Estimation of population counts combining official and mobile phone data

D. Salgado (coord.), Statistics Spain (INE), david.salgado.fernandez@ine.es
C. Alexandru, INS, ciprian.alexandru@insse.ro
M. Debusschere, Statistics Belgium (StatBel), marc.debusschere@economie.fgov.be
M.E. Esteban, Statistics Spain (INE), elisa.esteban.segurado@ine.es
O. Nurmi, Statistics Finland, ossi.nurmi@stat.fi
B. Oancea, INS and University of Bucharest, bogdan.oancea@insse.ro
P. Piela, Statistics Finland, pasi.piela@stat.fi
R. Radini, ISTAT, radini@istat.it
B. Sakarovitch, INSEE, benjamin.sakarovitch@insee.fr
S. Saldaña, Statistics Spain (INE), soledad.saldana.diaz@ine.es
L. Sanguiao, Statistics Spain (INE), luis.sanguiao.sande@ine.es
M. Tennekes, Statistics Netherlands (CBS), m.tennekes@cbs.nl
S. Williams, ONS, susan.williams@ons.gov.uk
M. Zwick, DESTATIS, markus.zwick@destatis@de

Abstract

Beyond doubt, mobile phone data stand as one of the most promising Big Data sources for the production of official statistics. In consonance, in the recent ESSnet on Big Data participated by 22 partners of the European Statistical System (ESS) a work package was completely devoted to the access to these data, the development of statistical methodology, the analysis of IT tools and of quality issues to make this promising information source become a regular resource in the production of official statistics.

We offer a summary of the works conducted in this work package, going from the intricate issue of accessing diverse forms of mobile phone data (microdata/aggregated data) over setting up an inferential framework to use aggregated mobile phone data in combination with official data to produce population counts, to the development of some IT tools for providing a proof of concept and first analytical results upon real data. All these enter as relevant factors in the quality assessment of the final estimates.

As explained in the results of the ESSnet, although we have been able to collect enough real data as to conduct the analytical study, the access to mobile phone data is still an open question which needs further work within the ESS and the European Union. A first set of conclusions and guidelines for partners of the ESS have been obtained.

Regarding the statistical methodology, unable to use traditional survey sampling techniques, we have

explored the use of hierarchical statistical models as in ecological sampling to propose a generic inferential framework for the counts of diverse target populations (commuters, resident tourists, inbound tourists, general population,. . .).

The analysis is completed providing software tools to implement this methodological proposal, showing a proof of concept with both simulated and real data, and assessing the quality of the final estimates.

Keywords: mobile phone data, access, production framework, hierarchical model, R package, quality indicator

1. Overview

The recent ESSnet on Big Data has aimed at "the integration of Big Data in the regular production of official statistics, through pilots exploring the potential of selected Big Data sources and building concrete applications". The work package 5 has been devoted to the analysis of mobile phone data, i.e. to the analysis of the digital trace of mobile devices left in telecommunication networks so that essentially spatiotemporal information on individuals can be potentially used to produce official statistics.

As all work packages centred on concrete data sources, our WP deals successively with the access, the statistical methodology, the IT infrastructure, and diverse quality issues to use and integrate this new data source into the standard production of official statistics. Lessons for the ongoing research within the ESS are also learnt. This structure is fully motivated by the bottom-up hands-on approach followed in this whole ESSnet.

The current contribution to this edition of the European Conference on Quality in Official Statistics (Q2018) presents the main findings reached in this project. Detailed results can be found in the deliverables of the project (WP5, 2016a, 2017, 2018a,b,c). This document is structured following the project stages, namely, access (section 2), methodology (section 3), IT infrastructure (section 4), quality issues (section 5), and considerations for the future research (section 6).

2. Access

Mobile phone data (more appropriately, mobile network data) are not public data. They are generated, stored, and processed in the private complex information systems owned by mobile network operators (MNOs hereafter). Accessing these data is a highly intricate challenge full with subtleties of diverse nature. The whole first phase of the ESSnet project for WP5 was devoted to this question, after which a decision was taken together with Eurostat whether to go on with this work package or not, since

access was granted in very limited conditions.

Our first action was to take stock of the (then) current access to mobile phone data across the ESS (WP5, 2016a). We designed and administered a questionnaire in September, 2016. The concrete conditions for accessing these data vary not only from country to country but also from MNO to MNO. In all cases, access for standard production conditions has not been reached in any country and only one-off agreements have been reached for research purposes, mainly for this project.

A workshop was organised in Luxembourg in 2016 between NSIs and European MNOs to facilitate an exchange of ideas and opinions from both sides. Presentations and detailed minutes of this meeting are publicly available (WP5, 2016b). Although a generally valid description embracing all situations is unattainable, we can conclude that access to these data within the ESS will take further work hand in hand with European MNOs. Basically, it is the perception of risks that blocks the access to mobile phone data for the ESS. These risks take diverse forms:

- Legal issues.- Legal obstacles are immediately adduced to deny access to these data for NSIs. By and large, current national and European legislations, with the support from national Data Protection Authorities, are enough to provide legitimate access to NSIs. However, an agreement is still out of reach.

- Operational risks.- Given the volume and sensitivity of data, the concrete operational conditions to access these data are key. Currently, in-situ access and processing at MNOs' premises stand up as a viable solution. Again, no agreement has been reached so far (with the exception of INSEE in France for research purposes).

- Costs.- Surprisingly, the golden principle by which Official Statistics does not pay for their data (otherwise this public service would be extremely endangered) is not widely known or even not clearly accepted. However, in connection with the preceding point, there exist costs associated to the operational procedures to access mobile phone data (as per the data collection costs present in currrent statistical production), but a detailed account of these costs is still unknown for their consideration in NSIs' budgets. More work hand in hand with MNOs is needed.

- Collision of interests.- In those MNOs investing in the statistical exploitation of their data, there seems to be a clear perception of collision of interests between

the private and public sectors. Basically, if NSIs produce this information, their new line of business seems to be under serious risk. In WP5 we have the conviction that this will not be the case, rather on the contrary, not only is there ample room for both public and private interests to coexist but more importantly the mutual collaboration will reinforce the quality of all statistical outputs.

All in all, we identify **mutual trust** as the key ingredient to unblock this situation. The construction of this mutual trust needs a close collaboration upon concrete projects which enable both NSIs and MNOs to empirically test how mobile-phone-data-based statistical products can be jointly provided to society.

3. Methodology

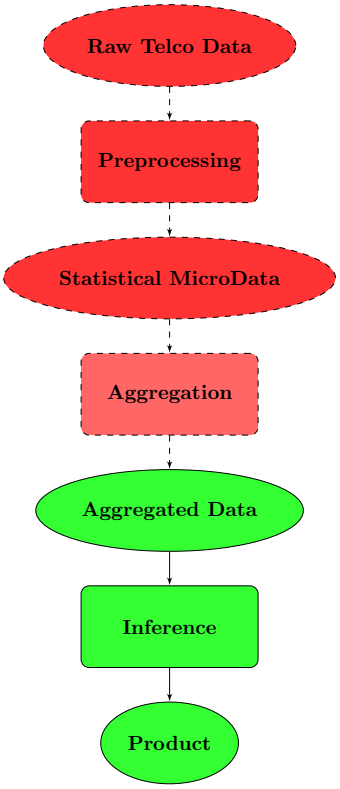
Based upon our experiences in accessing data and looking for a starting point to find and propose statistical methods to process this sort of data, we started our deliverable on methodology by revisiting the definition of Big Data *for Official Statistics*. Instead of the three Vs widely known and cited to introduce Big Data, we claim that for Official Statistics there exist more important features to be taken into account:

- Data refer to **third people** and not to data holders;
- Data are **central in data holders' economic activity**;
- Data **lack statistical metadata** (since they are generated for very different purposes).

The first two characteristics clearly lie behind some of the aforementioned issues to access data. The third feature is an essential trait for methodological considerations. Notice how administrative data also share these characteristics. Thus, we claim that existing tools for the use of admin data in official statistical production, with due modifications, are still valid for mobile phone data.

Although we have not been able to produce a concrete statistical output based on an end-to-end process, we do provide a set of elements for the construction of a production framework of official statistics based on these data. A schematic representation of the whole process is depicted in figure 1. With oval elements we represent different forms of data whereas with rectangular shapes we represent process steps. In red we depict those elements and phases of the process under no control by NSIs whereas in green we depict those elements and phases under control by NSIs in the project.

Figure 1. Sequence of large processing steps of mobile phone data.



Data accessed in the current process are in aggregated form (exceptions under limited conditions are INSEE, ISTAT, and CBS).

The process starts with the raw telecommunication data generated by the mobile device-antennae interaction in the telecommunication network. This form of data cannot be exploited for statistical purposes and thus needs preprocessing to turn them into statistical microdata. These are data at the mobile device level basically gathering information about their identification, time attributes, and spatial attributes together with some extra variables depending on the network (event types, duration, ...). They can be aggregated to produce the number of detected individuals at each territorial cell in a sequence of time instants. The final step infers from these aggregated data the estimates for the target population at stake. This is the ultimate goal of the whole process: to produce high-quality official statistics for a target population.

Concrete methodological proposals were provided for different elements of this process:

- The generation of data and the identification of error sources can be undertaken using the two-phase life-cycle model by Zhang (2012) (see Reid et al. (2017) for a wider adaptation to admin data). We claim that the Total Survey Error paradigm, duly adjusted, is still valid for mobile phone data. Especially, we underline the

increasing relevance of statistical methods to go from events to statistical units (from network events such as calls, SMS/MMS, Internet connections, pings, ... to individuals in the case of mobile phone data).

- The assignment of spatial attributes to network events is a key step in the generation of statistical microdata. This step strongly depends on data availability. Three scenarios of increasing complexity were explored:
  - Using only the location of the antennae, a Voronoi tessellation of the territory was computed to assign a geolocation to each event. Since this tessellation technique does not take into account both the directional character of antennae and the overlapping nature of the covered territorial cells, a first conclusion is the need for more sophisticated methods.
  - Taking into account the directional character of antennae and depending on the morphology of the territory and the estimated population density of subscribers, so-called Best Service Areas were computed and then used to locate the events (hence the individuals). Although offering much better results than Voronoi techniques, the non-overlapping assumption still poses some limitations. These areas are to be computed by MNOs since not all information is under NSIs' control.
  - Finally, considering both the directional and overlapping character of the antennae cells, a Bayesian approach to estimate the probability that an event (hence a mobile device) is located in a given territorial cell was followed. The probabilities are constructed using the prior information of each antennae and the geographical partition of the territory at stake. The likelihoods are computed using the signal strength. Should we have access to more variables, more sophisticated computations could be undertaken with this same structure.
- Once pseudonymised ID, spatial, and time attributes have been computed and assigned to each unit, a data model must be constructed containing diverse information elements for further producing statistical outputs of interest. These range from stay and movement sections (conceptually we are indeed observing sequences of stay and movement periods) to country of residence, anchor points (work/home/...), usual environment, trips, ...These are to be complemented with official data such as geographical administrative units, cell grids, etc.

- The inference exercise (possibly including the aggregation of microdata as an initial step) connecting aggregated data at each cell with the target population cannot follow the traditional probability sampling scheme, since no probabilistic sample selection can be undertaken. An alternative inference model must be used. Important points to consider are:

  – We argue that the concept of representativity must be duly understood and not to request from new data sources something which is not already present in traditional sources and probability sampling. Representativity is not a mathematical concept. We should pursue unbiased estimates with as lower variance as possible. Certainly, there will be selection biases and new elements such as model checking and model assessment will be needed.

  – We have adapted a hierarchical model already used by ecologists to solve the so-called species abundance problem to estimate population counts. The main working assumptions are:

    * At t_0 individuals are assumed to be physically in the territorial cell of auxiliary admin/survey data.
    * Mobility patterns of individuals do not depend on the concrete MNO they are subscribed to.

  The key parts in the specification of the model to estimate the number of individuals $N_i(t_n)$ at each each cell i and time period t_n are:

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

$$N_i^{\text{MNO}}(t_0) \simeq \text{Binomial}\left(N_i(t_0), p_i(t_0)\right), \quad i = 1, \ldots, I, \tag{1b}$$

  where $p_{ij}(t_0, t_n)$ are detection probabilities of individuals moving from cell i to cell j. The random variables $N_i(t_0)$ and $p_{ij}(t_0, t_n)$ are further specified according to prior probability distributions with their corresponding (hyper-) parameters modelled using our available data (from official population registers, survey data, and mobile network data).
  This is not intended to provide a definitive solution for the inference stage of the process, but to set up the first elements for an inferential framework in which the official statistician can adapt the model to the concrete inference

exercise at stake. We have followed a Bayesian approach to fit the model.

4. IT tools and infrastructure

Not having full access to data to undergo an end-to-end statistical process to produce a concrete output, we have concentrated on the IT part of the aforementioned elements for a production framework. Three main outputs have been provided:

- As stated above in relation to the access at MNOs' premises, we have provided a general description of an IT platform to access mobile phone data in situ.
- We have developed an R package called `mobloc` (Tennekes, 2018) implementing the Bayesian approach to geolocate network events based on the signal strength.
- We have developed an R package called `pestim` (Oancea et al., 2018) implementing the aforementioned statistical model to estimate population counts.

Both packages are freely available at WP5's Github page (WP5, 2018d).

5. Quality

Quality must be an ultimate goal in the production of official statistics and also for new data sources. Challenges arise when using mobile phone data which can be derived from the new changes in the production process depicted above. We have focused basically on two aspects of quality in the project. On the one hand, we have made a first incursion on how the European Statistics Code of Practice (CoP hereafter) is going to be affected according to our preceding proposals. On the other hand, we have made proposals to deal with the accuracy dimension of quality in the context of the new inference model for the production of official statistics using mobile network data.

Regarding the CoP, we have briefly analysed principle by principle suggesting how each one will potentially be affected by the use of mobile phone data for the production of official statistics. In summary, we have identified the three main factors as sources of change: (i) MNOs will be an **active part of the production process**; (ii) we will need a change of **inferential paradigm** from design-based to model-based (even Bayesian), and (iii) there will be an unprecedented **higher** degree of spatial and temporal **breakdown** in outputs. These three factors will affect the CoP in a cutting-cross way.

As the accuracy dimension is the most traditional measure of quality in Statistics, we have focused on producing accuracy measures in the context of the inference stage

depicted above. Having followed a Bayesian approach, since the output of the modelling exercise is a posterior distribution for the number of individuals $N_i(t_n)$ in each cell i and time period t_n, we have the conditions to produce any statistical indicator at will. In this line, the traditional confidence intervals and coefficients of variation can now be replaced by credible intervals (at least three alternative versions can be computed) and posterior coefficients of variation (which now can be also computed using such robust measures as the posterior interquartile range, the posterior median and similar robust indicators).

As a novel element, we now need to check and assess the goodness of fit of the model to make sure that final estimates are not provided upon a useless model (hence starting from inappropriate prior hypotheses). The core element here is the posterior predictive distribution by which we check whether we can reproduce the input data (mobile network data) firstly estimating the hyperparameters and then using the model to generate replicated mobile phone data.

6. The future research

The results in this work package constitute a first step and more research is needed. Different recommendations can be provided for this future research:

- The linear structure access-methodology-IT-quality has proven not to be the most efficient strategy to conduct this research, since the access to mobile phone data is currently blocked. We recommend to follow a parallel double-track with access issues on one track and methodology-IT-quality on the other track. They must advance in parallel working on simulated data as much as possible until real data can be used in optimal conditions.

- A track of research on how to simulate mobile phone data must be initiated, including simulations of the whole population. This will enable us to advance on the second track and to test both hypotheses and statistical models.

- The geolocation of network events must be further investigated including accuracy issues to build a full data model providing service for the production of any kind of statistics (population, tourism, transport, ...). The construction of this model must contain the spatiotemporal interpolation of data.

- The inference framework must be enriched with more hypotheses and more sophisticated and realistic models.

- The Quality Assurance Framework must be revised in terms of potential new indicators. The Total Survey Error paradigm must be adapted to this new data source in the search for the identification of all error sources.

Hopefully, in the forthcoming second ESSnet on Big Data we will go on with the lines of work initiated here.

7. References

Oancea, B., Salgado, D. and Sanguiao, L. (2018). *pestim: Population estimations using mobile phone data*. R package version 0.1.0. Available at https://github.com/MobilePhoneESSnetBigData/pestim.

Reid, G., Zabala, F. and Holmberg, A. (2017). Extending TSE to administrative data: A quality framework and case studies from Stats NZ, *Journal of Official Statistics* **33**(2): 477–511.

Tennekes, M. (2018). *mobloc: Mobile phone location algorithms and tools*. R package version 0.1.0. Available at https://github.com/MobilePhoneESSnetBigData/mobloc.

WP5 (2016a). Deliverable 5.1: Current status of access to mobile phone data in the ESS, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5_Deliverable_1.1.pdf.

WP5 (2016b). Workshop on Public-Private Partnerships for Mobile Phone Data for use in Official Statistics, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Meetings.

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

WP5 (2018a). Deliverable 5.3: Proposed elements for a methodological framework for the production of official statistics with mobile phone data, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.3.pdf.

WP5 (2018b). Deliverable 5.4: Some IT elements for the use of mobile phone data in the production of official statistics, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/65/WP5.Deliverable1.4.pdf.

WP5 (2018c). Deliverable 5.5: Some quality aspects and future prospects for the production of official statistics with mobile phone data, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/3/30/WP5_Deliverable_5.5_Preliminary_draft_version.pdf.

WP5 (2018d). Mobile Phone ESSnet Big Data, https://github.com/MobilePhoneESSnetBigData.

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