A hierarchical model to estimate population counts from aggregated mobile phone data

Dept. Methodology and Development of Statistical Production. Statistics Spain (INE)

29 Jan, 2018

Contents

1 General introduction **1**

2 The complete model **2**

3 Computations **3**

4 A toy example **3**

5 First conclusions **9**

References **10**

1 General introduction

This document contains a hierarchical model to estimate population counts using a combination of aggregated mobile phone data and official data both at a given time instant and along a sequence of time periods. This work completes the preceding proposal which sets the model for the initial time instant estimates (cf. ESSnet on Big Data WP5 (2017)). Thus, in the subsequent we show how to extend the estimates along a sequence of time instants, not just the initial one.

Again we focus only on the methodological aspects of the proposal. Software programming aspects will be dealt with in another document. We provide a full view of the proposal to offer a complete view of this methodological approach. We directly concentrate on the model leaving the general framework and motivation for the corresponding deliverables.

Now the approach can be graphically represented by figure 1:

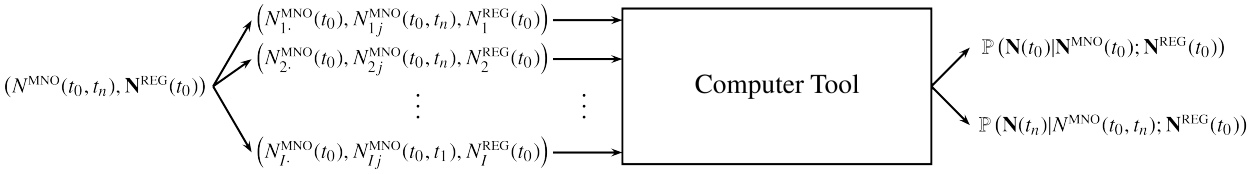


Figure 1: Schematic diagram for the output intended using mobile phone and official population data.

From the preceding stage of the whole production process with mobile phone data we get as input data for the final inference stage the number of individuals $N^{\text{MNO}}_{ij}(t_0,t_n)$ moving from territorial cell i to territorial cell j in the time interval (t_0,t_n) according to the mobile telecommunication network. These data will be combined with official data coming from, say, the population register or another source. At the end we intend to provide as outputs (i) the probability distribution of actual individuals in each territorial cell i at the initial time t_0 and (ii) the probability distribution of actual individuals at the time instants t_n for $n = 1, 2, \ldots$.

The two basic working assumptions are as follows:

1. To combine both aggregated mobile phone and official data we assume that a given time instant t_0 both population figures in each territorial cell can be equated to some extent. For the time being for ease of simplicity, we will take N_i^{Reg} as a fixed quantity without a prior distribution representing uncertainty in its knowledge. Therefore N_i^{Reg} will be fixed external parameters in the model.
2. The movements of individuals across the territory (from cell to cell) is assumed to be independent of being subscribers of a given Mobile Network Operator or another.

We propose a hierchical model supporting these hypotheses. The approach is not to think of the model as a definitive closed proposal, but as an initial step for a general methodological framework for the inference stage in which we can progressively introduce more complexities as the analysis on real data allows us to extract conclusions (e.g. introducing geospatial correlations among the territorial cells, dealing with the selection bias in relation with some demographic profiles – elderly people, children, ...).

2 The complete model

Let $p_{ij}(t_0, t_n)$ denote the probability for an individual to move from cell i to cell j in the time interval (t_0, t_n). Let $N_{ij}^{\text{MNO}}(t_0, t_n)$ the number of individuals moving from cell i to cell j according to the network. As usual, we denote $N_{i\cdot}^{\text{MNO}}(t_0) = \sum_{j=1}^{I} N_{ij}^{\text{MNO}}(t_0, t_n)$. The complete model which we propose is specified by:

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

$$\mathbf{p}_{i\cdot}(t_0,t_n) \simeq \text{Dir}\left(\alpha_{i1}(t_0,t_n),\dots,\alpha_{iI}(t_0,t_n)\right), \quad \mathbf{p}_{i\cdot}(t_0,t_n) \perp \mathbf{p}_{j\cdot}(t_0,t_n), \quad i \neq j = 1,\dots,I \tag{1b}$$

$$\alpha_{ij}(t_0,t_n) \simeq f_{\alpha ij}\left(\alpha_{ij}; \frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right), \quad i = 1,\dots,I \tag{1c}$$

$$N_i^{\text{MNO}}(t_0) \simeq \text{Bin}\left(N_i(t_0), p_i(t_0)\right), \qquad N_i^{\text{MNO}}(t_0) \perp N_j^{\text{MNO}}(t_0), \quad i \neq j = 1,\dots,I \tag{1d}$$

$$N_i(t_0) \simeq \text{Po}\left(\lambda_i(t_0)\right), \qquad N_i(t_0) \perp N_j(t_0), \quad i \neq j = 1,\dots,I \tag{1e}$$

$$p_i(t_0) \simeq \text{Beta}\left(\alpha_i(t_0), \beta_i(t_0)\right), \qquad p_i(t_0) \perp p_j(t_0) \quad i \neq j = 1,\dots,I \tag{1f}$$

$$(\alpha_i(t_0), \beta_i(t_0)) \simeq \frac{f_{ui}\left(\frac{\alpha_i}{\alpha_i+\beta_i}; \frac{N_i^{\text{MNO}}(t_0)}{N_i^{\text{REG}}(t_0)}\right) \cdot f_{vi}\left(\alpha_i+\beta_i; N_i^{\text{REG}}(t_0)\right)}{\alpha_i+\beta_i}, \quad (\alpha_i(t_0),\beta_i(t_0)) \perp (\alpha_j(t_0),\beta_j(t_0)), \quad i \neq j = 1,\dots,I \tag{1g}$$

$$\lambda_i(t_0) \simeq f_{\lambda i}(\lambda_i; N_i^{\text{REG}}(t_0)) \quad (\lambda_i(t_0) > 0, \quad \lambda_i(t_0) \perp \lambda_j(t_0)), \quad i = 1,\dots,I, \tag{1h}$$

where

- $[\cdot]$ denotes the nearest integer function;
- $f_{\alpha ij}$ stands for the probability density function of the parameters α_{ij}. The notation $f_{\alpha ij}\left(\alpha_{ij}; \frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}\right)$ is meant to indicate that $\frac{N_{ij}^{\text{MNO}}(t_0,t_n)}{N_{i\cdot}^{\text{MNO}}(t_0)}$ should be taken as the mode of the density function;
- f_{ui} stands for the probability density function of the parameter u (see the preceding document on the model) in cell i with mode $\frac{N_i^{\text{MNO}}(t_0)}{N_i^{\text{REG}}(t_0)}$;
- f_{vi} stands for the probability density function of the parameter v (see the preceding document on the model) in cell i with mode $N_i^{\text{REG}}(t_0)$;

- $f_{\lambda i}$ stands for the probability density function of the parameter λ (see the preceding document on the model) in cell i with mode $N_i^{\mathrm{REG}}(t_0)$.

Equations (1d) to (1h) are taken from the model in the preceding document just adding explicitly the time dependence. Therefore, their interpretation remains the same. Equations (1a), (1b), (1c) take care of the time evolution of the estimates. Their interpretation is straightforward.

Equation (1a) states that the number of individuals in a cell i at time t_n equals the initial number of individuals plus those arriving from other cells in the given time interval minus those leaving for other cells in the same time interval. The number of individuals arriving and leaving are estimated using the transition probability among cells.

Next, to estimate these transition probabilities we model them for a given cell i as a multivariate random variable with a Dirichlet distribution with parameters $\alpha_{i1}, \ldots, \alpha_{iI}$. These parameters, in turn, are given unimodal prior distributions $f_{\alpha_{ij}}$ with mode in $\frac{N_{ij}^{\mathrm{MNO}}}{N_{i\cdot}^{\mathrm{MNO}}}$ according to our second working assumption.

3 Computations

Ultimately we need to compute the probability functions $\mathbb{P}\left(N_i(t_n)\middle|N^{\mathrm{MNO}}(t_0,t_1)\right)$ for each cell i, which will allow us to choose an estimator as, e.g., the posterior mean, posterior median or posterior mode.

The computation is conducted empirically in three steps:

1. The initial population value $N_i(t_0)$ is generated for all cells $i = 1, \ldots, I$ according to the model using $N_{i\cdot}^{\mathrm{MNO}}(t_0)$ as input data and choosing weakly informative priors f_{ui}, f_{vi} and $f_{\lambda i}$.
2. A transition probability matrix $[p_{ij}(t_0,t_n)]$ is generated according to the model using $N^{\mathrm{MNO}}(t_0,t_n)$ as input data and choosing weakly informative priors $f_{\alpha ij}$.
3. These generated quantities are used in formula (1a) to generate $N_i(t_1)$ for all cells $i = 1, \ldots, I$.

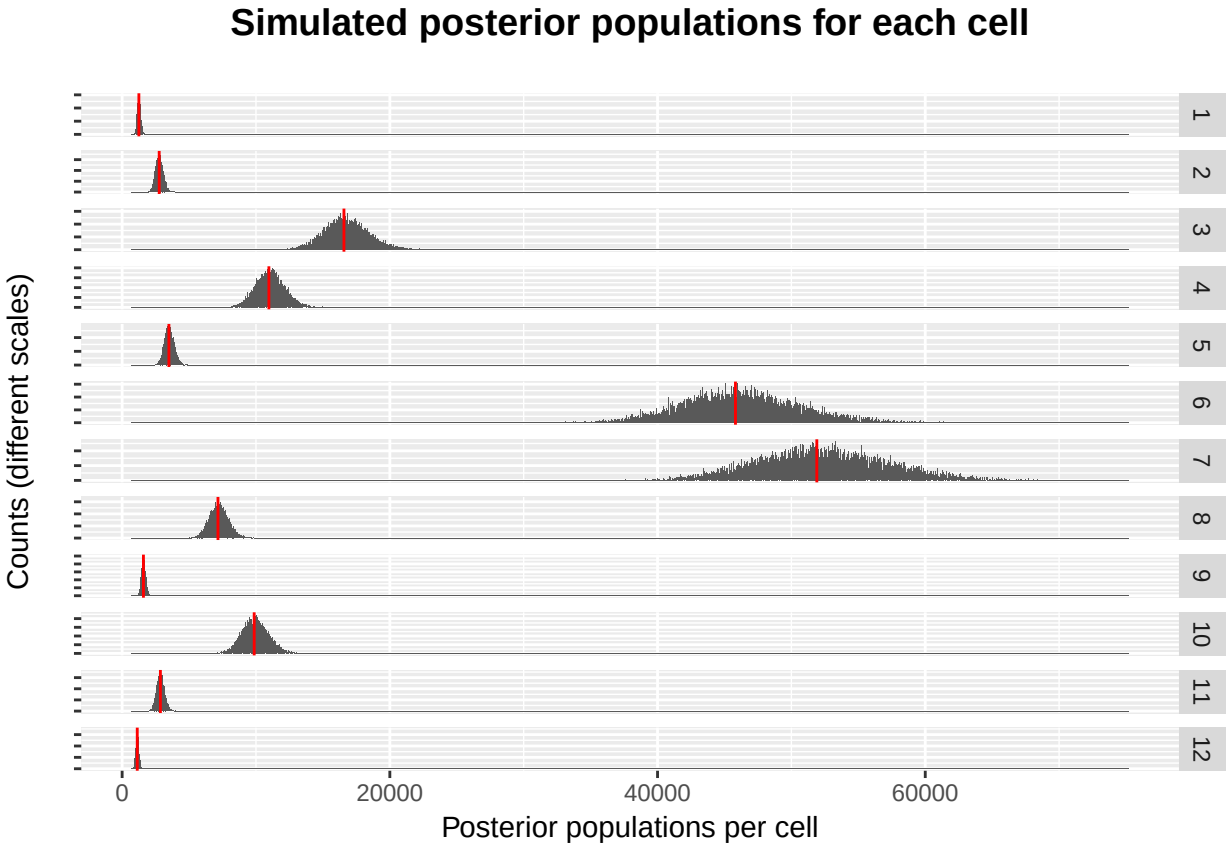
Following these steps we can generate an empirical posterior distribution of values $N_i(t_n)$ for each cell i. Then we can use these distributions to provide a point estimate according to its mean, median, or mode.

4 A toy example

Again let us simulate a toy population to illustrate this approach. We generate a population with a territory divided into 12 cells. At 4 succesive time intervals the individuals follow their own trajectories so that the population in each cell evolves according to actual transition matrices $N_{ij}(t_{n-1},t_n)$, $n = 1,2,3,4$. At the initial time instant t_0 we also have the population register figures for each cell $N_i^{\mathrm{Reg}}(t_0)$ which can be equated to the initial real population $N_i(t_0)$, although they are not completely equal due to nonsampling errors. Data used in this illustration can be found in URL.

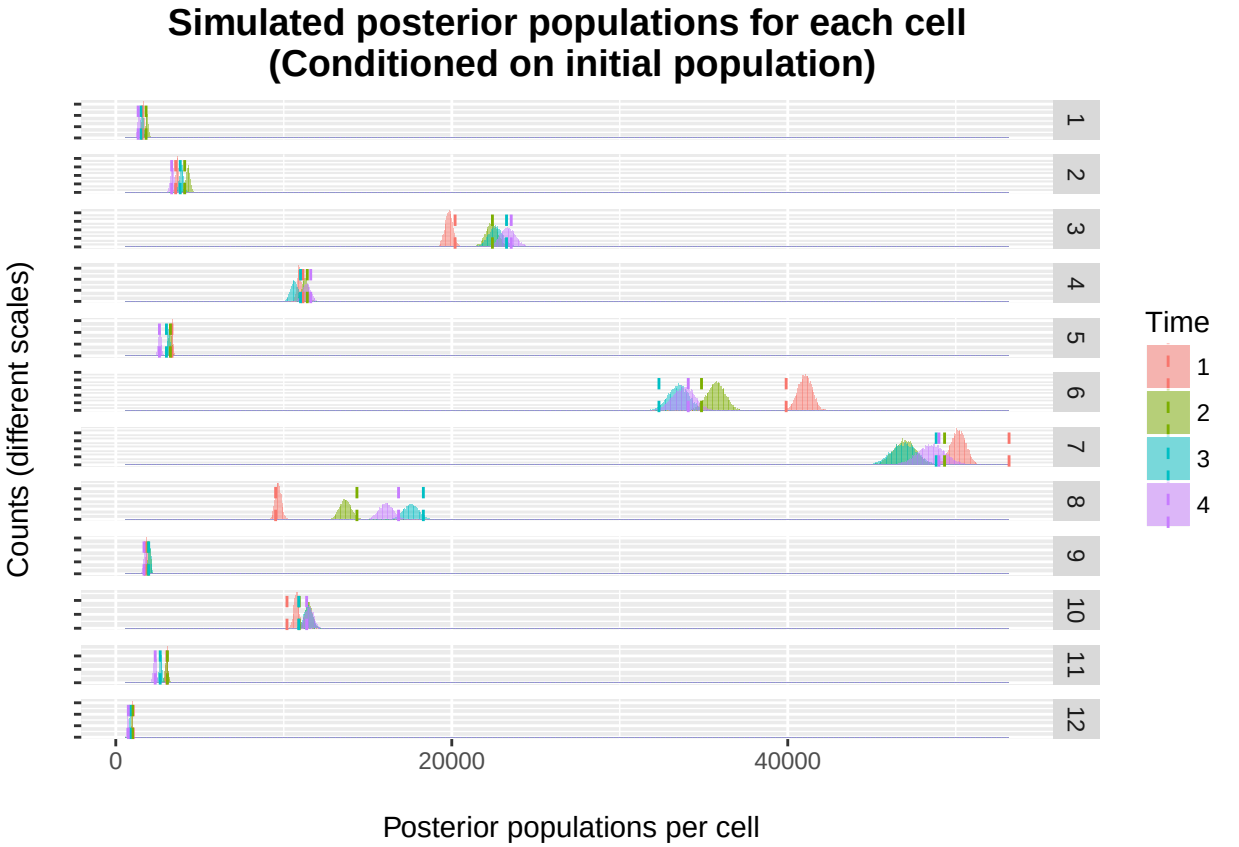
We also generate the transition matrices for each time interval $n = 1,2,3,4$ for the individuals $N_{ij}^{\mathrm{MNO}}(t_{n-1},t_n)$ detected with the network.

We proceed as sketched above. Firstly, we generate the estimates $N_i(t_0)$ for the initial time instant, as conducted in the preceding document.



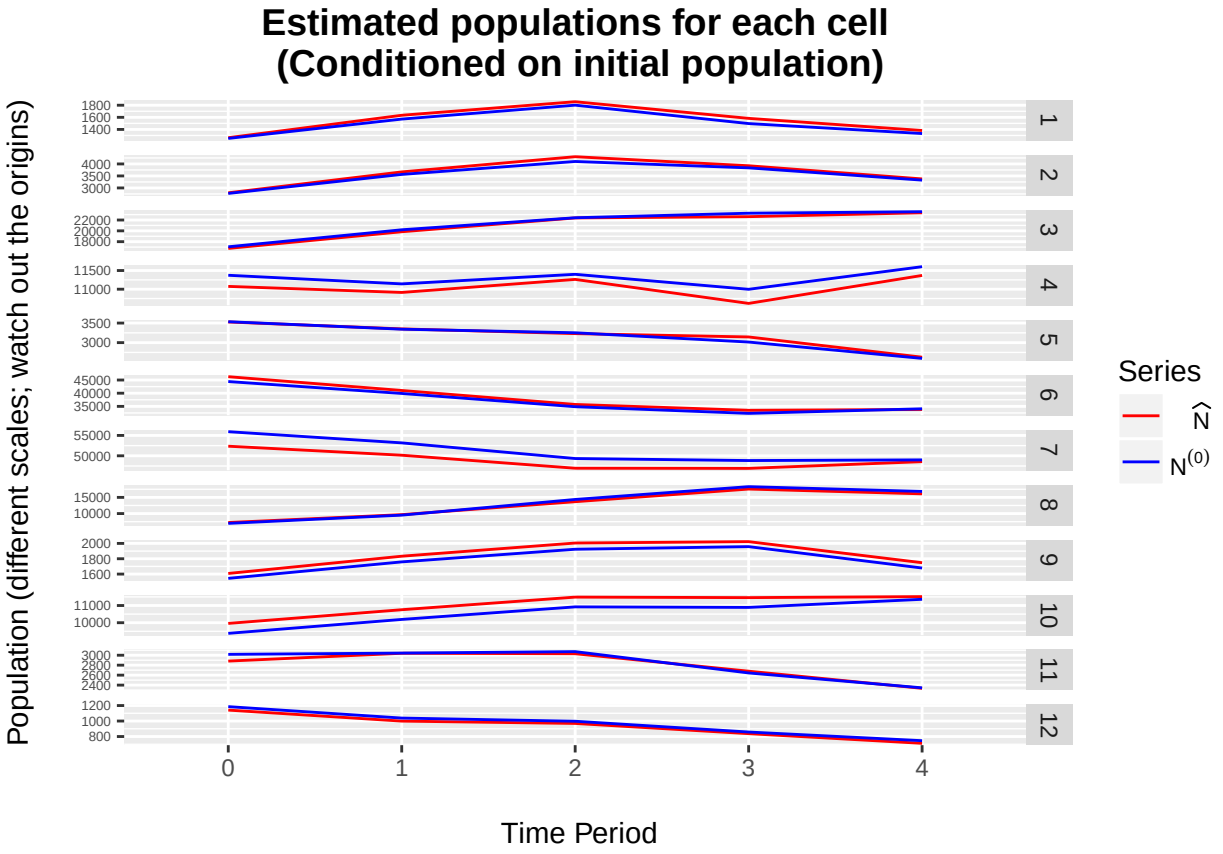
Now we can provide two kind of time evolutions for the population in each cell. On the one hand, we can generate simulated populations conditioned upon their estimated initial size $\widehat{N}_i(t_0)$. On the other hand, we can provide these simulated populations unconditioned upon their estimated initial size but starting from the input data themselves (thus uncertainty in the initial population estimate is included).

In the first case, taking the mean of the preceding populations as estimates for the initial population of each cell we obtain the following evolving simulated populations in each cell (in dashed lines the assumed true values of the simulated population):

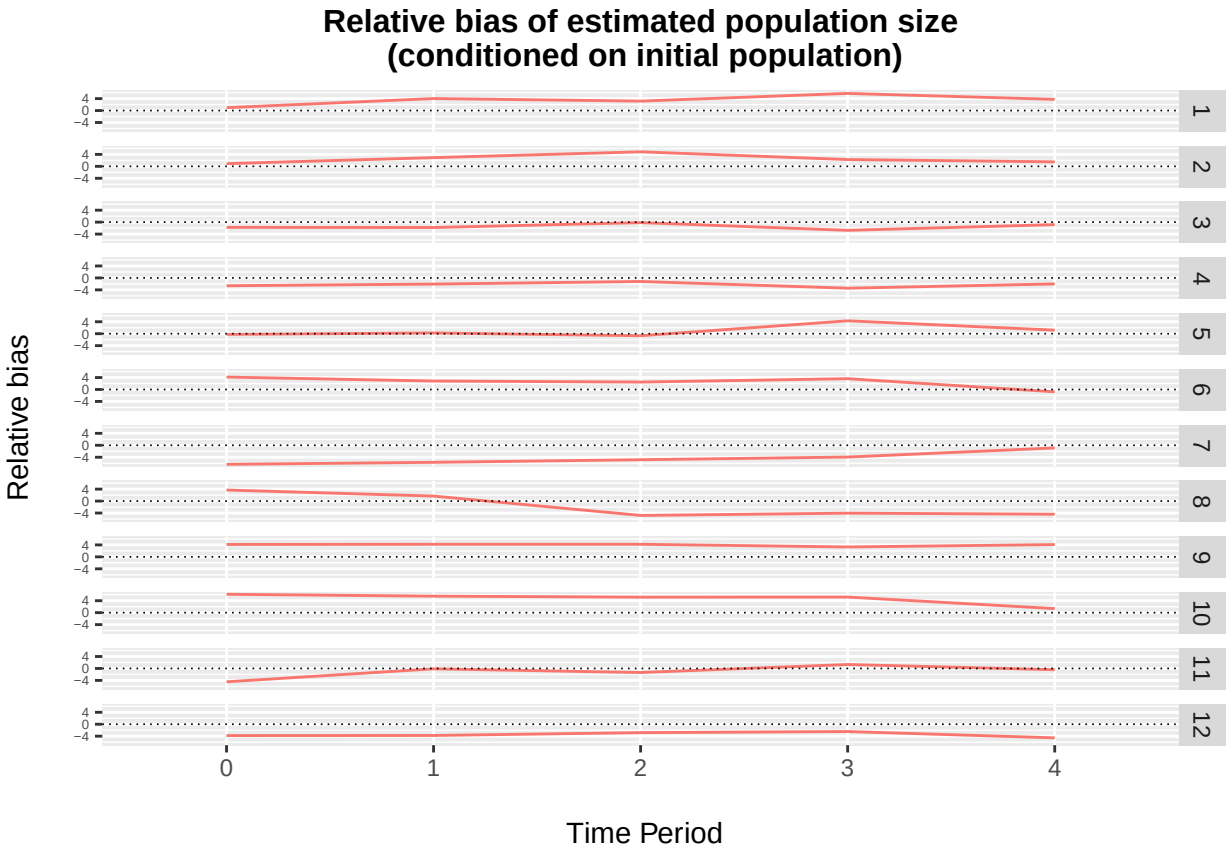


By comparing both figures we can detect how the uncertainty in the estimation of the initial population size propagates along the evolving estimates of the population size of each cell. In particular, in cells 6 and 7 we can observe how the uncertainty in the simulated initial populations provide rather inaccurate estimates for later time periods.

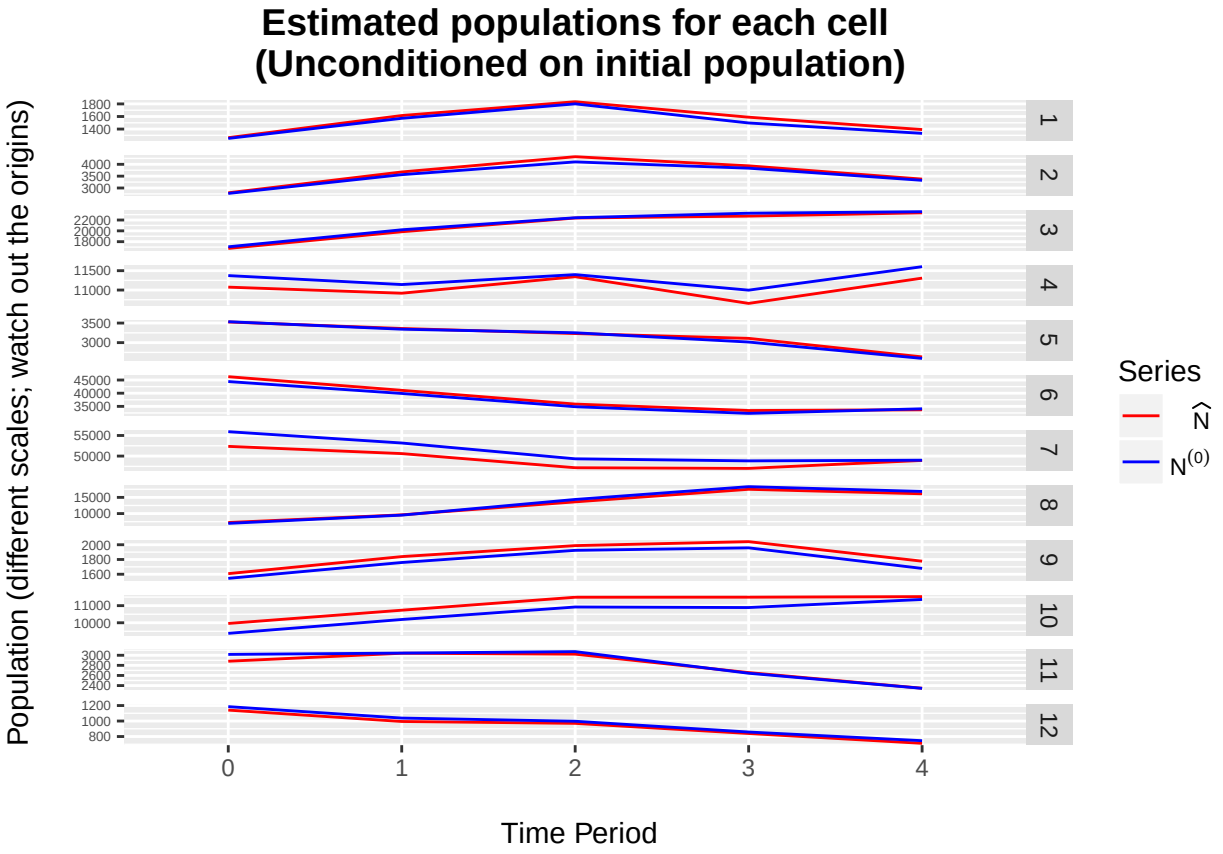
Using again the posterior mean as estimator of the population in each cell, we have the following evolutions:



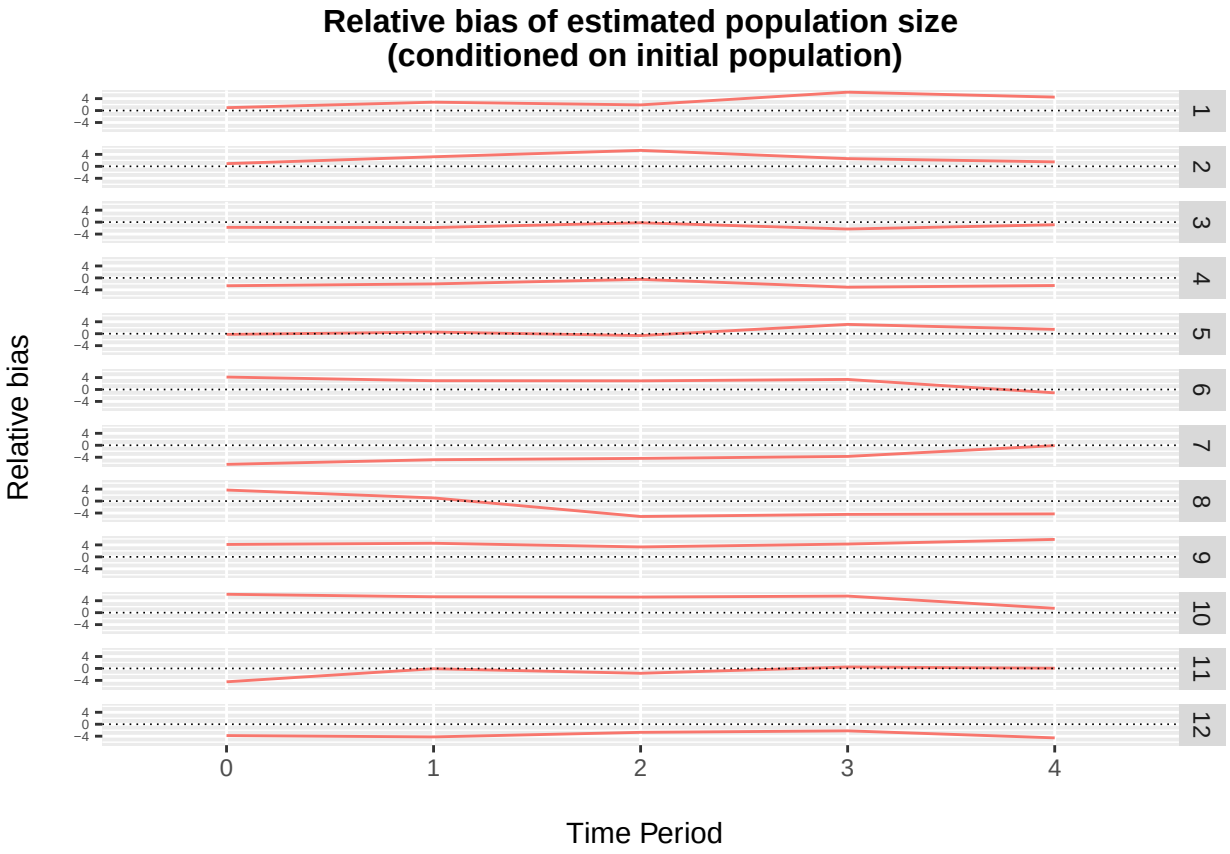
In terms of the relative bias, this comparison can be depicted as follows:



Complementarily, we can produce estimates unconditioned on the initial estimate of the population in each cell. Now the uncertainty in the estimation of the initial population of each cell is incorporated into the estimation process for later time periods.



In terms of the relative bias, this comparison can be depicted as follows:



5 First conclusions

Apparently, the model seems to work on these toy data. However there are many details to be tackled on before claiming such a conclusion:

1. The model must be checked against larger and more complex data (more cells and more time intervals). The process of data generation must be explicitly spelled out to assess about how realistic they are. This will be undertaken in the deliverable.
2. It is expected that the higher the complexity of the data is, the more computationally demanding the model will be. We will have to check the computational capabilities of the programmed functions.
3. No mention to accuracy (apart from the relative bias) and its estimation has been included. This will be an essential part of the quality assessment and will be undertaken in the last deliverable.
4. As the data complexity increases, we will certainly need a visualization tool to include population estimates of all cells for all time periods (possibly maps, grids, heat maps, etc.).
5. The model ultimately must be tested against real data compiled in the SGA-1 by WP members. In this case we do not have the ground truth figures to compare with and we must rely on the preceding performance and expert knowledge.
6. Notice that the model itself (not the data) can gain in complexity by adding new methodological elements:
    - The register population N_i^{Reg} is not modelled at all. We can also assume a prior distribution for these variables reflecting the uncertainty (due to nonsampling errores) in the register figures.

- No spatial correlation is introduced in the model between the cells. These correlations can introduce more realistically this feature.
- No time correlation is introduced in the model between time periods. Time series techniques (or space-time modelling techniques) can in principle improve the model.
- No bias selection analysis has been conducted. Techniques as Heckman bias correction technique can be also another element to take into consideration.

Thus, this proposal is indeed a first simple model which can be further potentially adapted to more realistic situations as the analysis with real data will suggest.

References

ESSnet on Big Data WP5. 2017. "A Simple Hierarchical Model to Estimate Population Counts from Aggregated Mobile Phone Data."