Inference with mobile network data

Keywords: *Mobile phone data, inference, target population, sample representativity*

1 Introduction

Mobile network data, aka mobile phone data, stand as a promising data source for the production of official statistics. Several results already show their potential. However, the configuration of an end-to-end industrialised statistical production process using this source in combination with other data still needs further work to achieve usual quality standards in Official Statistics.

The statistical production process is a complex process, which entails the need to deal with many highly interrelated different aspects. Some of these have been recently approached in the first ESSnet on Big Data (2016-2018) [1] and are currently under further research in the second ESSnet on Big Data (2018-2020). Indeed, a protoprocess has been recognised with key stages to be developed:

- Access to raw telecommunication data.- The data ecosystem in a telecommunication network is rather complex ranging from signaling data in the radio network (very close to the antenna-device system) to the Call Detail Records (CDRs) mainly for billing purposes.
- Preprocessing of raw telecommunication data.- These data are not suitable for direct statistical exploitation and need some non-negligible preprocessing. The three main attributes to be obtained as a result of this procedure are the pseudonymised ID of each mobile device and the time and spatial attributes about the recorded event. Apart from these fundamental attributes, some others also need to be taken into accont (as roamer identification, event types...).
- Construction of a statistical microdata base.- Once fundamental variables for each network event have been computed, a database with data for each mobile device must be constructed in which different pieces of statistical information must be included. These are the country of residence, diverse anchor points (home, working-time, second-home...), usual environment, trips, stay and movement sections...
- Aggregation.- Next, some form of aggregation must be undertaken into different territorial cells (not to be necessarily coincident with the cells in the geolocation stage). This stage shall provide basically the number of mobile devices of the target population (general population, inbound tourists, outbound tourists, commuters...) per cell and time period under analysis.
- Inference to the target population.- This stage shall establish the connection between the data and the target population under analysis. In other words, this is directly connected to the representativity issue regarding the data. Since no probability sampling is under use, this is a non-negligible issue intimately linked to the quality of the final estimates to be disseminated by the NSI as official statistics.

- Dissemination.- Finally, outputs must be duly disseminated in time and form. The challenge now is the unprecedented degree of spatial and time breakdown of the information. Appropriate visualization and access channels must be provided according to users' needs.

This schematic view of the production process is in full agreement with the Reference Architecture and Methodology Framework with Mobile Network Data proposed by Ricciato [2]. For example, the so-called data-, convergence-, and statistics- layers in this Reference Framework can be clearly distinguished in this protoprocess.

In this work, we propose a generic inferential framework with hierarchical models based on a preliminary proposal by WPMPD [1] and on the similar approach combining administrative registers by Bryant and Graham [3]. By and large, this is in line with our general claim that there exist several common points with the use of administrative data as e.g. the conceptualization of these data for official statistics and the identification of error sources along the data generation process using the two-phase life-cycle model.

2 Methods

We propose to adapt the approach by Bryant and Graham [3] for the use of mobile phone data. This approach makes use of hierarchical models to account both for the combination of different data sources in a natural way and for the uncertainty around data and parameters used in the estimation process. We will adapt a Bayesian approach in a fully pragmatic way. The ultimate goal in Bayesian inference is to provide the so-called posterior distribution $\mathbb{P}\left(\mathbf{Y}\middle|\mathbf{X},\mathbf{Z}\right)$ 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.

Following closely the work by Bryant and Graham [3], we will model both the observation process (of mobile network data, mainly) and the dynamics of the system under analysis (a human population in our cases). Thus, we introduce the next notation for the data involved in our inference exercise:

- $\mathbf{Q}$.- We shall denote the target variables by $\mathbf{Q}$ instead of $\mathbf{Y}$. This follows Bryant's and Graham's notation to denote the target *demographic account* in estimating population counts. This should be understood in a generic sense as referring to any possible target population (general population, inbound tourists, ...).
- $\mathbf{X}$.- This will denote the observed data in our inference exercise, i.e. mainly mobile phone data.
- $\mathbf{Z} \rightsquigarrow \mathbf{Z}_{obs}, \mathbf{Z}_{sys}$.- 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 subscript to distinguish between the observation process and the system process.

The model is represented schematically in figure 1, which illustrates the interrelationship between all data. This model produces the posterior distribution $\mathbb{P}\left(\mathbf{Q},\boldsymbol{\theta}_{\mathbf{X}},\boldsymbol{\theta}_{\mathbf{Q}}\middle|\mathbf{X},\right.$

$\mathbf{Z}_{sys}, \mathbf{Z}_{obs}$), whose marginal $\mathbb{P}\left(\mathbf{Q}\middle|\mathbf{X}, \mathbf{Z}_{sys}, \mathbf{Z}_{obs}\right)$ constitutes the desired output. This figure entails

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

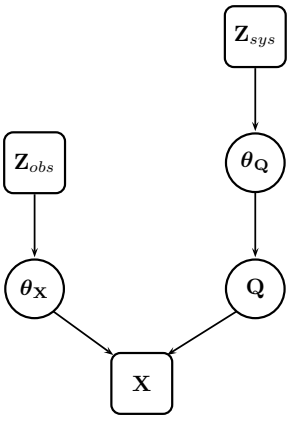

Figure 1: Hierarchical model for the inference on the demographic account $\mathbf{Q}$ using mobile phone data $\mathbf{X}$, auxiliary population registers $\mathbf{Z}_{sys}$ and national telecommunication regulator data $\mathbf{Z}_{obs}$. Round shapes denote unknown data whereas rectangular shapes denote known data.

This is the distribution to be ultimately generated by a simulation method (rejection algorithm, MCMC, Gibbs sampler,...). The model needs to be specified by providing each conditional distribution according to the modelling assumptions.

3 Results

As of this writing, we are in the interim between both ESSnets on Big Data. A first model for a **closed** population was proposed by WPMPD [1]. The different observed data sets involved in the model are the number $N_{it}^{MNO}(\rightsquigarrow \mathbf{X})$ of detected mobile devices per cell i and time period t, the population register $N_i^{Reg}(\rightsquigarrow \mathbf{Z}_{sys}, \mathbf{Z}_{obs})$ (assumed to be in a coarser time scale), and the MNO penetration rates $R_i(\rightsquigarrow \mathbf{Z}_{obs})$ per cell i and time period t (also assumed to be in a coarser time scale). It is important to notice that this population size determination based on mobile network data builds upon the estimates provided by the population register (hence the use of N_i^{Reg} as auxiliary covariates).

To specify the model we make two general assumptions: (i) there exists a short time period t_0 in which both the target population N_{i,t_0} and the official population N_i^{Reg} of each cell i can be assimilated and (ii) the mobility patterns are uncorrelated with the specific MNO individuals are subscribed to. Thus, from the first assumption we have at the initial time period t_0

$$\begin{aligned}
\left\{\mathbb{P}\left(\mathbf{X}\middle|\mathbf{Q}, \boldsymbol{\theta}_{\mathbf{X}}\right)\right\} \qquad & N_i^{\mathrm{MNO}}(t_0) \overset{\text{indep}}{\simeq} \text{Binomial}\left(N_i(t_0), p_i(t_0)\right), \quad i = 1, \ldots, I \\
\left\{\mathbb{P}\left(\mathbf{Q}\middle|\boldsymbol{\theta}_{\mathbf{Q}}\right)\right\} \qquad & N_i(t_0) \overset{\text{indep}}{\simeq} \text{Poisson}\left(\lambda_i(t_0)\right), \quad i = 1, \ldots, I \\
\left\{\mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{Q}}\middle|\mathbf{Z}_{sys}\right)\right\} \qquad & \lambda_i(t_0) \overset{\text{indep}}{\simeq} f_{\lambda i}(\lambda_i; N_i^{\mathrm{Reg}}) \quad \lambda_i(t_0) > 0, \quad i = 1, \ldots, I \\
\left\{\mathbb{P}\left(\boldsymbol{\theta}_{\mathbf{X}}\middle|\mathbf{Z}_{obs}\right)\right\} \qquad & p_i(t_0) \overset{\text{indep}}{\simeq} \text{Beta}\left(\alpha_i(t_0), \beta_i(t_0)\right), \quad i = 1, \ldots, I \\
& (\alpha_i(t_0), \beta_i(t_0)) \overset{\text{indep}}{\simeq} \frac{f_{ui}\left(\frac{\alpha_i}{\alpha_i+\beta_i}; R_{it}\right) \cdot f_{vi}\left(\alpha_i + \beta_i; N_i^{\mathrm{REG}}\right)}{\alpha_i + \beta_i}, \quad i = 1, \ldots, I.
\end{aligned}$$

The second assumption is implemented as

$$N_i(t_n) = \left[N_i(t_0) + \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ji}(t_0, t_n) N_j(t_0) - \sum_{\substack{j=1 \\ j \neq i}}^{I} p_{ij}(t_0, t_n) N_i(t_0) \right], \quad i = 1, \ldots, I$$

$$\mathbf{p}_{i\cdot}(t_0, t_n) \overset{\text{indep}}{\simeq} \text{Dirichlet}\left(\alpha_{i1}(t_0, t_n), \ldots, \alpha_{iI}(t_0, t_n)\right), \quad i = 1, \ldots, I$$

$$\alpha_{ij}(t_0, t_n) \overset{\text{indep}}{\simeq} f_{\alpha ij}\left(\alpha_{ij}; \frac{N_{ij}^{\text{MNO}}(t_0, t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right), \quad i, j = 1, \ldots, I.$$

The prior densities $f_{\lambda i}$, f_{ui}, f_{vi}, and $f_{\alpha ij}$ are chosen as weakly informative unimodal densities on their second argument. Using the Spanish population register and the penetration rates from the national telecommunication regulator we have simulated aggregate data $N_{it_0}^{MNO}, N_{it_0}$ for the initial time period t_0 for a Spanish municipality. The population dynamics among the 12 cells is simulated with a stochastic displacement inversely proportional to their mutual physical distance for a one-week period at intervals of 15 min. This model has been implemented in the R package `pestim` [4]. Working on simulated data enables us to compare estimates (posterior distributions) with population data (see figure 2).

Complementarily, hierarchical models also enable us to use accuracy measures like credible intervals and posterior variance estimation. These can be further complemented with model assessment through posterior predictive checking and related measures offering official statisticians the possibility to connect mobile phone data with target populations under an inferential framework under objective evaluation (see [1] for some proposals on these simulated data).

4 Conclusions

We can conclude that hierarchical models stand as a versatile tool to approach the representativity issue of mobile network data. However, further exploration and research needs to be done beyond this academic example. Human populations are never closed in practice and new specifications need to be introduced to account for net inward and/or outward flows. Moreover, a set of different specifications and priors need to be analysed in terms of their outputs with simulated data so that their application on real mobile network data (where population data are unknown) can be better understood.

This is ongoing work in the ESSnet on Big Data (2018-2020), whose latest results will be hopefully presented during this conference.

References

[1] Work Package on Mobile Phone Data. ESSnet on Big Data (2016-2018). https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Mobile_phone_data1.

[2] F. Ricciato (2018). Towards a Reference Methodological Framework for the processing of mobile network operator data for official statistics. International Conference Mobile Tartu 2018. 27-29 June, 2018. Tartu (Estonia). https://ec.europa.eu/eurostat/cros/system/files/rmf_mobiletartu2018_ricciato_printout.pdf.

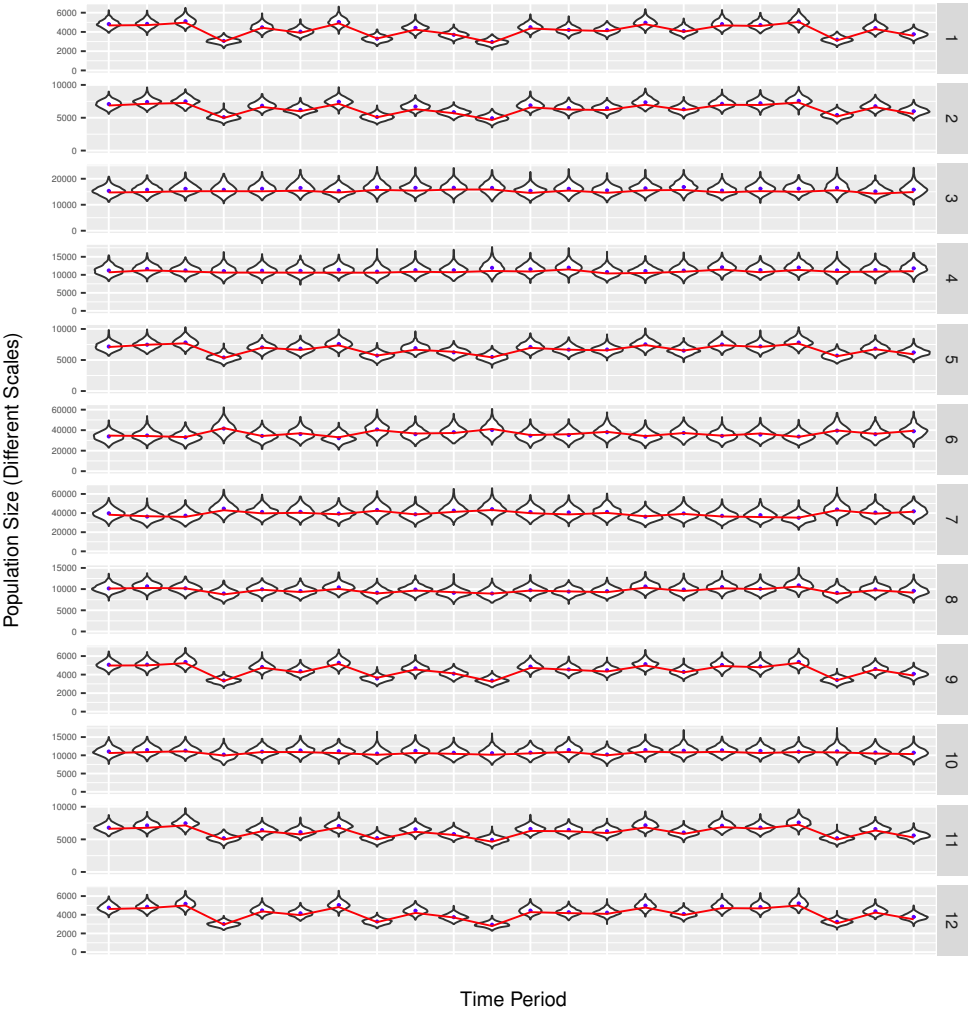

Figure 2: Posterior distributions per cell i and time period t for the population size in comparison with the (simulated) population data (in red).

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

[4] D. Salgado, B. Oancea, L. Sanguiao, and C. Alexandru (2018). pestim: Population Estimations Using Mobile Phone Data. R package version 0.1.0. https://github.com/MobilePhoneESSnetBigData/pestim.