right)}\times\frac{\mathbb{P}\left(\mathbf{X}|D_{d_1d_2},\mathbf{I}\right)}{\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2},\mathbf{I}\right)}},\end{aligned} \tag{4.12}$$

where we are using the same notation as above and where $\mathbf{X}$ denotes the variable(s) related to the estimated trajectories in terms of posterior location probabilities (see below). Notice that this computation runs parallel to the Bayesian approach with network event data. The difference now stems out from the computation of the pair-duplicity probabilities $\mathbb{P}\left(\mathbf{X}|D_{d_1d_2},\mathbf{I}\right)$ and $\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2},\mathbf{I}\right)$.

The idea to carry out this computation is very simple. Firstly, we find the probability distribution of the signed distance between the x- and y-axis position of each pair of devices. Let X_{dt} denotes the random variable for the x coordinate of a device d. This is a discrete random variable with support in the x coordinates of the centroid of all tiles. Let us define the signed distance Δ_{x,d_1d_2t} in the x axis as $\Delta_{x,d_1d_2t} = X_{d_1t} - X_{d_2t}$. Notice that this variable has support in the set of (signed) distances $\{\pm\delta_k\}$ for all differences δ_k between the x coordinates of the tile centroids. The probability distribution is given by

$$\mathbb{P}\left(\Delta_{x,d_1d_2t}=\pm\delta_k\right)=\sum_{\{i_1,i_2:x^c_{i_1}-x^c_{i_2}=\pm\delta_k\}}\gamma_{d_1i_1t}\gamma_{d_2i_2t},\tag{4.13}$$

where we have assumed independence in the geolocation of both devices. We proceed in a similar way and define the random variable Δ_{y,d_1d_2t} for each pair of devices d_1 and d_2 at each time instant t for axis y. The bottom line is straightforward. If a device d_1 corresponds to an individual with two devices (2:1), there will be another device d_2 such that their distance will be significatively close to 0 along their trajectories. We shall identify $\mathbf{X}$ with the statement that an appropriate measure of location of the probability distributions Δ_{x,d_1d_2t} and Δ_{y,d_1d_2y} will be in an interval around 0.

Thus, for each time instant t we shall compute the mode of both distributions Δ_{x,d_1d_2t} and Δ_{y,d_1d_2t}. Let us denote them by δ^*_{xt} and δ^*_{yt}. Next, we shall check whether $|\delta^*_{xt}|\leq\xi\cdot\max\{rd_{d_1t},rd_{d_2t}\}$ and $|\delta^*_{yt}|\leq\xi\cdot\max\{rd_{d_1t},rd_{d_2t}\}$, where rd_{d_1t},rd_{d_2t} stand for the radii of dispersion of devices d_1 and d_2, respectively (see 3.4). Then, we define

$$\hat{p}^{\text{mode}}_{d_1d_2}\equiv\mathbb{P}\left(\mathbf{X}|D_{d_1d_2},\mathbf{I}\right)=\frac{\#\{t=1,\dots,T:|\delta^*_{xt}|\leq\xi\cdot\max\{rd_{d_1t},rd_{d_2t}\},|\delta^*_{yt}|\leq\xi\cdot\max\{rd_{d_1t},rd_{d_2t}\}\}}{T}.\tag{4.14}$$

Notice that the mode can be substituted for other measures of distribution location such as mean or median. For the probability $\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2},\mathbf{I}\right)$ we simply assume

$$\mathbb{P}\left(\mathbf{X}|D^c_{d_1d_2},\mathbf{I}\right)=1-\mathbb{P}\left(\mathbf{X}|D_{d_1d_2},\mathbf{I}\right)\tag{4.15}$$

Notice that this equivalent to check whether $|\delta^*_{xt}|>\bar{\xi}\cdot\max\{rd_{d_1t},rd_{d_2t}\}$ or $|\delta^*_{yt}|>\bar{\xi}\cdot\max\{rd_{d_1t},rd_{d_2t}\}$ for the same threshold $\bar{\xi}=\xi$, and proceed analogously.

Thus, the pair-duplicity probability $p_{d_1d_2}$ is computed according to:

$$p_{d_1d_2}=1-\frac{1}{1+\text{OR}_{12}\times\frac{\hat{p}^{\text{mode}}_{d_1d_2}}{1-\hat{p}^{\text{mode}}_{d_1d_2}}}.\tag{4.16}$$

Again, once the pair-duplicity probabilities have been computed, the duplicity probability for each device d_n will be assigned as

$$p_{d_n}=\max_{m\neq n}\left(1-\frac{1}{1+\text{OR}_{nm}\times\frac{\hat{p}^{\text{mode}}_{d_nd_m}}{1-\hat{p}^{\text{mode}}_{d_nd_m}}}\right)\tag{4.17}$$

Let us consider extremely simple examples. Consider a set of only 2 devices carried by the same individual. Let us compute p_{d_1}. Ideally, both devices have the same spatial distribution $\{\gamma_{dti}\}$ (really this need not be so, since two devices of the same individual can connect to different antennas and thus generate different network events). Under this ideal assumption, the random variables Δ_{x,d_1d_2t} and Δ_{y,d_1d_2t} are degenerate with value 0 for all t. Then the modes δ^*_{xt} and δ^*_{yt} of their distributions will satisfy $|\delta^*_{xt}|\leq\xi\cdot rd_{d_1t}$ and $|\delta^*_{yt}|\leq\xi\cdot rd_{d_1t}$, for any value $\xi\geq0$ and all t, thus $\hat{p}^{\text{mode}}_{d_1d_2}=1$, $p_{d_1d_2}=1$ and $p_{d_1}=1$, as expected.

Let us consider the extreme opposite example with several devices each in a different city. Consider a device d_1. Under these conditions, the random variables Δ_{x,d_1d_mt} and Δ_{y,d_1d_mt} will have large absolute values with very high probability for all $d_m\neq d_1$, so that their modes will be certainly outside a small interval around 0. Thus, $\hat{p}^{\text{mode}}_{d_1 d_m} = 0$, $p_{d_1 d_m} = 0$ for all $d_m \neq d_1$, and $p_{d_1} = 0$, as expected.

For intermediate situations, the results will depend on the input spatial distributions $\{\gamma_{dti}\}$. Notice also that the robustness analysis conducted with network data is still valid here.

4.3.2. Illustrative examples

We make use of the network event data simulator to also illustrate the above proposal. In figure 4.4 we illustrate two extreme examples where the reader can appreciate the difference in the mass probability functions of two extreme situations of two devices (i) carried by the same individual and (ii) carried by different individuals.

Figure 4.4: Probability mass functions over time of the signed distance along the x axis for two devices carried by the same individual (left) and two different individuals (right).

To compute the pair-duplicity probabilities $p_{d_n d_m}$ we need to select priors $\mathbb{P}\left(D_{d_n d_m} | \mathbf{I}\right)$. We argue as above. If N_D is the total number of mobile devices and N^{ext} is the estimated number of individuals according to an external source, the number of pairs of devices of class 2:1 will be $N_D - N^{\text{ext}}$ whereas the total number of pairs would be $\binom{N_D}{2}$. Thus, we choose

$$\mathbb{P}\left(D_{d_n d_m} | \mathbf{I}\right) = \frac{N_D - N^{\text{ext}}}{\binom{N_D}{2}} = \frac{2 \cdot \left(N_D - N^{\text{ext}}\right)}{N_D \cdot (N_D - 1)}. \tag{4.18}$$

Thus, the odds ratio OR_{nm} will be readily computed. To analyse the performance of different thresholds r for the pair-duplicity probabilities we can make use of ROC analysis. This is illustrated in figure 4.5 for thresholds $r = \xi \times \max\{rd_{d_1}, rd_{d_2}\}$, with $\xi = 0.50, 0.75, 1.00, 1.25$, where rd_d stands for the radius of dispersion of mobile device d (see section 3.4). Two location measures for the distance distributions have been tested: mode and mean.

These curves clearly show our ability to discern whether a given pair corresponds to the same individual or not.

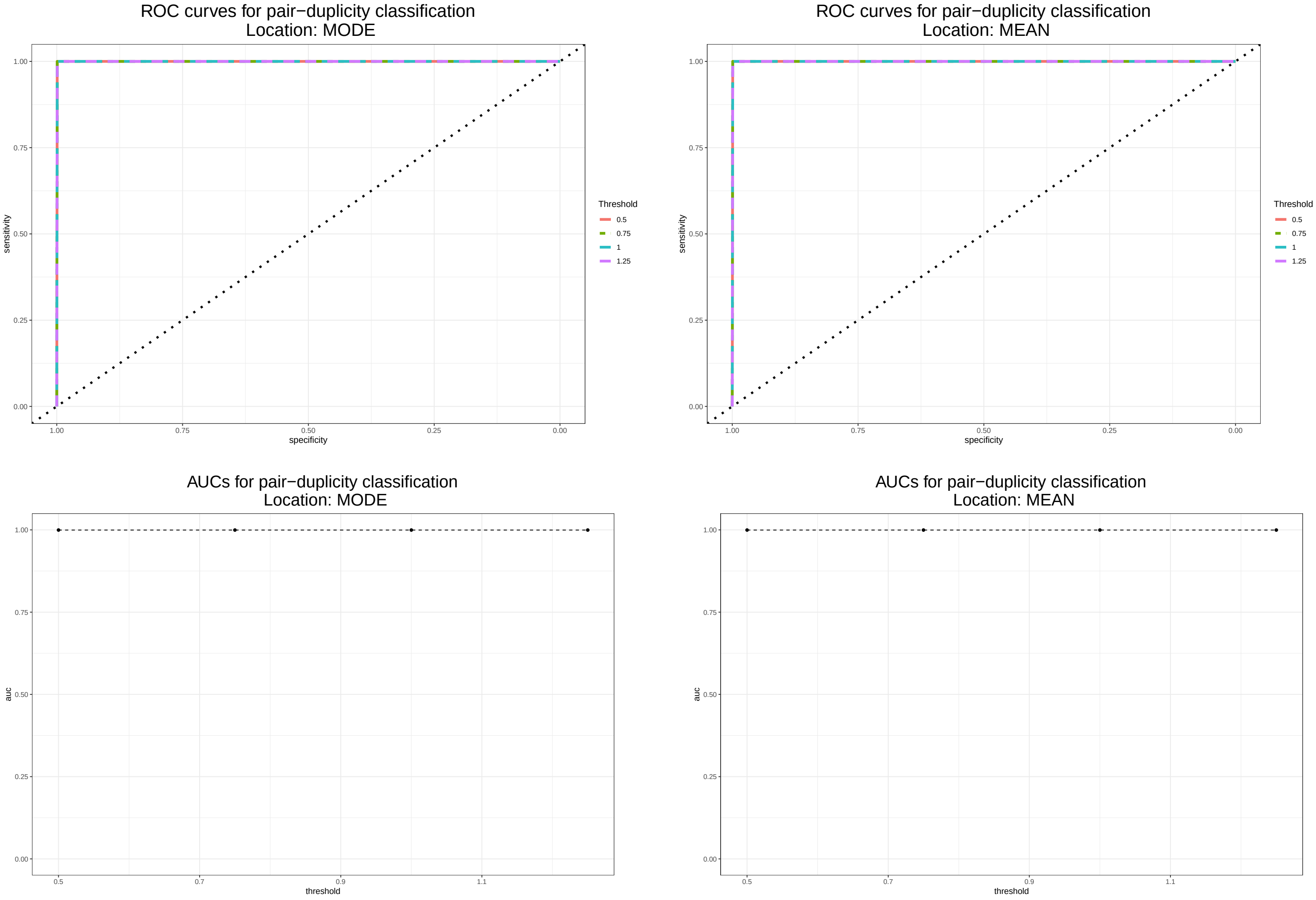

Figure 4.5: ROC curves and AUC of pair-duplicity for different thresholds ξ.

Once pair-duplicity probabilities $p_{d_1 d_2}$ are computed, the next step is to find the device-duplicity probabilities p_d. Again, we use ROC analysis to show our capacity to discern whether a given device belongs to the 1:1 or 2:1 class. This is represented in figure 4.6.

4.4. Computational issues

The proposals included above require pairwise comparison for all devices. In practice, indeed, this is too high a computational burden which can be simplified. We propose to undertake some preprocessing tasks selecting for each device d those potential candidates upon which we shall apply the above methodological solutions.

Two proposals are formulated:

- Define a distance measure between network events variables $\mathbf{E}_t$. For example, if the network event variables are the antenna identification variables corresponding to each event at time t, then we can think of $d(E_{d_1 t}, E_{d_2 t})$ as the physical distance between the

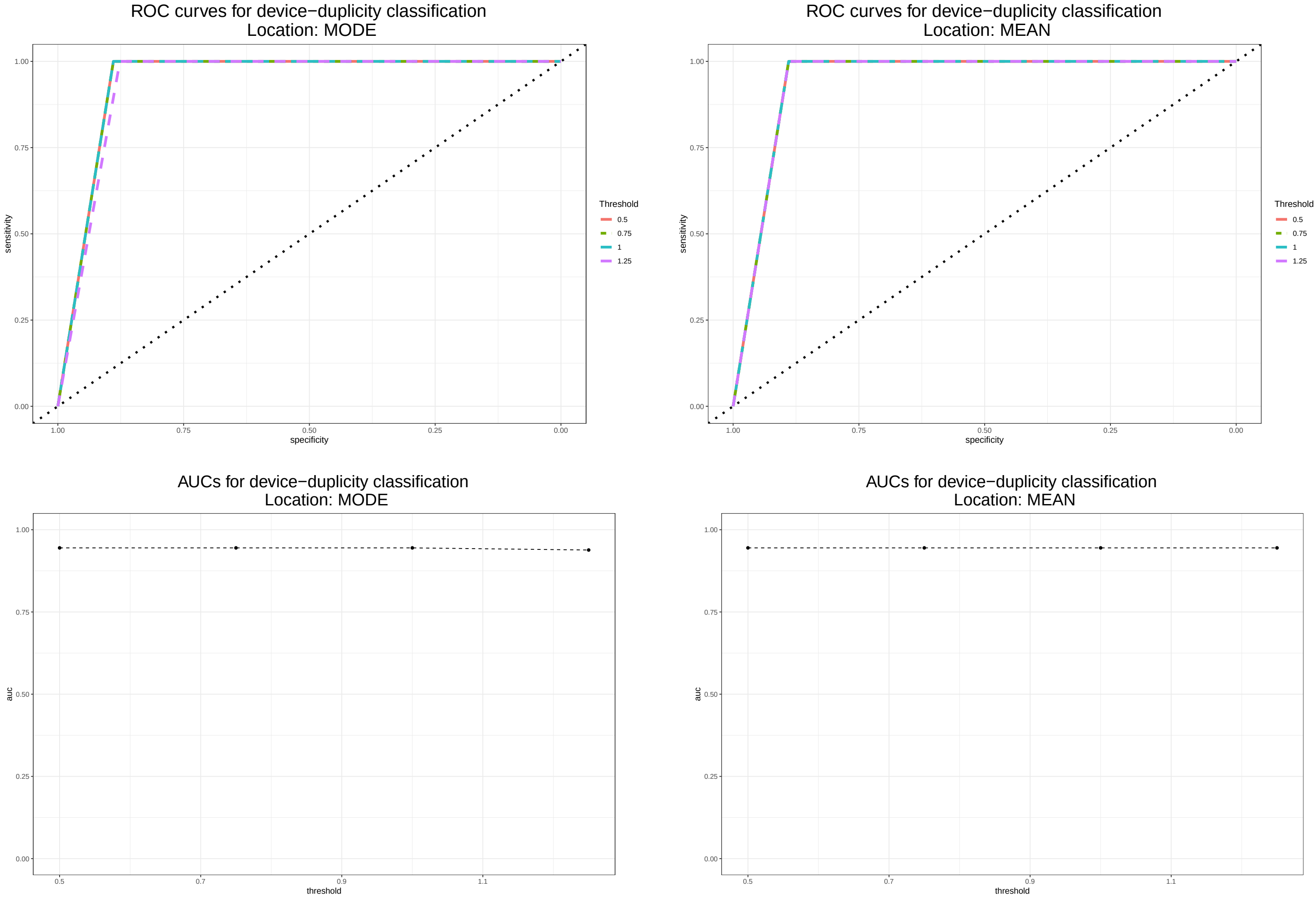

Figure 4.6: ROC curves and AUC of device-duplicity for different thresholds ξ.

antennas. Thus, given a device d, only those devices d' where $d(E_{dt}, E_{d't}) < \bar{\xi}$ for a given number of time instants t will be considered for the methods depicted above. Furthermore, this distance can be used for the computation of the log odds ratio λ_{nm} in the Bayesian approach.

- For the trajectory approach, take a reduced random sample of time instants to filter those critical devices for which the comprehensive comparison for the whole time span is to be undertaken. The rest will be discarded for the pairwise in-depth comparison.

Nonetheless, as in other modules, the computational issue is not fully approached in this document. Already with the low-scale synthetic data produced by the data simulator, we can already claim that both sparsity and parallelization will need to be introduced in the whole end-to-end process.

5

Statistical filtering

This chapter is devoted to the identification of the target population in the mobile network data set and derived data sets (posterior location probabilities, for example). In practical terms this amounts to identifying domestic tourists, inbound tourists, outbound tourists, commuters, etc. in our data sets. We refer to this as *statistical filtering*, where we use the term *statistical* to distinguish this filtering exercise from the preprocessing steps in which, e.g., machine-to-machine data are previously filtered out. Notice that the latter rests mostly on technological issues and definitions, whereas the former is a clearly statistical analytical exercise.

As in the whole deliverable, we shall be focusing on geolocation data, i.e., on movement data discarding interaction information (e.g. calls among subscribers) or Internet traffic (e.g. usage of mobile apps). In a fully-fledged production environment in real conditions, the ideal scenario would be to use as much information as possible. Thus, we shall concentrate on analyses upon the geolocation data, i.e. upon the network event data and location probabilities derived thereof.

Regretfully, given the problems in accessing real mobile network data, and the current status of development of the network event data simulator, the contents of this chapter are not so far developed as the preceding ones. The current movement patterns for individuals (hence also for mobile devices) in the data simulator are restricted to random walks and random walks with drift, both with intermixing periods of stops (stays, i.e. no movement at all) for the whole population. In this sense, we lack synthetic data to test concrete proposals, as with the geolocation of data. We would need more complex and realistic individual movement patterns and elements (Lévy flights, home/work locations, usual environments, etc.). For this reason, we will limit ourselves to provide generic proposals to be tested in the future both on real data and on synthetic data after a further development of the simulator.

5.1. The proposed approach

As part of the end-to-end process depicted in chapter 2, our proposed approach for the statistical filtering of target populations is strongly based on the geolocation outputs obtained from the preceding process modules. Different aspects are to be taken into account:

- As before, 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 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 hidden Markov models, as described in chapter 3, 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.

- The spatial resolution issue is dealt with by using the reference grid introduced in chapter 3. 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).

- The algorithms to be developed to statistically filter the target population will be mainly based on quantitative measures of movement data. In particular, from the HMMs fitted to the data (especially the location probabilities) we shall derive a 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 usual environment, home/work location, second home location, leisure activity times and locations, etc. A problematic aspect with this new data source is that traditional statistical definitions will need some revision or refinement. For example, in the home detection problem, which is an intermediate problem in the identification of target populations, census data (or similar official data) are commonly used to calibrate or validate estimates. The notion of home obtained from traditional sources is mainly an administrative concept arising from the use of administrative registers. In this way, e.g., a University student may be registered in her family home whereas she spends nine months in a college. What definition of home should then be used? This has introduced subtleties like the distinction between residential and present population in official statistics.

  In this line of thought, an important input for target population identification algorithms is the establishment of a clear-cut definition for each statistical concept involved, so that the algorithms are designed to cover these definitions.

- 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 movement 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 to validate the algorithm for the whole population needs to be rigorously assessed.

In the next section we will provide a generic view of quantitative measures, together with some concrete illustrative examples, upon the trajectories assigned to the geolocated data (location probabilities) obtained from the application of an HMM. Thus, the starting point will be the construction of a trajectory for each device.

In our model introduced in chapter 3 the state of the HMM was defined in terms of the tile where the device is positioned. Thus, the concept of space-time 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. If auxiliary information for each tile is available, instead of the geographical centroid of each tile, another "statistical"centroid can be used (e.g. using land use information and/or official population density figures). It is obvious again that the smaller the tiles, the more precise the estimation procedures.

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

$$T^{*}_{dt_0:t_N} = \text{argmax}_{T_{dt_0:t_N}} \mathbb{P}\left(T_{dt_0:t_N} \middle| \mathbf{E}_{dt_0:t_N}\right), \tag{5.1}$$

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

$$T^{*}_{dt_0:t_N} = \left(\text{argmax}_{T_{dt_0}} \gamma_{dt_0}, \text{argmax}_{T_{dt_1}} \gamma_{dt_1}, \ldots, \text{argmax}_{T_{dt_N}} \gamma_{dt_N}\right), \tag{5.2}$$

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.

5.2. Quantitative measures of movement data

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 movements (see e.g. Long and Nelson, 2013).

- Pattern and cluster methods.- These seek to identify spatial–temporal 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.

- Individual–group dynamics.- This set of measures compile methods focusing on individual device movement 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.

We include an illustrative example with a set of indicators. We shall compute them on the simulated scenario with 207 devices in a territory with an irregular polygon shape and a bounding box of 10km × 10km and 28 (25 omnidirectional and 3 directional) antennas. The indicators are strongly inspired on those used in animal trajectory analysis (see McLean and Skowron Volponi, 2018, and references therein).

1. Number of coordinates (`nCoord`).- This is the observed number of coordinates along the path, thus coincidental with the time extension of the HMM.
2. Path length (`length`).- This is the total length of the path, i.e.

$$\text{Length} = \sum_{t=1}^{T} \ell_t,$$

where $\ell_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$.

3. Path distance (`distance`).- This is the net distance between the initial and final fixes (points) in the path, i.e.

$$\text{distance} = \sqrt{(x_T - x_0)^2 + (y_T - y_0)^2}.$$

4. Path duration (`duration`).- This is the total duration of the path, i.e.

$$\text{duration} = t_N - t_0.$$

5. Mean velocity (`meanVelocity`).- This is the global mean velocity of the device along the path, i.e.

$$\text{meanVelocity} = \frac{1}{\text{duration}}(x_T - x_0, y_T - y_0).$$

Notice that it is a vector, thus we compute both the x- and y- dimensions.

6. Radius of gyration (`Rg`).- This is the radius of gyration of the path according to the formula

$$Rg = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_t^2 + y_t^2\right)}.$$

It provides a view of the extension of the territory range covered by the path.

7. Path straightness (`straightness`).- This is the index $\frac{\text{distance}}{\text{length}}$, which provides a first-order magnitude of the tortuosity of the path, with values between 0 (extremely tortuous) and 1 (a straight line).

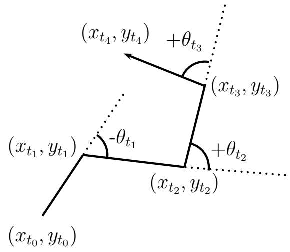

Figure 5.1: Turning angles of a path.

8. Turning angles (`turningAngles`).- These are the angles θ_t denoting the change of direction at each time instant t. See figure 5.1.

9. Directional change (`directionalChange`).- This is a measure of the speed of angular change of direction, defined as

$$\text{directionalChange}_{t_i} = \frac{\theta_{t_i}}{t_i - t_{i-1}}.$$

10. r Index (`r`).- This is another more complex measure of the tortuosity of the path, defined as

$$r = \left| \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} e^{i\bar{\theta}_t} \right|$$

where $\bar{\theta}_t$ denotes the turning angle (see figure 5.1) at time t of the rediscretized path obtained by sampling the path at equal-length steps.

11. Maximum expected displacement (`EmaxA` and `EmaxB`).- These two related indicators provide a measure of the maximum expected movement according to

$$E^a_{max} = \frac{\xi}{1-\xi},$$

where $\xi \equiv \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \cos\left(\bar{\theta}_t\right)$, with $\bar{\theta}_t$ being the turning angles of the rediscretised path obtained by sampling the path at equal-length steps.

The related indicator E^b_{max} is defined as

$$E^b_{max} = \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \bar{\Delta}_t \times \frac{\xi}{1-\xi},$$

where $\bar{\Delta}_t$ is the step length at time t of the rediscretised path.

12. Path sinuosity (`sinuosity` and `sinuosity2`).- The original path sinuosity index is defined as

$$\text{sinuosity} = 1.18 \times \frac{\sigma_\theta}{\sqrt{\bar{\Delta}}},$$

where $\sigma_\theta = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\theta_t - \bar{\theta})^2}$, $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ and $\bar{\Delta} = \frac{1}{T}\sum_{t=1}^{T}\Delta_t$. A second version using rediscretised paths is given by:

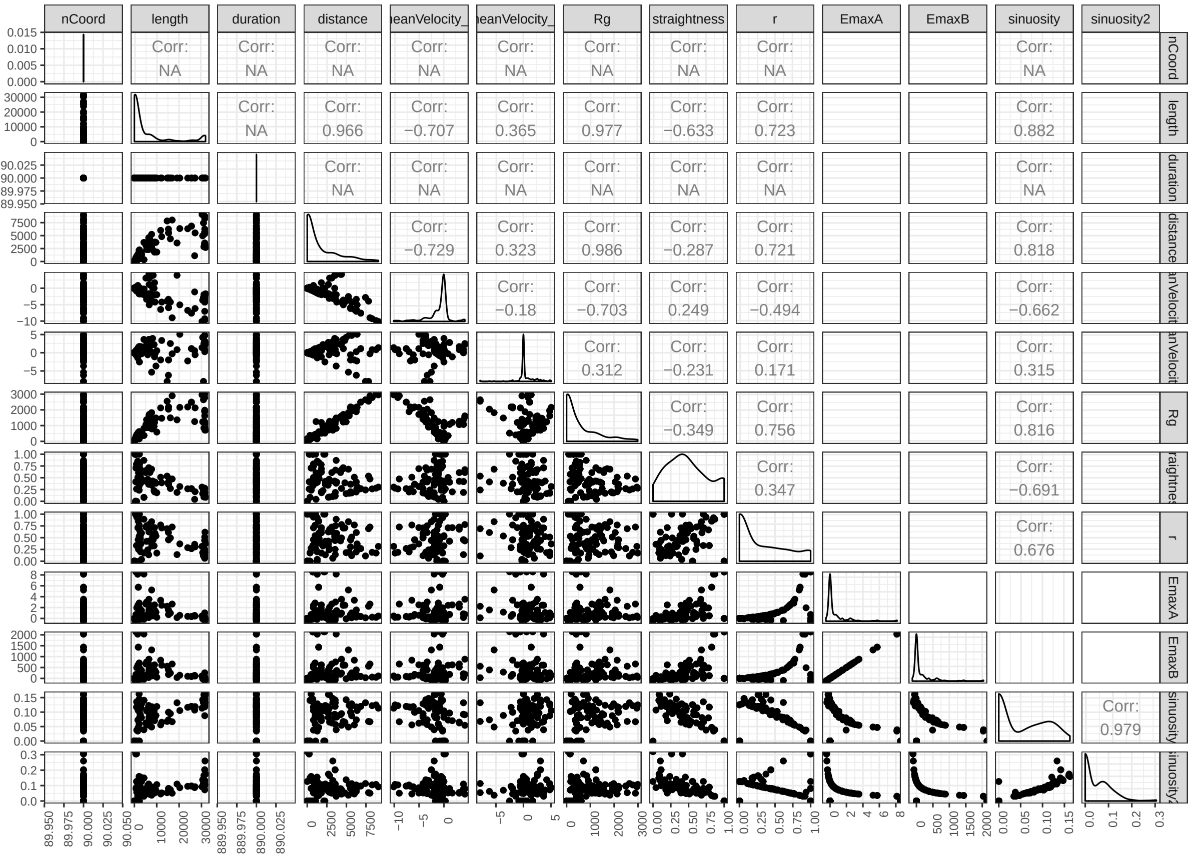

Figure 5.2: Path indicators (NA correlations indicate static devices; blank correlations indicate infinite values involved, i.e. straight line paths).

$$\text{sinuosity2} = \frac{2}{\sqrt{\bar{\Delta} \times \left(\frac{1+\xi}{1-\xi}\right) + \left(\frac{\sigma_{\Delta}}{\bar{\Delta}}\right)^2}}.$$

We have computed these indicators on our simulated scenario producing the values represented in figures 5.2 and 5.3. This list of indicators is not exhaustive (even some alternative forms for them can be found in the literature as for $E_{max}^{a,b}$ or the sinuosity index). Our main argument is that filtering, comprising identification of usual environment, home/work detection, second home detection, etc., must be based on detailed algorithms using these indicators avoiding as much as possible extremely simplistic approaches such as a home is a location where devices are between 23:00 and 6:00 or similar. Ultimately, findings thereof should be connected to other sociodemographic variables producing thus novel insights.

As a simple example, for each given path we can identify the time instants where the observed speed is below a given threshold for a consecutive number of time intervals thus identifying potential home/work locations (see figure 5.4). Then, different indicators can be computed for this subpath so that further distinction between activities could be unravelled (shopping, sporting, etc.). Notice that the limit imposed by the spatial resolution of the HMM establishes a bound in this regard.

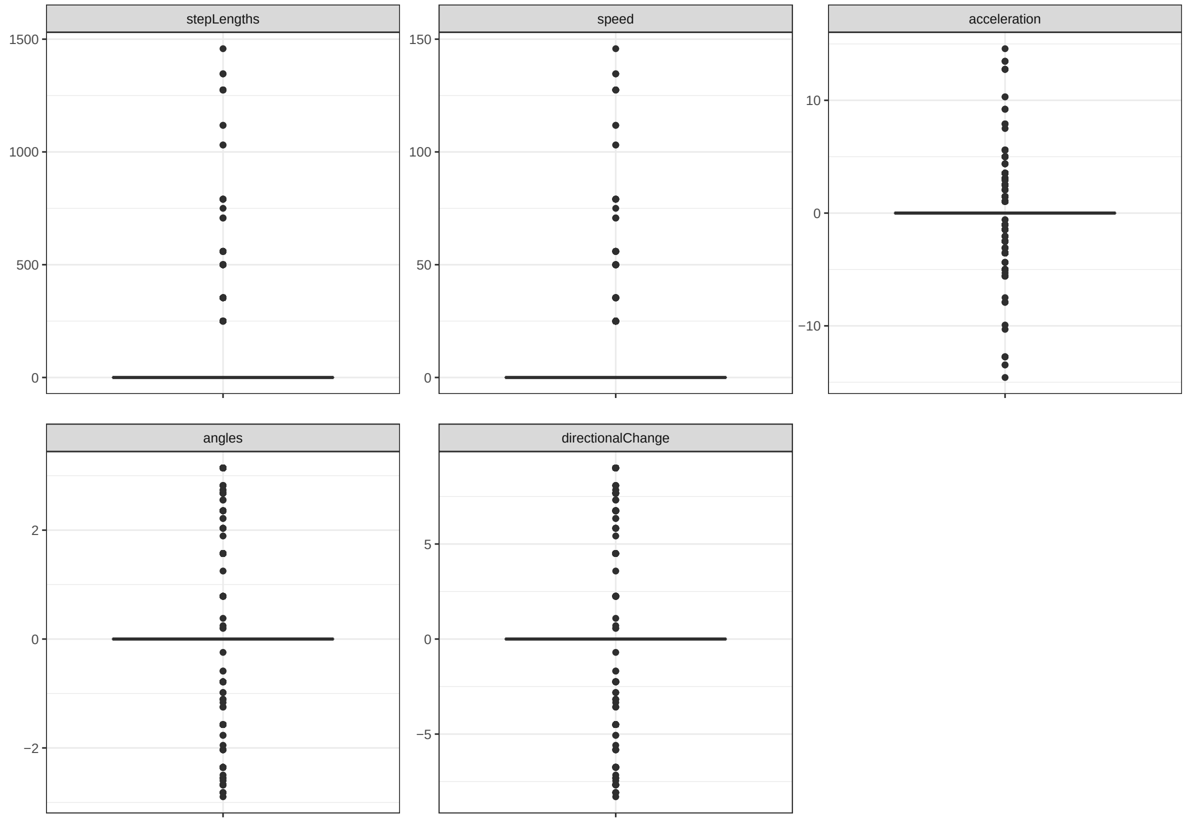

Figure 5.3: Path indicator distributions.

The reader immediately will realise how we need more complex and realistic movement patterns in the simulator to go deep into our analysis. In the example in figure 5.4 the movement pattern does not correspond to a realistic human movement whatsoever, so that no reasonable detection algorithm can be proposed using this data. This remains for further work in the future.

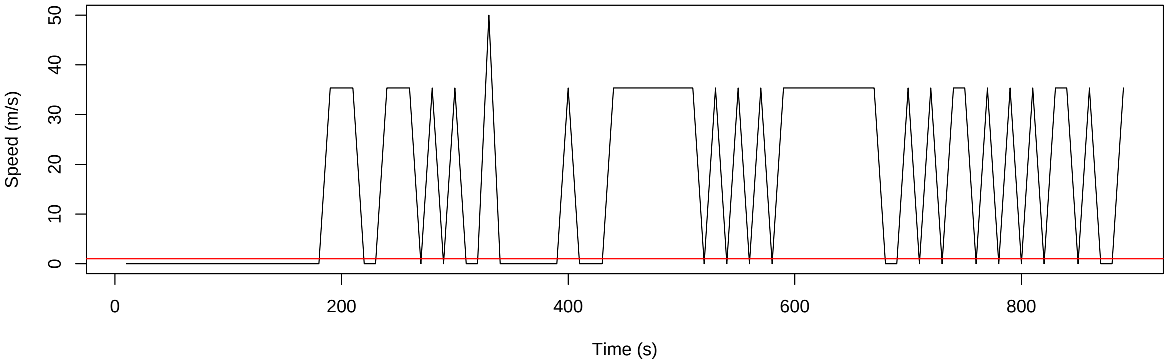

Figure 5.4: Time instants of a given device path with a speed below $1\ ms^{-1}$.

Finally, let us close this chapter by calling reader's attention on the positive feedback arising from this statistical filtering exercise. Once concepts such as usual environment, home/work location, second home location, etc. are computed, the definition of state for the HMM could be enhance thus incorporating more information into the geolocation estimation.

As a final suggestion widening the possibilities, instead of defining indicators such as above, deep learning techniques could be also tested to extract different characteristics of the trajectories.

6

Aggregation of detected individuals

This chapter is devoted to the statistical process of aggregating individuals detected by the network according to the preceding modules. Notice that this is the step where we go from the information at the device level (filtered location and duplicity probabilities) to the information at the territorial unit level.

The chapter is divided into section 6.1 with general remarks about our approach and section 6.2 with mathematical details on the construction of the probability distribution of the number of individuals detected by the network.

6.1. General remarks

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 chapter 4 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 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 in Lehmann and Casella, 1998, 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 scale.

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 - see chapter 3) and if the duplicity of a given device (carried by an individual with another device) is also probabilistic in nature (see chapter 4), then a priori it is impossible to provide a certain number of individuals[1] 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 to produce 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 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.

For the subsequent sections and the proposals contained therein, we shall concentrate of the estimation of present population counts. We shall understand as present population the collection of every single individual physically present in the geographic territory in the time interval under analysis. Alternative more rigorous definitions can be formulated but these would lead us to consider some form of statistical filtering 5, which we shall avoid at this point until this latter module is developed. In the present population we shall consider two types of individuals, namely subscribers and non-subscribers of the MNO network under analysis (we limit ourselves to the single-MNO scenario; the multiple-MNO scenario lies beyond the scope of this project and is left for future research). Subscribers are those individuals detected by the network whose total number will be estimated per territorial unit and time interval in this chapter.

6.2. Probability distribution of the number of detected individuals

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 $\mathbf{e}_i^{(2)} = \frac{1}{2} \times \mathbf{e}_i$, 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 proposal devised in chapter 4. Should we consider a more general situation, the generalization is obvious, although more computationally demanding.

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

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

$$\mathbb{P}\left(\mathbf{T}_{dt}=\mathbf{e}_i^{(1)}|\mathbf{E}_{1:D}\right) = \gamma_{dti}\times(1-p_d) \tag{6.1a}$$

$$\mathbb{P}\left(\mathbf{T}_{dt}=\mathbf{e}_i^{(2)}|\mathbf{E}_{1:D}\right) = \gamma_{dti}\times p_d \tag{6.1b}$$

where p_d is the device duplicity probability introduced in chapter 4. Notice that this is a categorical or multinoulli random variable. Finally, we define the multivariate random variable $\mathbf{N}_t^{\text{net}}$ providing the number of individuals N_{ti}^{net} detected by the network at each tile $i=1,\dots,N_T$ at time instant t:

$$\mathbf{N}_t^{\text{net}}=\sum_{d=1}^{D}\mathbf{T}_{dt}. \tag{6.2}$$

The sum spans over the number of devices filtered as members of the target population according to chapter 5. 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 $p_d=0$ 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 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 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 $p_d\neq0$ 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 movement 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 movement patterns and simulated correlated trajectories (probably connected to the sharing of usual environments, home/work locations, etc.).

The issue about the size of the tile also makes us consider the computation of the distribution of the number of detected individuals at a coarser territorial degree. Let us consider a coarser territorial breakdown composed of combination of tiles called, say, regions. We shall denote them as $\bar{T}_r = \bigcup_{i\in\mathcal{I}_r} T_i$, where $\mathcal{I}_r$ denotes the set of tile indices composing region r. If the independence assumption still holds (because the size of the region is still small enough), then we can reproduce the whole derivation above just by defining the location probability $\bar{\gamma}_{dtr}$ at region r as

$$\bar{\gamma}_{dtr} = \sum_{i\in\mathcal{I}_r} \gamma_{dti}. \tag{6.3}$$

The subsequent elaboration to build the final Poisson-multinomial-distributed number of detected individuals is completely similar. Notice again that there exists a limitation in the sum of device-level distributions put by the size of the underlying region breakdown.

Let us illustrate this approach with an example generated with the mobile network event simulator. We consider a toy scenario with a population of 186 subscribers with 218 mobile devices in a territory with a bounding box of 10km × 10km divided into 10 regions.

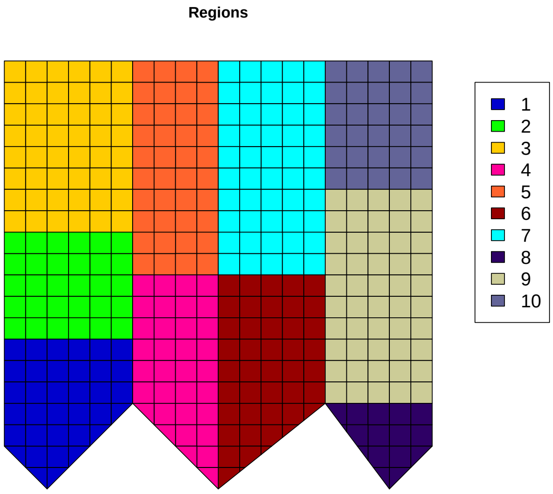

Figure 6.1: Territorial division into regions for a toy scenario.

The simulator runs a scenario where every individual moves during 890s according to a random walk with a drift across the territory. Network events (or no event if no connection is active) are recorded every 10s. We have available information from the network to apply the simplified radio wave propagation model in chapter 3 (basically, power and loss exponent for each antenna). Applying the hidden Markov model described in chapter 3 we compute the posterior location γ_{dti} for each device at each time instant t and each tile i. We apply the methodology in section 4.2 to compute the duplicity probability p_d for each device d. No statistical filtering algorithm is applied since we want to estimate the number of individuals detected by the network. The goal is to build the multivariate distribution for the number $\mathbf{N}_t^{(\mathrm{net})}$ of individuals detected by the network conditional upon the data coming from the preceding modules.

The procedure follows the description above. We first build the categorical variables $\mathbf{T}_{dt}$ and use Monte Carlo to generate by convolution many instances of $\mathbf{N}_t^{(\mathrm{net})}$, providing an empirical computation of the probability distribution. The result is depicted in figure 6.2.

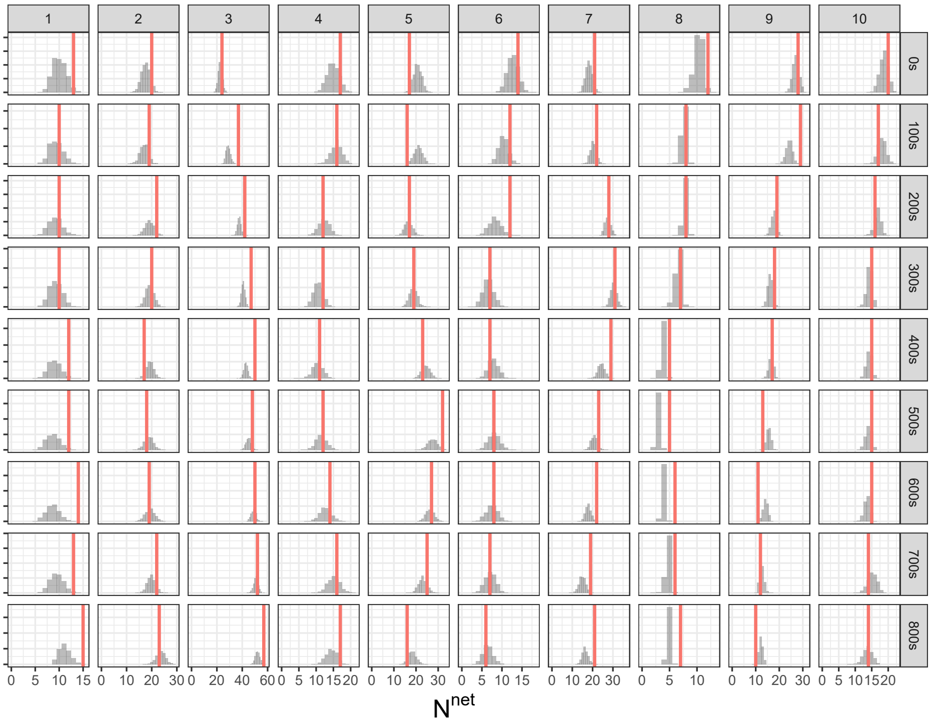

Figure 6.2: Marginals of the Poisson-multinomial distribution for $\mathbf{N}_t^{\mathrm{net}}$. True values in red.

Once we have an empirical approximation to the Poisson-multinomial distribution, we can produce a point estimation (e.g. mean) for the number of detected individuals as well as an accuracy measurement (e.g. standard deviation) and compare these estimates with the simulated ground truth.

By and large, the estimates behave fairly well, with some exceptions. Regions 3 and 8 show some noticeable departure from true values, as well as partially region 7. We have identified different factors potentially affecting this estimation procedure:

- If the duplicity probabilities p_d are pathologically computed, this will affect the estimation of individuals detected by the network. Currently, we have very little number of individuals and these probabilities are computed with high accuracy. More complex scenarios need to be analysed.

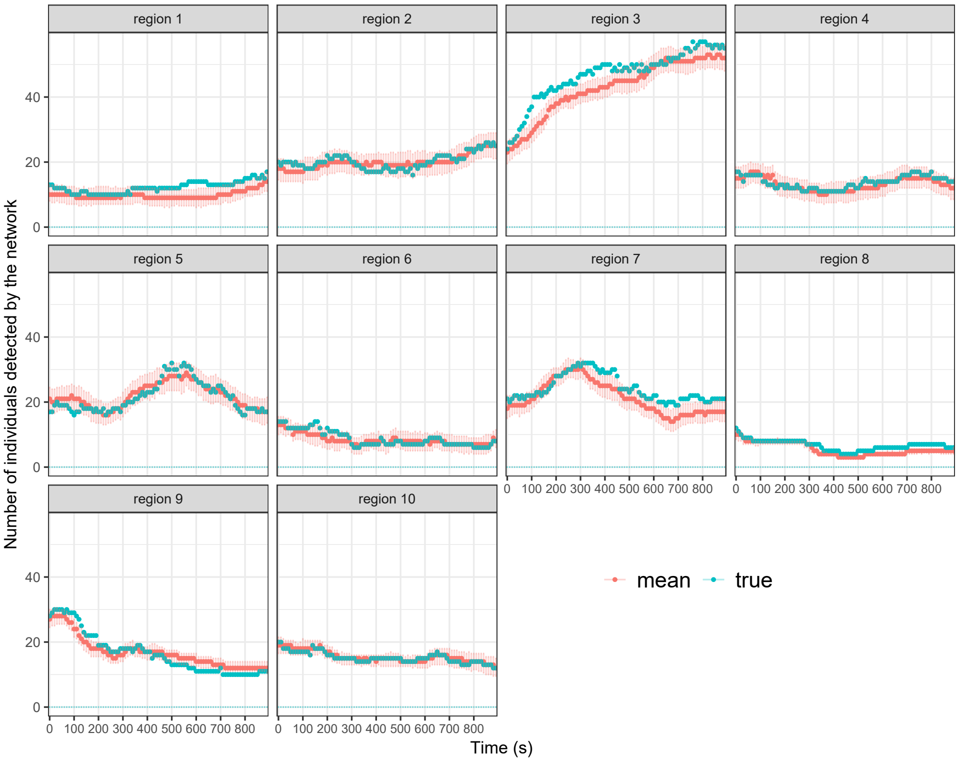

Figure 6.3: Comparison of point estimates (mean) with the simulated ground truth. Error bars represent equal-tailed credible intervals ($\alpha = 0.95$).

- Regions 3 and 8 show a higher variation in the radius of dispersion rd_d associated with the posterior location probabilities (see section 3.4). See figure 6.4. However, the true number of individuals per region and time period is too low to reach definitive conclusions (an estimation error of $1-2$ individuals in a population count of less than 10 individuals is high in relative terms, but it cannot be lower). See figure 6.5. Except for region 3, the absolute error ($|\hat{N}_r^{\text{net}} - N_r^{\text{true}}|$) is more or less constant ($1, 2$ individuals), whereas for region 3 the general trend is similar but we see more variation.

As of this writing, further research and more simulations with different scenarios, network configurations, and populations of devices and of individuals need to be conducted. Nonetheless, within the numeric range of true values investigated so far, the estimates are fairly accurate, thus providing a deduplication and aggregation method for the individuals detected by the network (subscribers).

6.3. Probability distribution for the number of detected individuals moving between tiles and regions

We close this chapter with a generalization of the above procedure to construct the probability distribution for the number of detected individuals moving from tile i to tile j in the time interval $(t-1, t)$. Although the proposal generalizes the above method, this is still under research because the preliminary analysis is showing that the level of accuracy is worst with conditional probabilities than with marginal probabilities (see below).

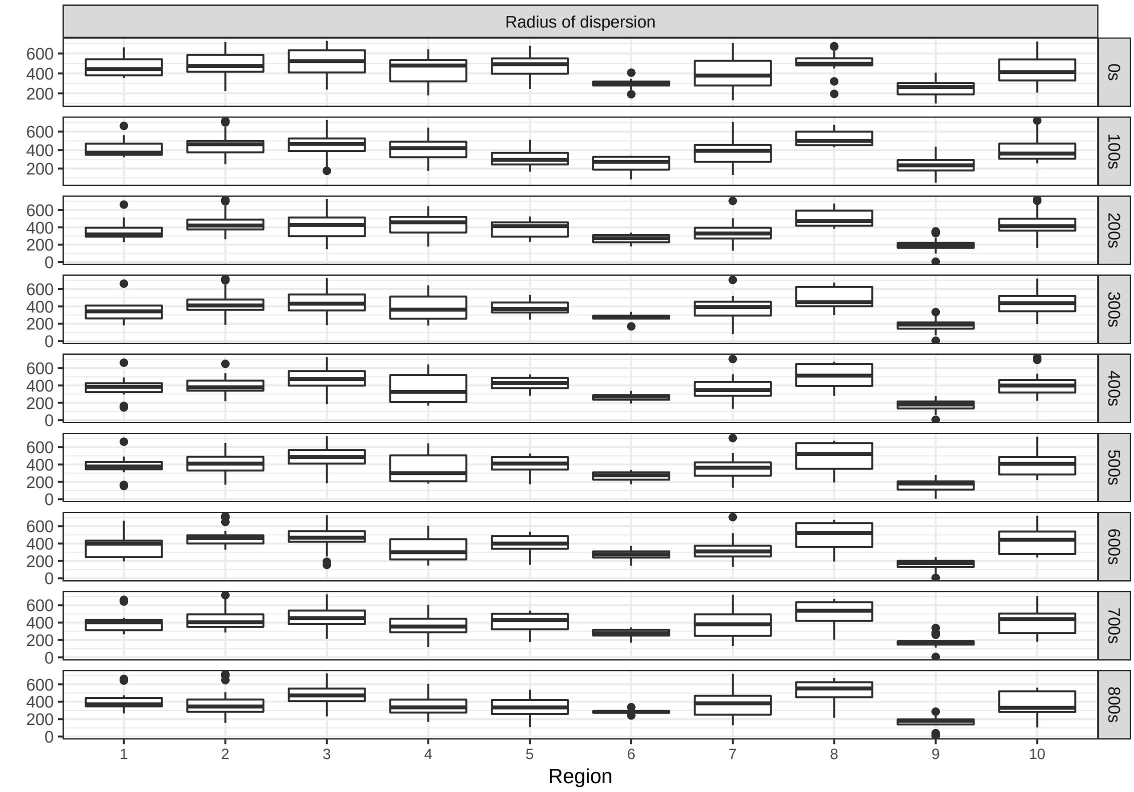

Figure 6.4: Distribution of radius of dispersion associated to the posterior location probabilities per region.

The reasoning is completely similar to that of section 6.2. We begin by defining matrices $E_{ij}^{(1)} = E_{ij}$ and $E_{ij}^{(2)} = \frac{1}{2} \cdot E_{ij}$, where E_{ij} are the Weyl matrices in $\mathbb{R}^{N_T} \times \mathbb{R}^{N_T}$. Next, we define the matrix random variable $E_{dt} \in \{E_{ij}^{(1)}, E_{ij}^{(2)}\}_{i,j=1\dots,N_T}$ with probability mass function given by

$$\mathbb{P}\left(E_{dt} = E_{ij}^{(1)}\right) = \gamma_{d(j|i)t} \times (1 - p_d), \tag{6.4a}$$

$$\mathbb{P}\left(E_{dt} = E_{ij}^{(2)}\right) = \gamma_{d(j|i)t} \times p_d, \tag{6.4b}$$

where $\gamma_{d(j|i)t}$ stands for the conditional location probability $\gamma_{d(j|i)t} \equiv \frac{\gamma_{dji,t}}{\gamma_{dit-1}}$ (see chapter 3. Notice that, although matricial, this is still a categorical or multinoulli random variable. Then, we can define the transition matrix of counts of individuals detected by the network by

$$\mathbf{N}_t^{(\text{net})} = \sum_{d=1}^{D} E_{dt}, \tag{6.5}$$

which, as before, distributes according to a multinomial-Poisson distribution. Again, we shall use Monte Carlo techniques to deal with it. Notice that $\mathbf{N}_t^{(\text{net})}$ is indeed an origin-destination matrix.

Should the transition of interest be between regions, then we shall use the aggregated conditional probabilities

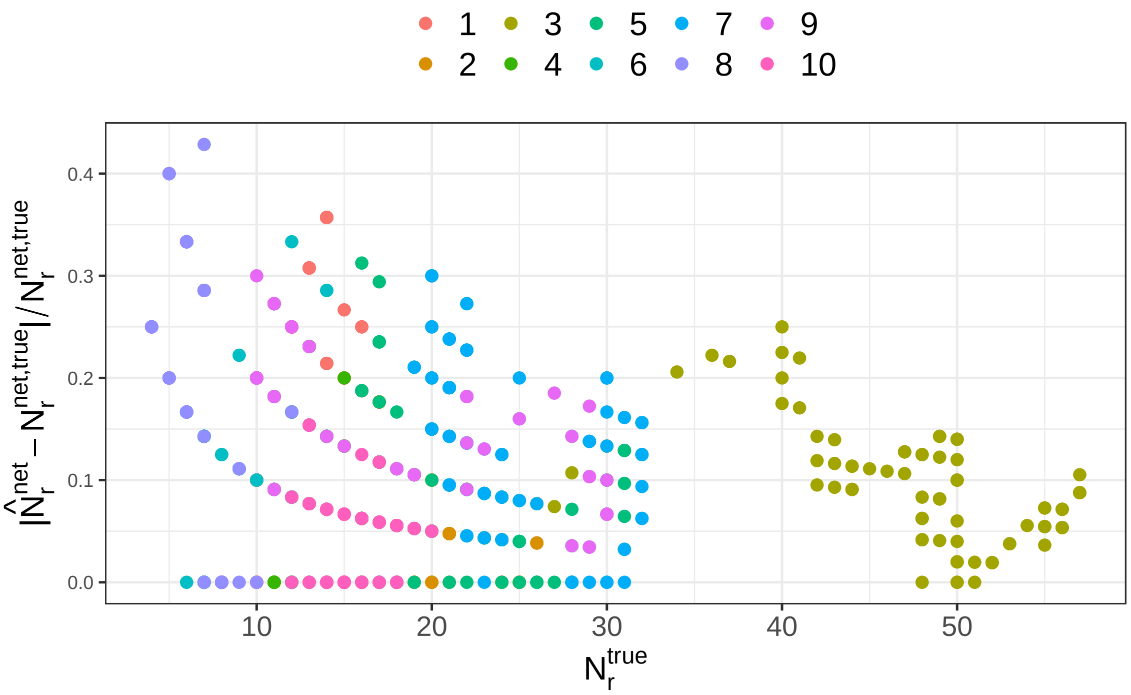

Figure 6.5: Absolute estimation error per region vs. true values.

$$\bar{\gamma}_{d(r|s)t} = \sum_{j \in \mathcal{I}_r} \sum_{i \in \mathcal{I}_s} \gamma_{dt(j|i)}. \tag{6.6}$$

The matricial random variable $\mathbf{N}_t^{(\text{net})}$ will be used later on to estimate the aggregated transition estimates of individuals in the target population.

7

Inference

This chapter deals with the final process step connecting the probability distribution of the number of individuals detected by the network of a given MNO with the number of individuals in the target population. Since in our case we are not conducting statistical filtering we are indeed estimating the present population. We shall be following the approach already initiated in the ESSnet on Big Data I (WP5.3, 2018), in which we built a hierarchical model inspired by ecological sampling to estimate species abundance (Royle and Dorazio, 2014). We introduce some improvements, mainly the multivariate nature of the problem, a clear distinction between the observation process and the system process and a neat connection with preceding modules. Also, we make an effort in showing how to build the hierarchy according to the assumptions and the available data.

In section 7.1 we provide a general description of the approach. In section 7.2 we show how to introduce the uncertainty in the observation process through the use of hierarchical models. In section 7.3 we show how to introduce the state process together with its uncertainties. In section 7.4 we introduce the dynamical element of the estimation problem. In section 7.5 we include some illustrative examples with synthetic data produced by the network event data simulator. We put off the mathematical content to appendix B.

7.1. The approach

The bottom line of our approach is already described in WP5 of the ESSnet on Big Data I (WP5.3, 2018), which is inspired by the approach to estimate the species abundance in Ecology (Royle and Dorazio, 2014). The use of hierarchical models (Gelman et al., 2013) stands as a versatile tool not only to produce final point estimates, but also (and especially) to account for underlying uncertainties and to combine different data sources. In the proposed inference model the observed data, the target process, and its underlying parameters must be given a joint probability distribution $\mathbb{P}\left(\text{data}, \text{process}, \text{parameters}\right)$. The hierarchical model allows us to decompose this joint distribution as (Royle and Dorazio, 2014)

$$\mathbb{P}\left(\text{data}, \text{process}, \text{parameters}\right) = \mathbb{P}\left(\text{data}\middle|\text{process}, \text{parameters}\right) \cdot \mathbb{P}\left(\text{process}\middle|\text{parameters}\right) \cdot \mathbb{P}\left(\text{parameters}\right),$$

which can be conveniently interpreted as the combination of three components:

- an observation process (data given the underlying dynamical process driving the target variable N);
- a state process (the underlying process for the target variable N modelled in terms of its parameters);

- assumptions about the parameters driving not only the state process but possibly also the observation process.

Graphically this can be represented by figure 7.1, where $\mathbf{Z}_{obs}$ and $\mathbf{Z}_{sys}$ stand for the elements of the auxiliary information related to the observation process and the system state process, respectively, In the next sections we shall clearly show how $\mathbf{Z}_{obs}$ basically amounts to the penetration rates, i.e. to the telco market information in a generic way, and $\mathbf{Z}_{sys}$ refers to the register-based population density. In shaded gray color, we show those variables which are known, i.e. for which we have data.

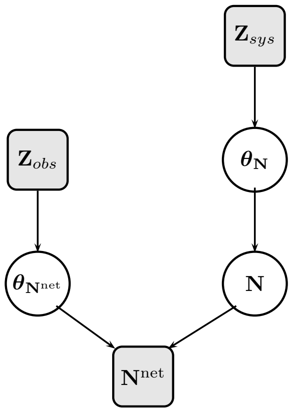

Figure 7.1: Generic structure of the hierarchical model for inference.

The bottom line of the whole approach can be introduced straightforwardly by considering one region r at a given time instant (univariate approach). We shall concentrate first on the observation process. 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

$$N_r^{\text{net}} \simeq \text{Bin}\left(N_r, p_r\right), \tag{7.1}$$

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 of an individual of being a subscriber of the given mobile telecommunication network. This assumption will be further discussed below. As a first approximation, we may think of p_r as a probability related to the penetration rate or the market share of the MNO in region r. For the time being, we shall consider this as an external parameter taken from the national telecommunication regulator.

Following Thompson (2012), under model (7.1), we can straightforwardly write $\mathbb{E}\left[N_r^{\text{net}}\right] = N_r \cdot p_r$, so that we can suggest a first naive point estimator given by

$$\widehat{N}_r^{\text{naive}} = \frac{\widehat{N}_r^{\text{net}}}{p_r}, \tag{7.2a}$$

where $\widehat{N}_r^{\text{net}}$ stands for a point estimator for the number of individuals detected by the network coming from preceding modules. Some proposals follow this approach but using number of devices instead of number of individuals detected by the network and thus including different correction factors for the detection probability p_r. Estimates derived thereof can serve as a first approximation. Notice that to put (7.2a) in a rigorous footing linked to the probability distribution (7.1), it is necessary that $N_r \geq N_r^{\text{net}}$. This would have been possibly violated if we had used number of devices instead of number of individuals detected by the network.

To account for the uncertainty of this estimation, we can reason about the variance of $\widehat{N}_r^{\text{naive}}$ in the same line:

$$\mathbb{V}\left[\widehat{N}_r^{\text{naive}}\right] = \frac{\mathbb{V}\left[\widehat{N}_r^{\text{net}}\right]}{p_r} = N_r \cdot \frac{1-p_r}{p_r},$$

which can be estimated unbiasedly by

$$\widehat{\mathbb{V}}\left[\widehat{N}_r^{\text{naive}}\right] = N_r^{\text{Nnet}} \cdot \frac{1-p_r}{p_r^2}. \tag{7.2b}$$

Notice, however, that the uncertainty in the detection probability p_r is not accounted for. To account for this uncertainty and to introduce a mechanism to consider the underlying assumptions in a systematic, we make use of a hierarchy of models, hence the hierarchical approach.

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.

$$\mathbb{P}\left(N_r|N_r^{\text{net}}\right) \propto \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \binom{N_r}{N_r^{\text{net}}}(1-p_r)^{N_r - N_r^{\text{net}}} p^{N_r^{\text{net}}} & \text{if } N_r \geq N_r^{\text{net}}, \end{cases}$$

which, after normalization, provides the normalized posterior

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

where $\text{negbin}\,(k;p,r) \equiv \binom{k+r-1}{k}p^k(1-p)^r$ 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 (the posterior mode) we can provide as point estimator

$$\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{N_r^{\text{net}}}{p_r} - N_r^{\text{net}} \right\rfloor, \tag{7.2c}$$

which is a more rigorous version of the naive estimator (7.2a). 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[1]

$$ppRB = \frac{\mathbb{E}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}}|\widehat{N}_r^{\text{net}}\right]}{\widehat{N}_r^{\text{net}}} \tag{7.3a}$$

$$ppRV = \frac{\mathbb{V}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}}|\widehat{N}_r^{\text{net}}\right]}{(\widehat{N}_r^{\text{net}})^2} \tag{7.3b}$$

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

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 $\mathbb{P}\left(N_r^{\text{net}}|\mathbf{E}\right)$ ($\mathbf{E}$ denoting the network event information) and study their mean values and dispersion. In all this reasoning, the detection probabilities p_r are considered as fixed external parameters without errors (no uncertainty), an assumption which we shall drop in the next section.

7.2. Introducing uncertainty in the observation process

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 $\mathbb{P}\left(N_r^{\text{net}}|\mathbf{E}\right)$. 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

$$\mathbb{P}\left(N_r, N_r^{\text{net}}|\mathbf{E}\right) \propto \text{negbin}\left(N_r - N_r^{\text{net}}; 1 - p_r, N_r^{\text{net}} + 1\right) \times \mathbb{P}\left(N_r^{\text{net}}|\mathbf{E}\right). \tag{7.4}$$

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 $p_{kr} = p_r + \text{noise}$. At a first stage, we propose to implement this idea by modeling $p_r \simeq \text{Beta}\left(\alpha_r, \beta_r\right)$ and choosing the hyperparameters α_r and β_r according to the penetration rates or the market shares and the official population data.

Let us denote by P_r^{net} the penetration rate at region r of the network, i.e. $P_r^{\text{net}} = \frac{N_r^{(\text{devices})}}{N_r}$. 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 computed in chapter 4 and of the posterior location probabilities $\bar{\gamma}_{dr}$ in region r for each device d. Notice that we have dropped out the time subscript for ease of notation, since we are currently focusing on a given time instant. We define

$$\Omega_r^{(1)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr} \cdot (1 - p_d)}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}, \tag{7.5a}$$

$$\Omega_r^{(2)} = \frac{\sum_{d=1}^{D} \bar{\gamma}_{dr} \cdot p_d}{\sum_{d=1}^{D} \bar{\gamma}_{dr}}. \tag{7.5b}$$

The deduplicated penetration rate is defined as

$$\tilde{P}_r^{\text{net}} = \left(\Omega_r^{(1)} + \frac{\Omega_r^{(2)}}{2}\right) \cdot P_r^{\text{net}}. \tag{7.5c}$$

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 $P_r^{\text{rm}} = \frac{16}{15}$. The devices are assumed to be neatly detected by the HMM (i.e. $\tilde{\gamma}_{dr} = 1 - O(\epsilon)$) and duplicities are also inferred correctly ($p_d = O(\epsilon)$ for $d^{(1)}$ and $p_d = 1 - O(\epsilon)$ 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 $\bar{P}_r^{\text{net}} = \frac{13}{15} + O(\epsilon)$, 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.

Let us now denote by N_r^{reg} the population of region r according to an external population register. Then, we fix

$$\begin{align}
\alpha_r + \beta_r &= N_r^{\text{reg}}, \tag{7.6a}\\
\frac{\alpha_r}{\alpha_r + \beta_r} &= \tilde{P}_r^{\text{net}}, \tag{7.6b}
\end{align}$$

which immediately implies that

$$\begin{align}
\alpha_r &= \tilde{P}_r^{\text{net}} \cdot N_r^{\text{reg}}, \tag{7.7a}\\
\beta_r &= \left(1 - \tilde{P}_r^{\text{net}}\right) \cdot N_r^{\text{reg}}. \tag{7.7b}
\end{align}$$

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 chapter 4, to estimate population counts N_r using network-based population counts N_r^{net} we need some extra parameter(s). Otherwise, it is impossible. Detection probabilities are indeed these extra 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 NSIs.

- 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:

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

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 (7.7) we get as a MAP estimate:

$$\begin{aligned}\widehat{N}^{\text{MAP}} &= N_r^{\text{net}} + \left\lceil \frac{N_r^{\text{reg}}}{\tilde{P}_r} - N_r^{\text{net}} - \frac{N_r^{\text{net}}}{N_r^{\text{reg}}}\frac{1}{\tilde{P}_r^{\text{net}}} \right\rceil, \qquad (7.8) \\ &\approx N_r^{\text{net}} + \left\lceil \frac{N_r^{\text{reg}}}{\tilde{P}_r} - N_r^{\text{net}} - 1 \right\rceil\end{aligned}$$

which is very similar to all estimates (7.2) 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 (7.8) 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

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

which is similar to preceding expressions.

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

$$\begin{aligned}
N_r^{\text{net}} &\simeq \text{Bin}\left(N_r, p_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.9a)}\\
p_r &\simeq \text{Beta}\left(\alpha_r, \beta_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.9b)}\\
\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r+\beta_r}\right), \alpha_r+\beta_r\right) &\simeq \text{N}\left(\mu_{\beta r}(\beta_0, \beta_1; \bar{P}_r^{\text{net}}), \tau_\beta^2\right) \times \text{Gamma}\left(1+\xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.9c)}\\
\left(\log\beta_0, \beta_1, \tau_\beta^2, \xi\right) &\simeq \text{f}_\beta\left(\log\beta_0, \beta_1, \tau_\beta^2\right) \times \text{f}_\xi(\xi), && \text{(7.9d)}
\end{aligned}$$

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

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.

Under the specifications (7.9), after some tedious computations, we can show that the multivariate distribution for the number of individuals $\mathbf{N}$ in the target population conditional on the number of individuals $\mathbf{N}^{\text{net}}$ detected by the network is given by

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto \int_{\mathbb{R}^R} d^R\mathbf{y}\, \omega_{\text{obs}}\left(\mathbf{y}; \bar{\mathbf{P}}^{\text{net}}\right) \prod_{r=1}^{R} \frac{\text{negbin}\left(N_r - N_r^{\text{net}}; 1-p(y_r), N_r^{\text{net}}+1\right)}{p(y_r)}, \qquad \text{(7.10)}$$

where

- $\text{negbin}(k; p, r)$ stands for the probability mass function of the negative binomial distribution for variable k and parameters p and r;
- $p(y_r) \equiv \frac{e^{y_r}}{1+e^{y_r}}$;
- $\omega_{\text{obs}}(\mathbf{y}; \mathbf{P}^{\text{net}}) = \int_{\Omega_\beta} \text{d}\log\beta_0 \text{d}\beta_1 \text{d}\tau_\beta^2\, \text{f}_\beta\left(\log\beta_0, \beta_1, \tau_\beta^2\right)\, \text{n}\left(\mathbf{y}; \boldsymbol{\mu}(\beta_0, \beta_1; \bar{\mathbf{P}}^{\text{net}}), \boldsymbol{\Sigma}_\beta\right)$ where

- $\mathrm{n}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ stands for the probability density function of the multivariate normal distribution for variable $\mathbf{x}$ and mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.
- $\mu_{\beta r}(\beta_0,\beta_1;\bar{P}_r^{\text{net}}) = \log\left(\beta_0\left[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}\right]^{\beta_1}\right)$.
- $\boldsymbol{\Sigma}_\beta = \tau_\beta^2\,\mathbb{I}_{R\times R}$.

In this derivation, again the assumption $\alpha_r,\beta_r \gg 1$ is taken for granted.

In rigour, we should have included $\mathbf{P}^{\text{net}}$ as conditioning random variables together with $\mathbf{N}^{\text{net}}$, but we have opted to keep the notation as simple as possible. To have an expression which can be computed we need to further specify the prior f_β. As a first example, let us consider $\beta_0 = \beta_1 = 1$ and $\tau_\beta^2 \to 0^+$. This amounts to having certainty about the values of α_r and β_r, as above, so that $\omega_{\text{obs}}(\mathbf{y};\bar{\mathbf{P}}^{\text{net}}) = \prod_{r=1}^R \delta(y_y - \log\bar{P}_r^{\text{net}})$, where $\delta(\cdot)$ stands for the Dirac delta function. Upon normalization expression (7.10) reduces to

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; 1-\tilde{P}_r^{\text{net}}, N_r^{\text{net}}+1\right). \tag{7.11}$$

The marginal distribution for region r reduces to (7.9), which was also obtained above through a direct reasoning.

If we choose another prior $\mathrm{f}_\beta\left(\log\beta_0,\beta_1,\tau_\beta^2\right)$, then we need to compute numerically:

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto \int \mathrm{d}\mathbf{y}\;\omega_{\text{obs}}(\mathbf{y};\tilde{\mathbf{P}}^{\text{net}})\cdot h(\mathbf{N};\mathbf{y}_k,\mathbf{N}^{\text{net}}), \tag{7.12}$$

where

- $h(\mathbf{N};\mathbf{y}_k,\mathbf{N}^{\text{net}}) = \prod_{r=1}^R \frac{\text{negbin}\left(N_r-N_r^{\text{net}};1-p(y_r),N_r^{\text{net}}+1\right)}{p(y_r)}$.
- The vector variable $\mathbf{y}\in\mathbb{R}^R$ is a multidimensional variate generated according to the continuously compound multivariate distribution

  $$\int \mathrm{d}\log\beta_0\mathrm{d}\beta_1\mathrm{d}\tau_\beta^2\; N(\boldsymbol{\mu}(\beta_0,\beta_1;\bar{\mathbf{P}}^{\text{net}}),\boldsymbol{\Sigma}_\beta|\log\beta_0,\beta_1,\tau_\beta^2)\times \mathrm{f}_\beta(\log\beta_0,\beta_1,\tau_\beta^2),$$

  with $\mu_{\beta r} = \log\left(\beta_0\left[\frac{\bar{P}_r^{\text{net}}}{1-\bar{P}_r^{\text{net}}}\right]^{\beta_1}\right)$ and $\boldsymbol{\Sigma}_\beta = \tau_\beta^2\,\mathbb{I}_{R\times R}$.

7.3. Introducing the state process

Now we introduce the state process. The system is a human population and we make a common modelling hypothesis to represent the number of individuals N_r in region r of the target population as a Poisson-distributed random variable in terms of the population density, i.e.

$$N_r \simeq \text{Poisson}\left(A_r\sigma_r\right), \tag{7.13}$$

where σ_r stands for the population density of region r and A_r denotes the area of region r. We choose to model N_r in terms of the population density to make an auxiliary usage of some results already found in the literature (Deville et al., 2014), as we shall see below.

Similarly to the observation process, we introduce the following hierarchy:

$$
\begin{aligned}
N_r &\simeq \text{Poisson}\left(A_r\sigma_r\right), \quad \text{for all } r = 1, \ldots, R, && (7.14a)\\
\sigma_r &\simeq \text{Gamma}\left(1+\zeta_r, \theta_r\right), \quad \text{for all } r = 1, \ldots, R, && (7.14b)
\end{aligned}
$$

where the hyperparameters will express the uncertainty about the register-based population. Notice that the modes of the gamma distributions are at $\sigma_r = \zeta_r \cdot \theta_r$ and the variances are given by $\mathbb{V}\left(\sigma_r\right) = (\zeta_r * 1) \cdot \theta_r^2$. We shall parametrise these gamma distributions in terms of the register-based population densities σ_r^{reg} as

$$
\begin{aligned}
\zeta_r \cdot \theta_r &= \sigma_r^{\text{reg}} + \Delta\sigma_r,\\
\sqrt{(\zeta_r+1)\cdot\theta_r^2} &= \epsilon_r \cdot \sigma_r^{\text{reg}},
\end{aligned}
$$

where ϵ_r can be viewed as the coefficient of variation for σ_r^{reg} and $\Delta\sigma_r$ can be interpreted as the bias for σ_r^{reg}. This parametrization implies that

$$
\begin{aligned}
\theta_r(\Delta\sigma_r, \epsilon_r) &= \frac{\sigma_r^{\text{reg}}}{2}\left(1+\frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}\right)\left[\sqrt{1+\left(\frac{2\epsilon_r}{1+\frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}}\right)^2}-1\right],\\
\zeta_r(\Delta\sigma_r, \epsilon_r) &= \frac{2}{\sqrt{1+\left(\frac{2\epsilon_r}{1+\frac{\Delta\sigma_r}{\sigma_r^{\text{reg}}}}\right)^2}-1}. && (7.15)
\end{aligned}
$$

Under assumptions (7.14) and assuming $\alpha_r, \beta_r \gg 1$, as above, we get

$$
\mathbf{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r+\beta_r}\cdot Q(\theta_r), N_r^{\text{net}}+1+\zeta_r\right) \tag{7.16}
$$

where $Q(\theta_r) \equiv \frac{A_r\theta_r}{1+A_r\theta_r}$. The interpretation of this hierarchy is also simple. It is just a Poisson-gamma model in which the gamma parameters have been chosen so that we account for the uncertainty in the register-based population figures N_r^{reg}.

The MAP estimators derived from the distribution (7.16) under identities (7.7) are

$$
\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{(1-\tilde{P}_r^{\text{net}})\cdot Q(\theta_r)}{1-(1-\tilde{P}_r^{\text{net}})\cdot Q(\theta_r)}\left(N_r^{\text{net}}+\zeta_r\right)\right\rfloor. \tag{7.17}
$$

Accuracy indicators such as posterior variance or credible intervals are to be computed from the distribution (7.16).

Expression (7.16) contains the uncertainty of both the observation and the state processes. In the limiting case $\epsilon_r \to 0$, i.e. having certainty about the state process, and with equations (7.7), we have the Poisson limit of the negative binomial distribution (see section B.3) so that

$$
\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{poisson}\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r^{\text{net}})\cdot A_r\sigma_r^{\text{reg}}\right). \tag{7.18}
$$

The MAP estimator is trivially $\hat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor(1-\bar{P}_r)A_r\sigma_r^{\text{reg}}\right\rfloor$, which can be readily read as the sum of the detected individuals by the network and the individuals not detected by the network accounted for by the population register.

On the contrary, when $\epsilon_r \to \infty$ (i.e. having no information at all about the state process), we have $Q(\theta_r) = 1$ and $\zeta_r = 0$ so that

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) = \prod_{r=1}^{R} \text{negbin}\left(N_r - N_r^{\text{net}}; 1 - \bar{P}_r, N_r^{\text{net}} + 1\right), \tag{7.19}$$

which is the same expression as (7.11), as expected, since having no information about the state process is equivalent to having only the observation process.

We can also introduce more levels in the hierarchy, as with the observation process:

$$\begin{aligned}
N_r &\simeq \text{Poisson}\left(A_r \sigma_r\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.20a)}\\
\sigma_r &\simeq \text{Gamma}\left(\zeta + 1, \frac{e^{\theta_r}}{\zeta}\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.20b)}\\
\theta_r &\simeq N\left(\log\left(\gamma_0\left[\sigma_r^{\text{reg}}\right]^{\gamma_1}\right), \tau_\gamma^2\right), \quad \text{for all } r = 1, \ldots, R, && \text{(7.20c)}\\
\left(\log\gamma_0, \gamma_1, \tau_\gamma^2, \zeta\right) &\simeq \text{f}_\gamma\left(\log\gamma_0, \gamma_1, \tau_\gamma^2\right) \times \text{f}_\zeta(\zeta), && \text{(7.20d)}
\end{aligned}$$

The interpretation of this hierarchy is also simple. It is just a Poisson-gamma model in which the gamma parameters have been chosen so that the mode is at $\exp(\theta_r)$ with an uncertainty degree provided by ζ. Notice that the smaller ζ, the more degree of uncertainty about the value of θ_r. The mode is correlated with the register-based population density σ_r^{net} through a linear regression.

Under the specifications (7.20) and also including the observation process, again after some tedious computation, we can show that the multivariate distribution for the number of individuals $\mathbf{N}$ in the target population conditional on the number of individuals $\mathbf{N}^{\text{net}}$ detected by the network is given by

$$\begin{aligned}
\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right) \propto\; & \int_{\mathbb{R}^R} d^R\mathbf{y}\, \omega_{\text{obs}}\left(\mathbf{y}; \bar{\mathbf{P}}^{\text{net}}\right) \prod_{r=1}^{R} \frac{\text{negbin}\left(N_r - N_r^{\text{net}}; 1 - p(y_r), N_r^{\text{net}} + 1,\right)}{p(y_r)}\\
\times\; & \int_{\mathbb{R}^R} d^R\mathbf{z}\, \omega_{\text{state}}\left(\mathbf{z}; \boldsymbol{\sigma}^{\text{reg}}\right) \prod_{r=1}^{R} \text{negbin}\left(N_r; q\left(\frac{A_r e^{z_r}}{\zeta}\right), 1 + \zeta\right),
\end{aligned} \tag{7.21}$$

where

- $\text{negbin}(k; p, r)$ stands for the probability mass function of the negative binomial distribution for variable k and parameters p and r;
- $q\left(\frac{A_r e^{z_r}}{\zeta}\right) \equiv \frac{\frac{A_r e^{z_r}}{\zeta}}{1+\frac{A_r e^{z_r}}{\zeta}}$;
- $\omega_{\text{state}}(\mathbf{z}; \boldsymbol{\sigma}^{\text{reg}}) = \int_{\Omega_{\gamma,\zeta}} \text{d}\log\gamma_0 \text{d}\gamma_1 \text{d}\tau_\gamma^2 \text{d}\zeta\, \text{f}_\gamma\left(\log\gamma_0, \gamma_1, \tau_\gamma^2\right) \times \text{f}_\zeta(\zeta)\, n\left(\mathbf{z}; \boldsymbol{\mu}(\gamma_0, \gamma_1; \boldsymbol{\sigma}^{\text{net}}), \boldsymbol{\Sigma}_\gamma\right)$ with
  - $\text{n}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the probability density function of the multivariate normal distribution for variable $\mathbf{x}$ and mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$.
  - $\mu_{\gamma r}(\gamma_0, \gamma_1; \sigma_r^{\text{reg}}) = \log\left(\gamma_0\left[\sigma_r^{\text{reg}}\right]^{\gamma_1}\right)$.

- $\mathbf{\Sigma}_\gamma = \tau_\gamma^2\, \mathbb{I}_{R\times R}$.

Notice how this expression reveals both factors arising from the observation and the state processes, respectively. When $\beta_0, \beta_1, \gamma_0, \gamma_1 \rightarrow 1$, $\zeta \rightarrow \zeta^*$, and $\tau_\beta^2, \tau_\gamma^2 \rightarrow 0^+$ (i.e. when having fully accurate information about the parameters α_r, β_r and θ_r), we have $\omega_\beta(\mathbf{y}) = \delta(\mathbf{y} - \boldsymbol{\mu}_\beta)$ and $\omega_\beta(\mathbf{z}) = \delta(\mathbf{z} - \boldsymbol{\mu}_\gamma)$ so that after normalization equation (7.21) reduces to

$$\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\mathrm{net}}\right) = \prod_{r=1}^{R} \mathrm{negbin}\left(N_r - N_r^{\mathrm{net}}; (1-\bar{P}_r)\cdot Q_r(\zeta^*), N_r^{\mathrm{net}} + \zeta^* + 1\right), \tag{7.22}$$

where we have denoted $Q_r(\zeta) \equiv q(\frac{A_r \sigma_r^{\mathrm{reg}}}{\zeta})$, which is indeed again equation (7.16). The general computation of expression (7.21) can be undertaken using Monte Carlo techniques, as before. We can write

$$\mathbb{P}(\mathbf{N}|\mathbf{N}^{\mathrm{net}}) \quad = \quad \frac{1}{K_1K_2}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} h(\mathbf{y}_{k_1}) g(\mathbf{z}_{k_2}, \zeta_{k_2}), \tag{7.23}$$

where

- $h(\mathbf{y}) = \prod_{r=1}^{R} \frac{\mathrm{negbin}\left(N_r - N_r^{\mathrm{net}}; 1-p(y_r), N_r^{\mathrm{net}}+1\right)}{p(y_r)}$.
- $g(\mathbf{z},\zeta) \equiv \prod_{r=1}^{R} \mathrm{negbin}\left(N_r; q(\frac{A_r e^{z_r}}{\zeta}), 1+\zeta\right)$.
- $\mathbf{y}_k$ are random variates generated by the distribution

  $$\int \mathrm{d}\log\beta_0 \mathrm{d}\beta_1 \mathrm{d}\tau_\beta^2\; N(\boldsymbol{\mu}(\beta_0, \beta_1; \bar{\mathbf{P}}^{\mathrm{net}}), \mathbf{\Sigma}_\beta | \log\beta_0, \beta_1, \tau_\beta^2) \times \mathrm{f}_\beta(\log\beta_0, \beta_1, \tau_\beta^2),$$

  with $\mu_{\beta r} = \log\left(\beta_0 \left[\frac{\bar{P}_r^{\mathrm{net}}}{1-\bar{P}_r^{\mathrm{net}}}\right]^{\beta_1}\right)$ and $\mathbf{\Sigma}_\beta = \tau_\beta^2\, \mathbb{I}_{R\times R}$.
- $(\mathbf{z}_k, \zeta_k)$ are random variates generated by the distribution

  $$\mathrm{f}_\zeta(\zeta) \times \int \mathrm{dlog}\gamma_0 \mathrm{d}\gamma_1 \mathrm{d}\tau_\gamma^2\; N(\boldsymbol{\mu}(\gamma_0, \gamma_1; \boldsymbol{\sigma}^{\mathrm{reg}}), \mathbf{\Sigma}_\gamma | \log\gamma_0, \gamma_1, \tau_\gamma^2) \times \mathrm{f}_\gamma(\log\gamma_0, \gamma_1, \tau_\gamma^2),$$

  with $\mu_{\gamma r} = \log\left(\gamma_0 \left[\sigma_r^{\mathrm{reg}}\right]^{\beta_1}\right)$ and $\mathbf{\Sigma}_\gamma = \tau_\gamma^2\, \mathbb{I}_{R\times R}$.

7.4. The dynamical aspects

So far, we have produce estimates at a given time instant t_0, which we shall consider this as an initial time when individuals are assumed to be physically at the georeference of the register-based population (i.e. reported home locations). Now we show how to find the network-based estimates N_{rt} for times $t > t_0$. Currently, we consider only **closed** populations, i.e. neither individuals nor devices enter into or leave the territory under analysis along the whole time period. This important restriction is posed to introduce progressively the different methods in order to get a thorough assessment of every single aspect of the procedure. It will have to be lifted in future work (e.g. considering sink and source tiles in the reference grid).

Our reasoning tries to introduce as less assumptions as possible. Thus, we begin by considering a balance equation. Let us denote by $N_{(r|s)t}$ the number of individuals moving from region s to region r in the time interval $(t-1, t)$. Then, we can write

$$\begin{aligned} N_{rt} &= N_{rt-1} + \sum_{\substack{r_t=1\\ r_t\neq r}}^{N_T} N_{(r|r_t)t} - \sum_{\substack{r_t=1\\ r_t\neq r}}^{N_r} N_{(r_t|r)t} \\ &= \sum_{r_t=1}^{N_T} \tau_{(r|r_t)t} \cdot N_{r_t t-1}, \end{aligned} \tag{7.24}$$

where we have defined $\tau_{(r|s)t} = \frac{N_{(r|s)t}}{N_{st-1}}$ (0 if $N_{st-1} = 0$). Notice that $\tau_{(r|s)t}$ can be interpreted as an aggregate transition probability from region s to region r at time interval $(t-1, t)$ in the target population.

We make the assumption that individuals detected by the network move across regions in the same way as individuals in the target population. Thus, we can use $\tau^{\text{net}}_{(r|s)t} \equiv \frac{N^{\text{net}}_{(r|s)t}}{N^{\text{net}}_{st-1}}$ to model $\tau_{(r|s)t}$. In particular, as our first choice we shall postulate $\tau_{(r|s)t} = \tau^{\text{net}}_{(r|s)t}$.

Everything is thus reduced to estimate $N^{\text{net}}_{(r|s)t}$, which can be accomplished using the techniques explained in section 6.3.

Finally, we mention two points:

- 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 (7.24) to compute the probability distribution at later times $t > t_0$.
- Again, 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).

7.5. Illustrative example

We shall provide an example of the inference step using data from the network data simulator. As input for this step we need the posterior location probabilities γ_{dti}, the joint posterior location probabilities γ_{dijt}, the duplicity probabilities p_d for each device $d = 1, \ldots, D$, and a point estimation for the number of individuals N^{net}_r detected by the network in region r (because we are focusing on the estimation for the target population conditional of the number of individuals detected by the network, i.e. only on this inference module). As auxiliary information we shall use the penetration rates P^{net}_r and the register-based population N^{reg}_r in each region r. Taking advantage of the simulator, we shall compare the estimates with the true values from the simulation.

We shall focus only on the initial time estimation. The dynamical aspects depend very sensitively on the estimation of the number of individuals detected by the network moving from one region to another (see section 6.3), thus in the preliminaries tests we have conducted the problems with the conditional probabilities detected there are inherited at this stage. New scenarios with larger number of individuals and devices (thus, more computational cost) is needed.

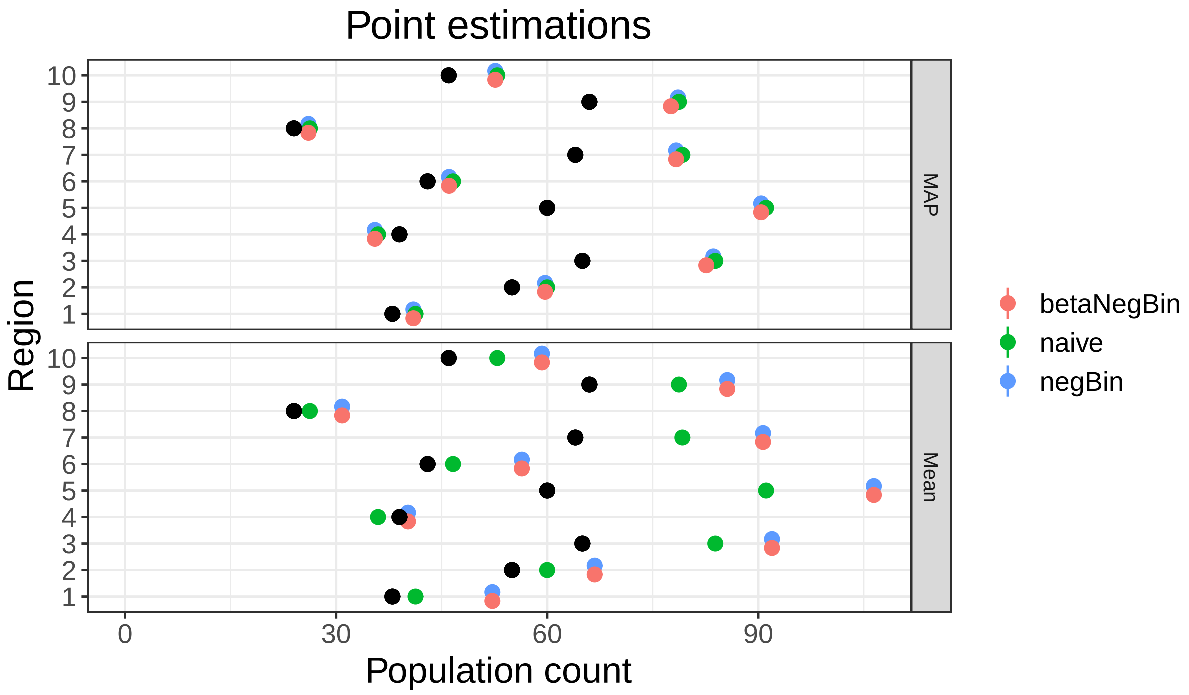

Figure 7.2: Point estimates based on naive, MAP, and posterior mean estimators. True values represented in black.

At the initial time, we shall make the assumption that both the network-based and register-based populations can be assimilated, since individuals stay physically at the geolocation collected in the population register. In this sense, it is meaningful to use the register-based population as auxiliary information. Notice, however, that due structural metadata regarding the differences between present population and residential population needs to be put in place. We assume that non-resident individuals (with foreign SIM cards) have been filtered out of the mobile network data set.

The scenario generated by the simulator is the same as that used in chapter 6. We have 218 devices and 500 individuals moving around the geographical territory divided into 10 regions. The point estimates for the number of individuals (not devices) detected by the network are taken from the aggregation module. The goal is to estimate the number of individuals N_r in each region r at the initial time t_0. The penetration rate in each region is computed as $P_r^{\text{net}} = \frac{N_r^{(\text{devices})}}{N_r}$ and the register-based population counts are denoted by N_r^{reg}.

Firstly, we shall provide point estimates using the naive estimator and MAP estimates from the posterior distributions introduced in preceding sections. When only the observation process is taken into account, we get the results depicted in figure 7.2. The estimates are fairly accurate, except for a number of regions. This is one of the reasons why we need to quantify accuracy and quality indicators are necessary.

To account for the accuracy of the estimation we shall use credible intervals, which can be computed from the posterior distributions (hence the naive estimator should be discarded). We can compute both the high-density and equal-tailed intervals (Gelman et al., 2013). This is represented in figure 7.3. Notice that all true values are inside the credible intervals, except for region 5 for negative binomial model (we shall analyse this below). We remind that the

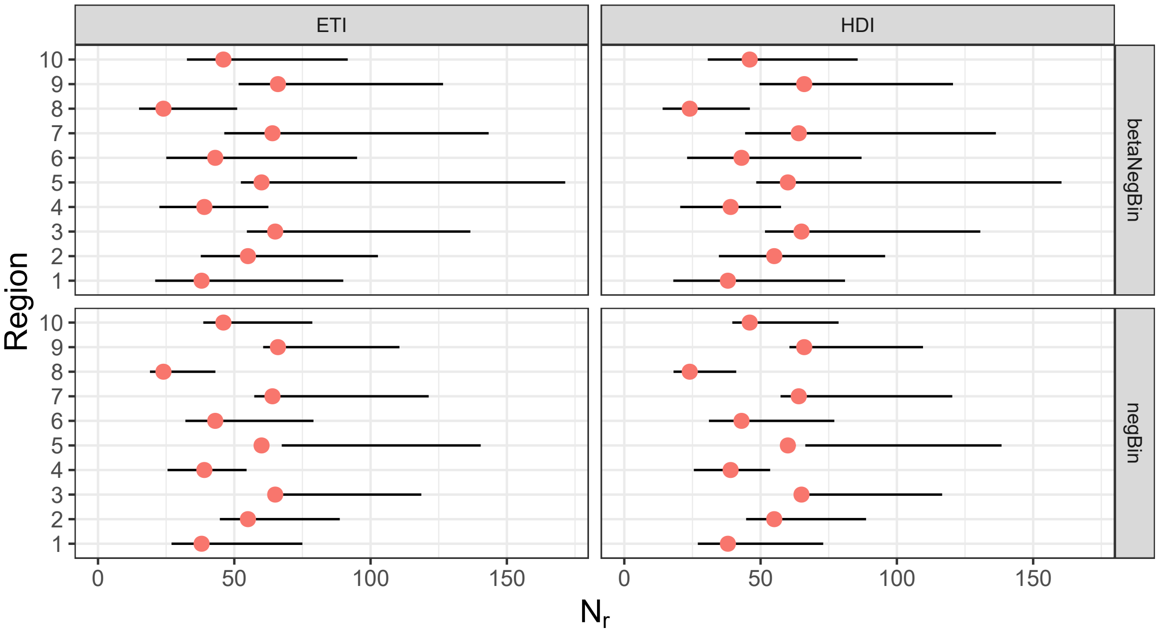

Figure 7.3: Credible intervals for the number of individuals. Confidence level = 0.95.

credible intervals can be safely interpreting as containing true values with probability 0.95 (in contraposition to frequentist confidence intervals).

To get acquainted with the posterior distributions in relation with the true values, we depict them in figure 7.4 together with the true values for each region. We can clearly see how regions 3 and 5 pose some problems. Notice that these estimations have been conducted with auxiliary information having no error, i.e. $N_r^{\text{reg}} = N_r$ and $P_r^{\text{net}} = \frac{N_r^{\text{devices}}}{N_r}$. Next, we conduct two analyses. We shall relate the errors in the estimates with known parameters from the simulation and we shall investigate how estimates change when noise is introduced in the auxiliary data.

To identify and understand the factors behind the quality of the estimates, firstly we investigate the relationship between the relative error in the target population estimates and the relative error in the number of individuals detected by the network, which is taken as an input from the preceding module. This relation is represented in figure 7.5. We can clearly see how the errors in the target population estimates are correlated with errors in the numbers of individuals detected by the network. The correlation is positive so that an overestimation of the latter will produce an overestimation of the former. This reinforces the idea of maintaining a set of performance indicators for each module of the end-to-end process, so that quality frameworks cannot be only output-oriented. This vision is reinforced when we also analyse the relationship between the errors in the target population estimates and in the deduplicated penetration rates. We represent this relation in figure 7.6. Again, a positive correlation exists so that an overestimation in the deduplicated penetration rates will produce an overestimation in the target population counts. We also see how regions 3 and 5 are outliers, partially explaining why their estimates are of lower poor quality.

This positive correlation is also apparently detected with the dispersion of the posterior distribution (computed numerically for this discussion), as shown in figure 7.7. The different

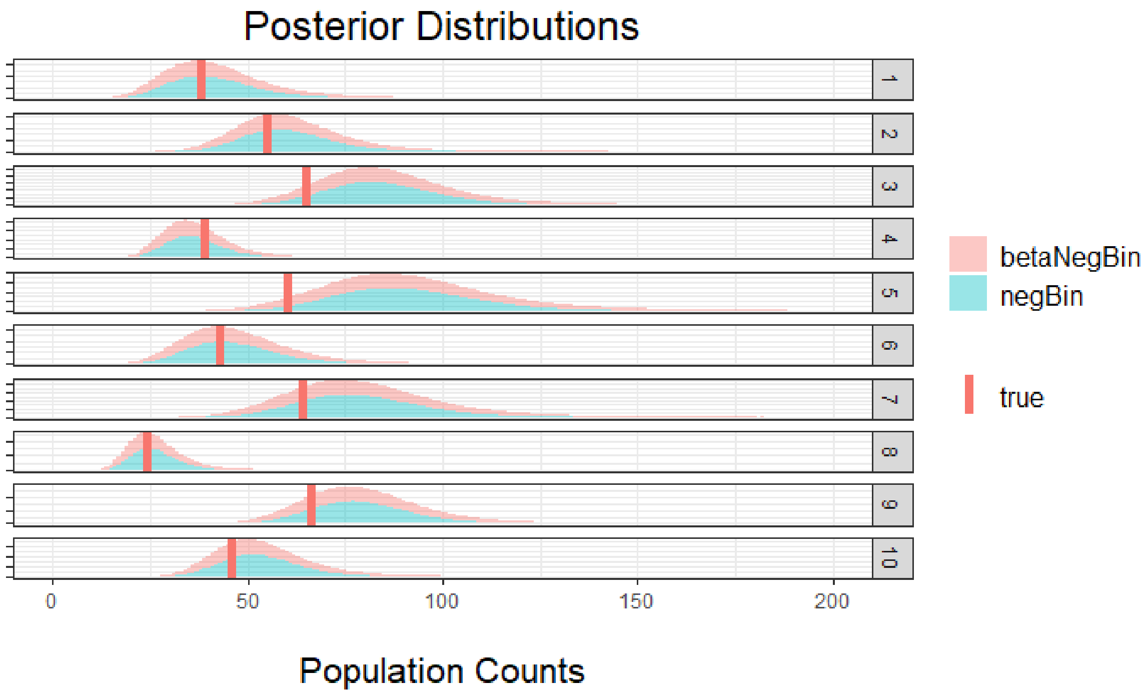

Figure 7.4: Posterior distributions for the number of individuals.

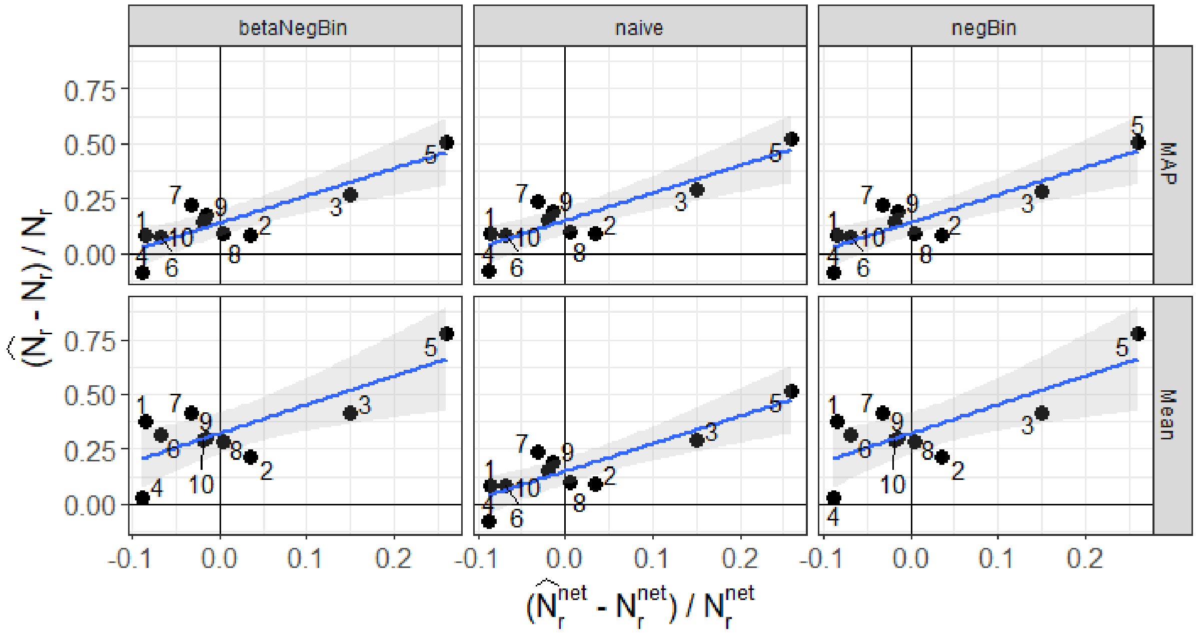

Figure 7.5: Relation between relative errors in target population estimates and estimates of the number of individuals detected by the network.

degree of robustness of the point estimations with the MAP and the posterior mean is visible, as expected. The outlying region 5 is, however, strongly affecting this regression. Dropping out region 5 we see the benefits of using the MAP estimators instead of the posterior mean estimator: errors in target population estimates remain constant despite the larger dispersion.

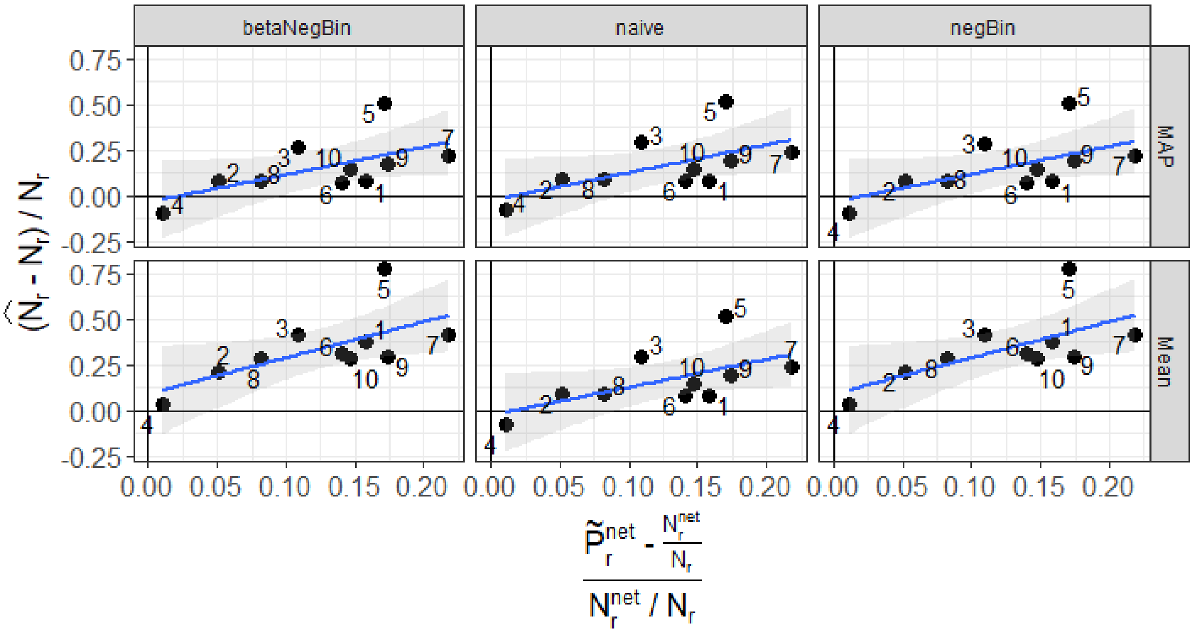

Figure 7.6: Relation between relative errors in target population estimates and estimates of the deduplicated penetration rates.

Another two relations of interest regard, on the one hand, the distance $d(\mathbf{cp}_d, \mathbf{r}_d^*)$ between the true position $\mathbf{r}_d^*$ of each device d and the center of probability of geolocation $\mathbf{cp}_d$ (see chapter **??**) and, on the other hand, the radius of dispersion rmsd_d of each device d. This linear distances are referred to the square root of the area A_r of each region as illustrative linear extension of the regions. This is depicted in figure 7.8 as a regression between the relative error of the target population estimates and the cp-tp distance averaged over each region. Each device d has been assigned to a region r according to the maximum posterior location probability γ_{drt_0} (thus, this assignment comes also with some noise). First, we observe that devices are in average correctly geolocated into each corresponding region ($\bar{d}_r(\mathbf{cp}, \mathbf{r}^*)/\sqrt{A_r} < 1$). Second, and consequently we also observe a virtually null relation between the distance cp-tp and the errors in the estimates, since the devices are correctly geolocated.

Finally, we complete the analysis by introducing uncertainty in the register-based population variables N_r^{reg}. Our goal is to investigate how this uncertainty affects the target population estimates. For simplicity, we shall concentrate only on MAP estimates. Firstly, we analyse the dependence of the point estimates on the parameters $\Delta\sigma_r$, which can be interpreted as a bias in the register-based population densities. We see in figure 7.9 a displacement produced by the density bias in all estimates (computed for a coefficient of variation $\epsilon_r = 0.01$ for all regions r). This same behaviour can be seen in the same figure for the coefficient of variation ϵ_r of the density (computed for unbiased values of σ_r^{reg}). Similarly, we obtain a negative correlation between the bias of the penetration rates and the relative errors of point estimates (see figure 7.10).

We repeat this same analysis investigating the effects of biases in the auxiliary information on the credible intervals. This is represented in figures 7.11 and 7.12. We observe both the

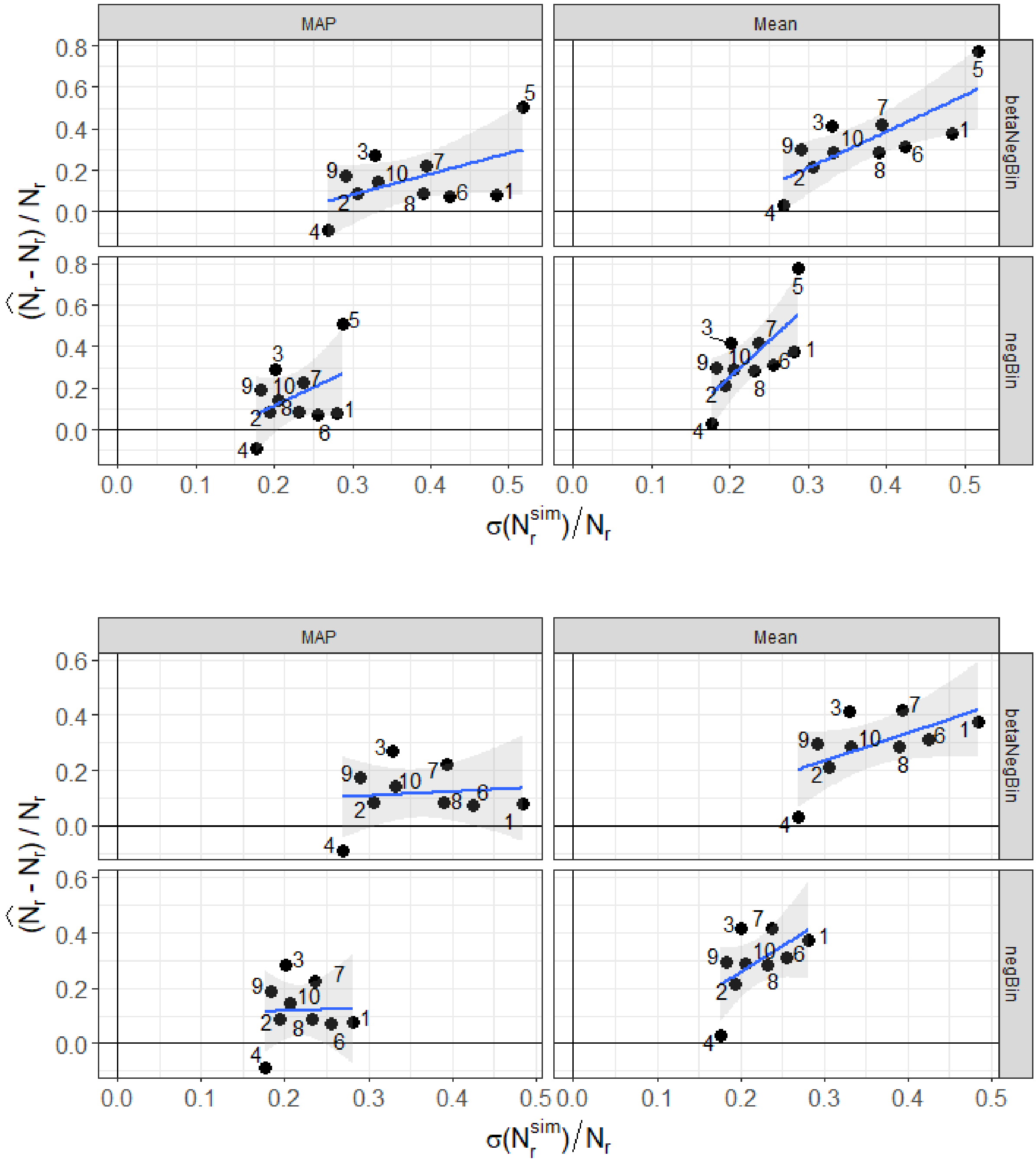

Figure 7.7: Relation between relative errors in target population estimates and dispersion of the posterior distribution. Graphs below are shown without outlying region 5.

displacement of the intervals as well as the increase of their length, i.e. of the uncertainty on the final estimates of the target population.

As a conclusion, we have seen that:

- Point estimates (both MAP and mean) are sensitive to estimation errors in the number of individuals detected by the network.

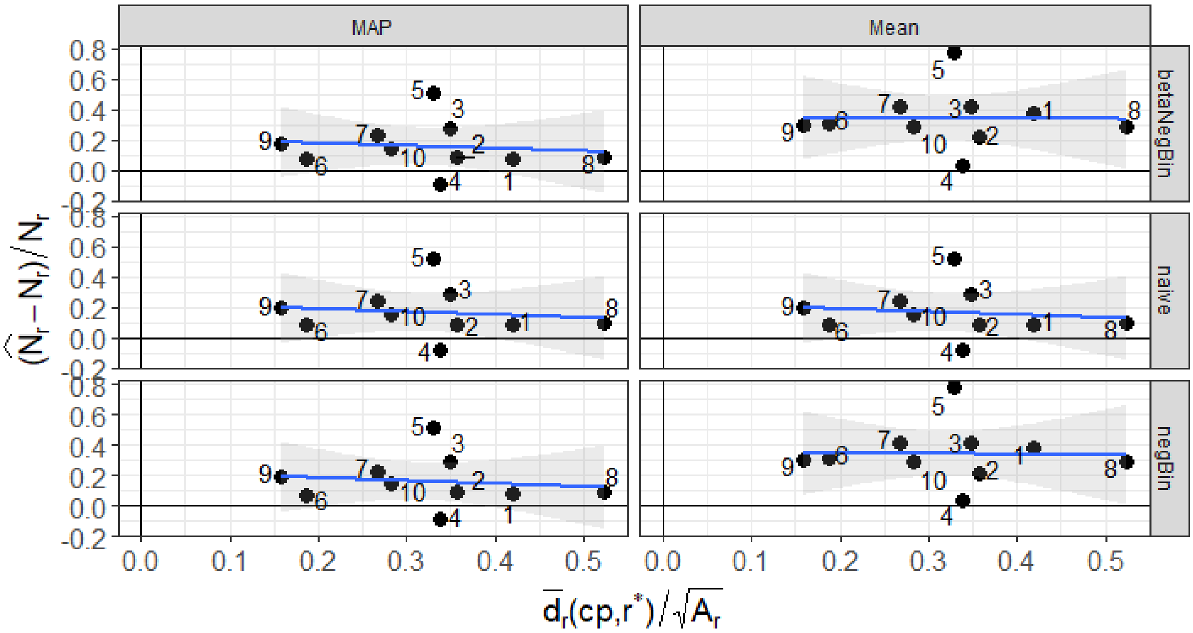

Figure 7.8: Relation between the relative error in target population estimates and the distance between the true position and the center of location probability in terms of the linear extension of each region.

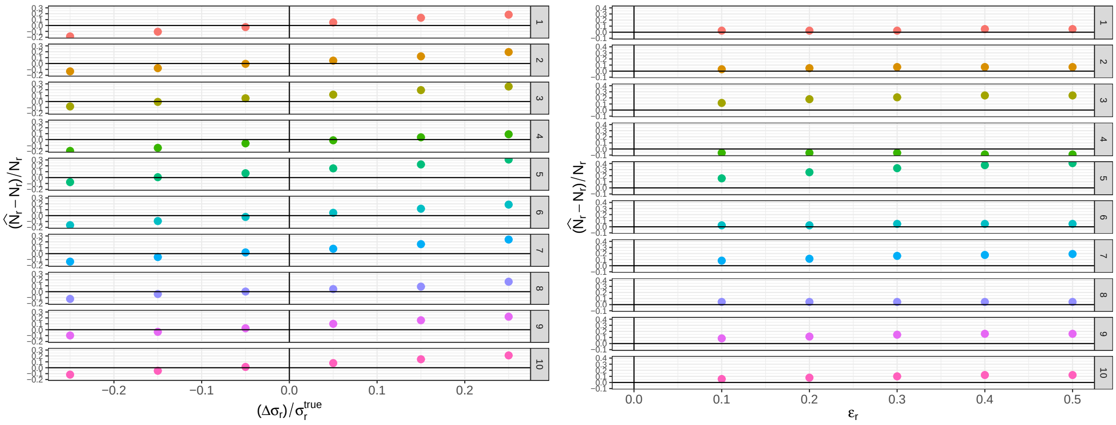

Figure 7.9: Relation between the relative error in target population estimates and the bias and coefficient of variation of register-based population densities.

- Point estimates (both MAP and mean) are sensitive to uncertainties in the penetration rates.
- MAP point estimates show more robustness to dispersion in the posterior distribution.
- Point estimates are very little sensitive to the geolocation errors provided the devices are detected correctly within the corresponding region.
- Credible intervals show clearly the uncertainty in the estimation and comprise in general true values.

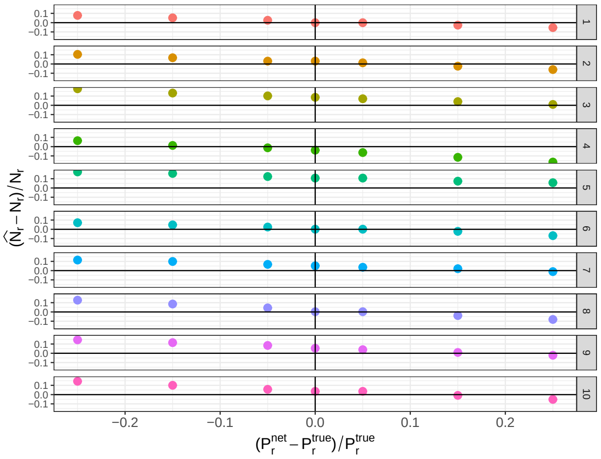

Figure 7.10: Relation between the relative error in target population estimates and the bias of the penetration rates.

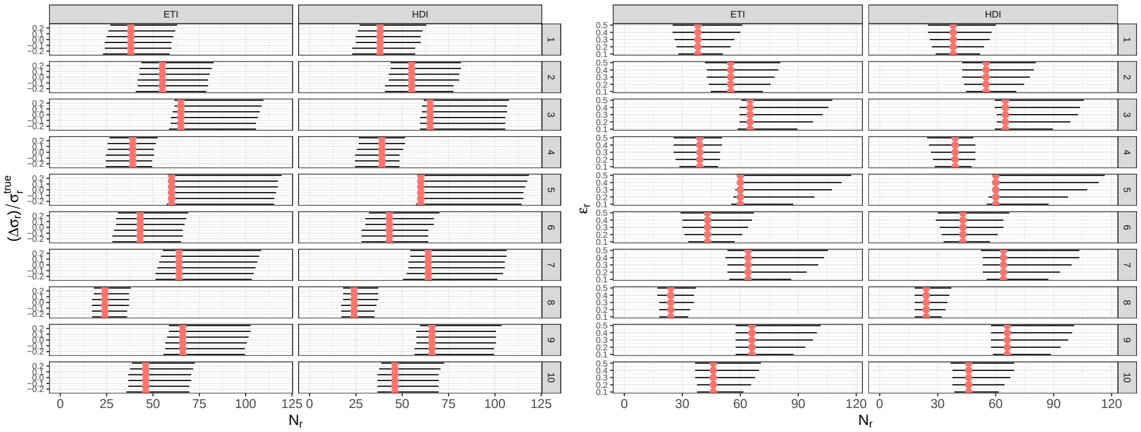

Figure 7.11: Credible intervals for different values of the bias and coefficient of variation of register-based population densities. True values represented in red. Confidence level 0.95.

- The bias in the register-based population density produces a bias in the MAP point estimate in the same direction.

- The coefficient of variation in the register-based population density produces a proportional bias in the MAP point estimate in the same direction.

- The bias in the penetration rates produces a bias in the MAP point estimates in the opposite direction.

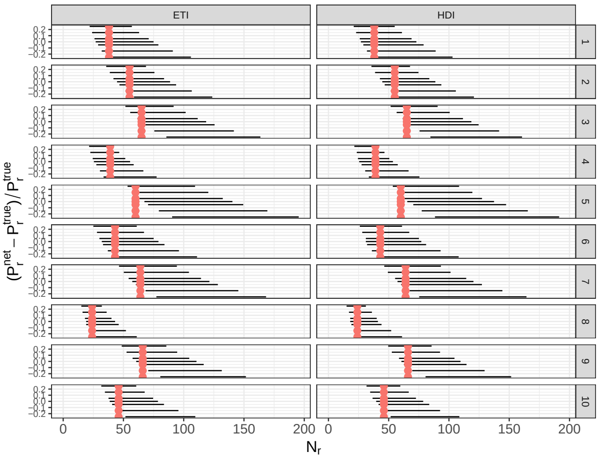

Figure 7.12: Credible intervals for different values of the penetration rates. True values represented in red. Confidence level 0.95.

8

Future prospects

This chapter is devoted to gather some reflections and conclusions for future work. As a conclusion from the preceding chapters we can claim that the ESS Reference Methodological Framework for Mobile Network Data already contains the first methodological elements implementing the modular and evolvable principles of production enabling us to detach the rapidly changing technological layers generating this data from the statistical analyses. The current proposal identifies the following production modules:

1. Geolocation.- This module assigns location probabilities to each tile for each device as well as consecutive joint location probabilities. These probabilities condense the information coming from the network thus allowing us to detach technology and statistics. These probabilities will provide the basic data input for the rest of modules (apart from external auxiliary information).

2. Device multiplicity.- This module aims at deduplicating devices carried by the same individual thus solving the problem of population coverage in terms of number of devices.

3. Statistical filtering.- This module aims at identifying devices of individuals belonging to the target population under analysis: domestic tourists, inbound tourists, outbound tourists, commuters, etc.

4. Aggregation.- This module aims at producing probability distributions for the number of individuals (not devices) detected by the network.

5. Inference.- This module aims at producing probability distributions for the number of individuals of the target population.

Probability theory underlies each module so that uncertainty (hence accuracy indicators) can be incorporated from the onset. Apart from mathematical and statistical details to be further improved and developed, it is important to identify key strategic issues arising from the current proposal included in this document. There exist several strategic issues derived from our proposal driving us to some considerations for the future work:

1. Access.- Although the use of the network data simulation has allowed us to develop and test different methodological elements, thus detaching the intricate issue of accessing mobile network data and reaching agreements with the MNOs, this does not mean that access and the development of the ESS Reference Methodological Framework should not be related. Rather on the contrary, we claim that any advancement in the methodological realm should be analysed searching for its counterpart for the access issue. For example, according to the present proposal, aggregated data sets of the number of devices per

time period and territorial unit makes our deduplication approach unattainable. Thus, the module on device multiplicity should also part of the negotiations with MNOs as part of the raw telco data preprocessing. Another example (not explored so far) is the identification of the network data to reach an acceptable regime of accuracy in the final estimates. In this sense, the use of probability theory allows us to measure the uncertainty and the network event simulator potentially enables us to prepare different scenarios with different parametrizations to identify the sufficient network data.

2. Scalability.- As pointed out across the document, scalability is not dealt with in the current methodological proposal. Sparsity and parallelization are two key ingredients which need intense research to implement these proposals into real production conditions. As a matter of fact, this should also be part of the agreements with MNOs.

3. Open Populations.- Only closed populations have been considered so far in our proposal. Open population with tiles (or territorial units) where individuals appear and disappear (airports, ports, train stations, border points, etc.) must be considered.

4. Realistic movement patterns.- In connection with the points above, the network event data simulator needs further development not only to include open populations but also to implement more realistic movement patterns for the individuals, introducing usual environments, home/work locations and similar elements common in human behaviour.

5. Cross-cutting methods.- There exists an intense pressure on this data source to produce clear use cases and concrete statistical products showing the advantages of using mobile network data especially to justify the tension regarding confidentiality and privacy issues. This pressure should be managed with care not to repeat mistakes in the past of Official Statistics. In our view, this pressure brings the risk of pushing the community through a product-oriented approach, as in the traditional production process (which drove us to the stove-pipe model and the inefficient silos). The process must result in a high-quality statistical product, but the focus should be on the development of cross-cutting statistical methods and techniques allowing subject-matter experts to produce novel insights. Currently, the efforts have been concentrated on geolocation, but there is still much promising information related to social interactions in this data source.

6. Statistical disclosure control.- Needless to say, privacy and confidentiality issues are at the front of the difficulties to access this data source for the production of official statistics. This is a missing module in the present methodological proposal which needs attention in the short term. Statistical disclosure control is a methodological branch of official statistical production which guarantees privacy-preserving and low risk of reidentification of statistical units in the disseminated statistics already with survey data. There exists proposals in the literature (see Xu et al., (2017)) empirically showing how the aggregation of mobility data does not preserve privacy. Further research needs to be conducted to identify conditions assuring privacy. The network event data simulator may again arise as a versatile tool to carry out this research.

In a more technical level, we can already detect and suggest directions of improvement for the current proposal:

1. State.- The concept of state in the HMM approach should be further developed and include velocity, mode of transport and some other potentially rich variable to model the movement of mobile devices.

2. Emission models.- More complex radiowave propagation models need to be explored investigating the gain in accuracy in the geolocation.

3. Transition models.- More complex transition models should also explored, especially in relation with more complex definitions of state.

4. Random number of devices D.- The total number of devices D in a network can change even for a closed population (e.g. due to churning). Thus, this should be treated as random, with the corresponding uncertainty assessment. More complex HMM state definitions should account for this phenomenon.

5. Duplicity change.- In this proposal we have considered the device duplicity as a fixed condition. More realistic conditions in which an individual may acquire or stop having a second device in the time interval under analysis should also be explored.

6. Deduplication and record linkage.- A detailed revision of the record linkage literature should be undertaken to check whether existing techniques can be reused in this context[1].

7. Detection algorithm.- Once more realistic movement patterns are available in the simulator, a set of parametrizable algorithms should be constructed to identify different target populations and key concepts such as usual environment and home/work locations. Attention should especially be paid to those devices of territorial units where coverage service by the network is noticeably different (e.g. rural areas) in order to avoid bias.

8. Aggregation and deduplication.- More alternatives for the interrelation between the deduplication procedure and the aggregation method should be investigated.

9. Inference models.- Further development of the hierarchical models needs to be explored to investigate the gain in accuracy. In particular, for all models posterior predictive distributions as a measure of model assessment should be explored together with a comparison of marginals probability distributions with true values from the simulator.

10. Anchor points.- The inference step does not make use of anchor points (home/work location, usual environment, etc.). This possibility should be explored by using this information in a level of the hierarchy.

11. Market shares.- In the same spirit, only penetration rates have been used, but the probabilities of detection p_{rt} could also be possibly modelled in terms of market shares, which can be computed using the same mobile network data. This also needs exploration.

12. External source-based and target populations.- For the estimation at time t_0 we make the assumption of assimilating register-based and target populations. This assimilation should be further explored using more external sources (e.g. from labour force market using figures for night workers). This could be possibly modelled using more complex relationships.

13. Validation.- Official data for some locations (hospitals, tourist accommodations, etc.) should be used whenever possible to validate final estimates.

In summary, the methodological framework proposed in this document does not intend to provide closed and definitive methods to produce estimations for the number of individuals in a target population together with its accuracy, but to put in place a first concrete substantiation of

[1] We acknowledge this suggestion by the internal Review Board of the present project.

the ESS Reference Methodological Framework for Mobile Network Data. Much work remains to be done in the future, hopefully in an international collaboration in the Official Statistics community. Results are encouraging since we achieve modularity and evolvability, as well as rigorous accuracy indicators.

Appendix A

A hidden Markov model for the geolocation of mobile devices

In this appendix we fully specify the construction and adaptation of a hidden Markov model (HMM) to estimate the geolocation of a mobile device d in a reference grid. We want to underline that this methodology is not proposing a close statistical model but a generic framework potentially adaptable to different situations depending on the available data.

As background literature for this methodology the reader may consult Rabiner (1989), Bishop (2006), and Murphy (2012). The state (latent) variables $T_d(t)$ will be the reference grid tile where the device d is located at time t (see section 3.1). The observed variables are the identification variables $E_d(t)$ of the antenna connecting to the mobile device d at time t (see also 3.1). Notice that the random variables $T_d(t)$ and $E_d(t)$ are discrete, thus we construct a discrete HMM. We shall use a variant of the standard HMM by introducing latent variables (states) not producing observed variables as a natural interpretation of time instants without an antenna-device connection. In another words, some of the observed variables will be missing.

We proceed stepwise in a standard way:

1. Time discretization.
2. Construction of the emission model.
3. Construction of the transition model.
4. Computation of the likelihood function.
5. Parameter estimation (likelihood maximization).
6. Application of the forward-backward algorithm.

A.1. Time discretization

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 instances, and (iii) an upper bound v_{max} for the velocity of the individuals in the population. For reasons which will be clear later on, we impose that in the time interval Δt, the device d at most can displace from one tile to a contiguous tile. This will allow us to choose a parsimonious model for the transition probabilities, thus reducing statistical complexity. Notice that this is a choice, although a convenient choice.

Under this condition, we can trivially set $\Delta t \lesssim \frac{l}{v_{\max}}$. For example, if $v_{\max} = 150\text{km/h} \approx 42\text{ms}^{-1}$, then $\Delta t \lesssim \frac{100}{42} \approx 2\text{s}$.

If in the dataset the device d is detected at longer time periods, e.g. once in a minute, then we artificially introduce missing values at intervals Δt between every two observed values. We will show below that this artificial non-response allow us to work with parsimonious models easier to estimate. Nonetheless, a systematization of this time padding procedure is needed to cover a wider range of time scopes.

Notice that we have used an a priori value for $v_{\max}$, but we can also possibly make an estimation using the observed values $E_d(t)$ and geometrical considerations about the respective coverage areas and their mutual distance.

Additionally, each observed time instance t is approximated to its closest multiple of Δt. Thus, we will have as input data a sequence of time instants as multiples $t_n = \Delta t \cdot n, (n \geq 0)$ and a randomly alternate sequence of missing values and of antenna IDs E_{t_n} (hereafter for ease of notation we drop out any reference to mobile device d since we are only focusing on one device).

A.2. Construction of the emission model

The emission model is specified by the HMM emission probabilities $b_{ia} = \mathbb{P}\left(E_{t_n} = a \middle| T_{t_n} = i\right)$, where a stands for the antenna ID and i denotes the tile index. We assume time homogeneity. Now we borrow the use of the simplified radio propagation model from the static analysis (see section 3.2) so that we can compute numerically these probabilities. Again, this is a modelling choice and several options could be possibly considered:

- Should we have rich raw telco data to use more complex radio propagation models, we could immediately improve the accuracy with a more sophisticated computation of the emission probabilities. In case of lacking data for these models, we could resort to geometrical considerations as with the Voronoi tessellation. The ideal recommendation is to work together with MNOs to identify the more feasible data set for the computation of these likelihoods. Ultimately, this will also depend on the chosen final accuracy in our estimates.

- A cautious reader may rapidly suggest that the emission probabilities can also be modelled in terms of unknown parameters to be estimated later on. In theory, this is always possible as in many other applications of HMMs. However, in our case we suggest to deal with the emission probabilities independently as a separate (sub)module in the whole process allowing us to detach the more technological stages directly dependent on raw telco data from the more statistical upper layers involving population count estimation. In this way, the joint work by MNOs and NSIs around the sensitive telco data is focused on this step paving the way for the functional modularity of the statistical process thus providing a concrete proposal for the implementation of the RMF.

- The computational cost of the emission probabilities is fixed in time. If N_A denotes the number of antennas in the geographical territory under analysis and the grid size is N_T, at most we need to compute $N_T \times N_A$ emission probabilities to conform the matrix $B = [b_{ia}]$, $i = 1, \dots, N_T$, $a = 1, \dots, N_A$. This is done once and for all t (assuming time homogeneity).

- Notice that having the numerical values of the emission probabilities will allow us to simplify the computation of the likelihood for the HMMs reducing its parameter dependency only to the transition model.
- If missing values are to be used according to the preceding section,*for numerical convenience later on* the corresponding emission probabilities can be conveniently set to 1, i.e. $b_{i0} = \mathbb{P}\left(E_{t_n} = \cdot \middle| T_{t_n} = i\right) = 1$. This will greatly facilitate the expression of the HMM likelihood and its further optimization. Remind that this probability is not real and completely meaningless.

Up to this point we have as input data the sequence of observed and missing values $a_{t_n} \in \{0, 1, \ldots, N_A\}$ for $t_n = 0, 1, \ldots, T$. We already have the emission matrix B, too.

A.3. Construction of the transition model

Now we specify a model for the transition between tiles $\{T_i\}$. 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. We shall comment on this later on. 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 $t_n = 0, 1, \ldots$ The number of tiles will be denoted by N_T. We assume a square regular grid for simplicity.

Now, we make use of our preceding imposition by which an individual can at most reach a contiguous tile in time Δt. Thus,

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

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

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

The last set of conditions is row-stochasticity:

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

The model can be easily summarised in a graphical way as in figure A.1.

Now back to the original notation for tiles and using the usual notation for the transition matrix $A = [a_{ij}]$, with $a_{ij} = \mathbb{P}\left(T_{jt} \middle| T_{it}\right)$, conditions (A.1) 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.

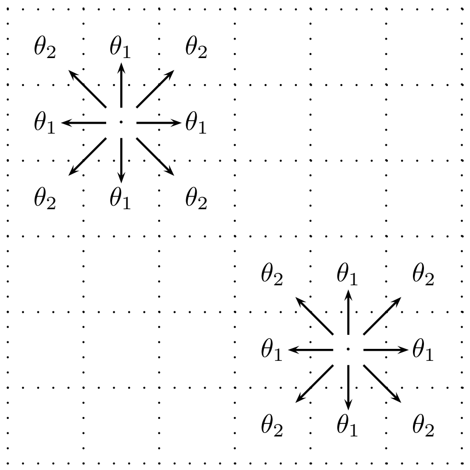
θ_2 θ_1 θ_2
θ_1 θ_1
θ_2 θ_1 θ_2
θ_2 θ_1 θ_2
θ_1 θ_1
θ_2 θ_1 θ_2

Figure A.1: Graphical representation of the rectangular isotropic model (A.1).

13	14	15	16
9	10	11	12
5	6	7	8
1	2	3	4

Figure A.2: Example with a regular square grid of dimensions 4×4.

As an illustrative example consider the 4×4 grid of figure A.2 (not highly sparse due to low dimensionality). The transition matrix can be specified as follows. Let

$$\begin{aligned}
D_1(\theta_1,\theta_2) &= \begin{pmatrix} 1-2\theta_1-\theta_2 & \theta_1 & 0 & 0 \\ \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 & 0 \\ 0 & \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 \\ 0 & 0 & \theta_1 & 1-2\theta_1-\theta_2 \end{pmatrix}, \\
D_2(\theta_1,\theta_2) &= \begin{pmatrix} 1-3\theta_1-2\theta_2 & \theta_1 & 0 & 0 \\ \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 & 0 \\ 0 & \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 \\ 0 & 0 & \theta_1 & 1-3\theta_1-2\theta_2 \end{pmatrix}, \\
M(\theta_1,\theta_2) &= \begin{pmatrix} \theta_1 & \theta_2 & 0 & 0 \\ \theta_2 & \theta_1 & \theta_2 & 0 \\ 0 & \theta_2 & \theta_1 & \theta_2 \\ 0 & 0 & \theta_2 & \theta_1 \end{pmatrix}, \\
O &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.
\end{aligned}$$

Then we can define the transition matrix A in blocks:

$$A(\theta_1,\theta_2)=\begin{bmatrix} D_1(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O & O \\ M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O \\ O & M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) \\ O & O & M(\theta_1,\theta_2) & D_1(\theta_1,\theta_2) \end{bmatrix} \tag{A.2}$$

Notice that $A(\theta_1,\theta_2)$ fulfills all restrictions (A.1). Indeed, in our proposed implementation, in order to seek future generalization, we will work with a generic block-tridiagonal matrix

$$A=\begin{bmatrix} D_{11} & D_{12} & O & O \\ D_{21} & D_{22} & D_{23} & O \\ O & D_{32} & D_{33} & D_{34} \\ O & O & D_{43} & D_{44} \end{bmatrix}, \tag{A.3}$$

where the restrictions leading to 0 have been included, and complemented with the rest of restrictions in matrix form $C \cdot \text{vec}(A') = \mathbf{b}$, where $\text{vec}(A')$ stands for the non-null elements of A. The rows of $[C|\mathbf{b}]$ encode each of the restrictions (A.1b), (A.1c), and (A.1d). For example, $a_{12}=\theta_1$ and $a_{21}=\theta_1$ produce a row like this

$$C_i{\cdot}\text{vec}(A')=\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}{\cdot}\begin{pmatrix} \cdots & 0 & a_{12} & 0 & \cdots & 0 & a_{21} & 0 & \cdots \end{pmatrix}^T = b_i = 0 \tag{A.4}$$

For the generic case of a grid with size N_T, we have

$$A(\theta_1,\theta_2)=\begin{bmatrix} D_1(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O & O & \cdots & O \\ M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) & O & \cdots & O \\ O & M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) & \cdots & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & M(\theta_1,\theta_2) & D_2(\theta_1,\theta_2) & M(\theta_1,\theta_2) \\ O & O & O & O & M(\theta_1,\theta_2) & D_1(\theta_1,\theta_2) \end{bmatrix}, \tag{A.5}$$

where now

$$
\begin{aligned}
D_1(\theta_1,\theta_2) &= \begin{pmatrix} 1-2\theta_1-\theta_2 & \theta_1 & 0 & 0 & \cdots & 0 \\ \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_1 & 1-3\theta_1-2\theta_2 & \theta_1 & \cdots & 0 \\ 0 & 0 & \theta_1 & 1-3\theta_1-2\theta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-2\theta_1-\theta_2 \end{pmatrix}_{N_T\times N_T}, \\
D_2(\theta_1,\theta_2) &= \begin{pmatrix} 1-3\theta_1-2\theta_2 & \theta_1 & 0 & 0 & \cdots & 0 \\ \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_1 & 1-4\theta_1-4\theta_2 & \theta_1 & \cdots & 0 \\ 0 & 0 & \theta_1 & 1-4\theta_1-4\theta_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1-3\theta_1-2\theta_2 \end{pmatrix}_{N_T\times N_T}, \\
M(\theta_1,\theta_2) &= \begin{pmatrix} \theta_1 & \theta_2 & 0 & 0 & \cdots & 0 \\ \theta_2 & \theta_1 & \theta_2 & 0 & \cdots & 0 \\ 0 & \theta_2 & \theta_1 & \theta_2 & \cdots & 0 \\ 0 & 0 & \theta_2 & \theta_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \theta_1 \end{pmatrix}_{N_T\times N_T}, \\
O &= \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}_{N_T\times N_T}.
\end{aligned}
$$

In our implementation, again we will consider a block-tridiagonal matrix together with a set of linear restrictions of the form $C \cdot \text{vec}\,(A') = \mathbf{b}$.

A.4. 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. The general expression for the likelihood is

$$
\begin{aligned}
L(\mathbf{E}) &= \sum_{i_0=1}^{N_T}\cdots\sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{t_0}=i_0\right)\prod_{n=1}^{N} \mathbb{P}\left(T_{t_n}=i_n|T_{t_{n-1}}=i_{n-1}\right)\mathbb{P}\left(E_{t_n}\middle|T_{t_n}=i_n\right) && \text{(A.6a)} \\
&= \sum_{i_0=1}^{N_T}\cdots\sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{t_0}=i_0\right)\prod_{n=1}^{N} a_{i_{n-1}i_n}(\boldsymbol{\theta})b_{i_n a_{t_n}} && \text{(A.6b)}
\end{aligned}
$$

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

The initial probability distribution $\mathbb{P}\left(T_{t_0}\right)$ has not been specified yet. With abundance of data, it has very little incidence on the final results. In our implementation two options are provided:

(i) the stationary state, i.e. the eigenvector of A with eigenvalue 1 and (ii) any distribution provided by the user. The difference will be ultimately in terms of computational cost (provided we have abundance of data).

A.5. Parameter estimation

The estimation of the unknown parameters $\boldsymbol{\theta}$ is conducted maximizing the likelihood. The restrictions coming from the transition model (A.1) 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:

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \\ & a_k \in [0,1], \end{aligned} \tag{A.7}$$

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 (A.1) not involving null rhs terms (restrictions (A.1b), (A.1c), and (A.1d).

Let us quantify the number of variables and restrictions in order to propose an abstract procedure possibly generalized to other situations. We illustrate this procedure for a square regular grid of size N_T. The number of zeroes in the transition matrix A can be computed as follows:

- There exist 4 rows in A corresponding to the 4 vertices in the grid. Each of these rows contains $N_T - 4$ zeroes.
- There exist 4 sets of $\sqrt{N_T} - 2$ rows in A corresponding to boundary tiles not being vertices. Each of these rows contains $N_T - 6$ zeroes.
- There exist $(\sqrt{N_T} - 2)^2$ rows in A corresponding to this same number of inner tiles. Each of these rows contains $N_T - 9$ zeroes.

Thus, the total number of zeroes in A is given by $4 \times (N_T - 4) + 4 \times (\sqrt{N_T} - 2) \times (N_T - 6) + (\sqrt{N_T} - 2)^2 \times (N_T - 9) = N_T^2 - 9 \cdot N_T + 12\sqrt{N_T} - 4$. The number of non-null components of $\mathbf{a}$ in problem (A.7) 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 movements. In the rectangular isotropic model considered above, we need to identify the number of entries (i) equal to θ_1, (ii) equal to θ_2, and (iii) in the diagonal (thus guaranteeing the row-stochasticity restriction). Using the same counting procedure as above, the number of entries equal to θ_1 will be given by $4 \times 2 + 4 \times (\sqrt{N_T} - 2) \times 3 + (\sqrt{N_T} - 2)^2 \times 4 = 4 \cdot N_T - 4\sqrt{N_T}$. Since θ_1 is a free parameters we get $4 \cdot N_T - 4\sqrt{N_T} - 1$ rows. For θ_2, we get $4 \times (\sqrt{N_T} - 1)^2 - 1$ rows. From the row-stochasticity restriction we get N_T rows. Thus, the matrix C will have dimensions $n_r \times d$, with $n_r = 4 \cdot N_T - 4\sqrt{N_T} - 1 + 4 \times (\sqrt{N_T} - 1)^2 - 1 + N_T = 9 \cdot N_T - 12\sqrt{N_T} + 2$. Notice that $d - n_r = 2$, as expected, since we have two free parameters θ_1 and θ_2.

For the illustrative example with $N_T = 16$ above we reduce the original 256 transition probabilities, to $d = 100$ non-null entries restricted with $n_r = 98$ constraints, thus leaving 2 free parameters, as expected.

The abstract optimization problem is thus

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \ , \\ & \mathbf{a} \in [0,1]^d, \end{aligned} \tag{A.8}$$

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 (A.8) then reduces to

$$\begin{aligned} \max \quad & \tilde{h}(a_{n_r+1}, \ldots, a_d) \\ \text{s.t.} \quad & 0 \leq \tilde{C} \cdot \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T \leq 1. \end{aligned} \tag{A.9}$$

We do not comment further here on the approach to solve this problem, since it can be undertaken with standard optimization techniques.

The solution $\mathbf{a}^*$ to problem (A.8) will be introduced in the transition probability matrix, which will thus be denoted by A^*.

A.6. 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 probabilities γ_{it} and $\gamma_{ij,t}$ (see equations (3.1)). No novel methodological content is introduced at this point. For our implementation, as suggested by Bishop (2006), we have used the scaled version of the algorithm.

These target probabilities are the main output of the geolocation estimation stage entering the next stage as input. It is important to remind that they are intended to stand as the mathematical objects allowing NSIs to detach the highly technological data ecosystem of MNOs from the statistical analyses for each statistical domain. Should these target probabilities fail to be enough to produce statistical outputs, we would need to come back to the preceding stepwise procedure, identify more target mathematical objects, and provide the algorithms to obtain them.

Appendix B

Hierarchical models to estimate the target population

B.1. The observation process

The probability mass function $\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right)$ coming from the observation process amounts to computing the multidimensional integrals

$$\prod_{r=1}^{R}\int_{\mathbb{R}}\mathrm{d}p_r\int_{\mathbb{R}^2}\mathrm{d}\alpha_r\mathrm{d}\beta_r\int_{\Omega_{\beta,\xi}}\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\tau_\beta^2\mathrm{d}\xi\,\mathbb{P}\left(N_r^{\text{net}}|N_r,p_r\right)\mathbb{P}\left(p_r|\alpha_r,\beta_r\right)\mathbb{P}\left(\alpha_r,\beta_r|\beta_0,\beta_1,\tau_\beta^2,\xi\right)\mathbb{P}\left(\beta_0,\beta_1,\tau_\beta^2,\xi\right), \tag{B.1}$$

where $\mathbb{P}(\cdot|\cdot)$ stands either for a density function or a mass probability function. These are taken from the specifications (7.9). To compute $\mathbb{P}\left(\alpha_r,\beta_r|\beta_0,\beta_1,\tau_\beta^2,\xi\right)$ we need to make the transformations

$$\begin{aligned} u_r &= \frac{\alpha_r}{\alpha_r+\beta_r}, \quad \text{for } r=1,\ldots,R, \\ v_r &= \alpha_r+\beta_r, \quad \text{for } r=1,\ldots,R. \end{aligned} \tag{B.2}$$

The variables u_r need a further transformation $y_r=\text{logit}u_r$. Then, we apply the approximation for the gamma function

$$\frac{\Gamma(x+n)}{\Gamma(x)}\approx x^n.$$

The final step is just to reorder and identify terms.

B.2. The state process

The part of the probability mass function $\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\text{net}}\right)$ coming from the state process also amounts to computing the multidimensional integrals

$$\prod_{r=1}^{R}\int_{\mathbb{R}^+}\mathrm{d}\sigma_r\int_{\mathbb{R}}\mathrm{d}\mu_{\gamma r}\int_{\Omega_{\gamma,\zeta}}\mathrm{d}\gamma_0\mathrm{d}\gamma_1\mathrm{d}\tau_\gamma^2\mathrm{d}\zeta\,\mathbb{P}\left(N_r|\sigma_r\right)\mathbb{P}\left(\sigma_r|\mu_{\gamma r},\zeta\right)\mathbb{P}\left(\mu_{\gamma r}|\gamma_0,\gamma_1,\tau_\gamma^2,\zeta\right)\mathbb{P}\left(\gamma_0,\gamma_1,\tau_\gamma^2,\zeta\right), \tag{B.3}$$

where $\mathbb{P}(\cdot|\cdot)$ stands either for a density function or a mass probability function. These are taken from the specifications (7.14). The integrals are computed readily. The final expression just needs term rearranging and their identification.

B.3. The Poisson limit of the negative binomial

We include details about the limit $\lim_{\zeta^*\to\infty}$ negbin $\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r)\cdot Q(\theta_r), N_r^{\text{net}} + \zeta_r + 1\right)$. Notice that

$$\lim_{\substack{\epsilon_r\to 0\\ \Delta\sigma_r\to 0}} (1-\bar{P}_r)\cdot Q_r(\theta_r)(N_r^{\text{net}} + \zeta_r + 1) = (1-\bar{P}_r)A_r\sigma_r^{\text{net}}. \tag{B.4}$$

Thus, we need to compute the Poisson limit of a negative binomial distribution of parameters p and r when $p\cdot r\to\mu$, which is

$$\binom{k+r-1}{k} p^k(1-p)^r \to \frac{\gamma(k+r)\cdot p^k}{k!\cdot\gamma(k)}(1-p)^r \to \frac{(p\cdot r)^k}{k!}(1-\frac{\mu}{r})^r \to e^{-\mu}\frac{\mu^k}{k!} \equiv \text{poisson}\,(k;\mu) \tag{B.5}$$

Back to our original parametrization, we get

$$\lim_{\substack{\epsilon_r\to 0\\ \Delta\sigma_r\to 0}} \text{negbin}\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r)\cdot Q(\theta_r), N_r^{\text{net}} + \zeta_r + 1\right) = \text{poisson}\left(N_r - N_r^{\text{net}}; (1-\bar{P}_r)A_r\sigma_r^{\text{net}}\right). \tag{B.6}$$

Bibliography

Ahas, R., A. Aasa, Ü. Mark, T. Pae, and A. Kull (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management 28*, 898–910.

American Planning Association (2018). Land based classification standards. `https://www.planning.org/lbcs/`.

Avouac, R., B. Sakarovitch, F. Sémécube, and Z. Smoreda (2019). A Bayesian approach to improve the estimation of population using mobile phone data. In preparation.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2015). *Hierarchical modelling and analysis of spatial data (2nd ed)*. CRC Press.

Barabási, A.-L. (2018). *Network Science*. Cambridge University Press.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion), in V.P. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*, pp. 203–242. Holt, Reinhart and Winston.

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. Wiley.

Bishop, C.M. (2006). *Pattern recognition and matching learning*. Springer.

Brown, J. and R. Churchill (2004). *Complex variables and applications (8th ed.)*. McGraw-Hill.

Bryant, J.R. and P.J. Graham (2013). Bayesian demographic accounts: subnational population estimation using multiple data sources. *Journal of Official Statistics 8* (3), 591–622.

Calabrese, F., L. Ferrari, and V. D. Blondel (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys 47*, 25:1-25:20.

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury Press.

Cassel, C.-M., C.-E. Särndal, and J. Wretman (1977). *Foundations of Inference in Survey Sampling*. Wiley.

Cochran, W. (1977). *Sampling Techniques* (3rd ed.). Wiley.

Daskalakis, C., G. Kamath, and C. Tzamos (2015). On the Structure, Covering, and Learning of Poisson Multinomial Distributions. Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS), pp. 1203–1217. `https://doi.org/10.1109/FOCS.2015.77`.

Deming, W. (1950). *Some theory of sampling*. Wiley.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. Stevens, A. Gaughan, V. Blondel, and A. Tatem (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences (USA) 111*, 15888– 15893.

Devroye, L. (1986). *Non-uniform random variable generation*. Springer.

DGINS (2018). Bucharest Memorandum. `http://www.dgins2018.ro/bucharest-memorandum/`.

Dobra, A., N.E. Williams and N. Eagle (2015). Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data. *PLOS ONE* 10, e0120449.

Douglass, R.W., D.A. Meyer, M. Ram, D. Rideout, and D. Song (2015). High resolution population estimates from telecommunications data. *EPJ Data Science* 4.

Doyle, J., P. Hung, R. Farrell, and S. Mcloone (2014). Population mobility dynamics estimated from mobile telephony data. *Journal of Urban Technology 21*, 109–132.

EPSG (2018). epsg.io – Coordinate Systems Worldwide. `https://epsg.io/28992`.

ESS (2011). European Statistics Code of Practice. `http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15`.

ESS (2018). ESSnet on Big Data I. `https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page1`.

ESS (2019). ESSnet on Big Data II. `https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page`.

Eurostat (2014). Methodological manual for tourism statistics (v3.1). `http://ec.europa.eu/eurostat/documents/3859598/6454997/KS-GQ-14-013-EN-N.pdf`.

Eurostat, NIT, University of Tartu, Statistics Estonia, Positium, IFSTTAT, and Statistics Finland (2014). Feasibility study on the use of mobile poistioning data for tourism statistics. `http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies`.

Gelfand, A. E. (2010). *Misaligned Spatial Data: The Change of Support Problem*, in A.E. Gelfand, P.J. Diggle, M. Fuentes, and P. Guttorp (eds.), Handbook of Spatial Statistics, CRC Press, 2010, 517-539.

Gelman, A., B. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian data analysis*. CRC Press.

Gezici, S. (2007). A Survey on Wireless Position Estimation. *Wireless Personal Communications, Springer Nature 44*, 263-282.

Graham, R., D. Knuth, and O. Patashnik (1996). *Concrete Mathematics (2nd ed.)*. Addison-Wesley.

Grimmet, G. and D. Stirzaker (2004). *Probability and random processes* (3rd ed.). Oxford Science Publications.

Groves, R. (1989). *Survey errors and survey costs*. Wiley.

Hájek, J. (1981). *Sampling from a finite population*. Marcel Dekker Inc.

Hansen, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science 2*, 180–190.

Hansen, M., W. Hurwitz, and W. Madow (1966). *Sample survey: methods and theory* (7th ed.). Wiley.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Hedayat, A. and B. Sinha (1991). *Design and Inference in Finite Population Sampling*. Wiley.

High-Level Group for the Modernisation of Official Statistics (2019). `https://www.unece.org/stats/mos.html`.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis 59*, 41–51.

INSPIRE (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). `https://inspire.ec.europa.eu/`.

ISO (2004). ISO 8601:2004. `https://www.iso.org/standard/40874.html`.

ISO (2007). ISO 19111:2007. `https://www.iso.org/standard/41126.html`.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models*. MIT Press.

Kriegel, H.-P., P. Kröger, J. Sander, and A. Zimek (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery 1* (3), 231–240.

Kruskal, W. and F. Mosteller (1979a). Representative sampling, i: Non-scientific literature. *International Statistical Review 47*, 13–24.

Kruskal, W. and F. Mosteller (1979b). Representative sampling, ii: scientific literature, excluding statistics. *International Statistical Review 47*, 111–127.

Kruskal, W. and F. Mosteller (1979c). Representative sampling, iii: the current statistical literature. *International Statistical Review 47*, 245–265.

Kruskal, W. and F. Mosteller (1980). Representative sampling, iv: The history of the concept in statistics. *International Statistical Review 48*, 169–195.

Lehamnn, E.L. and G. Casella (1998). *Theory of point estimation, 2nd ed*. Springer.

Lehtonen, R. and A. Veijanen (1998). Logistic generalized regression estimators. *Survey Methodology 24*, 51–55.

Lessler, J. and W. Kalsbeek (1992). *Nonsampling error in surveys*. Wiley.

Little, R. (2012). Calibrated bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics 28*, 309–334.

Loo, M. van der. (2019). Systematic approaches to data validation and data cleaning for statistical production. Seminar at Statistics Spain (INE). May 8, 2019.

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

Manly, B. and J. e. Navarro-Alberto (2014). *Introduction to ecological sampling*. CRC Press. McLean DJ, Skowron Volponi MA. trajr: An R package for characterisation of animal trajectories. Ethology. 2018;00:1–9. https://doi.org/10.1111/eth.12739.

McLean D.J. and M.A. Skowron Volponi (2018). trajr: An R package for characterisation of animal trajectories. *Ethology 124*, 440–448.

Meersman, F. D., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H. Reuter (2016). Assessing the quality of mobile phone data as a source of statistics. *Q2016 Conference paper, June 2016*.

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

Moro, E., D. Calacci, X. Dong, and A. Pentland (2019). Economical Segregation of Encounter Networks in Cities. NetMob 2019, Oxford, UK, July 2019.

Murphy, K. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Nemhauser, G. and L. Wolsey (1999). *Integer and combinatorial optimization*. Addison-Wesley.

NetMob (2017). Conference on the scientific analysis of mobile phone datasets. `http://netmob.org/`.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*, 558–625.

Okabe, A., B. Boots, K. Sugihara, and S.-N. Chiu (2000). *Spatial tessellations: concepts and applications of Voronoi diagrams (2nd ed)*. Wiley.

Pfeffermann, D. and C.R. Rao (2009). *Sample Surveys: Inference and Analysis, Volume 29A*. North Holland.

Pfeffermann, D. and C.R. Rao (2009). *Sample Surveys: Inference and Analysis, Volume 29B*. North Holland.

Positium (2016). Technical documentation for required raw data from mobile network operators for official statistics. Technical report, Positium.

Positium (2017). Common plan for methodology and data processing of mobile phone data from mobile network operators for official statistics. Technical report, Positium.

Positium (2018). `https://www.positium.com/`.

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

Rao, J. and I. Molina (2015). *Small area estimation (2nd ed)*. Wiley.

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

Ricciato, F., P. Widhalm, F. Pantisano, and M. Craglia (2017). 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.

Ricciato, F. (2018). Towards a Reference Methodological Framework forprocessing MNO data for Official Statistics. *15th Global Forum on Tourism Statistics. 28-30 November 2018. Cusco (Peru).* http://www.15th-tourism-stats-forum.com/pdf/Papers/S3/3_1_A_Reference_Methodological_Framework_for_processing_mobile_network_operatordata_for_official_statistics.pdf.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods (2nd ed).* Springer.

Robert, C. and G. Casella (2010). *Introducing Monte Carlo Methods with R*. Springer.

Royle, J. and R. Dorazio (2014). *Hierarchical modeling and inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press.

Salgado, D, M.E. Esteban, M. Novás, S. Saldaña, and L. Sanguiao (2018). Data Organisation and Process Design Based on Functional Modularity for a Standard Production Process. *Journal of Official Statistics 34*, 811–833.

Saltzer, J.H. and M.F. Kaashoek (2009). *Principles of Computer System Design*. Morgan Kaufmann.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*, 99–119.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. Springer.

Seynaeve, G., C. Demunter, F. D. Meersman, Y. Baeyens, M. Debusschere, P. Dewitte, P. Lusyne, F. Reis, H. Reuter, and A. Wirthmann (2016). When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics. *14th Global Forum on Tourism Statistics (Venice, Italy, Nov. 2016).*

Smith, T.M.F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society A 139*, 183–204.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Open Street Map Foundation. https://www.openstreetmap.org.

ESS (2017). ESSnet on Big Data. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php.

JAGS (2018). http://mcmc-jags.sourceforge.net/.

Stan (2018). Stan. http://mc-stan.org/.

Tennekes, M., Y.A.P.M. Gootzen, and S.H. Shah (2019). Deriving geographic location of mobile devices from network data. In preparation.

Thompson, S. (2012). *Sampling*. Wiley.

I. Ucar, M. Gramaglia, M. Fiore, Z. Smoreda, and E. Moro (2019). Netflix or Youtube? Regional income patterns of mobile service consumption. NetMob 2019, Oxford, UK, July 2019.

UNECE (2013). Generic Statistical Business Process Model v5.0. https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

UNECE (2016). Generic Statistical Data Editing Models. https://statswiki.unece.org/display/VSH/GSDEMs.

Valliant, R., A. Dorfmann, and R. Royall (2000). *Finite population sampling and inference. A prediction approach.* Wiley.

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

Wilysis (2018). Network Cell Info Lite app. https://play.google.com/store/apps/details?id=com.wilysis.cellinfolite&hl=en_419.

WP5 of ESSnet on Big Data I (2016). Deliverable 5.1. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5_Deliverable_1.1.pdf.

WP5 of ESSnet on Big Data I (2017). Guidelines for the access to mobile phone data within the ESS. Deliverable 5.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.2.pdf.

WP5 of ESSnet on Big Data I (2018). Proposed Elements for a Methodological Framework for the Production of Official Statistics with Mobile Phone Data. Deliverable 5.3. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4d/WP5_Deliverable_5.3_Final.pdf.

WPI of ESSnet on Big Data II (2019). Data Simulator - A simulator for network event data. Deliverable I.2. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/b/b9/WPI_Deliverable_I2_Data_Simulator_-_A_simulator_for_network_event_data.pdf.

WPK of ESSnet on Big Data II(2020). Report describing the methodological steps of using big data in official statistics with a section on the most important research questions for the future including guidelines. Deliverable K.10 (in construction). https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPK_Milestones_and_deliverables.

Xu, F., Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin (2017). Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web 1241–1250*. https://doi.org/10.1145/3038912.3052620.

Yates, F. (1965). *Sampling methods for censuses and surveys* (3rd ed.). Charles Griffins.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66(1)*, 41–63.