An inferential framework for official statistics based on mobile network data: Bayesian hierarchical models combining auxiliary information

David Salgado, Dept. Methodology and Development of Statistical Production, Statistics Spain (INE)
Bogdan Oancea, Dept. Innovative Tools in Statistics, Statistics Romania (INS) and Dept. Business Administration, University of Bucharest (Romania)

2018-11-05

Abstract

This is a draft internal document for WPI on mobile network data in which we explore a generic inferential framework for the production of official statistics based upon network data. This proposal elaborates further on the initial efforts conducted by the ESSnet on Big Data I and is strongly based on the work by Bryant and Graham [2013], where a similar approach is proposed to estimate population counts using a combination of administrative registers.

Contents

1 Overview

This is a draft internal document for the work package I on mobile network data of the ESSnet on Big Data II. Elaborating on the former ESSnet on Big Data I, we explore herein a generic inferential framework based upon Bayesian hierarchical models providing proposals for the construction of concrete statistical models. We formulate our proposal in general terms and leave computational details for the subsequent development of this task of track 3 on methodology in this work package.

The whole proposal is strongly based on the hierarchical model by Bryant and Graham [2013] to estimate population counts using a combination of administrative registers. We adapt these authors' ideas to the use of mobile phone data and show how our former model proposed by Salgado et al. [2018] is indeed contained in this general framework.

The goal is to introduce the main conceptual elements constituting the inferential framework as well as its structure. In this way we take first steps in the direction of providing solutions for the so-called representativity issue in mobile phone data (hopefully, also with Big Data sources in general).

2 Structure and elements of the framework

The ultimate goal in Bayesian inference is to provide the so-called posterior distribution $\mathbb{P}(\mathbf{Y}|\mathbf{X},\mathbf{Z})$ of the target random variable $\mathbf{Y}$ under analysis conditioned on the primary data $\mathbf{X}$ we have collected and on auxiliary covariates $\mathbf{Z}$ assisting in the inference exercise.

We already see a first change in the inferential paradigm with respect to traditional methods (probabilistic sampling): as an output we need to provide a probability distribution. In our view, this entails some immediate benefits for the production of official statistics:

a. We can provide point estimations by taking e.g. the mean, the median, the mode, or any other statistics built from this distribution.
b. We can also provide interval estimations by constructing so-called credible intervals, which would play a similar role to confidence interval in design-based inference. Now, we have an important good difference: credible intervals depend on the sampled data and not on any other possibly selected sample.
c. We can (and must) build quality indicators for the goodness of fit of the model $\mathbb{P}(\mathbf{Y}|\mathbf{X},\mathbf{Z})$ both at the output and at the input.

We will use hierarchical models using the same motivation as in the ESSnet on Big Data I [Salgado et al., 2018]. The use of hierarchical models enables us to incorporate elements from both the observation process and the system process in the inference exercise. In the inference model the observed data, the target process, and its underlying parameters must be given a joint probability distribution $\mathbb{P}(\text{data},\text{process},\text{parameters})$. The hierarchical model allows us to decompose this joint distribution as

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

Besides, if we are to include also the auxiliary covariates playing a conditioning role, we indeed have:

$$\begin{aligned}\mathbb{P}(\text{data},\text{process},\text{parameters}|\text{covariates}) =& \mathbb{P}(\text{data}|\text{process},\text{parameters}|\text{covariates})\cdot\\ & \mathbb{P}(\text{process}|\text{parameters},\text{covariates})\cdot\\ & \mathbb{P}(\text{parameters}|\text{covariates}).\end{aligned}$$

Following closely the work by Bryant and Graham [2013], we will model both the observation process and the dynamics of the system under analysis (a human population in our cases). Thus we introduce the next notation:

- $\mathbf{Q}$.- Instead of $\mathbf{Y}$, we shall denote the target variables by $\mathbf{Q}$. This follows Bryant's and Graham's notation to denote the target *demographic account* in estimating population counts. We shall promote this to more general situations (see below).
- $\mathbf{X}$.- This will denote the observed data in our inference exercise.

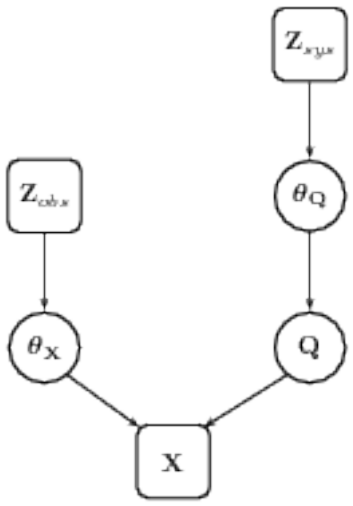

Figure 1: Model structure

- $\mathbf{Z} \rightsquigarrow \mathbf{Z}_{obs}, \mathbf{Z}_{sys}$.- This will denote the known auxiliary data which we use in the modelling exercise. We use the corresponding susbcript to distinguish between the observation process and the system process.
- $\boldsymbol{\theta} \rightsquigarrow \boldsymbol{\theta}_{\mathbf{X}}, \boldsymbol{\theta}_{\mathbf{Q}}$.- This will denote the set of parameters used in the modelling exercise. We also use the corresponding susbcript to distinguish between the observation process and the system process.

Below we comment on the role and conception of all these variables. First, we need to show the conceptual relationship between these variables in figure 1. Following Bryant and Graham [2013] we denote the observed variables by a square frame and unobserved variables by a round frame.

The goal of the model is to provide the joint posterior distribution conditioned on the available data and covariates $\mathbb{P}\left(\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{Q}}, \boldsymbol{\theta}_{\mathbf{X}} | \mathbf{X}, \mathbf{Z}_{obs}, \mathbf{Z}_{sys}\right)$, and more concretely, the marginal distribution $\mathbb{P}\left(\mathbf{Q} | \mathbf{X}, \mathbf{Z}_{obs}, \mathbf{Z}_{sys}\right)$.

The structure depicted in figure 1 entails

$$\mathbb{P}\left(\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{Q}}, \boldsymbol{\theta}_{\mathbf{X}} | \mathbf{X}, \mathbf{Z}_{obs}, \mathbf{Z}_{sys}\right) \propto \mathbb{P}\left(\mathbf{X} | \mathbf{Q}, \boldsymbol{\theta}_{\mathbf{X}}\right) \mathbb{P}\left(\mathbf{Q} | \boldsymbol{\theta}_{Q}\right) \mathbb{P}\left(\boldsymbol{\theta}_{Q} | \mathbf{Z}_{sys}\right) \mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{X}} | \mathbf{Z}_{obs}\right).$$

This is the distribution to be ultimately generated by a simulation method (rejection algorithm, MCMC, Gibbs sampler,...). We shall not make any further technical considerations in this line. This is the work to be developed during the project (track 3).

However, let us show some examples and then discuss some concrete points in the application of this approach to mobile phone data.

3 Population counts estimation using administrative data

In the work by Bryant and Graham [2013] the target variables are demographic variables about the population, which the authors formalise in the concept of a *demographic account* given by

$\mathbf{Q} = (\mathbf{n}, \mathbf{b}, \mathbf{d}, \mathbf{m}^{\mathrm{II}}, \mathbf{m}^{\mathrm{IO}}, \mathbf{m}^{\mathrm{EI}}, \mathbf{m}^{\mathrm{EO}})$, where each vector component is explained in the following table (see [Bryant and Graham, 2013] for more details):

Variable	**Notation**	**Definition**
Population	n_{rsat}	Number of people in region r of sex s and age group a at time t.
Births	b_{rsat}	Number of births in region r of sex s to women in age group a during period t.
Deaths	d_{rsat}	Number of deaths in region r of people of sex s and age group a during t.
Internal in-migration	m^{II}_{rsat}	Number of moves into region r from elsewhere inside the country by people of sex s and age group a during period t.
Internal out-migration	m^{IO}_{rsat}	Number of moves out of region r to elsewhere in the country by people of sex s and age group a during period t.
External in-migration	m^{EI}_{rsat}	Number of moves into region r from elsewhere outside the country by people of sex s and age group a during period t.
External out-migration	m^{EO}_{rsat}	Number of moves out of region r to elsewhere outside country by people of sex s and age group a during period t.

These vector components are linked through an accounting identity making explicit the population at two consecutive time periods $t-1$ and t. For $a = 1, 2, \ldots, \mathcal{A}-1$, this identity is:

$$n_{rsat} = n_{r,s,a-1,t-1} - d_{rsat} + m^{\mathrm{II}}_{rsat} - m^{\mathrm{IO}}_{rsat} + m^{\mathrm{EI}}_{rsat} - m^{\mathrm{EO}}_{rsat}.$$

For $a = 0$:

$$n_{rs0t} = \sum_{a} b_{rsat} - d_{rs0t} + m^{\mathrm{II}}_{rs0t} - m^{\mathrm{IO}}_{rs0t} + m^{\mathrm{EI}}_{rs0t} - m^{\mathrm{EO}}_{rs0t}.$$

For $a = \mathcal{A}$:

$$n_{rs\mathcal{A}t} = n_{r,s,\mathcal{A}-1,t-1} + n_{r,s,\mathcal{A},t-1} - d_{rs\mathcal{A}t} + m^{\mathrm{II}}_{rs\mathcal{A}t} - m^{\mathrm{IO}}_{rs\mathcal{A}t} + m^{\mathrm{EI}}_{rs\mathcal{A}t} - m^{\mathrm{EO}}_{rs\mathcal{A}t}.$$

Apart from the model parameters $\boldsymbol{\theta}_{obs}$ and $\boldsymbol{\theta}_{sys}$, in the work by Bryant and Graham [2013] the data sets entering into play are:

Model Variable	**Definition**
$\mathbf{X}$	Set of administrative registers: registered births and deaths, official estimates of resident population, Census data, International "permanent and long-term" arrivals and departures, tax system data, electoral roll, school roll
$\mathbf{Q}$	Demographic account, as stated above.

Model Variable	Definition
$\mathbf{Z}_{sys}$	Indicator variable taking values $0, 1$ depending on whether university students comprised more than $1/3$ of the population of that region and age group at the time of the 2006 census.
$\mathbf{Z}_{obs}$	None.

As naturally expected, the demographic account is not directly observed. However, we know $\mathbf{X}$ and the auxiliary covariates $\mathbf{Z}$.

Once the model is worked out (basically building every conditional probability in the expression $\mathbb{P}\left(\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{Q}}, \boldsymbol{\theta}_{\mathbf{X}} | \mathbf{X}, \mathbf{Z}_{obs}, \mathbf{Z}_{sys}\right)$), usual Bayesian computational techniques are used to provide this posterior probability.

From the conceptual point of view, we identify the following key points:

- Choice of data $\mathbf{X}$ and $\mathbf{Z}$.
- Choice of models $\mathbb{P}\left(\mathbf{X}|\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{X}}\right)$, $\mathbb{P}\left(\mathbf{Q}|\boldsymbol{\theta}_{Q}\right)$, $\mathbb{P}\left(\boldsymbol{\theta}_{Q}|\mathbf{Z}_{sys}\right)$, $\mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{X}}|\mathbf{Z}_{obs}\right)$.

If we want to adapt this approach to the use of mobile phone data, we must make the appropriate choices in these two points.

4 Population counts estimation using mobile network data

4.1 Considerations on the role of available data: choice of X and Z

In the administrative register case above, the observed data were divided into the observed admin data $\mathbf{X}$ and the auxiliary covariates $\mathbf{Z}_{sys}$, which contains indeed more information about the population from other sources. The target variables were chosen beyond doubt as the components of a demographic account.

For the use of mobile phone data, it makes sense to keep the same or similar target variables, also for other statistical domains (see e.g. below for tourism statistics). With high probability the level of details about the age group or sex cannot be trivially measured with this new data source (so far), thus we may be obliged to focus on a coarse-grained version of the demographic account (for example, estimating the number $n_{r \cdot \cdot t}$ of individuals of the population in region r at time t).

Now, a non-trivial question is how to divide the available observed information, i.e. the mobile phone data and both admin or survey data available at the statistical office. It seems natural that mobile phone data $\mathbf{N}^{MNO}$ must be a part of $\mathbf{X}$. The question is where to place the rest: in $\mathbf{X}$ or in $\mathbf{Z}$? We identify two extreme choices:

a. We make $\mathbf{X} = (\mathbf{N}^{MNO}, \mathbf{X}_k)$, i.e. both mobile phone data and all available data in administrative registers leaving $\mathbf{Z}$ as minimal as possible (like in the original proposal by Bryant and Graham [2013]).
b. We make $\mathbf{X} = \mathbf{N}^{MNO}$ and the rest of available information is considered as covariates $\mathbf{Z}$. This can be interpreted as an integration of the new source with already existing and validated statistics.

In our view, this choice amounts to deciding which problem we are facing. In the former case, we are measuring the population with all our data at hand appropriately combined. That is, we are providing population figures integrating all data sources either administrative or not. In the latter case, we are producing population figures from mobile phone data taking official population figures from administrative data for granted. In this sense, mobile phone data-based population is built on top of administrative data-based population.

Must we make a choice? Our answer is no. We should explore both options to explore what information a statistical office can provide. However, we must begin by one of them. In the subsequent, foreseeing the challenges coming from the different geographical and time scales of admin data and mobile phone data, we will concentrate on the second option, thus we shall take administrative data-based population figures as auxiliary covariates. Thus, we can interpret mobile phone data-based figures as complementary to official figures hopefully providing more geographical and time details.

4.2 Population statistics

In this statistical domain the main target variables of interest are given by the demographic account $\mathbf{Q}$ already explained above for administrative registers with components broken down into resident and non-resident population[1]. Now, to simplify both as a starting work and in prevision of the limited information (basically, time and position) in mobile phone data sets, we shall concentrate on the coarse-grained demographic account $\mathbf{Q}$ given by

$$\mathbf{Q} = (\mathbf{n}^{(r)}, \mathbf{n}^{(nr)}, \mathbf{b}^{(r)}, \mathbf{d}^{(r)}, \mathbf{m}_{\mathrm{II}}^{(r)}, \mathbf{m}_{\mathrm{II}}^{(nr)}, \mathbf{m}_{\mathrm{IO}}^{(r)}, \mathbf{m}_{\mathrm{IO}}^{(nr)}, \mathbf{m}_{\mathrm{EI}}^{(r)}, \mathbf{m}_{\mathrm{EI}}^{(nr)}, \mathbf{m}_{\mathrm{EO}}^{(r)}, \mathbf{m}_{\mathrm{EO}}^{(nr)}),$$

where each component $k_{rt}^{(\cdot)}$ is expressed in terms of two subscripts r (region) and t (time). Notice that we have assumed that there are no births and deaths of non-resident population in the territory under analysis. The components are then defined as:

Variable	**Notation**	**Definition**
Resident population	$n_{rt}^{(r)}$	Number of resident people in region r at time t.
Non-resident population	$n_{rt}^{(nr)}$	Number of non-resident people in region r at time t.
Births of resident population	$b_{rt}^{(r)}$	Number of births in region r to resident women during period t.
Deaths of resident population	$d_{rt}^{(r)}$	Number of deaths in region r of resident people during t.
Resident internal in-migration	$m_{rt}^{(r)\mathrm{II}}$	Number of moves into region r from elsewhere inside the country by resident people during period t.
Non-resident internal in-migration	$m_{rt}^{(nr)\mathrm{II}}$	Number of moves into region r from elsewhere inside the country by non-resident people during period t.

[1]For practical purposes we define as resident population those individuals included in the population register used as auxiliary covariate.

Variable	Notation	Definition
Resident internal out-migration	$m_{rt}^{(r)\mathrm{IO}}$	Number of moves out of region r to elsewhere in the country by resident people during period t.
Non-resident internal out-migration	$m_{rt}^{(nr)\mathrm{IO}}$	Number of moves out of region r to elsewhere in the country by non-resident people during period t.
Resident external in-migration	$m_{rt}^{(r)\mathrm{EI}}$	Number of moves into region r from elsewhere outside the country by resident people during period t.
Non-resident external in-migration	$m_{rt}^{(nr)\mathrm{EI}}$	Number of moves into region r from elsewhere outside the country by non-resident people during period t.
Resident external out-migration	$m_{rt}^{(r)\mathrm{EO}}$	Number of moves out of region r to elsewhere outside country by resident people during period t.
Non-resident external out-migration	$m_{rt}^{(nr)\mathrm{EO}}$	Number of moves out of region r to elsewhere outside country by non-resident people during period t.

Now the accounting identity is reduced to the set of relations:

$$n_{rt}^{(r)}+n_{rt}^{(nr)} = n_{r,t-1}^{(r)}+n_{r,t-1}^{(nr)}+b_{rt}^{(r)}-d_{rt}^{(r)}+m_{rt}^{(r)\mathrm{II}}+m_{rt}^{(nr)\mathrm{II}}-m_{rt}^{(r)\mathrm{IO}}-m_{rt}^{(nr)\mathrm{IO}}+m_{rt}^{(r)\mathrm{EI}}+m_{rt}^{(nr)\mathrm{EI}}-m_{rt}^{(r)\mathrm{EO}}-m_{rt}^{(nr)\mathrm{EO}}.$$

If we assume that changing the residence condition takes place at a larger time scale than τ_{MNO} we can indeed divide this accounting identity in two similar parts:

$$\begin{aligned} n_{rt}^{(r)} &= n_{r,t-1}^{(r)} + b_{rt}^{(r)} - d_{rt}^{(r)} + m_{rt}^{(r)\mathrm{II}} - m_{rt}^{(r)\mathrm{IO}} + m_{rt}^{(r)\mathrm{EI}} - m_{rt}^{(r)\mathrm{EO}}, \\ n_{rt}^{(nr)} &= n_{r,t-1}^{(nr)} + m_{rt}^{(nr)\mathrm{II}} - m_{rt}^{(nr)\mathrm{IO}} + m_{rt}^{(nr)\mathrm{EI}} - m_{rt}^{(nr)\mathrm{EO}}. \end{aligned}$$

As Bryant and Graham [2013] expressed, we also assume (conditionally) independent parameters:

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{Q}}|\mathbf{Z}_{sys}\right) =& \mathbb{P}\left(\boldsymbol{\theta}_{N}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{B}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{D}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{II}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{IO}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{EI}^{(r)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{EO}^{(r)}|\mathbf{Z}_{sys}\right)\times \\ & \mathbb{P}\left(\boldsymbol{\theta}_{N}^{(nr)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{II}^{(nr)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{IO}^{(nr)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{EI}^{(nr)}|\mathbf{Z}_{sys}\right)\mathbb{P}\left(\boldsymbol{\theta}_{EO}^{(nr)}|\mathbf{Z}_{sys}\right) \end{aligned}$$

As auxiliary covariates $\mathbf{Z}_{sys}$ we will use, as agreed above, the population register. The system model is then completed with the specification:

$$
\begin{aligned}
\mathbb{P}\left(\mathbf{Q}|\boldsymbol{\theta}_{\mathbf{Q}}, \mathbf{Z}_{sys}\right) \propto & \mathbb{P}\left(\mathbf{n}^{(r)}|\boldsymbol{\theta}_{N}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{n}^{(nr)}|\boldsymbol{\theta}_{N}^{(nr)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{b}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{B}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{d}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{D}^{(r)}, \mathbf{Z}_{sys}\right) \times \\
& \mathbb{P}\left(\mathbf{m}_{\mathrm{II}}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{M_{\mathrm{II}}}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{m}_{\mathrm{II}}^{(nr)}|\mathbf{n}^{(nr)}, \boldsymbol{\theta}_{M_{\mathrm{II}}}^{(nr)}, \mathbf{Z}_{sys}\right) \times \\
& \mathbb{P}\left(\mathbf{m}_{\mathrm{IO}}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{M_{\mathrm{IO}}}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{m}_{\mathrm{IO}}^{(nr)}|\mathbf{n}^{(nr)}, \boldsymbol{\theta}_{M_{\mathrm{IO}}}^{(nr)}, \mathbf{Z}_{sys}\right) \times \\
& \mathbb{P}\left(\mathbf{m}_{\mathrm{EI}}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{M_{\mathrm{EI}}}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{m}_{\mathrm{EI}}^{(nr)}|\mathbf{n}^{(nr)}, \boldsymbol{\theta}_{M_{\mathrm{EI}}}^{(nr)}, \mathbf{Z}_{sys}\right) \times \\
& \mathbb{P}\left(\mathbf{m}_{\mathrm{EO}}^{(r)}|\mathbf{n}^{(r)}, \boldsymbol{\theta}_{M_{\mathrm{EO}}}^{(r)}, \mathbf{Z}_{sys}\right) \mathbb{P}\left(\mathbf{m}_{\mathrm{EO}}^{(nr)}|\mathbf{n}^{(nr)}, \boldsymbol{\theta}_{M_{\mathrm{EO}}}^{(nr)}, \mathbf{Z}_{sys}\right) \times \\
& I\left(\mathbf{f}(\mathbf{Q})\right),
\end{aligned}
$$

where I is an indicator function for the fulfillment of the accounting identity $\mathbf{f}(\mathbf{Q}) = 0$ (both for the resident and non-resident components). Now, to provide a full specification we need to find adequate models for each component of the demographic account in terms of their parameters and for each parameter (see next section).

The observation model is specified along similar lines:

$$
\mathbb{P}\left(\mathbf{X}, \boldsymbol{\theta}_{\mathbf{X}}|\mathbf{Q}, \mathbf{Z}_{obs}\right) \propto \mathbb{P}\left(\mathbf{X}|\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{X}}\right) \mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{X}}|\mathbf{Z}_{obs}\right).
$$

What variables do we observe with mobile phone data? To build our observation model we minimally assume that mobile phone data have at least four attributes for each network event in the data set:

- Pseudonymized ID, identifying each SIM card.
- Time attribute, providing the time instant (second) at which the event took place.
- Spatial attribute, providing a location at which the event took place. Usually the geolocation of network events needs some processing of the raw telco data (see Salgado et al. [2018]).
- Roaming indicator, providing information about whether the SIM card corresponds or not to a roamer. We assume that roaming in both directions (inbound and outbound) are provided (MNOs do have this information for billing processes).

Indeed, it is technologically feasible to provide the country of origin/destination of each roamer but here we will not deal with that level of detail. This would imply to further specify finer models for the non-resident components of the demographic account. We shall assume that non-resident people can be identified as roamers in the mobile data sets. This is motivated by the hypothesis that non-resident people visit (either by work or by leaisure) a foreign territory with their own national SIM card and only under long-term conditions close to become a resident citizen, these people will buy a SIM card in the territory under analysis.

As observed data sets we must firstly identify variables in the mobile phone data set. Trivially we decompose the observed variables $\mathbf{X}$ as

$$
\begin{aligned}
\mathbf{X} =& (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_7, \mathbf{X}_8, \mathbf{X}_9, \mathbf{X}_{10}) \\
=& (\mathbf{n}^{(NR)\mathrm{MNO}}, \mathbf{n}^{(R)\mathrm{MNO}}, \\
& \mathbf{m}_{\mathrm{II}}^{(NR)\mathrm{MNO}}, \mathbf{m}_{\mathrm{II}}^{(R)\mathrm{MNO}}, \mathbf{m}_{\mathrm{IO}}^{(NR)\mathrm{MNO}}, \mathbf{m}_{\mathrm{IO}}^{(R)\mathrm{MNO}}, \mathbf{m}_{\mathrm{EI}}^{(NR)\mathrm{MNO}}, \mathbf{m}_{\mathrm{EI}}^{(R)\mathrm{MNO}}, \mathbf{m}_{\mathrm{EO}}^{(NR)\mathrm{MNO}}, \mathbf{m}_{\mathrm{EO}}^{(R)\mathrm{MNO}}),
\end{aligned}
$$

where their components are given by:

Variable	Notation	Definition
Non-roaming population	$n_{rt}^{(NR)\text{MNO}}$	Number of mobile devices of non-roaming people in region r at time t.
Roaming population	$n_{rt}^{(R)\text{MNO}}$	Number of mobile devices of roaming people in region r at time t.
Non-roaming internal in-migration	$m_{rt}^{(NR)\text{MNO, II}}$	Number of mobile devices of non-roaming people moving into region r from elsewhere inside the country during period t.
Roaming internal in-migration	$m_{rt}^{(R)\text{MNO, II}}$	Number of mobile devices of roaming people moving into region r from elsewhere inside the country during period t.
Non-roaming internal out-migration	$m_{rt}^{(NR)\text{MNO, IO}}$	Number of mobile devices of non-roaming people moving out of region r to elsewhere in the country during period t.
Roaming internal out-migration	$m_{rt}^{(R)\text{MNO, IO}}$	Number of mobile devices of roaming people moving out of region r to elsewhere in the country during period t.
Non-roaming external in-migration	$m_{rt}^{(NR)\text{MNO, EI}}$	Number of mobile devices of non-roaming people moving into region r from elsewhere outside the country during period t.
Roaming external in-migration	$m_{rt}^{(R)\text{MNO, EI}}$	Number of mobile devices of roaming people moving into region r from elsewhere outside the country during period t.
Non-roaming external out-migration	$m_{rt}^{(NR)\text{MNO, EO}}$	Number of mobile devices of non-roaming people moving out of region r to elsewhere outside country during period t.
Roaming external out-migration	$m_{rt}^{(R)\text{MNO, EO}}$	Number of mobile devices of roaming people moving out of region r to elsewhere outside country during period t.

There is also an accounting identity for MNO data:

$$\begin{aligned} n_{rt}^{(NR)\text{MNO}} + n_{rt}^{(R)\text{MNO}} =& n_{r,t-1}^{(NR)\text{MNO}} + n_{r,t-1}^{(R)\text{MNO}} \\ &+ m_{rt}^{(NR)\text{MNO, II}} + m_{rt}^{(R)\text{MNO, II}} - m_{rt}^{(NR)\text{MNO, IO}} - m_{rt}^{(R)\text{MNO, IO}} \\ &+ m_{rt}^{(NR)\text{MNO, EI}} + m_{rt}^{(R)\text{MNO, EI}} - m_{rt}^{(NR)\text{MNO, EO}} - m_{rt}^{(R)\text{MNO, EO}}. \end{aligned}$$

This identity can also be split up into two parts under the assumption that in the time scale of observation, roamers do not transform into non-roamers in the territory under analysis:

$$\begin{aligned} n_{rt}^{(NR)\text{MNO}} =& n_{r,t-1}^{(NR)\text{MNO}} + m_{rt}^{(NR)\text{MNO, II}} - m_{rt}^{(NR)\text{MNO, IO}} + m_{rt}^{(NR)\text{MNO, EI}} - m_{rt}^{(NR)\text{MNO, EO}}, \\ n_{rt}^{(R)\text{MNO}} =& n_{r,t-1}^{(R)\text{MNO}} + m_{rt}^{(R)\text{MNO, II}} - m_{rt}^{(R)\text{MNO, IO}} + m_{rt}^{(R)\text{MNO, EI}} - m_{rt}^{(R)\text{MNO, EO}}. \end{aligned}$$

Conditional on the observation parameters $\boldsymbol{\theta}_X$ and the target account $\mathbf{Q}$, the generic observation model is then specified by

$$\begin{aligned}\mathbb{P}\left(\mathbf{X}|\mathbf{Q},\boldsymbol{\theta}_{\mathbf{X}}\right) \propto & \mathbb{P}\left(\mathbf{n}^{(NR)\mathrm{MNO}}|\mathbf{n}^{(r)},\boldsymbol{\theta}_{\mathbf{X}}\right)\mathbb{P}\left(\mathbf{n}^{(R)\mathrm{MNO}}|\mathbf{n}^{(nr)},\boldsymbol{\theta}_{\mathbf{X}}\right)\times \\ & \mathbb{P}\left(\mathbf{m}_{\mathrm{II}}^{(NR)\mathrm{MNO}}|\mathbf{m}_{\mathrm{II}}^{(r)},\boldsymbol{\theta}_{\mathbf{X}}\right)\mathbb{P}\left(\mathbf{m}_{\mathrm{II}}^{(R)\mathrm{MNO}}|\mathbf{m}_{\mathrm{II}}^{(nr)},\boldsymbol{\theta}_{\mathbf{X}}\right) \\ & \mathbb{P}\left(\mathbf{m}_{\mathrm{IO}}^{(NR)\mathrm{MNO}}|\mathbf{m}_{\mathrm{IO}}^{(r)},\boldsymbol{\theta}_{\mathbf{X}}\right)\mathbb{P}\left(\mathbf{m}_{\mathrm{IO}}^{(R)\mathrm{MNO}}|\mathbf{m}_{\mathrm{IO}}^{(nr)},\boldsymbol{\theta}_{\mathbf{X}}\right)\times \\ & \mathbb{P}\left(\mathbf{m}_{\mathrm{EI}}^{(NR)\mathrm{MNO}}|\mathbf{m}_{\mathrm{EI}}^{(r)},\boldsymbol{\theta}_{\mathbf{X}}\right)\mathbb{P}\left(\mathbf{m}_{\mathrm{EI}}^{(R)\mathrm{MNO}}|\mathbf{m}_{\mathrm{EI}}^{(nr)},\boldsymbol{\theta}_{\mathbf{X}}\right)\times \\ & \mathbb{P}\left(\mathbf{m}_{\mathrm{EO}}^{(NR)\mathrm{MNO}}|\mathbf{m}_{\mathrm{EO}}^{(r)},\boldsymbol{\theta}_{\mathbf{X}}\right)\mathbb{P}\left(\mathbf{m}_{\mathrm{EO}}^{(R)\mathrm{MNO}}|\mathbf{m}_{\mathrm{EO}}^{(nr)},\boldsymbol{\theta}_{\mathbf{X}}\right)\times \\ & I\left(\mathbf{g}(\mathbf{Q}^{\mathrm{MNO}})\right),\end{aligned}$$

where I denotes again an indicator variable for the fulfillment of the accounting identity $\mathbf{g}(\mathbf{Q}^{\mathrm{MNO}})$ for mobile phone data.

Regarding the parameters $\boldsymbol{\theta}_{\mathbf{X}}$, we will also impose conditional independence. As auxiliary covariates $\mathbf{Z}_{obs}$ we will make use of the information provided by the national telecommunication regulator about the penetration rates (market shares) $R_{r0}^{(NR)}$ and roaming rates $R_{r0}^{(R)}$ computed compiling data from the different MNOs operating in the territory under analysis.

Again, to accomplish the full specification we need to provide concrete models for $\mathbf{X}$ and $\boldsymbol{\theta}_{\mathbf{X}}$.

4.2.1 Full model specification for the system process

We need to provide a model for each of the components of the broken-down demographic account $\mathbf{Q}$. Let us begin with $n_{rt}^{(r)}$. Along the same lines as Bryant and Graham [2013], we propose:

$$\begin{aligned} n_{rt}^{(r)}|\lambda_{rt}^{(r,N)} &\overset{\text{indep}}{\simeq} \text{Poisson}\left(\lambda_{rt}^{(r,N)}\right) \\ \lambda_{rt}^{(r,N)}|\xi^{(r,N)},\mu_{rt}^{(r,N)} &\overset{\text{indep}}{\simeq} \text{Gamma}\left(\xi^{(r,N)},\frac{\xi^{(r,N)}}{\mu_{rt}^{(r,N)}}\right) \\ \mathbb{P}\left(\xi^{(r,N)}\right) &= \frac{\zeta_0^{(r,N)}}{(\zeta_0^{(r,N)}+\xi^{(r,N)})^2} \\ \log\left(\mu_{rt}^{(r,N)}\right)-\log\left(\mu_{r,t-1}^{(r,N)}\right)|\beta_r^{(r,N)},\sigma_r^{(r,N)2} &\overset{\text{indep}}{\simeq} \mathrm{N}\left(\beta_r^{(r,N)},\sigma_r^{(r,N)2}\right), \quad t=1,2,\ldots,T; r=1,2,\ldots,R \\ \log\left(\mu_{r0}^{(r,N)}\right)|\gamma_0^{(r,N)},\gamma_1^{(r,N)},\tau^{(r,N)2},N_{r0}^{\mathrm{REG}} &\overset{\text{indep}}{\simeq} \mathrm{N}\left(\gamma_0^{(r,N)}+\gamma_1^{(r,N)}\cdot\log\left(N_{r0}^{\mathrm{REG}}\right),\tau^{(r,N)2}\right) \\ \mathbb{P}\left(\beta_r^{(r,N)},\sigma_r^{(r,N)2}\right) &\overset{\text{indep}}{\propto} 1 \\ \mathbb{P}\left(\gamma_0^{(r,N)},\gamma_1^{(r,N)},\tau^{(r,N)2}\right) &\propto 1 \end{aligned}$$

This model implements the assumption that at the initial time instant $t=0$ we can assimilate the expected value of $n_{r0}^{(r)}$ to the number of individuals in the administrative register N_{r0}^{REG}. In later time periods, independently for each region r, the system evolves according to a random walk model (weakly informative behaviour).

For the parallel non-resident component $n_{rt}^{(nr)}$ we propose a similar model:

$$
\begin{aligned}
n_{rt}^{(nr)} | \lambda_{rt}^{(nr,N)} &\overset{\text{indep}}{\simeq} \text{Poisson}\left(\lambda_{rt}^{(nr,N)}\right) \\
\lambda_{rt}^{(nr,N)} | \xi^{(nr,N)}, \mu_{rt}^{(nr,N)} &\overset{\text{indep}}{\simeq} \text{Gamma}\left(\xi^{(nr,N)}, \frac{\xi^{(nr,N)}}{\mu_{rt}^{(nr,N)}}\right) \\
\mathbb{P}\left(\xi^{(nr,N)}\right) &= \frac{\zeta_0^{(nr,N)}}{(\zeta_0^{(nr,N)} + \xi^{(nr,N)})^2} \\
\log\left(\mu_{rt}^{(nr,N)}\right) - \log\left(\mu_{r,t-1}^{(nr,N)}\right) | \beta_r^{(nr,N)}, \sigma_r^{(nr,N)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\beta_r^{(nr,N)}, \sigma_r^{(nr,N)2}\right), \quad t = 1, 2, \ldots, T; r = 1, 2, \ldots, R \\
\log\left(\mu_{r0}^{(nr,N)}\right) | \gamma_0^{(nr,N)}, \gamma_1^{(nr,N)}, \tau^{(nr,N)2}, Z_{r0}^{\text{REG}} &\overset{\text{indep}}{\simeq} \text{N}\left(\gamma_0^{(nr,N)} + \gamma_1^{(nr,N)} \cdot Z_{r0}^{(nr,N,\text{REG})}, \tau^{(nr,N)2}\right) \\
\mathbb{P}\left(\beta_r^{(nr,N)}, \sigma_r^{(nr,N)2}\right) &\overset{\text{indep}}{\propto} 1 \\
\mathbb{P}\left(\gamma_0^{(nr,N)}, \gamma_1^{(nr,N)}, \tau^{(nr,N)2}\right) &\propto 1
\end{aligned}
$$

Now the difference stems from the use of another auxiliary information $Z_{r0}^{(nr,N,\text{REG})}$ potentially correlated with the total number of non-resident people in region r at the initial time period 0 (tourists, availability of tourist accomodation in region r –e.g number of beds, hotels, hostels,...–, one-day visitors,...). If this auxiliary information is not available or is not reliable enough at this space scale, we can pose instead the less informative specification:

$$
\begin{aligned}
\mu_{rt}^{(nr,N)} | \alpha_r^{(nr,N,R)}, \alpha_t^{(nr,N,T)} &\overset{\text{indep}}{\simeq} \alpha_r^{(nr,N,R)} + \alpha_t^{(nr,N,T)} \\
\alpha_r^{(nr,N,R)} | \beta^{(nr,N,R)}, \sigma^{(nr,N)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\beta^{(nr,N,R)}, \sigma^{(nr,N,R)2}\right) \\
(\alpha_t^{(nr,N,R)} - \alpha_{t-1}^{(nr,N,R)}) | \beta^{(nr,N,T)}, \sigma^{(nr,T)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\beta^{(nr,N,T)}, \sigma^{(nr,N,T)2}\right) \\
\mathbb{P}\left(\alpha_0^{(nr,N,R)}\right) &\propto 1 \\
\mathbb{P}\left(\beta^{(nr,N,R)}, \sigma^{(nr,N,R)2}\right) &\propto 1 \\
\mathbb{P}\left(\beta^{(nr,N,T)}, \sigma^{(nr,N,T)2}\right) &\propto 1 \\
\mathbb{P}\left(\alpha_0^{(nr,N,R)}\right) &\overset{\text{indep}}{\propto} 1
\end{aligned}
$$

We leave open the possibility to investigate both spatial and spatiotemporal models for $\mu_{rt}^{(\cdot)}$ by relating their values for different values of r and t.

Now, we must provide a model for births b_{rt} and deaths d_{rt}. When using mobile phone data, we implicitly assume that we are measuring at very fine time scales. The models will depend on the information about the registered births and deaths. If births and deaths are registered at the same time scale and assuming this information is very accurate, we can safely model b_{rt} and d_{rt} as degenerate variables $b_{rt} = b_{rt}^{\text{REG}}$ and $d_{rt} = d_{rt}^{\text{REG}}$.

If births and deaths are only provided at a larger time scale $\tau_{\text{REG}} \gg \tau_{\text{MNO}}$, then, as a working hypothesis, we shall assume they are distributed among the longer period τ_{REG} in a uniform way[2].

[2]For example, if $\tau_{\text{D}} = 1$ month and $\tau_{\text{MNO}} = 1$ hour, we assume that every hour births and deaths are equally distributed so that they sum up the monthly total.

With current birth and mortality rates in Europe, these working hypotheses should not impinge very much on the final result. Thus, let N_T denote the number of time intervals of length τ_{MNO} in the longer interval τ_{REG}. Let $p_T^{B)} = p_T^{(D)} = \frac{1}{N_T}$ be the constant probabilities for a person to be born or die in any of these intervals, respectively. We propose

$$\begin{aligned}(b_{r1},\ldots,b_{rT})\,|b_{r\tau_{\text{REG}}} &\overset{\text{indep}}{\simeq} \text{Multinomial}(b_{r\tau_{\text{REG}}}, p_T^B,\ldots,p_T^B)\\ (d_{r1},\ldots,d_{rT})\,|d_{r\tau_{\text{REG}}} &\overset{\text{indep}}{\simeq} \text{Multinomial}(d_{r\tau_{\text{REG}}}, p_T^D,\ldots,p_T^D)\end{aligned}$$

Next, we must provide models for the migration components. As in all demographic studies, this is the most volatile component of the demographic account and the most difficult to measure. At the time scales of mobile phone data traditional data sources cannot provide any reliable information, thus we will fully rely on mobile network data. Thus, we propose the following model for the resident component of the internal in-migration:

$$\begin{aligned}m_{rt}^{(r,II)}|\lambda_{rt}^{(r,II)} &\overset{\text{indep}}{\simeq} \text{Poisson}\left(\lambda_{rt}^{(r,II)}\left(\frac{n_{r,t-1}^{(r)}+n_{r,t}^{(r)}}{2}+\epsilon\right)\right)\\ \lambda_{rt}^{(r,II)}|\xi^{(r,II)},\mu_{rt}^{(r,II)} &\overset{\text{indep}}{\simeq} \text{Gamma}\left(\xi^{(r,II)},\frac{\xi^{(r,II)}}{\mu_{rt}^{(r,II)}}\right)\\ \mathbb{P}\left(\xi^{(r,II)}\right) &= \frac{\zeta_0^{(r,II)}}{(\zeta_0^{(r,II)}+\xi^{(r,II)})^2}\\ \log\left(\mu_{rt}^{(r,II)}\right) &= \alpha_r^{(r,II)}+\alpha_t^{(r,II)}\\ \alpha_r^{(r,II)}|\nu^{(r,II,R)},\tau^{(r,II,R)2} &\overset{\text{indep}}{\simeq} N\left(\nu^{(r,II,R)},\tau^{(r,II,R)2}\right)\\ (\alpha_t^{(r,II)}-\alpha_{t-1}^{(r,II)})|\nu^{(r,II,T)},\tau^{(r,II,T)2} &\overset{\text{indep}}{\simeq} N\left(\nu^{(r,II,T)},\tau^{(r,II,T)2}\right)\\ \mathbb{P}\left(\nu^{(r,II,R)},\tau^{(r,II,R)2}\right) &\propto 1\\ \mathbb{P}\left(\nu^{(r,II,T)},\tau^{(r,II,T)2}\right) &\propto 1\\ \mathbb{P}\left(\alpha_0^{(r,II)}\right) &\propto 1\end{aligned}$$

The small positive $\epsilon > 0$ is included to avoid zeros. For the non-resident components and the internal out-migration IO, external in-migration EI and external out-migration EO components, the models are completely similar (just changing the superscripts). Notice that, as pointed out by Bryant and Graham [2013], the formulas for the out-migration components offer a natural interpretation in terms of the "risk of exposure" to out-migrate as a fraction of the actual population of the cell. In the case of in-migration, we lack this interpretation but we keep the same structure since it seems to work in the case of administrative registers.

Notice that in all models there is ample room to increase the complexity in the time and spatial structure of the correlations. Regarding time, just a correlation with the immediately before time period has been considered in general, but more intricate time dependencies can be explored. Equally, regarding space, no spatial correlations have been considered. This is illustrated in the simple specification $\log\left(\mu_{rt}^{\text{II}}\right) = \alpha_r^{\text{II}} + \alpha_t^{\text{II}}$. Higher spatial complexity can be introduced at this point by involving correlations between different regions r.

4.2.2 Full model specification for the observation process

To propose an observation model we must first identify which parameters $\theta_{\mathbf{X}}$ can be used and which auxiliary covariates $\mathbf{Z}_{obs}$ are available. In the former case, the parameters will be given by the choice of models. The choice of auxiliary covariates is more subtle since it depends on the availability of data. We will focus on the national telecommunication regulator and will assume that they can provide information in a larger time scale $\tau_{\text{REG}} \gg \tau_{\text{MNO}}$ as a result of its supervising activity. In particular, we assume that they can provide the following data:

- National penetration rates (or market shares) $R_{r0}^{(NR)}$ for non-roamers for each region r and MNO. Basically these will be understood as detection probabilities for resident people (see below).
- Rates of roamers $R_{r0}^{(R)}$ with respect to the total of roamers in the territory for each region r and MNO. Basically these will be understood as detection probabilities for non-resident people (see below).

Let us firstly concentrate on the resident component of the total population:

$$
\begin{aligned}
n_{rt}^{(NR)\text{MNO}} | n_{rt}^{(r)}, p_{rt}^{(NR)} &\overset{\text{indep}}{\simeq} \text{Binomial}\left(n_{rt}^{(r)}, p_{rt}^{(NR)}\right) \\
p_{rt}^{(NR)} | \alpha_{rt}^{(NR)}, \beta_{rt}^{(NR)} &\overset{\text{indep}}{\simeq} \text{Beta}\left(\alpha_{rt}^{(NR)}, \beta_{rt}^{(NR)}\right) \\
\text{logit}\left(\frac{\alpha_{rt}^{(NR)}}{\alpha_{rt}^{(NR)} + \beta_{rt}^{(NR)}}\right) - \text{logit}\left(\frac{\alpha_{r,t-1}^{(NR)}}{\alpha_{r,t-1}^{(NR)} + \beta_{r,t-1}^{(NR)}}\right) | \beta_r^{(p,NR)}, \sigma_r^{(p,NR)2} &\overset{\text{indep}}{\simeq} N\left(\beta_r^{(p,NR)}, \sigma_r^{(p,NR)2}\right), \quad t = 1, 2, \dots \\
\text{logit}\left(\frac{\alpha_{r0}^{(NR)}}{\alpha_{r0}^{(NR)} + \beta_{r0}^{(NR)}}\right) | \gamma_0^{(p,NR)}, \gamma_1^{(p,NR)}, \tau^{(p,NR)2}, R_{r0}^{(NR)} &\simeq N\left(\gamma_0^{(p)} + \gamma_1^{(p)} \cdot \text{logit}\left(R_{r0}^{(NR)}\right), \tau^{(p)2}\right) \\
\log\left(\alpha_{rt}^{(NR)} + \beta_{rt}^{(NR)}\right) - \log\left(\alpha_{r,t-1}^{(NR)} + \beta_{r,t-1}^{(NR)}\right) | \delta_r^{(p,NR)}, \upsilon_r^{(p,NR)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\delta_r^{(p,NR)}, \upsilon_r^{(p,NR)2}\right) \\
\alpha_{r0} + \beta_{r0} | \xi^{(p,NR)}, \mu_{r0}^{(p,NR)} &\overset{\text{indep}}{\simeq} \text{Gamma}\left(\xi^{(p,NR)}, \frac{\xi^{(p,NR)}}{\mu_{r0}^{(p,NR)}}\right) \\
\mathbb{P}\left(\xi^{(p,NR)}\right) &= \frac{\zeta_0^{(p,NR)}}{(\zeta_0^{(p,NR)} + \xi^{(p,NR)})^2} \\
\log\left(\mu_{r0}^{(p,NR)}\right) | \epsilon^{(p,NR)}, \phi^{(p,NR)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\epsilon^{(p,NR)}, \phi^{(p,NR)2}\right) \\
\mathbb{P}\left(\beta_r^{(p,NR)}, \sigma_r^{(p,NR)2}\right) &\overset{\text{indep}}{\propto} 1 \\
\mathbb{P}\left(\gamma_0^{(p,NR)}, \gamma_1^{(p,NR)}, \tau^{(p,NR)2}\right) &\propto 1 \\
\mathbb{P}\left(\delta_r^{(p,NR)}, \upsilon_r^{(p,NR)2}\right) &\overset{\text{indep}}{\propto} 1 \\
\mathbb{P}\left(\epsilon^{(p,NR)}, \phi^{(p,NR)2}\right) &\propto 1
\end{aligned}
$$

These specifications rely implicitly on the simplifying assumption of every individual in the demographic account $\mathbf{Q}$ being possibly detected by the network. This is not clearly the case for children under, say, 10 years. We will first explore this observation model and then, upon an analysis of the

results, make it more complex in this respect[3]. The bottom line of the observation process is again to choose an initial time period in which we assimilate external data (from the regulator) to data from the network. In successive time periods, the observation process parameters evolve in a weakly informative trend in the model, since the regulator cannot provide information at this time scale.

For the case of roamers and non-resident population we can propose a similar model *mutatis mutandi*:

$$
\begin{aligned}
n_{rt}^{(R)\mathrm{MNO}} | n_{rt}^{(nr)}, p_{rt}^{(R)} &\overset{\text{indep}}{\simeq} \text{Binomial}\left(n_{rt}^{(nr)}, p_{rt}^{(R)}\right) \\
p_{rt}^{(R)} | \alpha_{rt}^{(R)}, \beta_{rt}^{(R)} &\overset{\text{indep}}{\simeq} \text{Beta}\left(\alpha_{rt}^{(R)}, \beta_{rt}^{(R)}\right) \\
\text{logit}\left(\frac{\alpha_{rt}^{(R)}}{\alpha_{rt}^{(R)}+\beta_{rt}^{(R)}}\right) - \text{logit}\left(\frac{\alpha_{r,t-1}^{(R)}}{\alpha_{r,t-1}^{(R)}+\beta_{r,t-1}^{(R)}}\right) | \beta_r^{(p,R)}, \sigma_r^{(p,R)2} &\overset{\text{indep}}{\simeq} N\left(\beta_r^{(p,R)}, \sigma_r^{(p,R)2}\right), \quad t = 1, 2, \ldots \\
\text{logit}\left(\frac{\alpha_{r0}^{(R)}}{\alpha_{r0}^{(R)}+\beta_{r0}^{(R)}}\right) | \gamma_0^{(p,R)}, \gamma_1^{(p,R)}, \tau^{(p,R)2}, R_{r0}^{(R)} &\simeq N\left(\gamma_0^{(p)} + \gamma_1^{(p)} \cdot \text{logit}\left(R_{r0}^{(R)}\right), \tau^{(p)2}\right) \\
\log\left(\alpha_{rt}^{(R)}+\beta_{rt}^{(R)}\right) - \log\left(\alpha_{r,t-1}^{(R)}+\beta_{r,t-1}^{(R)}\right) | \delta_r^{(p,R)}, \upsilon_r^{(p,R)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\delta_r^{(p,NR)}, \upsilon_r^{(p,NR)2}\right) \\
\alpha_{r0}+\beta_{r0} | \xi^{(p,R)}, \mu_{r0}^{(p,R)} &\overset{\text{indep}}{\simeq} \text{Gamma}\left(\xi^{(p,R)}, \frac{\xi^{(p,R)}}{\mu_{r0}^{(p,R)}}\right) \\
\mathbb{P}\left(\xi^{(p,R)}\right) &= \frac{\zeta_0^{(p,R)}}{(\zeta_0^{(p,R)}+\xi^{(p,R)})^2} \\
\log\left(\mu_{r0}^{(p,R)}\right) | \epsilon^{(p,R)}, \phi^{(p,R)2} &\overset{\text{indep}}{\simeq} \text{N}\left(\epsilon^{(p,R)}, \phi^{(p,R)2}\right) \\
\mathbb{P}\left(\beta_r^{(p,R)}, \sigma_r^{(p,R)2}\right) &\overset{\text{indep}}{\propto} 1 \\
\mathbb{P}\left(\gamma_0^{(p,R)}, \gamma_1^{(p,R)}, \tau^{(p,R)2}\right) &\propto 1 \\
\mathbb{P}\left(\delta_r^{(p,R)}, \upsilon_r^{(p,R)2}\right) &\overset{\text{indep}}{\propto} 1 \\
\mathbb{P}\left(\epsilon^{(p,R)}, \phi^{(p,R)2}\right) &\propto 1
\end{aligned}
$$

The models for the migration components of MNO data are similar. In all these cases, the interpretation of the models is straighforward, where the auxiliary rates provided by the regulator are used as proxies to the actual detection probabilities (which are unobserved).

4.3 Tourism statistics

This approach can also be followed for other statistical domains. In the case of tourism statistics, the target variables are commonly the number of tourists, thus the difference stems from the identification of tourists both in the mobile phone data set **X** and in the available information **Z**. In addition, we must propose also a *tourism account* **Q**.

For the *tourism account*, firstly we shall consider only tourist trips thus excluding same-day visits[4].

[3]For example, we could introduce an auxiliary covariate $Z_{r,obs}$ indicating whether there are schools in the region r or not.

[4]This methodology can be adapted *mutatis mutandi* to same-day visits.

According to European regulations we shall focus on *domestic*, *inbound*, and *outbound* tourism[5]. Then, as *tourism account* components we shall consider $\mathbf{Q} = (\mathbf{n}^{\mathrm{D}}, \mathbf{m}^{\mathrm{DI}}, \mathbf{m}^{\mathrm{DO}}, \mathbf{n}^{\mathrm{I}}, \mathbf{m}^{\mathrm{II}}, \mathbf{m}^{\mathrm{IO}}, \mathbf{n}^{\mathrm{O}}, \mathbf{m}^{\mathrm{OI}}, \mathbf{m}^{\mathrm{OO}})$, with the following definitions:

Variable	**Notation**	**Definition**
Total of domestic tourists	n_{rt}^{D}	Number of domestic tourists in region r at time t.
Domestic in-migration	m_{rt}^{DI}	Number of moves into region r during period t by domestic tourists.
Domestic out-migration	m_{rt}^{DO}	Number of moves out of region r during period t by domestic tourists.
Total of inbound tourists	n_{rt}^{I}	Number of domestic tourists in region r at time t.
Inbound in-migration	m_{rt}^{II}	Number of moves into region r during period t by domestic tourists.
Inbound out-migration	m_{rt}^{IO}	Number of moves out of region r during period t by domestic tourists.
Total of outbound tourists	n_{rt}^{O}	Number of domestic tourists in region r at time t.
Outbound in-migration	m_{rt}^{OI}	Number of moves into region r during period t by domestic tourists.
Outbound out-migration	m_{rt}^{OO}	Number of moves out of region r during period t by domestic tourists.

There are some evident accounting identities:

$$
\begin{aligned}
n_{rt}^{\mathrm{D}} &= n_{r,t-1}^{\mathrm{D}} + m_{rt}^{\mathrm{DI}} - m^{\mathrm{DO}}, \\
n_{rt}^{\mathrm{I}} &= n_{r,t-1}^{\mathrm{I}} + m_{rt}^{\mathrm{II}} - m^{\mathrm{IO}}, \\
n_{rt}^{\mathrm{O}} &= n_{r,t-1}^{\mathrm{O}} + m_{rt}^{\mathrm{II}} - m^{\mathrm{OO}}.
\end{aligned}
$$

As observed variables $\mathbf{X}$ we must investigate whether these or similar quantities can be directly or indirectly measured through the mobile telecommunication network. This is intimitely related to the preprocessing stage and data model construction described by Salgado et al. [2018], where variables like anchor points (place of residence, second-home anchor point, work-time anchor point, other regular anchor points, ...), usual environment, stay sections, trips, transit points, ... In this process we assume we can identify (needless to say, with an estimation error) which individuals are behaving as tourists and which others as non-tourists (i.e. we identify statistical units of our target population). As observed variables we can thus propose:

[5] *Domestic tourism* means visits within a Member State by visitors who are residents of that Member State; *inbound tourism* means visits to a Member State by visitors who are not residents of that Member State; *outbound tourism* means visits by residents of a Member State outside that Member State.

Variable	Notation	Definition
Total of domestic tourists	$n_{rt}^{\text{MNO, D}}$	Number of domestic tourists in region r at time t detected by the network.
Domestic in-migration	$m_{rt}^{\text{MNO, DI}}$	Number of moves into region r during period t by domestic tourists detected by the network.
Domestic out-migration	$m_{rt}^{\text{MNO, DO}}$	Number of moves out of region r during period t by domestic tourists detected by the network.
Total of inbound tourists	$n_{rt}^{\text{MNO, I}}$	Number of domestic tourists in region r at time t detected by the network.
Inbound in-migration	$m_{rt}^{\text{MNO, II}}$	Number of moves into region r during period t by domestic tourists detected by the network.
Inbound out-migration	$m_{rt}^{\text{MNO, IO}}$	Number of moves out of region r during period t by domestic tourists detected by the network.
Total of outbound tourists	$n_{rt}^{\text{MNO, O}}$	Number of domestic tourists in region r at time t detected by the network.
Outbound in-migration	$m_{rt}^{\text{MNO, OI}}$	Number of moves into region r during period t by domestic tourists detected by the network.
Outbound out-migration	$m_{rt}^{\text{MNO, OO}}$	Number of moves out of region r during period t by domestic tourists detected by the network.

As auxiliary covariates for the observation process $\mathbf{Z}_{obs}$ we again use the penetration rates $R_{r0}^{(NR)}$ and roamer rates $R_{r0}^{(R)}$ provided by the regulator.

As auxiliary covariates for the system process $\mathbf{Z}_{sys}$ we focus on the tourism statistics produced by NSIs according to the European regulations: estimates of total number of domestic $\widehat{N}^D$, inbound $\widehat{N}^I$, and outbound tourists $\widehat{N}^O$, on the one hand, and the tourist accomodation occupancy rates $\widehat{A}^D$ by country of residence (resident or non-resident), on the other hand.

The specifications of both the system and observation models can be worked out in the same spirit. This will be undertaken elsewhere, since details now seem a bit more complex thus models are also expected to gain in complexity (auxiliary information are estimates, probably land use will be needed, ...).

5 Discussion

This is a proposal for a general framework illustrated with a concrete example. Many details need further discussion:

1. This inferential framework is built on top of a set of data sets or databases entering into the proposals with different roles. On the one hand, we have the traditional survey or administrative data compiled and produced by NSIs. In general, administrative registers show a nearly complete coverage of the target population at stake or, at least, large enough

to provide valuable information at different scales of territorial disaggregation (let us think e.g. of the population register or the electoral roll). In this sense, administrative registers can be favoured in their use as auxiliary covariates in the framework because a priori we can reach more reliable levels of geographical dissagregation. In the case of survey data, this is hardly the case (otherwise the small area estimation problem would be trivial). Thus, these pose a limit in this level of breakdown. When auxiliary covariates cannot provide reliable information at that level of disaggregation, models should include some form on benchmarking regarding known aggregates. On the other hand, we make an intensive use of mobile phone data. This data ecosystem is very rich and complex (although highly structured) and it is compulsory to make clear what our starting point is to apply this inferential framework. We are building on our preceding efforts to propose an end-to-end statistical process with different steps. Starting from raw telecommunication data, firstly we need some non-negligible preprocessing to prepare these data for their further statistical analysis. In particular, both the space and time attributes of each event (call, SMS/MMS, Internet connection, ping, ...) must be undertaken using as much information at hand as possible. The geolocation of these network events is especially critical. After that, we basically have a set of events with a pseudonymised ID, space and time attributes, and some auxiliary variables (in special, the roaming/non-roaming character of the subscriber). Then, a comprehensive data model should be followed to build a database with this event information per subscriber together with variables like anchor points (place of residence, second-home anchor point, work-time anchor point, other regular anchor points, ...), usual environment, stay sections, trips, transit points, ... This involves the design, development, and implementation of algorithms to derive this information from the data. The observed data we make use as observed variables $\mathbf{X}$ in this inferential framework are indeed an aggregation of these derived variables. We take all this procedure for granted. We will deal with it elsewhere in the project.

2. A first critical issue arises regarding the different scales of the data sources to be combined. Survey and administrative data provide data typically at a monthly time scale at most. Regarding the geographical scale, survey data cannot reach a high level of breakdown without entering into reliability problems (hence the small area estimation problem), whereas administrative data in some cases (e.g. the population register) can do. Mobile network data can reach unprecedented scales of disaggregation both in time and in space. How to optimally combine both types of data sources is not trivial. In our models we have followed one of the assumptions introduced in our preceding work [Salgado et al., 2018] by which we identify a time instant at the mobile phone data scale in which we can assimilate data from both scales. This is also assumed in other approaches to estimate population size with mobile phone data. The general idea is to provide an initial estimation using auxiliary information from administrative registers and then let the system evolve at the smaller time scale (as e.g. hours). When new data from administrative registers are available, the system is reinitiated. This is not a very sophisticated way of combining both time scales, but it is a starting working assumption. Furthermore, this way of combining these scales partially motivates our choice of roles ($\mathbf{Z}$ or $\mathbf{X}$) for the administrative data (see next point).

3. We want to underline that the application of this inferential framework begins by choosing the role of available information either as observed data (variables $\mathbf{X}$ in the framework) or as auxiliary covariates (variables $\mathbf{Z}$ in the framework). This amounts to choosing between complementing already existing and validated official figures with the estimates from mobile network data or building these figures anew. Although here we have focused on the former option, both exercises should be put in practice to analyse both results. Ultimately, it will be

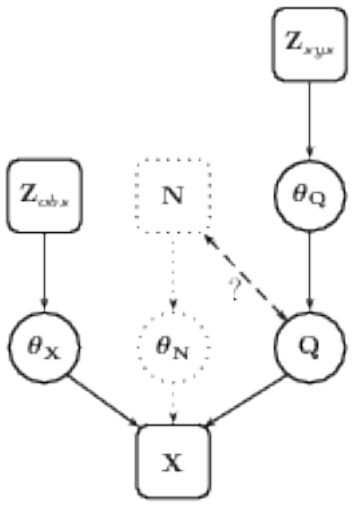

Figure 2: Model structure including simulated instrumental data

the domain expert criterion which should be considered. From a methodological point of view, we are exploring tools to produce statistics.

4. The demographic account we propose to consider provides information only about the geographic region r and the time slot t, aggregating on the other two dimensions (sex s and age group a). As mobile network data analysis makes advances enabling us to observe these variables, those will need to be included in the models.

5. In our model specifications, we have not taken into consideration the numerical consistency of nested levels of aggregation (both in time and in space). For the time being, for simplicity's ease we have preferred concentrating on a given level. Results will be analysed in this sense and, if necessary, models will be modified to guarantee this consistency.

6. The proposed models should be assessed using real data both for the system process and the observation process. Regarding the former, an exploratory analysis of real administrative data should be carried out and compared to the proposed models. Regarding the latter, in the absence of real data, semi-simulated data will be used as an instrumental tool. The simulation should not make use of the models in this proposal. It should implement an agent-based simulation starting from real data and simulating the population of subscribers of the MNO at hand and their sequences of stays and displacements. We illustrate this idea in figure 2. Notice that final estimates in the demographic account must be compared to the original simulated population.

7. Geostatistical considerations, in particular, spatial autocorrelations, can be easily introduced in this framework (with the corresponding increase in computational complexity). In general, we can formulate geospatial models by providing more complex variance structures in our specifications:

$$\log\left(\boldsymbol{\mu}^{(\cdot)}\right) \simeq \mathrm{N}\left(\mathbf{x}^T\boldsymbol{\beta}^{(\cdot)}, \sigma^{(\cdot)2}H^{(\cdot)}(\phi) + \tau^{(\cdot)2}\mathbb{I}\right),$$

where $H^{(\cdot)}(\phi)$ is a correlation matrix with the assumed spatial correlation structure. We will not delve into these models for the time being.

8. There are similarities and differences with the former model proposed by Salgado et al. [2018].

The former model can be thought of as another instance of the general approach depicted in figure 1, where the auxiliary covariates for the system process comes from the population register (as in the present model) and those for the observation process also comes from the regulator but at the same scale as the network, which is not reallistic. The most crucial difference arises from the use of transition probabilities in the former model where here we have not made use of this concept. In every time step t the network provides data to estimate transition probabilities from every cell to another and this is not considered in the models proposed in this document. Although the current proposal poses a rigorous Bayesian inferential framework, it still needs further refinement to include this valuable information.

9. Finally, the model for the external out-migration components is completeley ad-hoc and taken literally from the administrative data proposal. External data from admin or survey data should be explored at this point.

References

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

D. Salgado, M. Debusschere, O. Nurmi, P. Piela, E. Coudin, B. Sakarovitch, S. Hadam, M. Zwick, R. Radini, T. Tuoto, M. Tennekes, C. Alexandru, B. Oancea, M.E. Esteban, S. Salda na, L. Sanguiao, and S. Williams. Proposed elements for a methodological framework for the production of official statistics with mobile phone data. Deliverable 5.3 of Work Package 5 of ESSnet on Big Data, 2018. URL https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4d/WP5_Deliverable_5.3_Final.pdf.