WP5 SGA2 Internal Document No. 3:

A high-level methodological proposal for count statistics based on mobile phone data

May 24, 2017

WP5 on Mobile Phone Data

ESSNET ON BIG DATA, 2017

1. Introduction: definition of the problem

This document contains a high-level methodological proposal to produce concrete statistical outputs using the mobile phone data sets compiled during the SGA-1. The proposal tries to follow the bottom-up approach of the general ESSnet so that we go from the particular to the generic in the methodology with a view in the future assessment of the output quality. The document suggests concrete directions of work to follow in order to produce the desired statistical outputs.

Both accounting for the compiled data sets s to be used as inputs and to assess the question whether aggregated data may be enough to produce good statistical outputs, we propose as a general framework to follow a two-step procedure:

1. From the microdata $\mathbf{z}_k$ of each mobile device $k \in s$ in the data set s to aggregates $\dot{Y}_{s_i}$ for each cell $s_i \subset s$ such that $s = \bigcup_{i \in \mathcal{I}} s_i$.
2. From the aggregates $\dot{Y}_{s_i}$ to the estimators $\hat{Y}_{U_i}$ of the variable of interest y in each population domain U_i such that $U = \bigcup_{i \in \mathcal{I}} U_i$.

The cell aggregates $\dot{Y}_{s_i}$ are intendedly written to denote that (i) they are sampling aggregates and **not** estimators of variable y in the domain U_i (hence the subscript s_i) and (ii) although referring to the domain U_i they account only partially for the target variable Y_{U_i} since they are computed using data from a single MNO (hence the dot). For example, they can be the number of mobile phones in domain i from a concrete MNO. Notice also that the population partition $\{U_i\}_{i \in \mathcal{I}}$ of interest (most usually an administrative geographical partition of national territories) induces a parallel partition into the data set s.

With this scheme we cover both situations in which NSIs have access either to data at the mobile device level or to data at an aggregated level. However, most of our input data sets are at an aggregated level only, thus we will especially focus on the second step in this note with some minor references to the first step.

Furthermore, since the statistical outputs of interest are in the domains of tourism and/or mobility, and, as official statisticians, daytime population is a must, from a methodological point of view we can focus on *count statistics*, i.e. on the estimation of population sizes, either populations of people, tourists, commuters, etc.

With these ingredients we will try to solve the following problem:

Problem 1.1. Let

- a finite population U of interest of unknown size N;
- a partition into population domains U_i such that $U = \bigcup_{i \in \mathcal{I}} U_i$, mostly geographical areas;

- either a mobile phone (micro)data set s or a collection of partial aggregates $\dot{N}_{s_i}$ of count statistics of interest.

We want to construct:

- an estimator $\hat{N}$ for the population size N;
- an accuracy measure for $\hat{N}$ (estimated mean square error or similar);
- estimators $\hat{N}_{U_i}$ for the population domain sizes N_i;
- accuracy measures for $\hat{N}_i$ (estimated mean square errors or similar). □

Notice that daytime population, tourist population or commuter population is included in this formulation (although these two cases will be foreseeably more difficult because tourists and commuters must be filtered from the input data sets or provided somehow by MNOs in the partial aggregated data).

2. The proposal: general lines

The problem 1.1 is neither a problem of estimation in a finite population (Särndal et al., 1992) nor a classical problem of statistical inference (Casella and Berger, 2002). It is a problem of estimation of population size. One of the richest field in methods to estimate population size is ecology (Manly and Navarro-Alberto, 2015), where capture-recapture methods are partially known to official statisticians in frame construction and other tasks. However, although several methods to estimate a population size can be found in this discipline, we feel that we should consider more ingredients from other fields for concrete parts of the problem. For example (see below), geostatistical considerations seem to be naturally in place (Schabenberger and Gotway, 2005) and the selection bias correction technique by Heckman (Heckman, 1979) stands as an interesting technique. Let us consider the problem as a jigsaw puzzle of statistical methods so that we need to assemble it in the optimal way to produce high-quality statistics.

Apart from the concrete technique to use (see below), a first issue to debate is the role of official data to estimate the size of the population using mobile phone data. In particular, we have to consider two possible roles for the official population figures in the estimation process, whatever it is. On the one hand, we can use them to make some kind of calibration procedure so that the estimated population $\hat{N}$ is made coincidental with the official figure N_{official} (an example can be found by Deville et al. (2014)). On the other hand, we can use official data not to calibrate the estimators of the quantities of interest so that we can compare the outputs from both sources. Both alternatives have pros and cons which we must elucidate.

Obviously we cannot explore in depth all ecological sampling techniques. To be concrete we propose to focus on the next method[1].

We will begin by the simplest scenario to be later on sophisticated progressively. Let us assume a 1:1 correspondence between mobile phones and statistical units (people, tourists, commuters, ...). Let p denote the probability of detecting a statistical unit $k \in U$ with the mobile network. The number Y of detected statistical units (thus of different mobile phones included in the data set s) follows a binomial distribution $Y \simeq B(N, p)$. Then we consider the following estimator of the population size N (Thompson, 2012):

$$\hat{N} = \frac{Y}{p}.$$

If p were known, $\hat{N}$ would be unbiased: $\mathbb{E}\left[\hat{Y}\right] = \frac{N \cdot p}{p} = N$. Furthermore, we can compute its variance: $\mathbb{V}\left[\hat{N}\right] = \frac{\mathbb{V}[Y]}{p^2} = \frac{N \cdot p \cdot (1-p)}{p^2} = N \cdot \frac{(1-p)}{p}$, which can be estimated straightforwardly: $\hat{\mathbb{V}}\left[\hat{N}\right] = \hat{N} \cdot \frac{(1-p)}{p}$ (note that this variance estimator is indeed unbiased).

When we have only the partial aggregates $\dot{Y}_{s_i}$ as input data, then $Y = \sum_{i \in \mathcal{I}} \dot{Y}_{s_i}$. When having microdata $\mathbf{z}_k$ a procedure to compute $\dot{Y}_{s_i}$ must be put in place previously to apply this formula.

This is clearly an overly simplistic scenario. More realistic assumptions must be progressively introduced.

3. The proposal: more realistic assumptions

We consider progressively more realistic assumptions in this estimation setting.

3.1. Estimation in each domain

As stated in the formulation of problem 1.1, not only are we interested in the global population size N but also in that of each population domain U_i. In this case, the same estimation setting can be followed by relaxing the assumption of equal detection probability for all domains U_i so that possibly we can have diverse detection probabilities p_i. As a crude assumption we can approximate p_i by the market shares of the MNO in each domain U_i and treat them as fixed parameters in the scheme. This is the approach apparently followed by De Meersman et al. (2016) (apparently using the same national market share for all domains). Then we would have estimators $\hat{N}_i = \frac{Y_i}{p_i}$ with analogous considerations. This requires the detection probabilities p_i to be known for each

[1] Recently, as we progress in our work, we are discovering in the literature a wealth of variants with a high degree of mathematical sophistication which need further exploration for our purposes. For the time being, we just want to suggest a general direction of work.

domain U_i.

Notice that no calibration procedure to the official figures is practised so that a comparison between both estimates is meaningful.

3.2. Estimation of the probability of detection

It is natural to expect that the detection probabilities p_i are not known and must be somehow estimated (Thompson, 2012). It appears as more realistic to model p_i:

$$\Phi\left(p_i\right) = \mathbf{x_i}' \cdot \boldsymbol{\beta}_i + e_i, \tag{1}$$

where Φ is a linking function, $\mathbf{x}_i$ are suitably chosen auxiliary variables, $\boldsymbol{\beta}_i$ are parameters and e_i are error terms. The search for the model and these auxiliary variables stand as the central part in this approach. This is where official data must play their role (e.g. official population density figures can be a potential candidate, since we expect higher detectability in those domains where population density is larger[2]).

Once modelled, we can compute the predicted values $\hat{p}_i$ for each domain and our estimators would be given by $\hat{N}_i = \frac{Y_i}{\hat{p}_i}$. Now, $\hat{N}_i$ becomes approximately unbiased provided that the estimation of the detection probability is precise: $\mathbb{E}\left[\hat{N}_i\right] = N_i + N_i \cdot \frac{\mathbb{V}(\hat{p}_i)}{p_i^2} + O\left(N_i \cdot \mathbb{E}\left(\frac{\hat{p}_i - p_i}{p_i}\right)^3\right)$. For the variance, $\mathbb{V}\left[\hat{N}_i\right] = \frac{\mathbb{V}Y_i}{p_i^2} + \frac{\mathbb{E}^2 Y_i}{p_i^2}\frac{\mathbb{V}Y_i}{p_i^2} + O\left(\frac{\mathbb{E}^2 Y_i}{p_i^2}\frac{\mathbb{V}(\hat{p}_i - p_i)^2}{p_i^2}\right) = N_i \cdot \frac{1-p_i}{p_i} + N_i^2 \cdot \frac{\mathbb{V}\hat{p}_i}{p_i^2} + O_4$. Approximately unbiased estimation of this variance needs further work.

3.3. More than one mobile device per person

The assumption of having one mobile device per person is clearly unrealistic. To deal with the real situation we need further auxiliary information. Let us first consider a concrete methodological proposal and later on we will consider how to obtain the involved auxiliary information.

Now we must distinguish between the size N_{md} of the mobile device population U_{md} and the size $N = N_U$ of the population of interest U. In the preceding sections we are clearly estimating N_{md}, i.e. $\hat{N}_{md} = \sum_{i \in \mathcal{I}} \hat{N}_{md,i} = \sum_{i \in \mathcal{I}} \frac{Y_i}{\hat{p}_i}$, since Y_i denotes the number of detected mobile devices.

Now we prove that having the mobile penetration rates τ_i allow us to account for the multiple devices per person. Let

- $N_{md,k}$ be the size of the subpopulation $U_{md,k} \subset U_{md}$ of devices belonging to a subscriber using k mobile devices;

[2] This clearly follows the rationale of mobile network designs.

- N_k be the size of the subpopulation $U_k \subset U$ using k mobile devices;
- $P_k = \frac{N_k}{N}$.

Then we have

$$\frac{N_{md,k}}{N_{md}} = \frac{k \cdot N_k}{N_{md}} = k \cdot \frac{N_k}{N} \cdot \frac{N}{N_{md}} = \frac{k \cdot P_k}{\frac{N_{md}}{N}} = \frac{k \cdot P_k}{\tau},$$

where τ is the penetration rate of mobile phones in the population.

Assuming $k_{\max}$ mobile devices per person at most (work mobile, personal mobile, tablet with mobile connection, ...), the population size N can then be decomposed as

$$\begin{aligned} N &= \sum_{k=0}^{k_{\max}} N_k = N_0 + \sum_{k=1}^{k_{\max}} \frac{N_{md,k}}{k} \\ &= P_0 \cdot N + \frac{N_{md}}{\tau} \cdot \sum_{k=1}^{k_{\max}} P_k \Rightarrow N = \frac{N_{md}}{\tau}. \end{aligned} \tag{2}$$

Equation (2) could have been established by definition. The goal, however, is to show explicitly that this definition covers the multiple devices per person issue.

Thus, recovering the breakdown in domains U_i, we can write

$$\hat{N}_i = \frac{\hat{N}_{md,i}}{\hat{\tau}_i} = \frac{Y_i}{\hat{\tau}_i \cdot \hat{p}_i}. \tag{3}$$

The variance and its estimation depend very sensitively on the source of data to compute $\hat{\tau}_i$, and $\hat{p}_i$.

This connects with the data source to compute $\hat{\tau}_i$. Firstly, if information exists in the MNO's databases to discern those mobile devices belonging to the same person, this must obviously be pursued. Secondly, since $\tau_i = \frac{N_{md,i}}{N_i}$, the key variable is to have the total number of mobile devices $N_{md,i}^{\text{official}}$ with some degree of breakdown U_i. This figure is potentially collected by national market regulators or the corresponding Ministry or similar (e.g. the ITU). Then estimators (3) are indeed ratio estimators calibrated to official population figures N_i^{official}:

$$\hat{N}_i = \frac{\hat{N}_{md,i}}{N_{md,i}^{\text{official}}} \cdot N_i^{\text{official}} = \frac{Y_i}{\hat{p}_i \cdot N_{md,i}^{\text{official}}} \cdot N_i^{\text{official}}. \tag{4}$$

Notice that in this case we are using official figures so that a comparison of the total population N is somewhat restricted in meaning. However, the estimates $\hat{N}_i$ can be useful to analyse the distribution of the population.

3.4. Geostatistical considerations

It should be clear by now that the accurate estimation of detection probabilities p_i is crucial in this approach. Thus, we should search for the optimal model for these quantities. Being the domains U_i essentially geographical domains, it is natural to analyse the potential contributions from geostatistical techniques.

Firstly, in geostatistics three types of data (thus of analyses) can be identified (Schabenberger and Gotway, 2005), namely (i) geostatistical data, (ii) lattice data, and (iii) point patterns. The essential statistical object in geostatistics is the spatial stochastic process $Z(\mathbf{s})$, where $\mathbf{s} \in D \subset \mathbb{R}^d$. When data correspond to fixed points $\mathbf{s}_j \in D$, we deal with geostatistical data. When data are aggregated in fixed geographical domains $D_i \subset \mathbb{R}^d$, we deal with lattice data. Finally, when data are random points $\mathbf{s}_j$, we deal with point patterns.

When analysing mobile phone data, their details are fundamental to choose the correct technique. If data are aggregated it seems clear that we are dealing with lattice data. Thus, when dealing with the partial aggregates $\dot{Y}_{s_i}$, we will investigate techniques for lattice data. When data are at the mobile phone level (microdata), the situation is subtler. The precision of the spatial attribute in the data set is essential to decide. Should we have such as a precision as to deal with individual spatial attribute for each record at the point level (i.e. spatial coordinates (x_k, y_k) for each mobile phone k), point pattern techniques appear as more appropriate, since the location of each mobile phone within the geographical domain D of analysis is expected to be random. If spatial attributes are coarse-grained, say, in antenna or LA or network cell positions, apparently we are again in the lattice data case, although deeper analysis is needed.

A cautious reader may wonder why not to apply a geospatial model directly to the number of mobile phones $\dot{Y}_{s_i}$. In this sense, the use of geostatistical techniques in this scenario is somewhat peculiar or restricted because the variable Y_i does not contain a full measurement of the number of mobile phones in the domain U_i, as in standard situations (unless we have data from all MNOs). Thus conventional techniques cannot be directly applied to model Y_i.

In consonance also with our preceding proposal, we will focus on modeling p_i for the fixed set of domains U_i. That is, we are dealing with lattice data. Thus, model (1) can incorporate geostatistical information:

$$\Phi(p(\mathbf{s})) = \mathbf{x}'(\mathbf{s}) \cdot \boldsymbol{\beta}(\mathbf{s}) + W(\mathbf{s}) + e(\mathbf{s}), \tag{5}$$

where $W(\mathbf{s})$ incorporates the geostatistical modeling information as a geospatial model. There exists a wealth of methods to model spatial interaction accounting for different situations (Schabenberger and Gotway, 2005) and research work needs to be done to

find an appropriate geospatial model in each case.

Variance estimation cannot be undertaken unless a explicit model is formulated.

4. Some comments

4.1. From microdata to partial aggregates

The preceding proposal focuses upon the estimation using the partial aggregates $\hat{Y}_i$. These can be the total number of mobile phones of people in general, of inbound tourists, of commuters, etc. For those having access to microdata, the complementary task is to compute these partial aggregates from the microdata.

The situation here is complex. Firstly an agreement of standard definition of (inbound/outbound/domestic) tourist, commuter, etc. must be put in place. Then this has to be implemented to filter them out of the microdata sets and compute the partial aggregates.

The methodology is diverse depending on the statistical domain. In this point we propose to use the internal technical report by Positium as a guide.

4.2. Heckman correction

In the preparation for the meeting in Madrid, Bogdan, from INSSE, proposed to consider the selection bias correction by Heckman (1979) as one of the potential tools for the modelling exercises derived from these non-probability sampling techniques. This is an excellent suggestion which will be exposed during the meeting.

4.3. Admin data methodology

In official statistics production, estimation based on administrative records is necessarily based on non-probability inference. A first look will also be shared during the meeting to see if some technique can be useful for our purposes.

4.4. The misalignment problem

In revising the literature, we found an important side-result in the simulation exercise made by Ricciato et al. (2015). Basically they showed how the geographical breakdown into territorial cells affects the quality of the estimation. Further reading led us to the so-called *misalignment problem* (see e.g. Banerjee et al. (2015) and references therein) where we learnt that the inference in geospatial models is sensitive to the territorial disaggregation. This should be somehow taken into account into our problem. Further reading is still ongoing.

4.5. Frequentist vs. Bayesian: Calibrated Bayes

In the literature consulted so far, either for ecological sampling or for geostatistics, as everywhere in statistics, you also face the eternal dilemma of making frequentist or Bayesian inferences. Design-based sampling inference is frequentist and now, not able to use it, we must make up our mind. We propose not to spend very much energy for the time being in this question and be as pragmatic as possible after firstly providing the due details for the above proposal. If something in the Bayesian domain is found to work good, let us go ahead. If, on the contrary, some frequentist technique produces good estimates, let us follow this direction. This suggestion tries to follow the avoidance of the *file drawer problem* mentioned in our preceding internal document. Let us publicly share what we find.

Bibliography

S. Banerjee, B.P. Carlin, and A.E. Gelfand. Hierarchical modeling and analysis for spatial data (2nd ed). CRC Press.

G. Casella and R.L. Berger. Statistical inference. Duxbury.

F. De Meersman, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, H.I. Reuter (2016). Assessing the Quality of Mobile Phone Data as a Source of Statistics. Q2016 Conference paper, June 2016.

P. Deville, C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel, and A.J. Tatem (2014). Dynamic population mapping using mobile phone data. PNAS 111, 15888–15893.

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

B.F.J. Manly and J.A. Navarro Alberto (eds.) (2015). Ecological sampling. CRC Press.

F. Ricciato, P. Widhalm, M. Craglia, and F. Pantisano (2015). Extraction of population density distribution from network-based mobile phone data. Link.

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

O. Schabenberger and C.A. Gotway (2005). Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC.

S.K. Thompson (2012). Sampling. Wiley.