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

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

22 Nov, 2017

Contents

1 General introduction

This document contains the technical details of a simple hierarchical model to estimate the population counts of different territorial cells combining the information from aggregated mobile phone data and a population register or survey data. This approach follows the framework agreed during the internal meeting of the WP5 in Madrid in last June, i.e. the use of ecological sampling techniques to estimate population counts (see e.g. Manly and Navarro Alberto (2015)). In particular, we follow the work by Royle and Dorazio (2008). A general discussion about these techniques and its adequacy for our purposes will be included elsewhere (certainly in the deliverable) especially in connection with the current status of access to mobile phone data.

In this first proposal we envisage the inference exercise from both mobile phone and official data to population counts as a two-step process. Firstly using the official data we assume that they correspond to a given time instant t_0 so that at this time instant both mobile phone and official data are used to infer the population counts in each territorial cell. Then, for later time instants, we shall infer transition probabilities using only mobile phone data to study the spatial and time evolution of the population.

For the time being we will concentrate on mathematical aspects of the model for the first step and a preliminary prototyping implementation to assess the performance of the model. In separate documents we will consider the details about the second step and all the software implementation.

2 The hierarchical model

2.1 Setting up the model

Firstly we set the notation. We shall denote by $\mathbf{N}^{\mathrm{MNO}} = (N_1^{\mathrm{MNO}}, \ldots, N_I^{\mathrm{MNO}})^T$ the population counts according to the mobile devices reported by the mobile network operator in each territorial cell $i \in \mathcal{I} = \{1, \ldots, I\}$ (i.e. the aggregated mobile phone data). These can refer to general population counts, tourist counts, commuters counts, etc. This is considered as an input in the model. Along with the efforts to gain access to mobile phone data NSIs will have also to develop methodologies to obtain these aggregated data out of statistical microdata. Following the generic bottom-up approach of the ESSnet we will concentrate on the part of the process upon which we can carry out concrete analyses. The issue of the access to these microdata must be conveniently solved to follow an empirical approach on the processing of microdata. For the present project we will cover this part of the production process with the internal technical reports by Positium.

In the model, we will make use as auxiliary information of the official population register number of individuals $\mathbf{N}^{\mathrm{REG}} = (N_1^{\mathrm{REG}}, \ldots, N_I^{\mathrm{REG}})^T$ in each of the cells or some equivalent survey or administrative source. The goal is to provide estimates for the actual population counts $\mathbf{N} = (N_1, \ldots, N_I)^T$ combining both data sources. The interplay between official data and mobile phone data will also be discussed elsewhere (see also the internal documents of the WP5).

More rigorously, aiming at the quality assessment of the estimation procedure, we will produce a probability distribution for the number of individuals of the target population in each cell i using both the mobile phone and official population data as inputs (see figure 1). The posterior probability distribution $\mathbb{P}\left(\mathbf{N}|\mathbf{N}^{\mathrm{MNO}}; \mathbf{N}^{\mathrm{REG}}\right)$ will allow us to assess the uncertainty in the output estimates.

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

We propose the hierarchical model given by:

$$
\begin{aligned}
N_i^{\mathrm{MNO}} &\simeq \mathrm{Bin}\left(N_i, p_i\right), \qquad N_i^{\mathrm{MNO}} \perp N_j^{\mathrm{MNO}}, \quad i \neq j = 1, \ldots, I \\
N_i &\simeq \mathrm{Po}\left(\lambda_i\right), \qquad N_i \perp N_j, \quad i \neq j = 1, \ldots, I \\
p_i &\simeq \mathrm{Beta}\left(\alpha_i, \beta_i\right), \qquad p_i \perp p_j \quad i \neq j = 1, \ldots, I \\
(\alpha_i, \beta_i) &\simeq \frac{f_1(\frac{\alpha_i}{\alpha_i+\beta_i}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha_i + \beta_i; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{\alpha_i + \beta_i}, \qquad (\alpha_i, \beta_i) \perp (\alpha_j, \beta_j), \quad i \neq j = 1, \ldots, I \\
\lambda_i &\simeq f_3(\lambda_i; N_i^{\mathrm{REG}}, \mathbf{z}) \quad (\lambda_i > 0, \lambda_i \perp \lambda_j), \quad i = 1, \ldots, I.
\end{aligned}
\tag{1}
$$

2.2 Interpretating the model

The interpretation of the model is more or less straightforward. If in a territorial cell i there are N_i individuals and we have an independent detection probability p_i for each individual through the mobile telecommunication network, then we will detect N_i^{MNO} individuals according to the aggregated mobile phone data naturally following a binomial distribution.

Now, the number of individuals N_i in each cell can be understood as a Poisson random variable (potentially arising from an underlying birth-death Poisson process). These variables are pairwise independent and depend on unknown independent parameters λ_i. For the time being we will keep the model as simple as possible to test a first proof of concept.

Now, the detection probabilities p_i in our mobile phone setting differs from the usual ecological setting. In the latter, the field work (observation sites, visual techniques, ...) allows us to propose a model for these probabilities according to the measurement process. In the telecommunication setting, in principle, the measurement process in cell i is always successful provided that a subscriber interacting with the network is within the territorial cell i. Thus at the given instant of time t_0 the detection probabilities p_i amount to establish the proportion of individuals of interest at each cell i being detected by the MNO's cellular network. In other words, p_i are the proportions of individuals detected by the MNO at time t_0 in each cell i.

It is interesting to make a short reflection about these proportions p_i and the so-called local market shares. The latter are the number of subscribers of a given MNO in a cell i and they are sometimes considered as an important piece of information in performing the inference exercise from mobile phone data to the target population. We must stress that, in our view, it is not the concept of market share which is important but that of the actual proportion of individuals detected by the network. As an illustrative example, a call between a subscriber in a cell i and a non-subscriber in another cell j of a given MNO is certainly detected by the network in ***both cells***, thus potentially being part of the aggregated data N_i^{MNO} and N_j^{MNO}. This is a clear example of why having knowledge of the preprocessing and aggregation procedures from microdata is important for the final results.

We will model the detection probabilities p_i to account for the uncertainty we have in these quantities. Thus they are modelled as beta random variables with parameters α_i, β_i independently in each cell. The prior distribution of the beta distribution parameters α_i, β_i arises from the following reasoning. We assume that $\frac{\alpha_i}{\alpha_i+\beta_i}$ and $\alpha_i + \beta_i$ distribute independently according to $\frac{\alpha_i}{\alpha_i+\beta_i} \simeq f_1(\frac{\alpha_i}{\alpha_i+\beta_i}; \mathbf{N}^{\text{REG}}, \mathbf{z})$ and $\alpha_i + \beta_i \simeq f_2(\alpha_i+\beta_i; \mathbf{N}^{\text{REG}}, \mathbf{z})$, where f_1 and f_2 are respective weakly informative prior distributions for $\frac{\alpha}{\alpha+\beta}$ and $\alpha+\beta$. Notice that we have again made use of the auxiliary information coming from the population register ($\mathbf{N}^{\text{REG}}$) and any other auxiliary information $\mathbf{z}$. The quantity $\alpha_i/(\alpha_i + \beta_i)$ can be understood as a priori proportions of individuals detected by the MNO in cell i (e.g. should we have no information, then $f_1 = \text{Unif}[0, 1]$). The parameters $\alpha_i + \beta_i$ can be essentially understood as the population size of each cell N_i (thus with support in $(0, \infty)$) upon which the detection is executed at that time instant. For example, we may assume f_2 to be a gamma distribution with parameters $(N_i^{\text{MNO}} + 1, \frac{N_i^{\text{MNO}}}{N_i^{\text{REG}}})$. In this way, the most probable value for the sample size is N_i^{REG} in consonance with the preceding hypothesis for N_i.

Finally, the parameters λ_i are modeled with another weakly information prior f_3 which may incorporate the information we have from the population register or similar sources. Notice that the only a priori information incorporated is coming from this auxiliary source.

There is a clear abuse of notation by denoting both random variables and their realization in the same way (the context will make this clear). $\mathbf{N}^{\text{REG}}$ are treated as fixed parameters in the current model. By relaxing this and modelling also $\mathbf{N}^{\mathbf{REG}}$ we can pave the way to account for proposing more complex models for the uncertainty (possible non-sampling errors) in the official population figures for this estimation procedure.

Notice also that the cells are treated independently leaving the door open for geostatitical considerations naturally accounting for geospatial correlations among the cells. We will concentrate here on the preceding simple model to provide a proof of concept.

2.3 Getting the flavour of the model

To get a flavour of the model, let us make the following simplifying assumption. Let us suppose that the prior distributions f_1 and f_2 are degenerate so that equivalently we are assuming that we have full knowledge of the a priori proportion of detected individuals[1] $u = \frac{\alpha}{\alpha+\beta} = u^*$ and of the population cell size $N^* = \alpha + \beta$ whose proportion of subscribers is detected by our MNO.

Then it is straighforward to show that the unnormalized posterior probability density $\mathbb{P}\left(\lambda | N^{\mathrm{MNO}}; N^{\mathrm{REG}}\right)$ is given by

$$\begin{aligned}
\mathbb{P}\left(\lambda | N^{\mathrm{MNO}}; N^{\mathrm{REG}}\right) &\propto f_3(\lambda; N^{\mathrm{REG}}) \cdot \mathrm{Po}\left(N^{\mathrm{MNO}}; \lambda\right) \cdot \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \frac{B(u^* \cdot N^* + N^{\mathrm{MNO}}, (1-u^*) \cdot N^* + n)}{B(u^* \cdot N^*, (1-u^*) \cdot N^*)} \\
&\propto\propto f_3(\lambda; N^{\mathrm{REG}}) \cdot \mathrm{Po}\left(N^{\mathrm{MNO}}; \lambda\right) \cdot \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \cdot {u^*}^{N^{\mathrm{MNO}}} \cdot (1-u^*)^n \\
&\propto\propto f_3(\lambda; N^{\mathrm{REG}}) \cdot e^{-\lambda u^*} \cdot \frac{(\lambda u^*)^{N^{\mathrm{MNO}}}}{N^{\mathrm{MNO}}!}, \qquad (2)
\end{aligned}$$

where we have used the approximation $\frac{\Gamma(x+a)}{\Gamma(x)} \approx x^a$ (which can be proved using Stirling's approximation) and where $\mathrm{Po}(N; \lambda)$ denotes the probability function of a Poisson random variable N with parameter λ..

In the case of noninformative prior $f_3 \propto 1$ the posterior (2) corresponds to a gamma distribution for λ with parameters $N^{\mathrm{MNO}} + 1$ and u^*. The mode of this distribution (thus the most probable value for λ) is $\frac{N^{\mathrm{MNO}}}{u^*}$. In turn, the most probable value for N in the model is $\lfloor \lambda \rfloor = \lfloor \frac{N^{\mathrm{MNO}}}{u^*} \rfloor$. With the due rigorous proviso, u^* can be somehow understood as a sampling weight connecting the population of detected individuals through the mobile phone network with the target population.

Suppose now that we assume a prior gamma distribution $\lambda \simeq \Gamma(\alpha + 1, N^{\mathrm{REG}}/\alpha)$, where $\alpha > 0$. Then the posterior (2) is again a gamma distribution now with parameters $\Gamma(\alpha + N^{\mathrm{MNO}} + 1, u^* + \frac{\alpha}{N^{\mathrm{REG}}})$. The most probable value then for λ is $\frac{N^{\mathrm{MNO}}+\alpha}{u^* + \frac{\alpha}{N^{\mathrm{REG}}}}$ and for N is $\lfloor \frac{N^{\mathrm{MNO}}+\alpha}{u^* + \frac{\alpha}{N^{\mathrm{REG}}}} \rfloor$, which can be written as

$$\begin{aligned}
\widehat{N} &= \left\lfloor \frac{u^* \cdot N^{\mathrm{REG}}}{\alpha + u^* \cdot N^{\mathrm{REG}}} \cdot \frac{N^{\mathrm{MNO}}}{u^*} + \frac{\alpha}{\alpha + u^* \cdot N^{\mathrm{REG}}} \cdot N^{\mathrm{REG}} \right\rfloor \\
&\approx \frac{u^* \cdot N^{\mathrm{REG}}}{\alpha + u^* \cdot N^{\mathrm{REG}}} \cdot \left\lfloor \frac{N^{\mathrm{MNO}}}{u^*} \right\rfloor + \frac{\alpha}{\alpha + u^* \cdot N^{\mathrm{REG}}} \cdot N^{\mathrm{REG}} \qquad (3)
\end{aligned}$$

The estimate is thus an accurately approximate convex combination of both extremes: (i) having no auxiliary information at all about the population register and (ii) using only the information from the population register. The relative weight between these two components is provided by the parameter α.

The full Bayesian approach in the forthcoming sections incorporate our uncertainty in the knowledge of the hyperparameters (especially of $u = \frac{\alpha}{\alpha+\beta}$ and $v = \alpha + \beta$), since we do not know with certainty the values of the proportion of individuals and of the actual population size of each cell upon which the detection is executed.

3 Posterior distribution for the target population

The quantity of interest is the target population counts $\mathbf{N} = (N_1, \ldots, N_I)^T$ in each cell i. To account for uncertainty (thus for accuracy estimation and quality measures) we will follow a Bayesian approach. Notice

[1] For ease of notation we drop out the subscripts i regarding the cells, since they are independent.

that we can leverage the prior information we have by choosing the probability distribution f_1, f_2 and f_3. Should we choose very weakly informative priors, final estimates would not a priori differ very much from the frequentist approach. We shall not make considerations about the philosophical differences between both approaches. We will keep pragmatic and assess the final results according to our simulations. The choice of the Bayesian approach allows us to account for the inference and the assessment of the uncertainty, hence of quality, thus complying with the goals of the project.

Thus we must find the posterior distribution $\mathbb{P}\left(\mathbf{N}\middle|\mathbf{N}^{\mathrm{MNO}};\mathbf{N}^{\mathbf{REG}}\right)$ or equivalently the marginal distributions $\mathbb{P}\left(N_i\middle|\mathbf{N}^{\mathrm{MNO}};\mathbf{N}^{\mathbf{REG}}\right)$. We drop the subscripts since the treatment of each cell is equivalent and independent of each other. As usual, we will focus on the unnormalised version of the involved probabilities. We make use of the hierarchy:

$$\begin{aligned}\mathbb{P}\left(N\middle|N^{\mathrm{MNO}};N^{\mathbf{REG}}\right) &\propto \int_0^\infty d\lambda \quad \mathbb{P}\left(N|\lambda\right)\mathbb{P}\left(\lambda\middle|\mathbf{N}^{\mathrm{MNO}};N^{\mathbf{REG}}\right)\\ &\propto \int_0^\infty d\lambda \quad \mathbb{P}\left(\lambda\middle|N^{\mathrm{MNO}};N^{\mathbf{REG}}\right)\cdot e^{-\lambda}\cdot\frac{(\lambda)^N}{N!}\\ &\propto \int_0^\infty d\lambda \quad \mathbb{P}\left(\lambda\middle|N^{\mathrm{MNO}};N^{\mathbf{REG}}\right)\cdot \mathrm{Po}(N;\lambda),\end{aligned} \tag{4}$$

As expected, we need the posterior distribution for the hyperparameters, which moreover will allow us also to practise inference and simulations and to assess the quality of the model.

As we have stated in the preceding section notice that being N a Poisson random variable, the most probable value for N is given by $\lfloor\lambda\rfloor$. Thus the posterior distribution for the hyperparameter λ will allow us to provide a point estimator for N (indeed as many as we want: mode, mean, median, ...).

4 Posterior distribution for the hyperparameters

4.1 A first analytical expression

We will try to be analytical as far as possible, leaving numerical computations for later developments. To compute the posterior $\mathbb{P}\left(\lambda\middle|\mathbf{N}^{\mathrm{MNO}};\mathbf{N}^{\mathbf{REG}}\right)$ we proceed as always (see e.g. Gelman et al. (2013)):

$$\begin{aligned}\mathbb{P}\left(\lambda\middle|N^{\mathrm{MNO}};N^{\mathbf{REG}}\right) &\propto \mathbb{P}\left(N^{\mathrm{MNO}}|\lambda;N^{\mathbf{REG}}\right)\\ &\propto \int_0^\infty\int_0^\infty d\alpha d\beta\int_0^1 dp\sum_{n=N^{\mathrm{MNO}}}^\infty \mathbb{P}\left(N^{MNO}|p,N\right)\mathbb{P}\left(N|\lambda;N^{\mathrm{REG}}\right)\mathbb{P}\left(p|\alpha,\beta\right)\mathbb{P}\left(\alpha,\beta;N^{\mathrm{REG}}\right)\mathbb{P}\left(\lambda\right)\\ \propto \mathbb{P}\left(\lambda\right)\int_0^\infty\int_0^\infty d\alpha d\beta\int_0^1 dp\sum_{n=N^{\mathrm{MNO}}}^\infty &\binom{n}{N^{\mathrm{MNO}}}p^{N^{\mathrm{MNO}}}(1-p)^{n-N^{\mathrm{MNO}}}e^{-\lambda}\frac{(\lambda)^n}{n!}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\frac{f_1(\frac{\alpha}{\alpha+\beta};\mathbf{N}^{\mathrm{REG}})\cdot f_2(\alpha+\beta;\mathbf{N}^{\mathrm{REG}})}{\alpha+\beta}\end{aligned} \tag{5}$$

Simplifying we arrive at

$$\mathbb{P}\left(\lambda \middle| N^{\mathrm{MNO}}; N^{\mathbf{REG}}\right) \propto$$

$$\mathbb{P}(\lambda) \sum_{n=N^{\mathrm{MNO}}}^{\infty} \binom{n}{N^{\mathrm{MNO}}} e^{-\lambda} \frac{\lambda^n}{n!} \int_0^{\infty}\int_0^{\infty} d\alpha d\beta \frac{f_1(\frac{\alpha}{\alpha+\beta}; N^{\mathrm{REG}}) \cdot f_2(\alpha+\beta; N^{\mathrm{REG}})}{\alpha+\beta} \frac{B\left(\alpha + N^{\mathrm{MNO}}, \beta + n - N^{\mathrm{MNO}}\right)}{B(\alpha,\beta)}$$

$$\propto \mathbb{P}(\lambda) \sum_{n=N^{\mathrm{MNO}}}^{\infty} \binom{n}{N^{\mathrm{MNO}}} e^{-\lambda} \frac{\lambda^n}{n!} I_{N^{\mathrm{MNO}}, n-N^{\mathrm{MNO}}}(N^{\mathrm{REG}})$$

$$\propto \mathbb{P}(\lambda)\, \mathrm{Po}(N^{\mathrm{MNO}}; \lambda) \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} I_{N^{\mathrm{MNO}}, n}(N^{\mathrm{REG}})$$

$$\propto \mathbb{P}(\lambda) \cdot \mathrm{Po}(N^{\mathrm{MNO}}; \lambda) \cdot S\left(\lambda, N^{\mathrm{MNO}}, N^{REG}\right), \tag{6}$$

where we have defined

$$I_{N^{\mathrm{MNO}}, n}(N^{\mathrm{REG}}) = \int_0^{\infty}\int_0^{\infty} d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; N^{\mathrm{REG}}) \cdot f_2(\alpha+\beta; N^{\mathrm{REG}}}{\alpha+\beta} \frac{B\left(\alpha + N^{\mathrm{MNO}}, \beta + n - N^{\mathrm{MNO}}\right)}{B(\alpha,\beta)}, \tag{7}$$

$$S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} I_{N^{\mathrm{MNO}}, n}(N^{\mathrm{REG}}). \tag{8}$$

Everything is thus reduced to the computation of the expression $S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}})$ (the integral and the sum of the series).

4.2 Computation of $S(\lambda, N^{\mathbf{MNO}}, N^{\mathbf{REG}})$

We include here the different attempts conducted to compute $S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}})$ analytically.

4.2.1 Approach 1

In this approach we first compute the integral $I_{N^{\mathrm{MNO}}, n}(N^{\mathrm{REG}})$ and then we sum up the series. We perform the change of variables $u = \frac{\alpha}{\alpha+\beta}, v = \alpha + \beta$ so that the integral transforms into

$$I_{n,m}(N^{\mathrm{REG}}) = \int_0^{\infty} dv f_2(v) \int_0^1 du\; f_1(u) \frac{B(u\cdot v + n, (1-u)\cdot v + m)}{B(u\cdot v, (1-u)\cdot v)} \tag{9a}$$

$$= \int_0^{\infty} dv\; f_2(v) \frac{\Gamma(v)}{\Gamma(v+n+m)} \int_0^1 du\; f_1(u) \frac{\Gamma(u\cdot v + n)}{\Gamma(u\cdot v)} \frac{\Gamma((1-u)\cdot v + m)}{\Gamma((1-u)\cdot v)}$$

$$= \int_0^{\infty} dv\; f_2(v) \frac{\Gamma(v)}{\Gamma(v+n+m)} \int_0^v dt\; f_1(t/v) \frac{\Gamma(t+n)}{\Gamma(t)} \frac{\Gamma(v-t+m)}{\Gamma(v-t)} \tag{9b}$$

Expression (9a) will allow us to compute the integral via Monte Carlo methods. We will pursue the analytical computation using expression (9b). The inner integral can be computed expressing the integrand in terms of Stirling numbers of the first kind (see e.g. Graham, Knuth, and Patashnik (1996)):

$$\begin{aligned}
\int_0^v dt f_1(t/v)\frac{\Gamma(t+n)}{\Gamma(t)}\frac{\Gamma(v-t+m)}{\Gamma(v-t)} &= \int_0^v dt(t+n-1)\cdots(t+1)t\cdot(m-1+v-t)\cdots(1+v-t)(v-t) \\
&= \int_0^v dt f_1(t/v)\prod_{k=0}^{n-1}(t+k)\prod_{l=0}^{m-1}(v-t+l) \\
&= \int_0^v dt f_1(t/v)\sum_{k=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}t^k\sum_{l=0}^{m}\begin{bmatrix}m\\l\end{bmatrix}(v-t)^l \\
&= \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}v^{k+l+1}\int_0^1 f_1(x)x^k(1-x)^l dx \\
&= \sum_{k=0}^{n}\sum_{l=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)v^{k+l+1}, \qquad (10)
\end{aligned}$$

where $\begin{bmatrix}n\\k\end{bmatrix}$ denotes the unsigned Stirling numbers of the first kind and $\bar{B}(k+1,l+1)=\int_0^1 f_1(x)x^k(1-x)^l dx$ (notice that for $f_1=$ Unif(0,1) we have $\bar{B}=B$). Then we can write

$$I_{n,m}(N^{\mathrm{REG}}) = \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)\int_0^\infty dv f_2(v;\mathbf{N}^{\mathrm{REG}},\mathbf{z})\frac{\Gamma(v)}{\Gamma(v+n+m)}v^{k+l+1} \qquad (11)$$

Denoting

$$J_{n+m,k+l}(N^{\mathrm{REG}}) = \int_0^\infty dv\cdot f_2(v;N^{\mathrm{REG}})\cdot\frac{v^{k+l}}{\prod_{i=1}^{n+m-1}(v+i)}, \qquad (12)$$

we have

$$\begin{aligned}
I_{n,m}(N^{\mathrm{REG}}) &= \sum_{k=0}^{n}\sum_{l=0}^{m}\begin{bmatrix}n\\k\end{bmatrix}\begin{bmatrix}m\\l\end{bmatrix}\bar{B}(k+1,l+1)J_{n+m,k+l}(N^{\mathrm{REG}}) \\
&= \sum_{p=0}^{n+m}J_{n+m,p}(N^{\mathrm{REG}})\sum_{q=0}^{p}\begin{bmatrix}n\\q\end{bmatrix}\begin{bmatrix}m\\p-q\end{bmatrix}\bar{B}(q+1,p-q+1) \qquad (13)
\end{aligned}$$

Thus we have reduced the integral to the computation of $J_{n+m,p}(N^{\mathrm{REG}})$ and

$$a_{n,m}(p) = \sum_{q=0}^{p}\begin{bmatrix}n\\q\end{bmatrix}\begin{bmatrix}m\\p-q\end{bmatrix}\bar{B}(q+1,p-q+1).$$

These two quantities can be further expressed analytically in some cases (see the appendices). However we do not see an easy computer implementation of these expressions. Thus we opt for Monte Carlo computation. We focus on expression (9a) to conduct a Monte Carlo evaluation of the integral $I_{n,m}(N^{\mathrm{REG}})$. To this end consider the function $g_{n,m}(\mathbf{x}) = \frac{B(x_1\cdot x_2+n,(1-x_1)\cdot x_2+m)}{B(x_1\cdot x_2,(1-x_1)\cdot x_2)}$ and generate M bidimensional random variables $\mathbf{x}\in[0,1]\times\mathbf{R}^+$ according to the bidimensional distribution $f_1\times f_2$. Then, using $f(\mathbf{x})=f_1(x_1)f_2(x_2)$ as importance function, we can write as a first option

$$I_{n,m}(N^{\mathrm{REG}}) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} g_{n,m}(\mathbf{x}_i). \tag{14}$$

To accelerate the convergence we make use of stratified importance sampling. To introduce the stratification let us define $H_1 \cdot H_2$ strata as the rectangular domains $[a_{h_1-1}, a_{h_1}] \times [b_{h_2-1}, b_{h_2}]$, where $a_{h_1} = F_1^{-1}(h_1/H_1)$ ($h_1 = 1, \ldots, H_1$) and $b_{h_2} = F_2^{-1}(h_2/H)$ ($h_2 = 1, \ldots, H_2$), and F_i stands for the distribution function corresponding to the density function f_i. Defining the importance function in each stratum by $f_{h_1 h_2} = H_1 \cdot H_2 \cdot f_1 \cdot f_2$ truncated at $[a_{h_1-1}, a_{h_1}] \times [b_{h_2-1}, b_{h_2}]$ and taking equal-size strata $M_{h_1 h_2} = \frac{M}{H_1 H_2}$, then we finally write

$$I_{n,m}(N^{\mathrm{REG}}) = \lim_{M\to\infty} \frac{1}{M} \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \sum_{i_{h_1}=1}^{M/H_2} \sum_{i_{h_2}=1}^{M/H_1} g_{n,m}(\mathbf{x}_{i_{h_1} i_{h_2}}) \tag{15}$$

where the random values $\mathbf{x}_{i_{h_1} i_{h_2}}$ are generated with the corresponding density function $f_{h_1 h_2}$.

Once computed the integrals $I_{N^{\mathrm{MNO}},n}(N^{\mathrm{REG}})$, the series $S(\beta_0, N^{\mathrm{MNO}}, N^{\mathrm{REG}})$ is summed up with standard procedures (iteratively or with a tail-recursice algorithm) within a given tolerance τ.

4.2.2 Approach 2

In this second approach we reverse the order of computation. We first sum up the series and then we compute the integral. Thus we write

$$\begin{aligned}
S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) &= \int_0^\infty \int_0^\infty d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha+\beta; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{\alpha+\beta} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \frac{B(\alpha + N^{\mathrm{MNO}}, \beta + n)}{B(\alpha, \beta)} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha+\beta; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{(\alpha+\beta) \cdot B(\alpha, \beta)} \int_0^1 dx \; x^{\beta-1}(1-x)^{\alpha+N^{\mathrm{MNO}}-1} \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha+\beta; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{(\alpha+\beta) \cdot B(\alpha, \beta)} \int_0^1 dx \; e^{\lambda x} x^{\beta-1}(1-x)^{\alpha+N^{\mathrm{MNO}}-1} \\
&= \int_0^\infty \int_0^\infty d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha+\beta; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{\alpha+\beta} \frac{B(\alpha + N^{\mathrm{MNO}}, \beta)}{B(\alpha, \beta)} \cdot {}_1F_1(z; \beta, \alpha+\beta+N^{\mathrm{MNO}}) \\
&\equiv \int_0^\infty \int_0^\infty d\alpha d\beta \; \frac{f_1(\frac{\alpha}{\alpha+\beta}; \mathbf{N}^{\mathrm{REG}}, \mathbf{z}) \cdot f_2(\alpha+\beta; \mathbf{N}^{\mathrm{REG}}, \mathbf{z})}{\alpha+\beta} \Phi(\alpha, \beta; \lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}),
\end{aligned} \tag{16}$$

where we have defined $\Phi(\alpha, \beta; \lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) = \frac{B(\alpha+N^{\mathrm{MNO}}, \beta)}{B(\alpha,\beta)} \cdot {}_1F_1(\lambda; \beta, \alpha+\beta+N^{\mathrm{MNO}})$. Now to compute this integral let us change variables as in the preceding section so that:

$$\begin{aligned}
S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) &= \int_0^\infty dv \; f_2(v) \int_0^1 du \; f_1(u) \cdot \Phi(u \cdot v, (1-u) \cdot v; \lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) \\
&= \int_0^\infty dv \; f_2(v) \int_0^1 du \; f_1(u) \cdot \bar{\Phi}(u, v; \lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}})
\end{aligned} \tag{17}$$

Mutatis mutandi defining $\mathbf{x} = (u, v)^T$ we can apply the same Monte Carlo technique as in the preceding section to arrive at

$$S(\lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}) = \lim_{M\to\infty} \frac{1}{M} \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \sum_{i_{h_1}=1}^{M/H_2} \sum_{i_{h_2}=1}^{M/H_1} \bar{\Phi}(\mathbf{x}_{i_{h_1} i_{h_2}}; \lambda, N^{\mathrm{MNO}}, N^{\mathrm{REG}}), \tag{18}$$

where the random values $\mathbf{x}_{i_{h_1} i_{h_2}}$ are generated with the corresponding density function $f_{h_1 h_2}$.

Both approaches have been tested. Details about the computer implementation of these expressions will be provided elsewhere (corresponding to the deliverables about IT).

5 Generation of random variables

To conduct simulation studies and carry out the estimation on the number of individuals per cell we need to generate random variables according to the posterior distribution $\mathbb{P}\left(N_i|\mathbf{N}^{\mathrm{MNO}}; \mathbf{N}^{\mathrm{REG}}\right)$. We have two options: either we use (4) to compute an expression for the probability function of N_i or we make use of the hierarchy by generating random values of λ according to $\mathbb{P}\left(\lambda|\mathbf{N}^{\mathrm{MNO}}; \mathbf{N}^{\mathrm{REG}}\right)$ and then use these values to generate N_i according to $\mathrm{Po}(\lambda)$. We choose the latter. Thus we need to generate random values of the hyperparameter λ according to its posterior distribution (6):

$$\mathbb{P}\left(\lambda\middle|N^{\mathrm{MNO}}; N^{\mathbf{REG}}\right) \propto \mathbb{P}\left(\lambda\right) \cdot \mathrm{Po}(N^{\mathrm{MNO}}; \lambda) \cdot S\left(\lambda, N^{\mathrm{MNO}}, N^{REG}\right).$$

The unnormalized posterior density $\mathbb{P}\left(\lambda|\mathbf{N}^{\mathrm{MNO}}; \mathbf{N}^{\mathrm{REG}}\right)$ does not allow us to find easily the corresponding posterior distribution function to apply the inverse method to generate random variables (see e.g. Devroye (1986)). We then try to use the acceptance-rejection method. Indeed this method is appropriate to use with unnormalized probability functions.

Let us define $f(x) = \mathbb{P}\left(x|N^{\mathrm{MNO}}; N^{\mathrm{REG}}\right)$. As candidate distribution $g(x)$ we will use a Cauchy distribution $g(x) = \mathrm{Cauchy}(x; x_0 = \lambda^*, \sigma)$ truncated at $\mathbb{R}^+$ with $\lambda^* = \mathrm{argmax}_{\lambda\geq 0} f(\lambda)$ (i.e. the mode of $f(\lambda)$). We need to prove rigorously that f is majorized by this candidate distribution g. So far, computing tests have been satisfactory. Also we do not have a general recipe for the scale parameter σ. For the family of prior distributions $\lambda \simeq \Gamma(\alpha + 1, \alpha/N^{\mathrm{REG}})$ (with mode N^{REG} and variance $\frac{\alpha+1}{(\alpha/N^{\mathrm{REG}})^2}$) we choose $\sigma = N^{\mathrm{REG}}/\sqrt{\alpha}$. The parameter α is a measure of the a priori concentration of λ around the value N^{REG} (see next section).

Next we must find $c \in \mathbb{R}$ such that

$$\inf_{x\geq 0} \frac{c \cdot g(\beta)}{f(x)} \geq 1 \tag{19}$$

Taking the minimal c for sampling efficiency reasons we have

$$c = \sup_{x\geq 0} \frac{f(x)}{g(x)}.$$

To generate random values λ according to $\mathbb{P}\left(\lambda|\mathbf{N}^{\mathrm{MNO}}; \mathbf{N}^{\mathrm{REG}}\right)$ we generate values according to $g(\lambda)$, and values v according to $\mathrm{Unif}(0, 1)$ so that we accept those λ such that $v \leq \frac{f(\lambda)}{c \cdot g(\lambda)}$.

To generate random values N according to $\mathbb{P}\left(N|N^{\mathrm{MNO}}; N^{\mathrm{REG}}\right)$ we generate values λ and then the corresponding values N according to the Poisson distribution with parameter λ.

6 Illustrative examples: toy simulations

Let us illustrate these considerations with concrete examples. We will proceed from the simplest case to the actual data collected during the SGA-1 passing through some simulated data sets.

6.1 The prior distributions

For the model we will make use of the following priors.

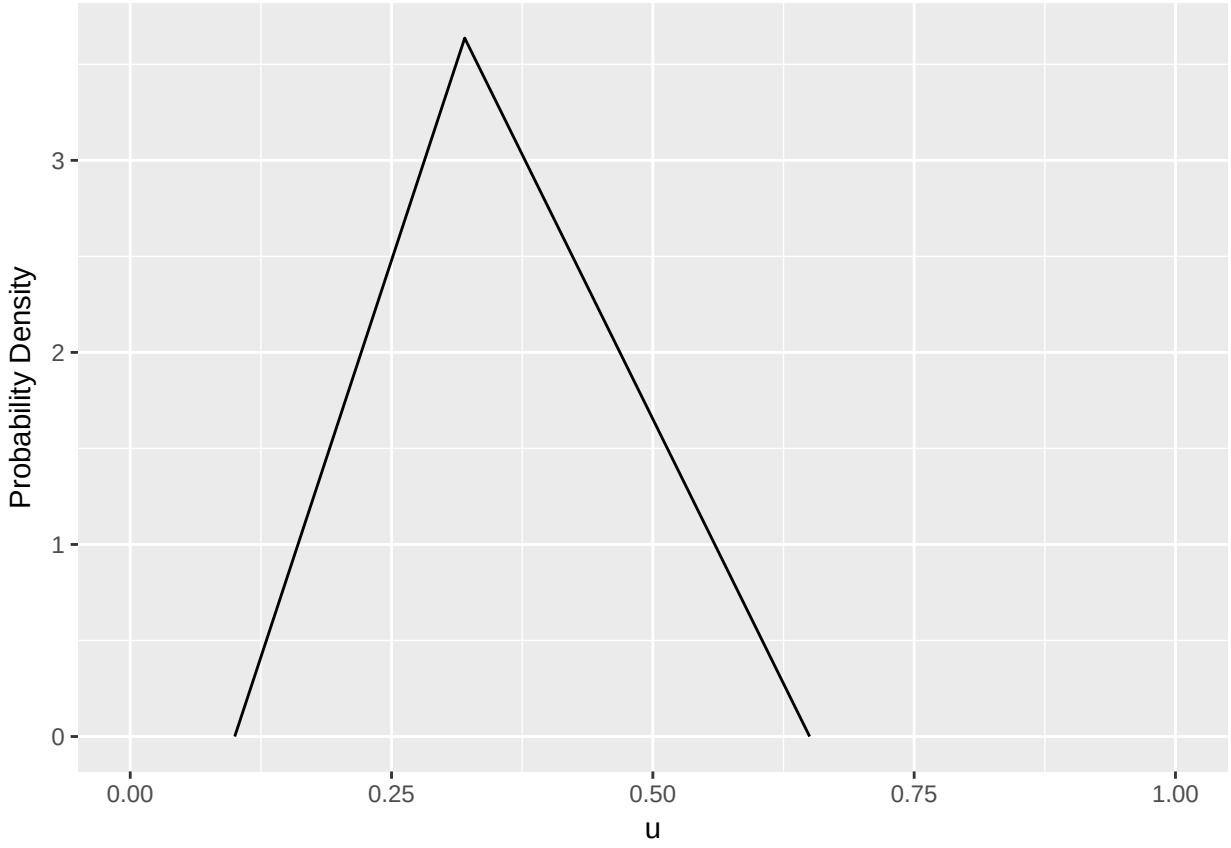
6.1.1 Uniform

This is a well known distribution. For $u = \frac{\alpha}{\alpha+\beta}$ the support can be $[0, 1]$ or any subinterval therein. For example, we can assume that the local market share in a territorial cell lies between a minimum value u_m and a maximum value u_M. This is a safe way to minimally incorporate some weak information into the model.

For $v = \alpha + \beta$ we can also use prior uniform distributions with support in large intervals $[v_m, v_M]$ of possible minimum and maximum values for the cell size.

6.1.2 Triangular

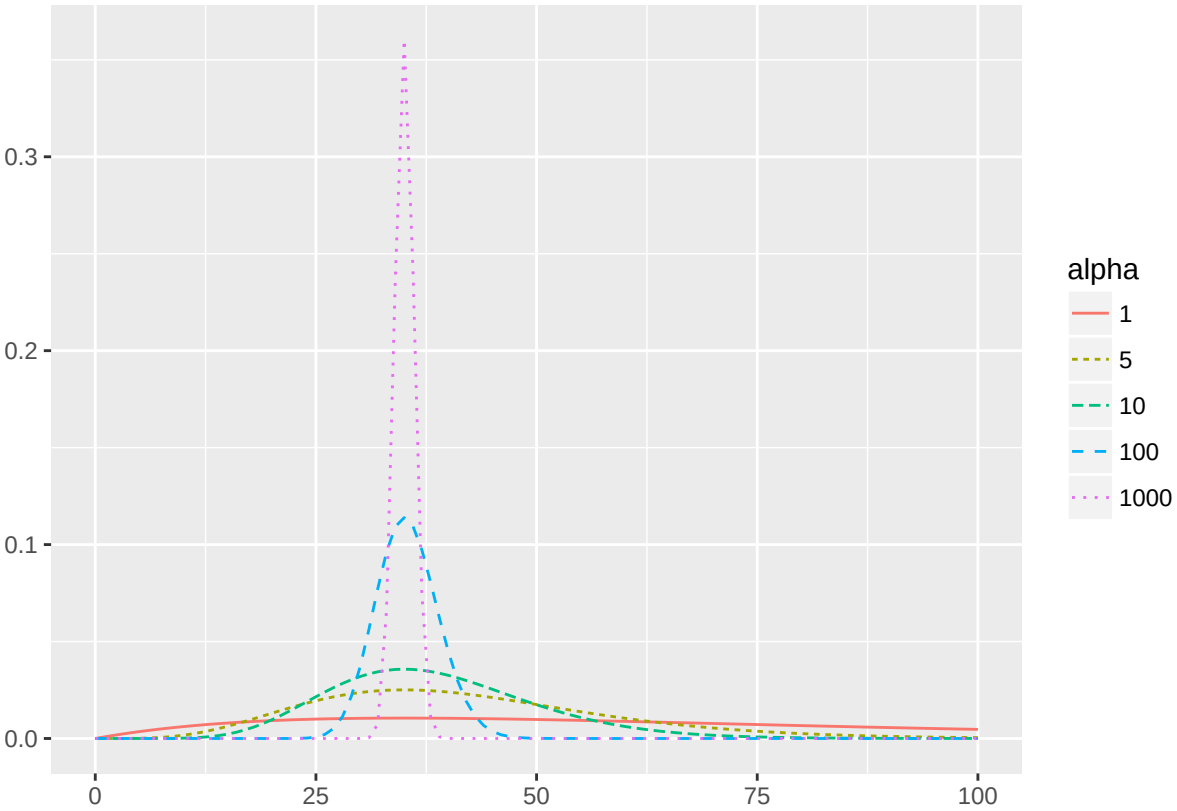
This can be considered a slight variant of the preceding priors. Instead of having a flat distribution along the support, we may single out a mode so that we have probability density functions as



Triangular distributions can be easily generated. They can be used for modelling the local market shares u, the cell size v and the hyperparameter λ.

6.1.3 Gamma distribution

The gamma distribution is another choice for modelling both the cell size v and the hyperparameter λ.The reasoning is common to both cases. We can assume a parameterization $\Gamma(\alpha+1, \xi^*/\alpha)$, where ξ^* stands for the mode of the modelled variable (v or λ) and $\alpha > 0$ determines the degree of concentration around the mode ξ^* (see figure below).



The choice of α must be guided by the data themselves. Thus we do not propose concrete methods until we practise the simulations.

Let us remind that we are combining both aggregated mobile phone data and official data at a given time instant to infer the actual population size in each cell. Official data for a concrete time instant will certainly not represent the exact population size of each cell since both data sources work at very different time scales. However, we can assume that appropriately choosing the time instant we can expect a high correlation between the actual size and official data (see e.g. De Meersman et al. (2016)).

In this sense, for the simulated data that we will generate, for a given actual true value N_i^0 we will simulate a population register value $N^{\text{REG}_i} \simeq \lfloor N(\mu = N_i^0, \sigma = 10\% \cdot N_i^0) \rfloor$. For the corresponding number of individuals detected through the mobile network, we will generate it assuming a proportion of detected individuals randomly between 15% and 40% as realistic figures (see e.g. ESSnet on Big Data WP5 (2017) to compare with market shares as an approximation to these figures).

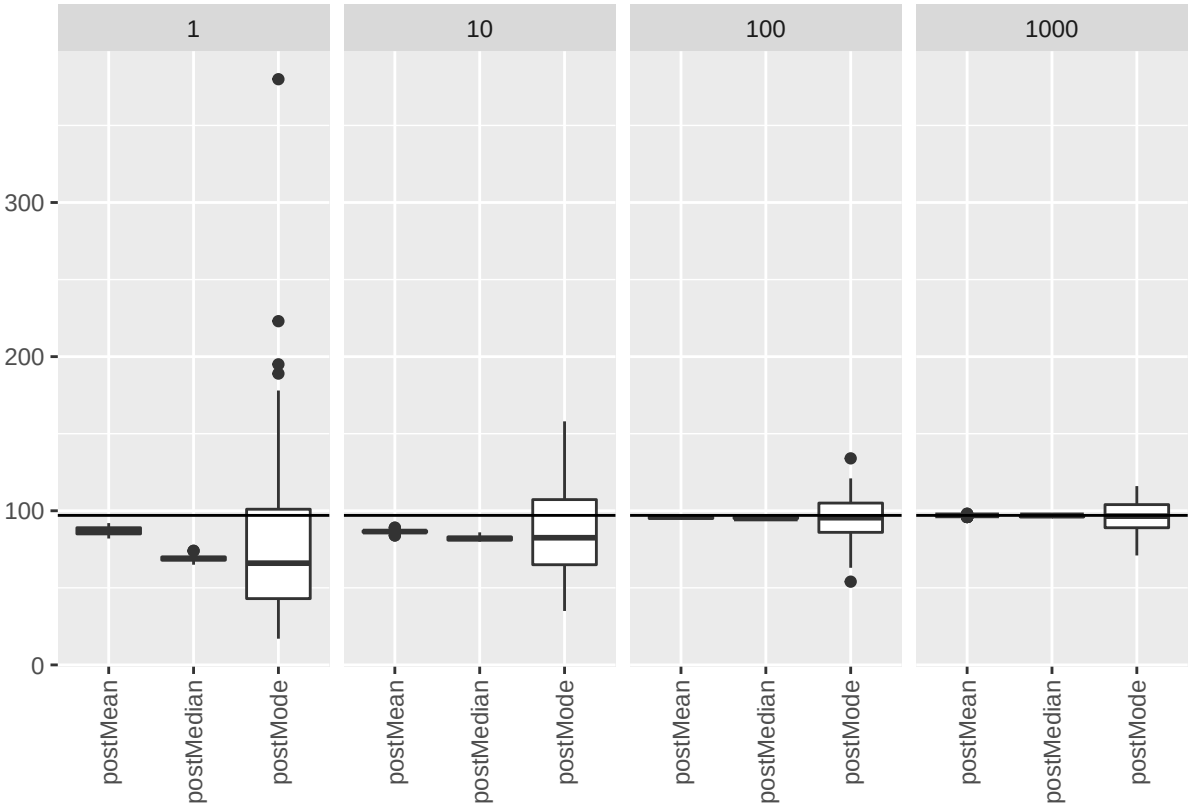
6.2 One cell

Since the treatment of all cells is independent of each other, it is fundamental to get acquainted with the estimation process for a single cell. We investigate different combinations of priors and numerical regimes for N^{MNO} and N^{REG}. In all cases we assume a priori $f_3 \simeq \Gamma\left(\alpha+1, \frac{N^{\text{Reg}}}{\alpha}\right)$.

Firstly, we will investigate how the estimates vary in different realizations of the estimation procedure. Let us consider a true population size of $N^{(0)} = 100$. The population register gives $N^{\text{Reg}} = 97$ assuming an error of 3%. Let us also consider the number of individuals detected by the mobile network as $N^{\text{MNO}} = 19$ assuming a proportion of detected individuals of around 20%.

For the prior distribution of the proportion of detected individuals we will assume a weakly informative distribution $f_u = \text{Unif}(u_m, u_M)$ with $u_m = 0$ and $u_M = 0.50$. For the prior distribution of the cell size we will assume a triangular distribution with parameters $v_m = 87$, $v_M = 107$, and $v^* = 97$, assuming an (unknown) error of 10% over the population register size. Later on we will study the effect of diverse choices on these priors.

We compute the estimates for values of $\alpha = 1, 10, 100, 1000$ to observe the effect of the amount of uncertainty in the population size (from more uncertainty to less uncertainty, respectively).



We can observe how the more precise the prior value of $\lambda = N^{\text{Reg}}$ is, the more precise the final estimate around N^{Reg} will result. Notice how this final estimate inherits the original difference between $N^{(0)}$ and N^{Reg}, as expected (we are not modelling the values coming from the population register).

In the subsequent sections we will just compute one single estimate and vary the diverse prior parameters to study the effect.

6.2.1 $f_1 \simeq \textbf{Unif}(u_m, u_M)$, $f_2 \simeq \textbf{Unif}(N_m, N_M)$

Let us now consider a range of values for the hyperparameters to observe the effects on the final estimate. We will choose $\alpha = 1$ as a weakly informative choice[2].

For the intervals (u_m, u_M) we will choose as centres of the intervals the natural value $N^{\text{MNO}}/N^{\text{Reg}}$. As radii, we will progressively shorten the intervals starting from $r_1 = min(N^{\text{MNO}}/N^{\text{Reg}}, 1 - N^{\text{MNO}}/N^{\text{Reg}})$ down to

[2]The value of α will be later on better motivated when having data from several cells.

0.005.

For the intervals (N_m, N_M) we will choose as centres of the intervals the natural value N^{Reg}. As radii, we will progressively shorten the intervals starting from $R_1 = \lfloor 0.25 \cdot N^{\text{Reg}} \rfloor$ down to 1.

The relative bias $\frac{\hat{N}-N^{\text{Reg}}}{N^{\text{Reg}}} \cdot 100$ for the posterior mean, median and mode estimates, respectively, for each pair of interval lengths $(u_M - u_m, N_M - N_m)$ are:

Table 1: Relative bias (%) for posterior mean estimates

	48	44	38	32	28	22	18	12	8	2
0.39	2.1	6.2	8.2	0.0	5.2	8.2	8.2	7.2	6.2	8.2
0.35	12.4	15.5	10.3	7.2	8.2	10.3	9.3	14.4	9.3	10.3
0.31	7.2	10.3	14.4	11.3	10.3	15.5	11.3	11.3	15.5	10.3
0.26	17.5	11.3	11.3	8.2	13.4	11.3	13.4	13.4	9.3	12.4
0.22	9.3	12.4	8.2	8.2	12.4	10.3	15.5	8.2	9.3	8.2
0.18	10.3	10.3	9.3	8.2	9.3	10.3	9.3	10.3	7.2	9.3
0.14	11.3	6.2	7.2	9.3	6.2	5.2	6.2	8.2	10.3	5.2
0.09	8.2	3.1	5.2	6.2	4.1	8.2	4.1	6.2	6.2	6.2
0.05	5.2	3.1	4.1	4.1	5.2	5.2	4.1	3.1	4.1	6.2
0.01	4.1	4.1	4.1	3.1	6.2	5.2	5.2	4.1	5.2	5.2

Table 2: Relative bias (%) for posterior median estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-14.4	-14.4	-11.3	-15.5	-13.4	-11.3	-11.3	-15.5	-13.4	-11.3
0.35	-6.2	-4.1	-8.2	-7.2	-10.3	-9.3	-9.3	-7.2	-11.3	-9.3
0.31	-7.2	-6.2	-1.0	-3.1	-9.3	-4.1	-6.2	-5.2	-1.0	-7.2
0.26	0.0	-5.2	-3.1	-5.2	-1.0	0.0	0.0	1.0	-3.1	-1.0
0.22	-2.1	2.1	-1.0	0.0	-1.0	0.0	2.1	-1.0	-2.1	-2.1
0.18	4.1	1.0	0.0	1.0	1.0	0.0	0.0	3.1	-1.0	0.0
0.14	3.1	1.0	0.0	4.1	-1.0	-1.0	1.0	2.1	3.1	-2.1
0.09	3.1	-2.1	-1.0	1.0	0.0	5.2	1.0	2.1	1.0	1.0
0.05	1.0	-1.0	0.0	1.0	1.0	0.0	-1.0	0.0	0.0	1.0
0.01	-1.0	1.0	-1.0	-1.0	2.1	1.0	2.1	0.0	1.0	0.0

Table 3: Relative bias (%) for posterior mode estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-17.5	9.3	-15.5	-22.7	6.2	-29.9	-11.3	-39.2	89.7	-33.0
0.35	89.7	16.5	40.2	-55.7	-39.2	22.7	-40.2	126.8	-58.8	-20.6
0.31	-3.1	-8.2	-24.7	32.0	-47.4	88.7	141.2	-21.6	-44.3	15.5
0.26	-7.2	-63.9	8.2	-22.7	-19.6	-40.2	86.6	45.4	-26.8	2.1
0.22	12.4	49.5	70.1	9.3	28.9	5.2	-47.4	-14.4	-44.3	29.9
0.18	1.0	-39.2	8.2	16.5	29.9	78.4	18.6	1.0	60.8	-1.0
0.14	-28.9	-34.0	90.7	52.6	79.4	1.0	3.1	-14.4	135.1	2.1
0.09	15.5	57.7	-26.8	-4.1	-13.4	13.4	39.2	29.9	-59.8	-40.2
0.05	34.0	-8.2	66.0	-9.3	-49.5	-36.1	67.0	33.0	-14.4	-39.2
0.01	-28.9	-46.4	3.1	-16.5	-9.3	1.0	-28.9	11.3	-21.6	-9.3

6.2.2 $f_1 \simeq \mathbf{Unif}(u_m, u_M)$, $f_2 \simeq \mathbf{triang}(N_m, N_M, N^{\mathbf{Reg}})$

We now carry out the same computation with the prior distribution for the actual population size f_2 being a triangular distribution. Its parameters will be N_m and N_M as in the preceding section and the mode as $N^* = N^{\mathrm{Reg}}$.

Table 4: Relative bias (%) for posterior mean estimates

	48	44	38	32	28	22	18	12	8	2
0.39	8.2	6.2	5.2	8.2	5.2	9.3	8.2	6.2	4.1	6.2
0.35	9.3	8.2	8.2	12.4	10.3	6.2	3.1	12.4	10.3	11.3
0.31	14.4	11.3	12.4	13.4	12.4	10.3	13.4	12.4	9.3	5.2
0.26	12.4	10.3	15.5	13.4	12.4	14.4	10.3	12.4	12.4	18.6
0.22	11.3	7.2	11.3	13.4	10.3	12.4	10.3	8.2	7.2	12.4
0.18	11.3	10.3	8.2	10.3	9.3	8.2	11.3	10.3	10.3	9.3
0.14	9.3	9.3	8.2	9.3	8.2	5.2	6.2	6.2	6.2	7.2
0.09	5.2	5.2	5.2	5.2	8.2	6.2	5.2	6.2	6.2	4.1
0.05	5.2	5.2	3.1	6.2	5.2	7.2	6.2	3.1	6.2	3.1
0.01	4.1	3.1	3.1	6.2	6.2	4.1	4.1	4.1	5.2	3.1

Table 5: Relative bias (%) for posterior median estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-10.3	-13.4	-15.5	-11.3	-14.4	-11.3	-11.3	-11.3	-15.5	-12.4
0.35	-7.2	-6.2	-11.3	-5.2	-9.3	-13.4	-10.3	-8.2	-8.2	-8.2
0.31	-1.0	-7.2	-7.2	-5.2	-6.2	-5.2	-4.1	-5.2	-8.2	-7.2
0.26	-1.0	-3.1	-2.1	-1.0	-3.1	-3.1	-2.1	-4.1	-3.1	5.2
0.22	0.0	-4.1	1.0	0.0	-4.1	2.1	1.0	-2.1	-2.1	1.0
0.18	3.1	1.0	1.0	2.1	-1.0	-2.1	3.1	-1.0	4.1	1.0
0.14	2.1	3.1	0.0	1.0	0.0	-1.0	0.0	-1.0	-1.0	-1.0
0.09	1.0	2.1	0.0	1.0	3.1	0.0	0.0	2.1	2.1	0.0
0.05	1.0	0.0	-2.1	2.1	1.0	2.1	2.1	-1.0	2.1	2.1
0.01	1.0	-2.1	-1.0	2.1	4.1	1.0	1.0	1.0	2.1	-1.0

Table 6: Relative bias (%) for posterior mode estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-11.3	37.1	47.4	-52.6	-66.0	-29.9	-59.8	74.2	-5.2	-68.0
0.35	10.3	-5.2	41.2	13.4	249.5	-38.1	-13.4	2.1	18.6	-64.9
0.31	20.6	-34.0	3.1	99.0	-23.7	-29.9	-19.6	162.9	-37.1	-50.5
0.26	-22.7	15.5	1.0	78.4	-25.8	-4.1	-53.6	6.2	50.5	-39.2
0.22	-16.5	-29.9	101.0	91.8	37.1	-52.6	9.3	-9.3	83.5	-56.7
0.18	30.9	-22.7	54.6	-40.2	-43.3	21.6	40.2	104.1	-30.9	-36.1
0.14	158.8	41.2	51.5	-28.9	-47.4	39.2	-30.9	32.0	-17.5	-6.2
0.09	48.5	7.2	23.7	0.0	47.4	-50.5	68.0	-40.2	-36.1	-36.1
0.05	-5.2	25.8	42.3	-14.4	-2.1	-30.9	52.6	41.2	14.4	15.5
0.01	55.7	-6.2	-43.3	92.8	9.3	-29.9	-14.4	-36.1	3.1	10.3

6.2.3 $f_1 \simeq \mathbf{Unif}(u_m, u_M)$, $f_2 \simeq \Gamma(a+1, \frac{N^{\mathbf{Reg}}}{a})$

Again we repeat the computation now with $f_2 \simeq \Gamma(a+1, \frac{N^{\mathrm{Reg}}}{a})$ and $\log_{10}(a) = -3, -2, \ldots, 2, 3$.

Table 7: Relative bias (%) for posterior mean estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	-13.4	-18.6	-13.4	5.2	6.2	5.2	4.1
0.35	-14.4	-9.3	-8.2	14.4	14.4	11.3	11.3
0.31	-16.5	-8.2	-7.2	7.2	13.4	8.2	10.3
0.26	-13.4	-16.5	-12.4	8.2	13.4	10.3	15.5
0.22	-10.3	-13.4	-10.3	18.6	8.2	11.3	12.4
0.18	-18.6	-13.4	-9.3	10.3	9.3	9.3	11.3
0.14	-19.6	-14.4	-6.2	10.3	6.2	6.2	7.2
0.09	-15.5	-10.3	-1.0	9.3	7.2	7.2	6.2
0.05	-10.3	-19.6	-7.2	8.2	4.1	3.1	5.2
0.01	-14.4	-19.6	-12.4	8.2	5.2	2.1	3.1

Table 8: Relative bias (%) for posterior median estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	-46.4	-51.5	-42.3	-11.3	-12.4	-12.4	-13.4
0.35	-49.5	-47.4	-41.2	-9.3	-6.2	-7.2	-7.2
0.31	-50.5	-42.3	-44.3	-7.2	0.0	-7.2	-6.2
0.26	-48.5	-49.5	-43.3	-7.2	-1.0	-4.1	-2.1
0.22	-45.4	-47.4	-42.3	3.1	-2.1	0.0	2.1
0.18	-48.5	-47.4	-42.3	0.0	-1.0	-1.0	1.0
0.14	-49.5	-46.4	-38.1	1.0	-1.0	-1.0	1.0
0.09	-50.5	-44.3	-27.8	-1.0	3.1	1.0	-1.0
0.05	-43.3	-52.6	-38.1	-1.0	0.0	-3.1	0.0
0.01	-51.5	-48.5	-40.2	-1.0	2.1	0.0	0.0

Table 9: Relative bias (%) for posterior mode estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	-81.4	-61.9	-57.7	277.3	-33.0	44.3	-60.8
0.35	-45.4	216.5	-61.9	-87.6	13.4	40.2	-27.8
0.31	-38.1	30.9	-78.4	-10.3	-41.2	-67.0	47.4
0.26	-49.5	-68.0	-43.3	-8.2	-37.1	64.9	113.4
0.22	-35.1	-86.6	-67.0	-44.3	-30.9	-10.3	-47.4
0.18	-81.4	171.1	118.6	-79.4	1.0	-57.7	-28.9
0.14	-88.7	-86.6	-29.9	-5.2	14.4	-37.1	28.9
0.09	-6.2	-43.3	-27.8	-19.6	-38.1	33.0	24.7
0.05	-17.5	-79.4	42.3	17.5	51.5	-33.0	20.6
0.01	76.3	-68.0	-84.5	-67.0	-22.7	29.9	4.1

6.2.4 $f_1 \simeq$ **Triang**(u_m, u_M, u^*), $f_2 \simeq$ **Unif**(N_m, N_M)

Once more, we repeat the computation now with $f_1 \simeq$ Triang and $f_2 \simeq$ Unif. The hyperparameters are chosen as before.

Table 10: Relative bias (%) for posterior mean estimates

	48	44	38	32	28	22	18	12	8	2
0.39	10.3	7.2	9.3	7.2	8.2	8.2	9.3	7.2	9.3	7.2
0.35	9.3	11.3	15.5	12.4	9.3	8.2	10.3	10.3	10.3	11.3
0.31	11.3	10.3	10.3	8.2	11.3	11.3	8.2	13.4	12.4	11.3
0.26	8.2	10.3	5.2	9.3	7.2	5.2	8.2	8.2	10.3	11.3
0.22	9.3	10.3	7.2	9.3	11.3	8.2	7.2	6.2	7.2	11.3
0.18	5.2	7.2	7.2	8.2	7.2	6.2	8.2	6.2	8.2	8.2
0.14	6.2	5.2	6.2	5.2	4.1	7.2	7.2	7.2	6.2	6.2
0.09	3.1	6.2	5.2	6.2	4.1	4.1	5.2	3.1	3.1	7.2
0.05	5.2	5.2	5.2	5.2	6.2	4.1	5.2	5.2	5.2	4.1
0.01	5.2	5.2	5.2	5.2	6.2	5.2	4.1	4.1	4.1	6.2

Table 11: Relative bias (%) for posterior median estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-3.1	-4.1	-2.1	-4.1	-5.2	-3.1	-2.1	-7.2	-3.1	-5.2
0.35	-3.1	-2.1	0.0	1.0	-3.1	0.0	-2.1	0.0	-2.1	0.0
0.31	-1.0	-1.0	1.0	0.0	1.0	2.1	0.0	2.1	4.1	-1.0
0.26	2.1	3.1	-3.1	0.0	-1.0	-4.1	1.0	-1.0	-2.1	1.0
0.22	1.0	3.1	1.0	2.1	3.1	1.0	1.0	-1.0	0.0	3.1
0.18	0.0	2.1	3.1	1.0	1.0	-1.0	1.0	0.0	3.1	3.1
0.14	1.0	0.0	1.0	0.0	-1.0	1.0	2.1	2.1	1.0	0.0
0.09	0.0	2.1	-1.0	2.1	0.0	-1.0	0.0	0.0	0.0	3.1
0.05	3.1	-1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	-1.0
0.01	1.0	1.0	1.0	0.0	3.1	1.0	-1.0	-1.0	2.1	2.1

Table 12: Relative bias (%) for posterior mode estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-5.2	201.0	-24.7	139.2	-27.8	42.3	5.2	87.6	27.8	-45.4
0.35	3.1	-41.2	60.8	134.0	-2.1	57.7	22.7	16.5	9.3	-30.9
0.31	15.5	-13.4	42.3	57.7	-18.6	0.0	-58.8	46.4	-40.2	-34.0
0.26	22.7	80.4	22.7	-38.1	29.9	-28.9	11.3	-61.9	-37.1	-9.3
0.22	-25.8	14.4	-44.3	14.4	-14.4	17.5	106.2	-7.2	-25.8	-3.1
0.18	-17.5	84.5	-16.5	15.5	2.1	85.6	2.1	24.7	119.6	-20.6
0.14	19.6	29.9	-14.4	-11.3	7.2	-25.8	51.5	19.6	-29.9	-19.6
0.09	-24.7	-41.2	-46.4	-34.0	4.1	-8.2	6.2	-5.2	5.2	24.7
0.05	-5.2	-40.2	9.3	41.2	74.2	-7.2	-16.5	-28.9	156.7	73.2
0.01	24.7	4.1	1.0	-30.9	70.1	-10.3	14.4	5.2	14.4	6.2

6.2.5 $f_1 \simeq \textbf{Triang}(u_m, u_M, u^*)$, $f_2 \simeq \textbf{Triang}(N_m, N_M, N^{\textbf{Reg}})$

Now both prior distributions are triangular with the same choice of hyperparameters are before.

Table 13: Relative bias (%) for posterior mean estimates

	48	44	38	32	28	22	18	12	8	2
0.39	10.3	10.3	7.2	10.3	10.3	11.3	9.3	9.3	9.3	11.3
0.35	9.3	12.4	9.3	13.4	9.3	11.3	11.3	11.3	11.3	12.4
0.31	9.3	9.3	10.3	13.4	7.2	11.3	12.4	10.3	10.3	10.3
0.26	10.3	9.3	9.3	10.3	8.2	9.3	7.2	10.3	12.4	9.3
0.22	8.2	9.3	5.2	9.3	7.2	5.2	6.2	9.3	10.3	11.3
0.18	9.3	7.2	9.3	6.2	7.2	9.3	10.3	8.2	8.2	8.2
0.14	6.2	6.2	7.2	5.2	6.2	3.1	6.2	4.1	6.2	5.2
0.09	5.2	2.1	5.2	7.2	5.2	4.1	6.2	6.2	7.2	6.2
0.05	4.1	4.1	4.1	5.2	5.2	2.1	6.2	7.2	4.1	4.1
0.01	5.2	3.1	5.2	4.1	5.2	4.1	5.2	5.2	5.2	5.2

Table 14: Relative bias (%) for posterior median estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-1.0	-2.1	-5.2	-3.1	-2.1	-2.1	-3.1	-5.2	-2.1	-5.2
0.35	-2.1	-1.0	-3.1	1.0	-1.0	1.0	-1.0	-1.0	-1.0	1.0
0.31	0.0	-1.0	1.0	3.1	0.0	0.0	1.0	-3.1	1.0	0.0
0.26	4.1	2.1	2.1	0.0	1.0	-1.0	-1.0	1.0	4.1	-2.1
0.22	2.1	2.1	0.0	1.0	1.0	-2.1	-1.0	2.1	5.2	1.0
0.18	4.1	1.0	2.1	0.0	1.0	3.1	2.1	0.0	2.1	3.1
0.14	-1.0	1.0	3.1	-1.0	1.0	-2.1	3.1	-1.0	0.0	2.1
0.09	1.0	-2.1	0.0	2.1	1.0	-1.0	1.0	2.1	1.0	1.0
0.05	1.0	0.0	1.0	1.0	-1.0	-1.0	2.1	3.1	1.0	1.0
0.01	1.0	-2.1	1.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0

Table 15: Relative bias (%) for posterior mode estimates

	48	44	38	32	28	22	18	12	8	2
0.39	-24.7	35.1	-22.7	38.1	337.1	-27.8	8.2	50.5	-34.0	49.5
0.35	25.8	2.1	105.2	125.8	51.5	74.2	-58.8	17.5	-33.0	-30.9
0.31	49.5	-16.5	-6.2	-49.5	-49.5	119.6	-39.2	29.9	-54.6	33.0
0.26	-22.7	-1.0	64.9	40.2	-11.3	29.9	-26.8	-41.2	10.3	-38.1
0.22	71.1	-56.7	-9.3	-11.3	17.5	59.8	-53.6	46.4	11.3	14.4
0.18	4.1	5.2	-15.5	-32.0	2.1	43.3	-10.3	13.4	56.7	-46.4
0.14	-30.9	-8.2	28.9	40.2	30.9	69.1	16.5	-66.0	-41.2	-29.9
0.09	-3.1	-15.5	44.3	-40.2	-39.2	-11.3	33.0	17.5	43.3	13.4
0.05	20.6	42.3	-34.0	58.8	2.1	-40.2	-24.7	-8.2	22.7	78.4
0.01	-1.0	14.4	32.0	12.4	-6.2	33.0	78.4	34.0	-24.7	-11.3

6.2.6 $f_1 \simeq \textbf{Triang}(u_m, u_M, u^*)$, $f_2 \simeq \Gamma(a+1, \frac{N^{\textbf{Reg}}}{a})$

Finally we combine a triangular distribution for f_1 and a gamma distribution for f_2.

Table 16: Relative bias (%) for posterior mean estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	-15.5	-16.5	-9.3	8.2	10.3	8.2	11.3
0.35	-14.4	-13.4	-12.4	17.5	12.4	11.3	14.4
0.31	-14.4	-16.5	-11.3	11.3	8.2	10.3	10.3
0.26	-13.4	-11.3	-8.2	10.3	8.2	10.3	7.2
0.22	-16.5	-12.4	-4.1	11.3	7.2	8.2	7.2
0.18	-12.4	-11.3	-9.3	10.3	8.2	7.2	6.2
0.14	-15.5	-15.5	-8.2	12.4	4.1	7.2	6.2
0.09	-11.3	-15.5	-6.2	10.3	4.1	7.2	4.1
0.05	-10.3	-14.4	-9.3	7.2	5.2	5.2	5.2
0.01	-13.4	-11.3	-15.5	12.4	4.1	5.2	5.2

Table 17: Relative bias (%) for posterior median estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	-48.5	-48.5	-40.2	-8.2	-2.1	-5.2	-2.1
0.35	-46.4	-46.4	-41.2	5.2	0.0	1.0	1.0
0.31	-46.4	-50.5	-43.3	-3.1	-1.0	-1.0	2.1
0.26	-45.4	-47.4	-36.1	-2.1	0.0	1.0	-1.0
0.22	-49.5	-46.4	-35.1	-2.1	-2.1	-1.0	0.0
0.18	-50.5	-49.5	-39.2	-1.0	0.0	1.0	0.0
0.14	-49.5	-47.4	-37.1	1.0	-2.1	3.1	1.0
0.09	-45.4	-49.5	-37.1	2.1	0.0	1.0	0.0
0.05	-50.5	-48.5	-42.3	-2.1	-1.0	2.1	1.0
0.01	-50.5	-46.4	-44.3	3.1	0.0	1.0	1.0

Table 18: Relative bias (%) for posterior mode estimates

	0.001	0.01	0.1	1	10	100	1000
0.39	155.7	-79.4	-9.3	141.2	68.0	-24.7	102.1
0.35	-61.9	-3.1	-42.3	119.6	10.3	4.1	53.6
0.31	-48.5	-43.3	-83.5	8.2	-7.2	-16.5	-29.9
0.26	-56.7	-79.4	-51.5	114.4	23.7	23.7	59.8
0.22	0.0	-75.3	112.4	-80.4	73.2	25.8	0.0
0.18	-11.3	81.4	-57.7	62.9	73.2	-22.7	-33.0
0.14	-80.4	-20.6	109.3	-39.2	-17.5	-9.3	-22.7
0.09	-62.9	23.7	-9.3	70.1	13.4	37.1	16.5
0.05	-44.3	-79.4	43.3	15.5	6.2	49.5	70.1
0.01	-77.3	158.8	-2.1	4.1	7.2	-40.2	-37.1

6.3 Several cells

Since by design the estimation in each cell is independent of each other, the process for a grid of cells only entails longer computation times. However, having more information allows us to seek for more objective criteria to set the prior information.

Regarding the prior distribution for the proportions of detected individuals u_i we can consider the ratios

$\frac{N_i^{\text{MNO}}}{N_i^{\text{Reg}}}$ as an initial guess as a highly probable value whose uncertainty will depend both on the process to obtain N_i^{MNO} (in the preprocessing and aggregation stages of the whole process for mobile phone data; not considered here) and on the process to compile the population register figures N_i^{Reg} (with non-sampling error considerations like measurement errors, processing errors, coverage, ...). Any of the three prior distributions (uniform, triangular, gamma) can be used to express this uncertainty around $\frac{N_i^{\text{MNO}}}{N_i^{\text{Reg}}}$ and, if necessary, even more complex alternative possibilities can be built.

Regarding the prior distribution for the local cell size v_i we can make exactly similar considerations around the value N_i^{Reg} for each cell i now only focusing on the process of construction of the population register.

Finally, for the prior distribution for the parameter λ_i we must clearly state how critical this choice is. If we choose $\alpha_i \gg 1$, we are expressing a high confidence on our population register as the true population at this instant of time. In this sense it seems advisable to be conservative and choose low values so that we do not artificially "force" the final estimates to be close to N_i^{Reg}. In the choice of α_i in the multiple cell case, however, we can make use of the grid construction and the distribution of N_i^{Reg} to propose some prior values. The variance of $\Gamma(\alpha_i + 1, \frac{N_i^{\text{Reg}}}{\alpha_i})$ is $\frac{\alpha_i+1}{\alpha_i^2} \cdot N_i^{\text{Reg}}$. Thus, under the assumption of having a regular grid over the population, we can equate $\frac{\alpha_i+1}{\alpha_i^2} \cdot N^{\text{Reg}} = \frac{1}{N_{cells}-1} \sum_{i=1}^{N_{cell}} \left(N_i^{\text{Reg}} - \bar{N}^{\text{Reg}}\right)^2$ to have a first value (upper bound) for $\alpha \leq \min_i \alpha_i$. In case of an irregular grid (in terms of population size), we can think of some clustering analysis to carry out the same procedure.

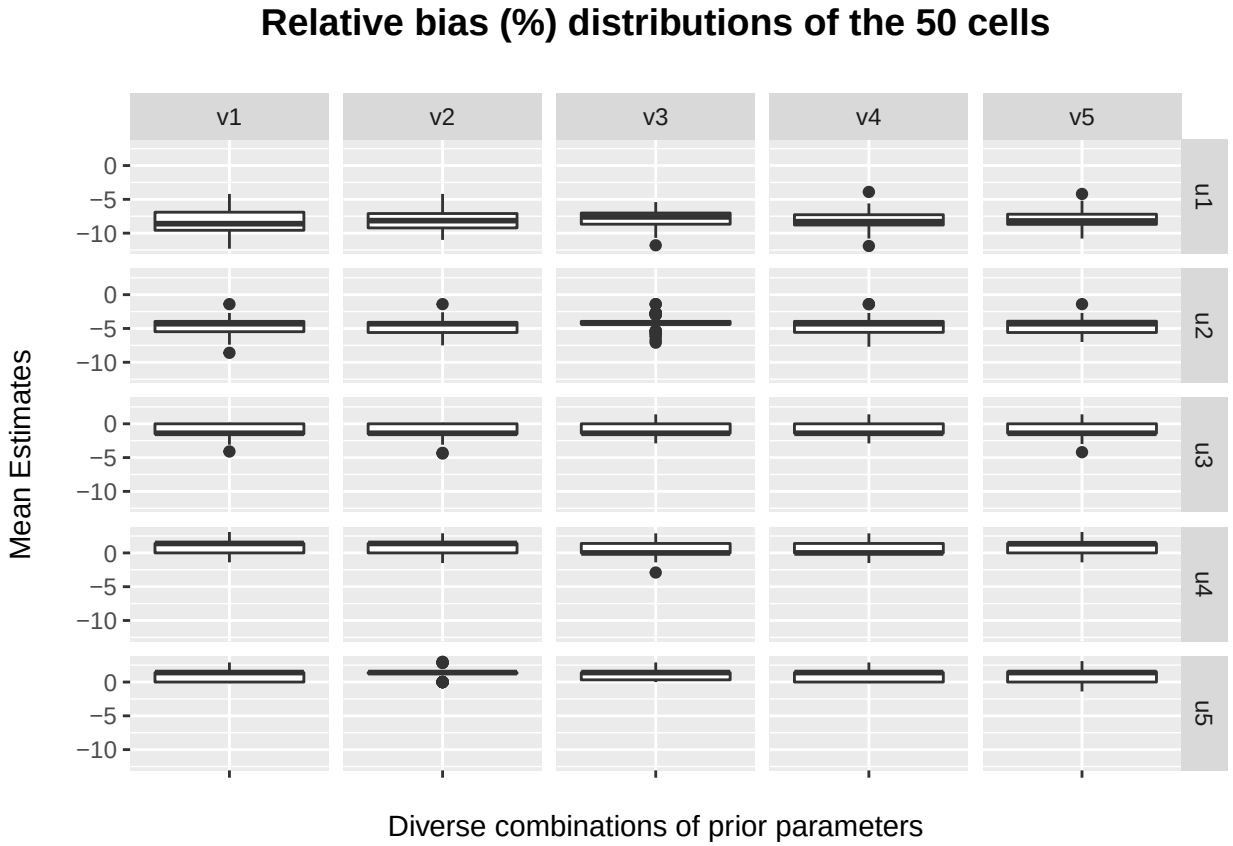
In any case, we claim that one can envisage objective estimation processes for the prior hyperparameters so that final estimates are not based upon dangerously subjective beliefs. In our view, this is another indication of why having knowledge about the preprocessing and aggregation phases in the whole process is important to arrive at final objective values to be considered official figures.

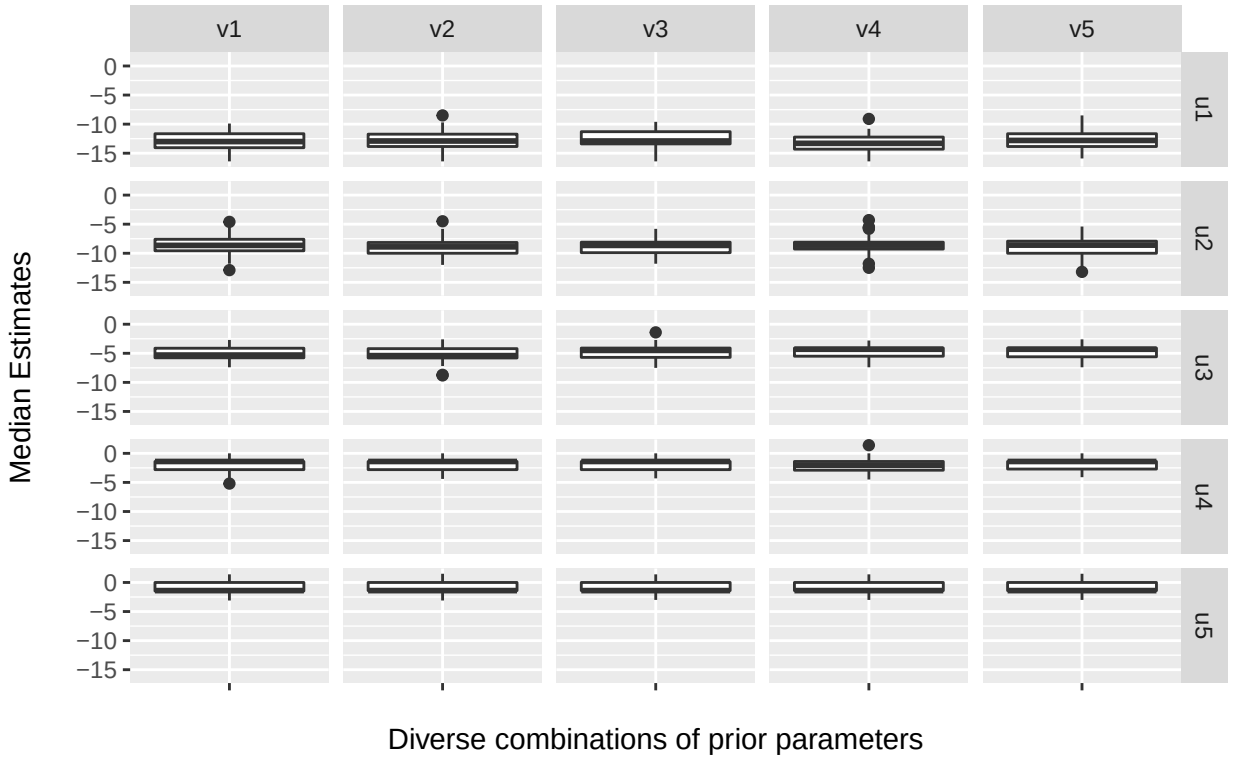
6.3.1 $f_1 \simeq \textbf{Unif}(u_{m,i}, u_{M,i})$, $f_2 \simeq \textbf{Unif}(N_{m,i}, N_{M,i})$

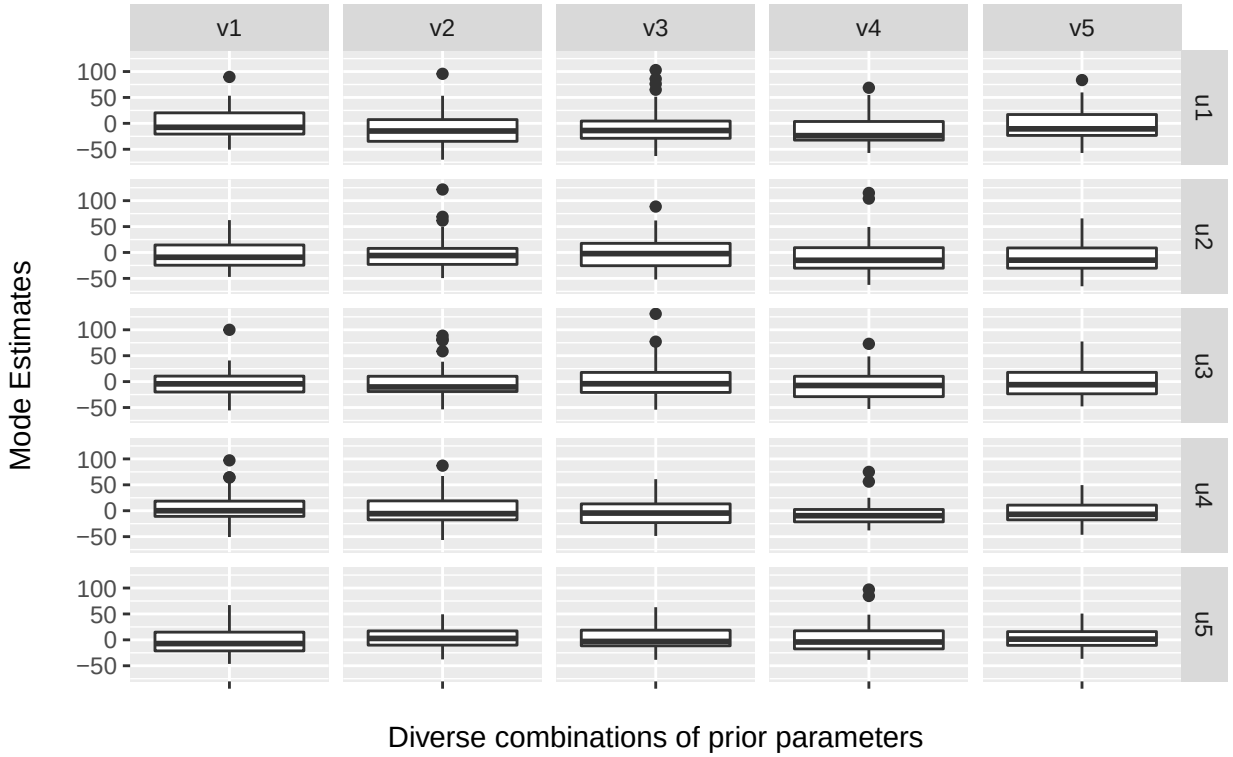
Let us now consider a range of values for the hyperparameters to observe the effects on the final estimate. For the intervals $(u_{m,i}, u_{M,i})$ we will choose as centres of the intervals the natural values $N_i^{\text{MNO}}/N_i^{\text{Reg}}$. As radii, we will progressively shorten the intervals starting from $r_{1,i} = min(N_i^{\text{MNO}}/N_i^{\text{Reg}}, 1 - N_i^{\text{MNO}}/N_i^{\text{Reg}})$ down to 0.005.

For the intervals $(N_{m,i}, N_{M,i})$ we will choose as centres of the intervals the natural values N_i^{Reg}. As radii, we will progressively shorten the intervals starting from $R_{1,i} = \lfloor 0.25 \cdot N_i^{\text{Reg}} \rfloor$ down to 1.

The distribution of the relative bias $\frac{\hat{N}_i - N_i^{\text{Reg}}}{N^{\text{Reg}}} \cdot 100$ for the posterior mean, median and mode estimates, respectively, for all pairs of interval lengths $(u_{M,i} - u_{m,i}, N_{M,i} - N_{m,i})$ and all cells ($N_c = 50$) are:

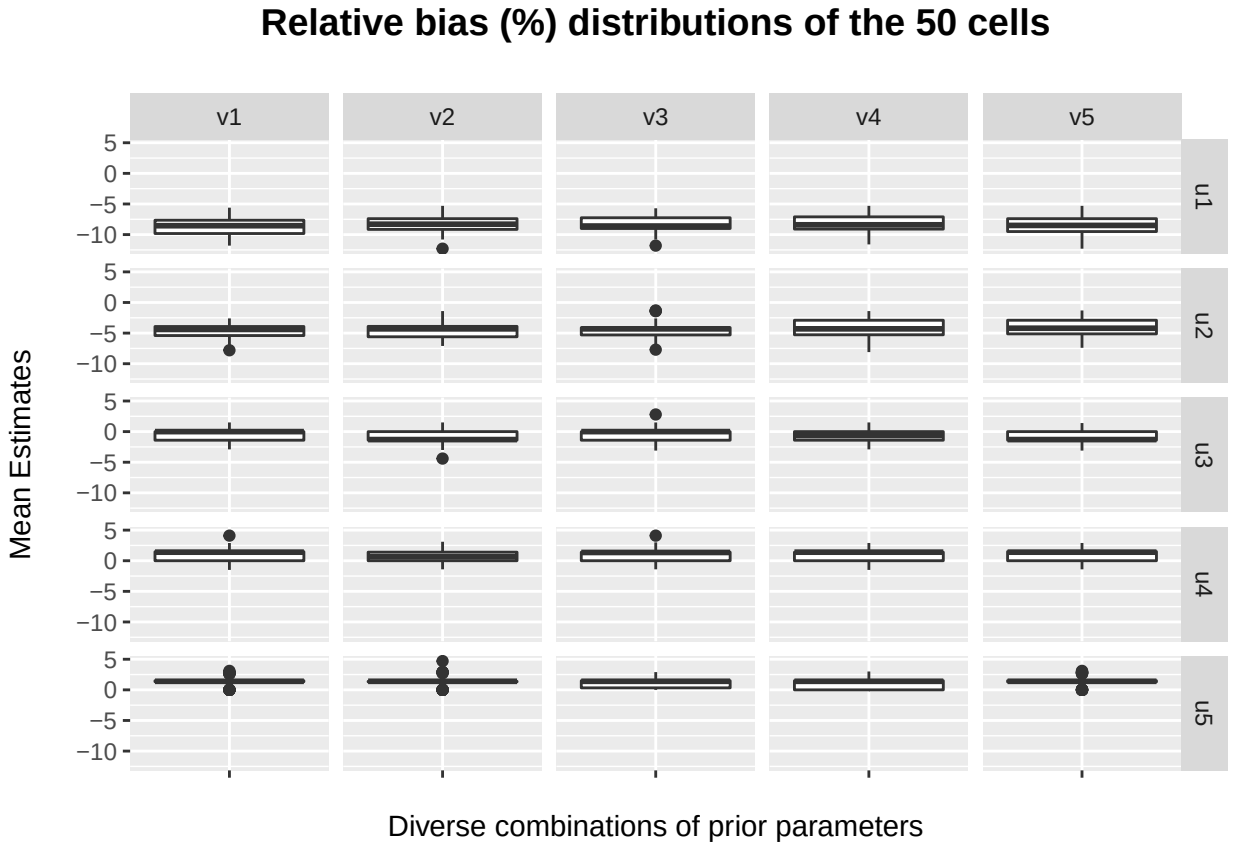

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
0
−5
−10
Mean Estimates
Diverse combinations of prior parameters

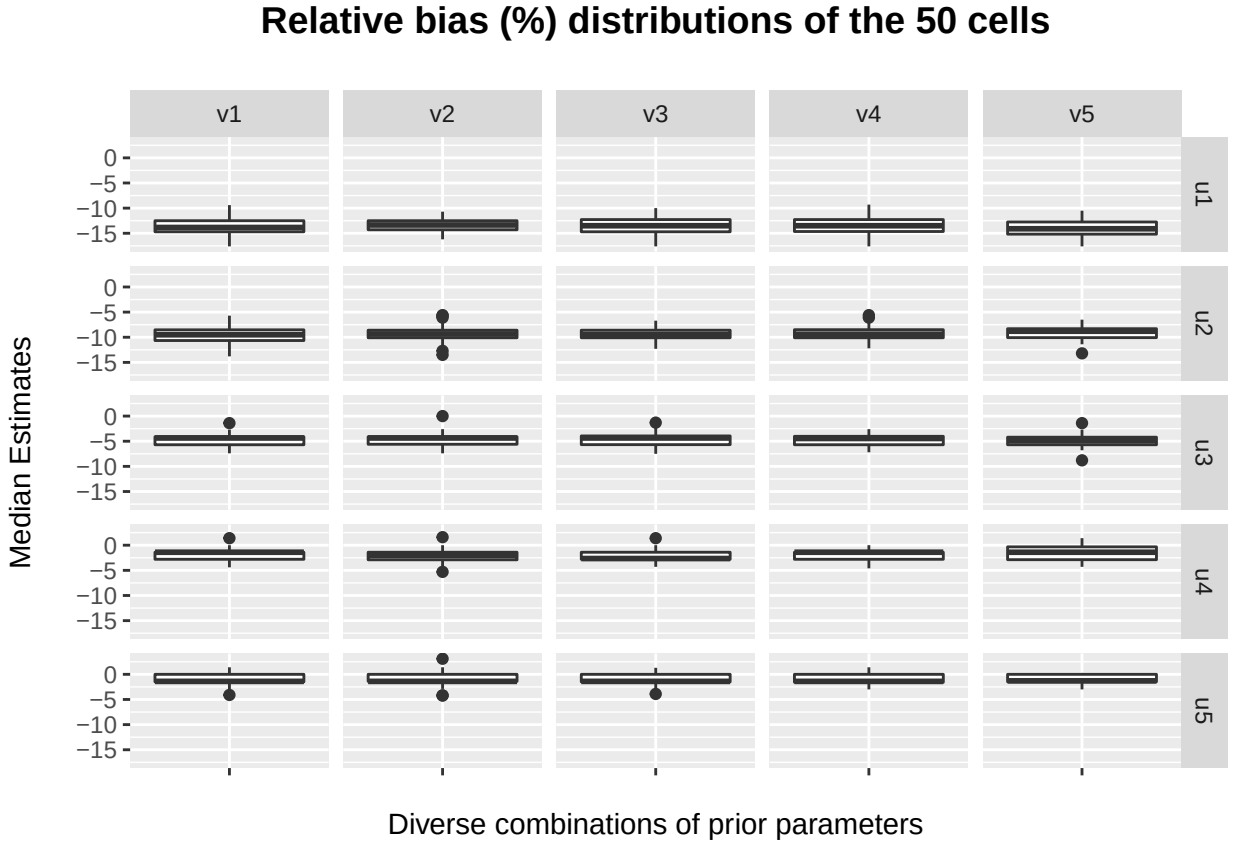

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
0
−5
−10
−15
Median Estimates
Diverse combinations of prior parameters

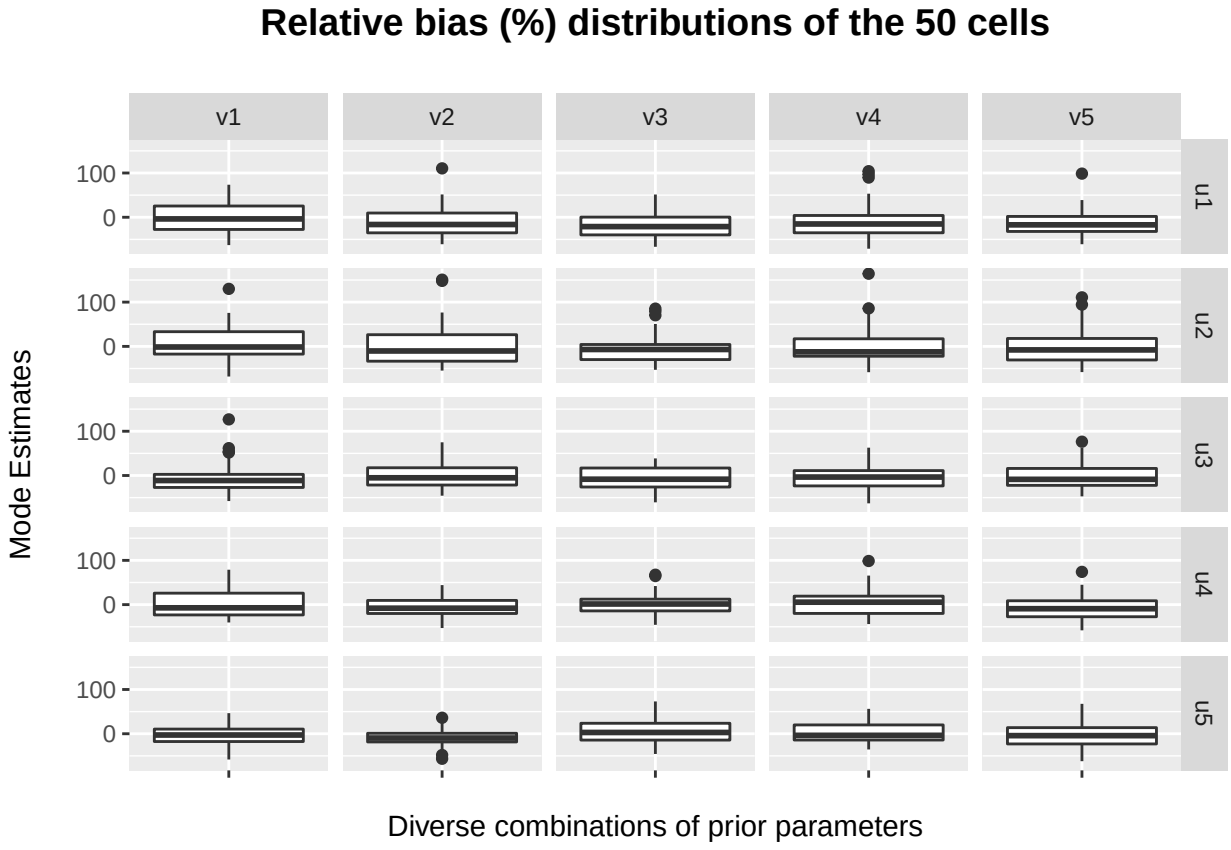


6.3.2 $f_1 \simeq \mathbf{Unif}(u_m, u_M)$, $f_2 \simeq \mathbf{triang}(N_m, N_M, N^{\mathbf{Reg}})$

We now carry out the same computation with the prior distribution for the actual population size f_2 being a triangular distribution. Its parameters will be N_m and N_M as in the preceding section and the mode as $N^* = N^{\mathrm{Reg}}$.

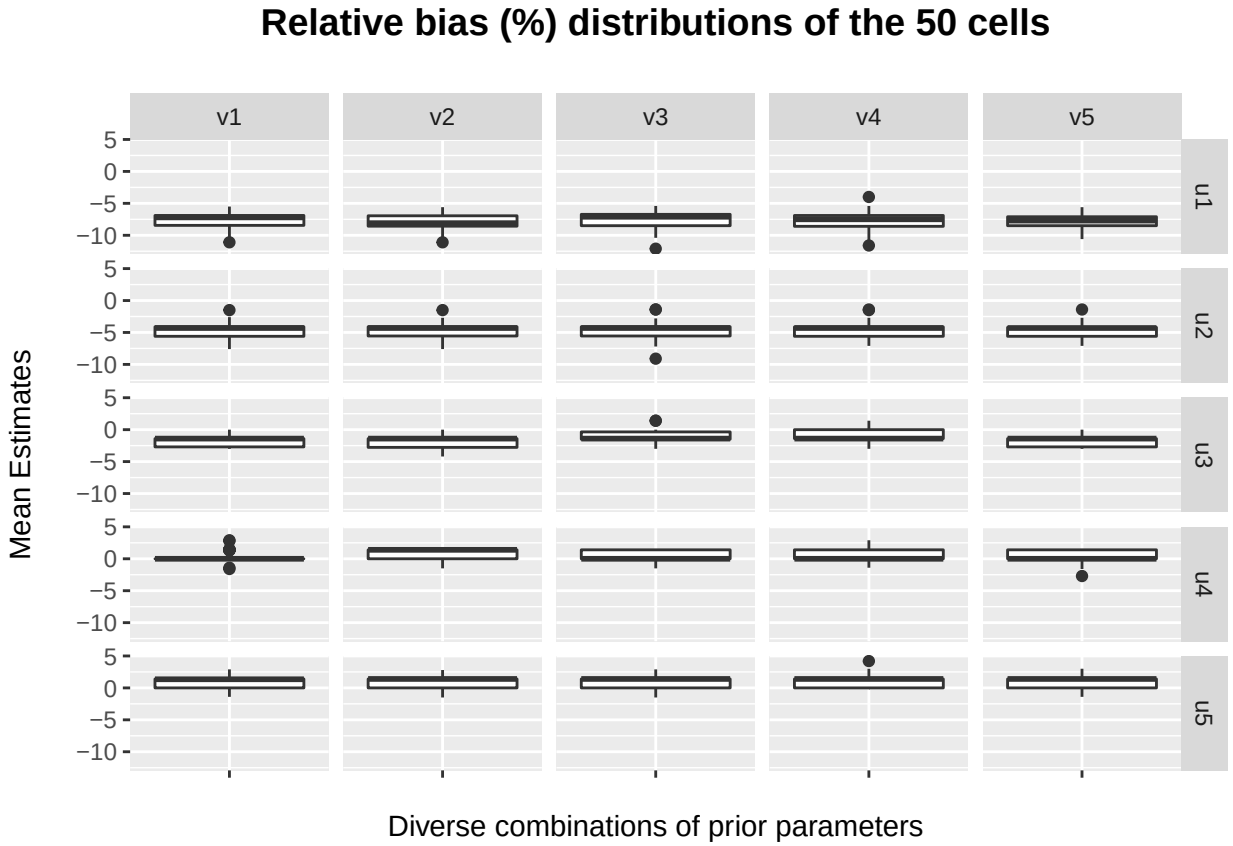

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
5
0
−5
−10
Mean Estimates
Diverse combinations of prior parameters

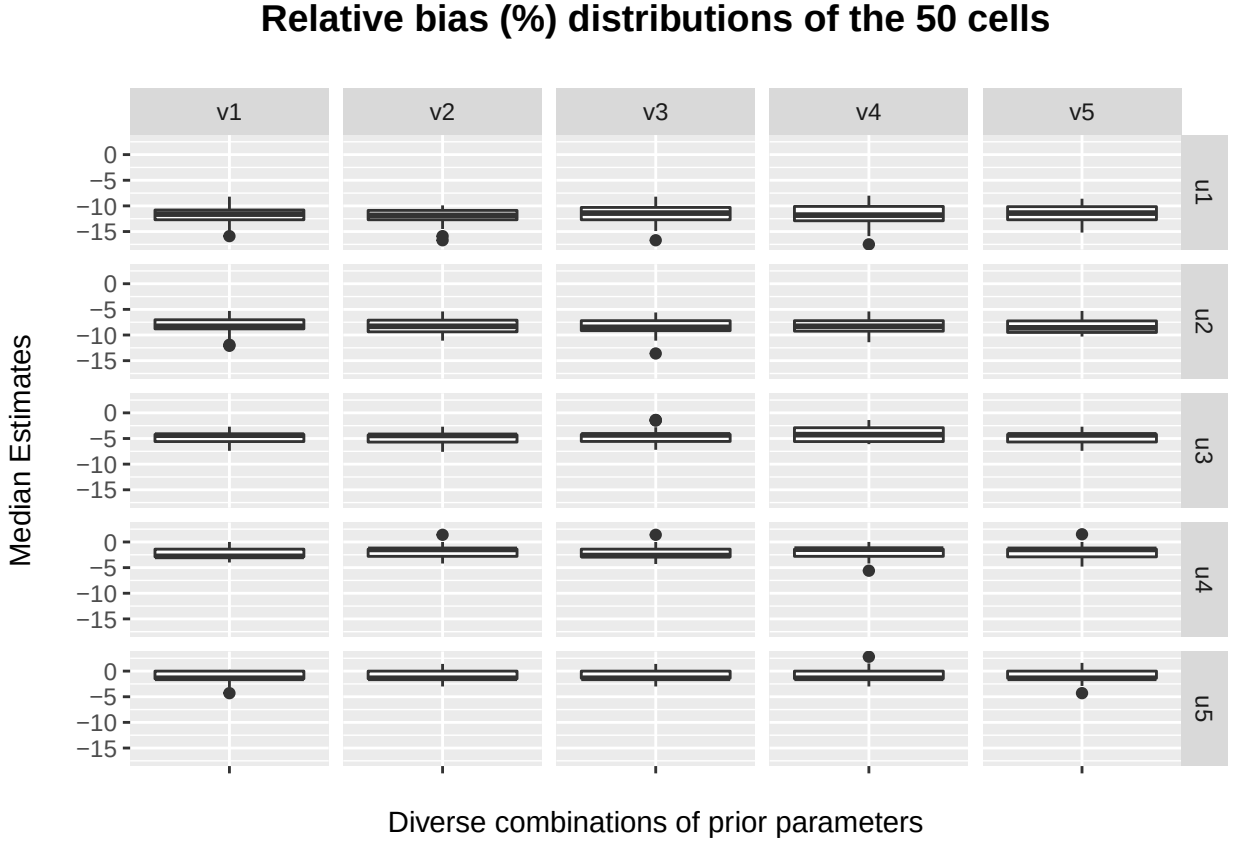

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
0
−5
−10
−15
Median Estimates
Diverse combinations of prior parameters

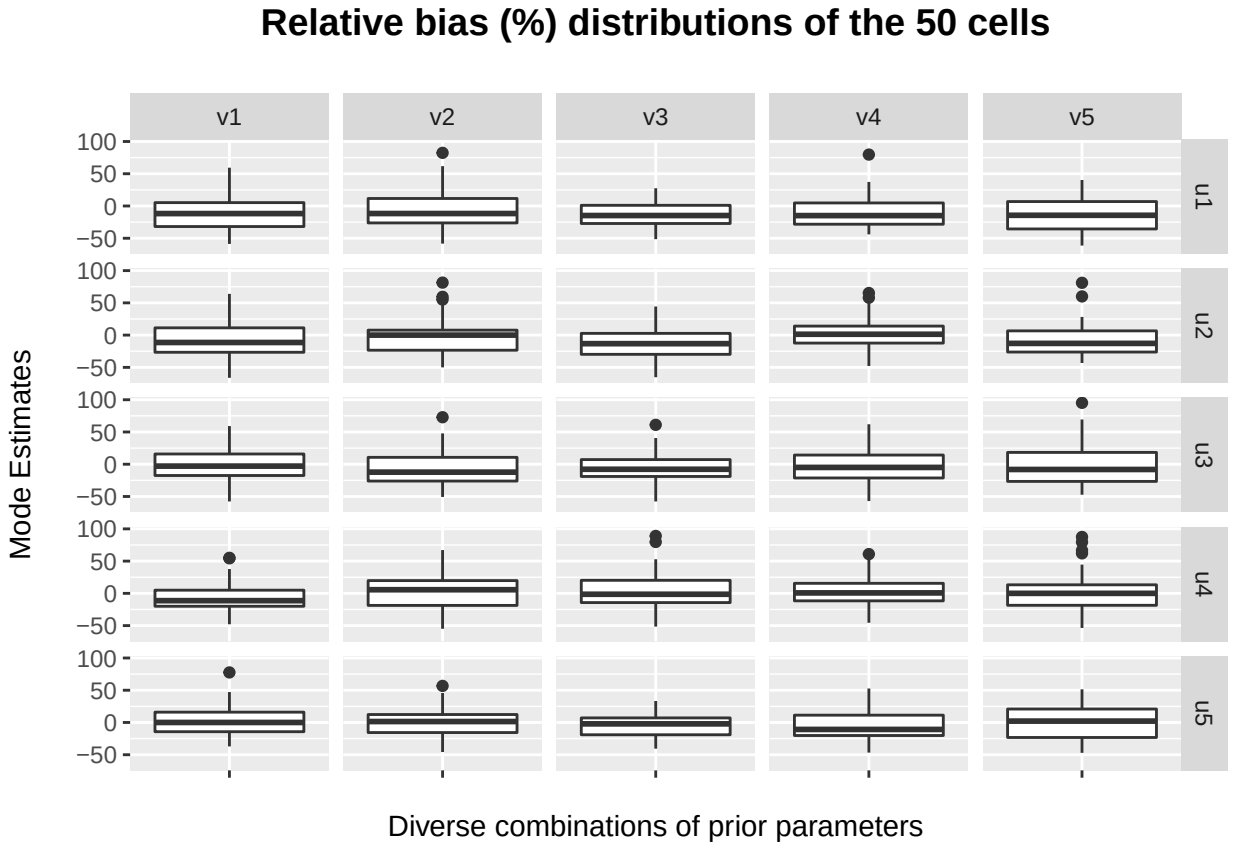


6.3.3 $f_1 \simeq \mathbf{Unif}(u_m, u_M)$, $f_2 \simeq \Gamma(a+1, \frac{N^{\mathbf{Reg}}}{a})$

Again we repeat the computation now with $f_2 \simeq \Gamma(a+1, \frac{N^{\mathrm{Reg}}}{a})$ and $\log_{10}(a) = -3, -2, \ldots, 2, 3$.

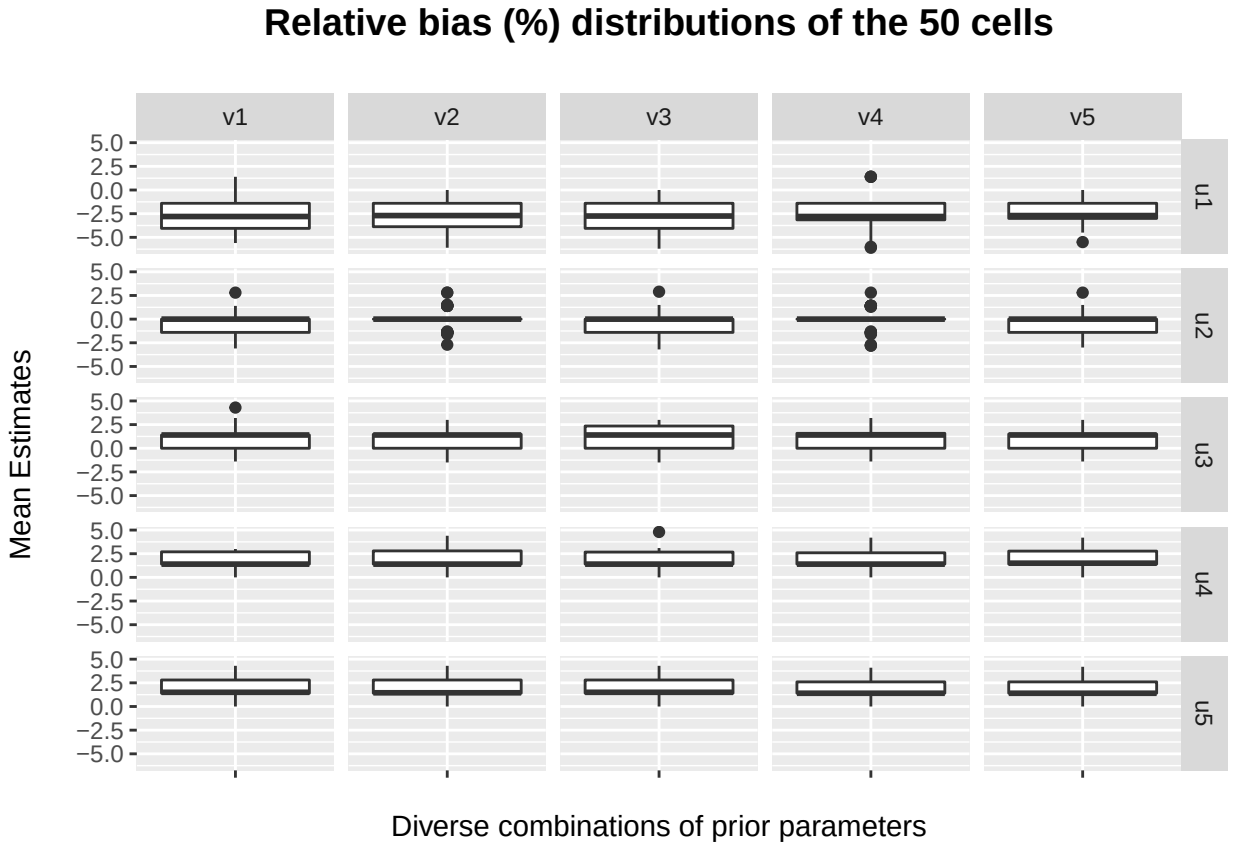

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
5
0
−5
−10
Mean Estimates
Diverse combinations of prior parameters

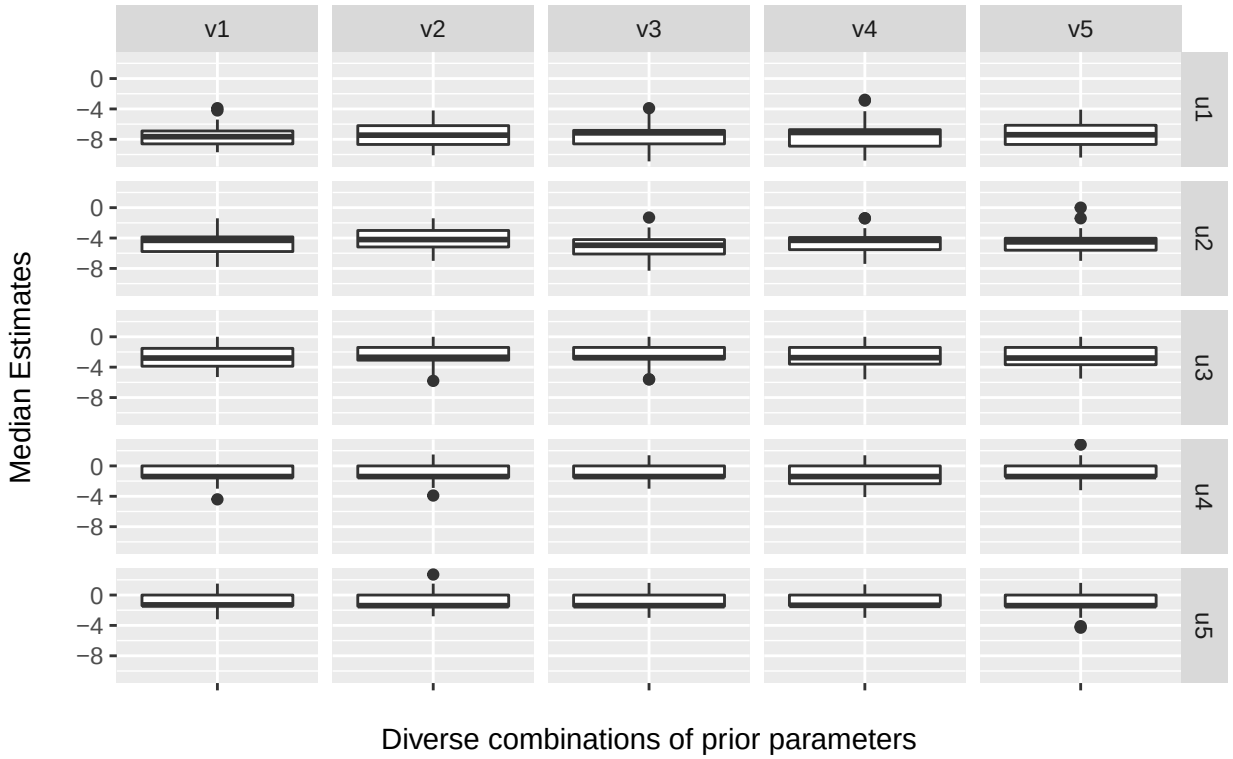

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
0
−5
−10
−15
Median Estimates
Diverse combinations of prior parameters

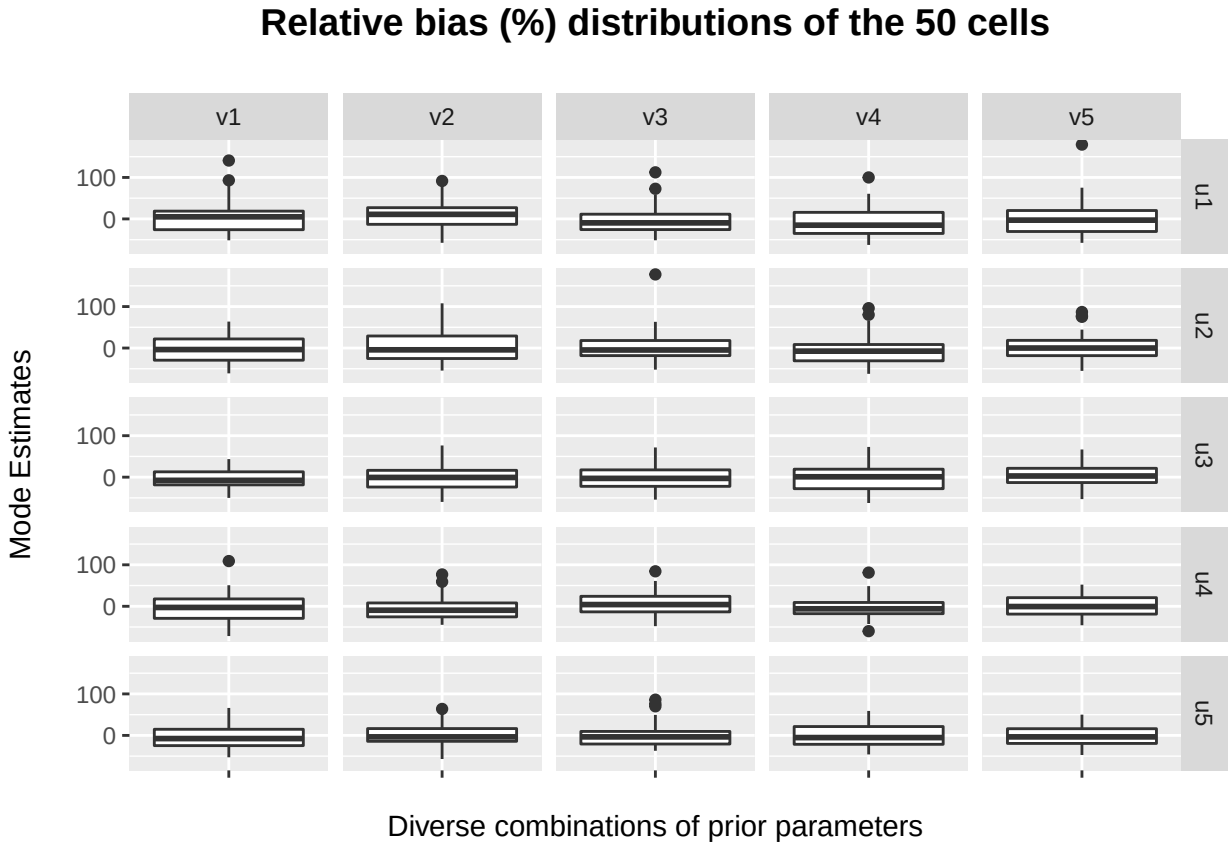


6.3.4 $f_1 \simeq \mathbf{Triang}(u_m, u_M, u^*)$, $f_2 \simeq \mathbf{Unif}(N_m, N_M)$

Once more, we repeat the computation now with $f_1 \simeq$ Triang and $f_2 \simeq$ Unif. The hyperparameters are chosen as before.

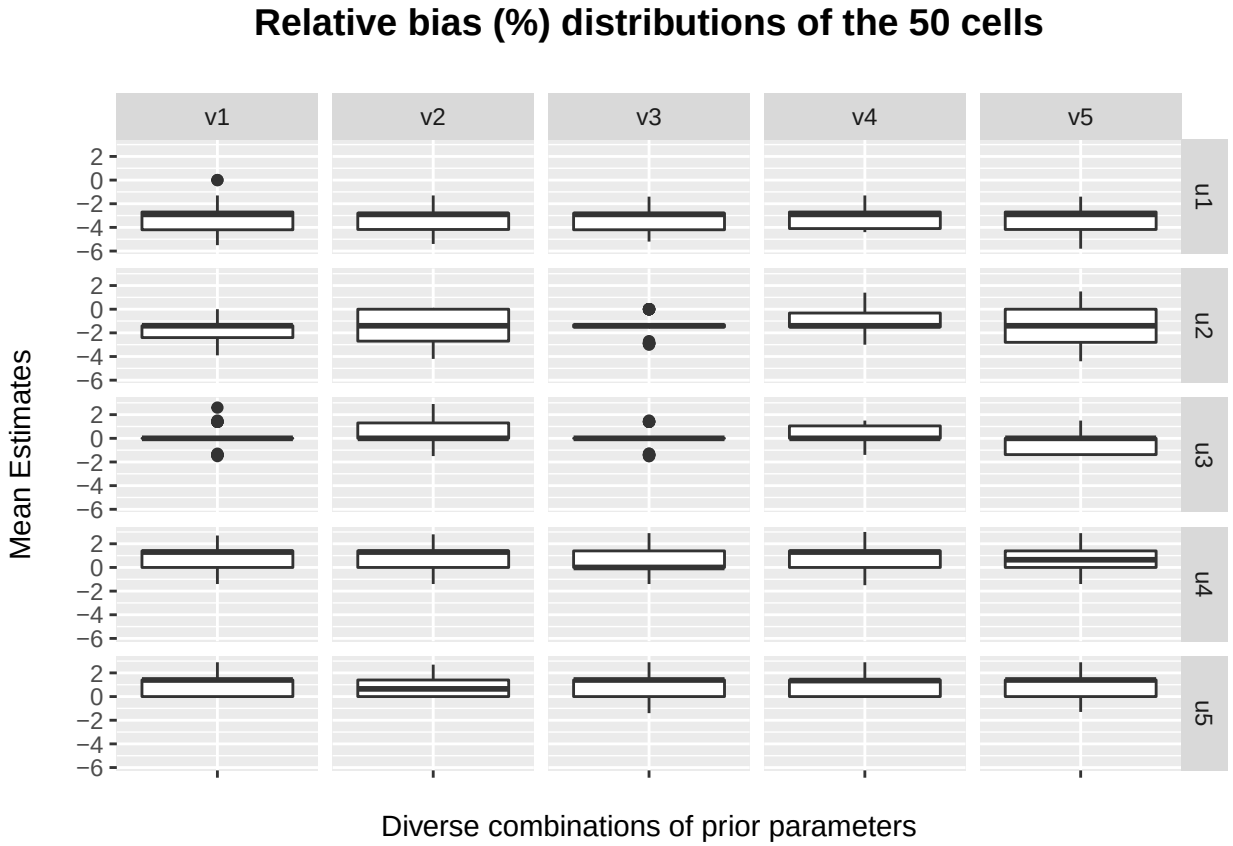

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
5.0
2.5
0.0
−2.5
−5.0
Mean Estimates
Diverse combinations of prior parameters

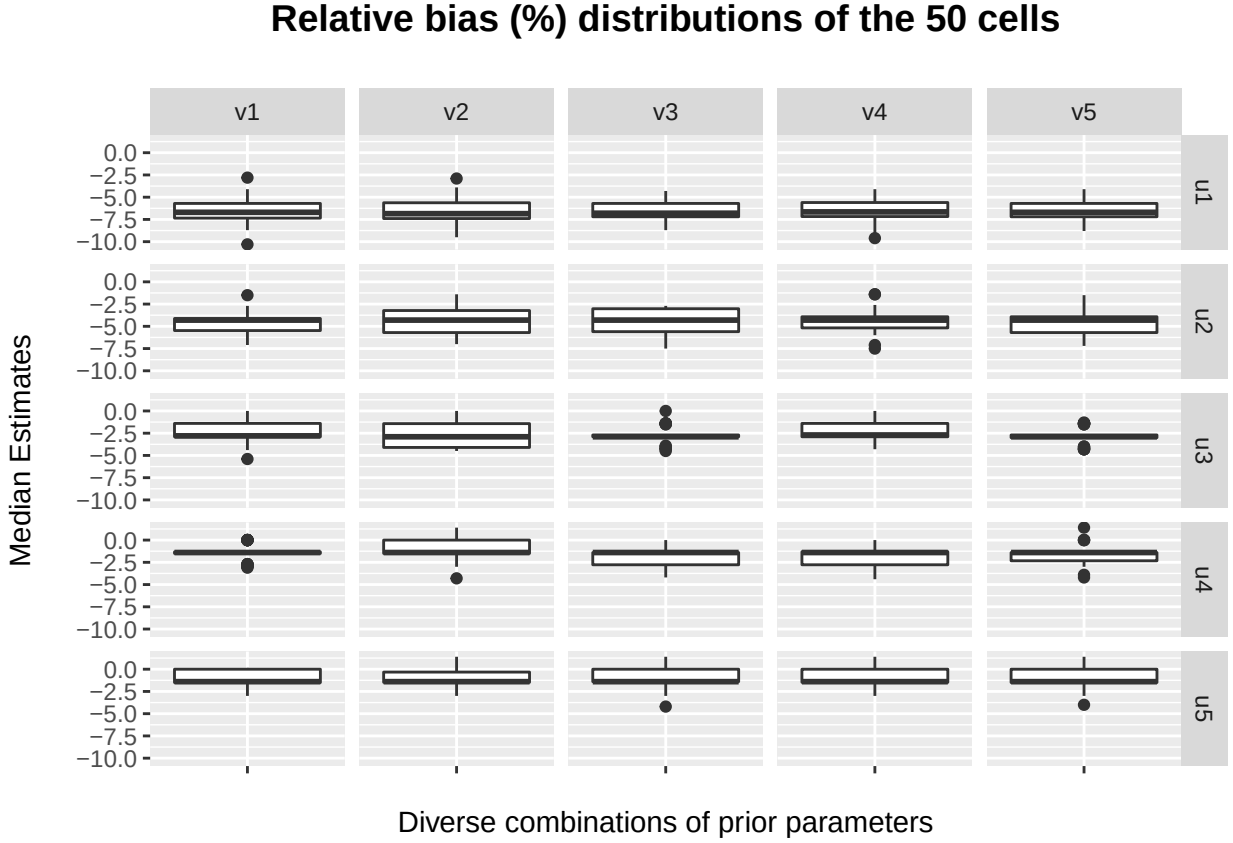

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
0
−4
−8
Median Estimates
Diverse combinations of prior parameters

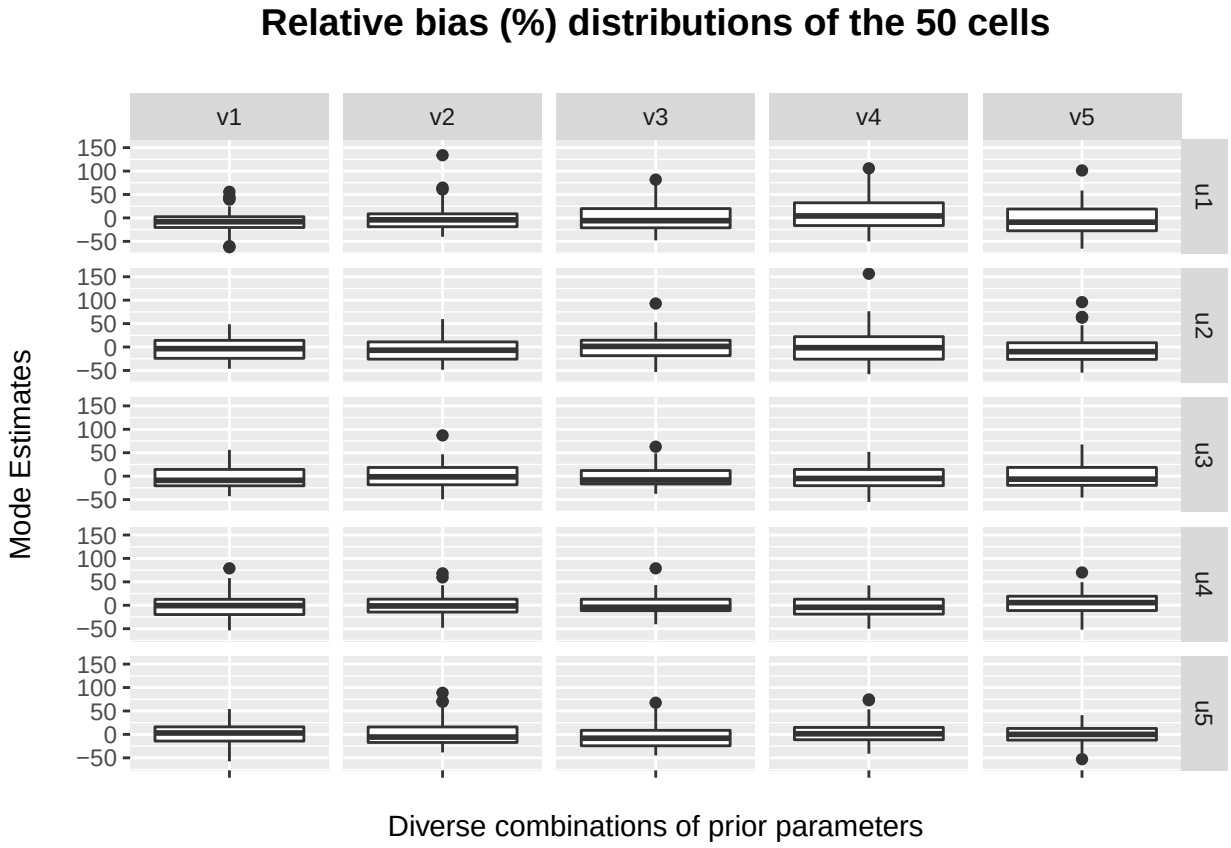


6.3.5 $f_1 \simeq \mathbf{Triang}(u_m, u_M, u^*)$, $f_2 \simeq \mathbf{Triang}(N_m, N_M, N^{\mathbf{Reg}})$

Now both prior distributions are triangular with the same choice of hyperparameters are before.

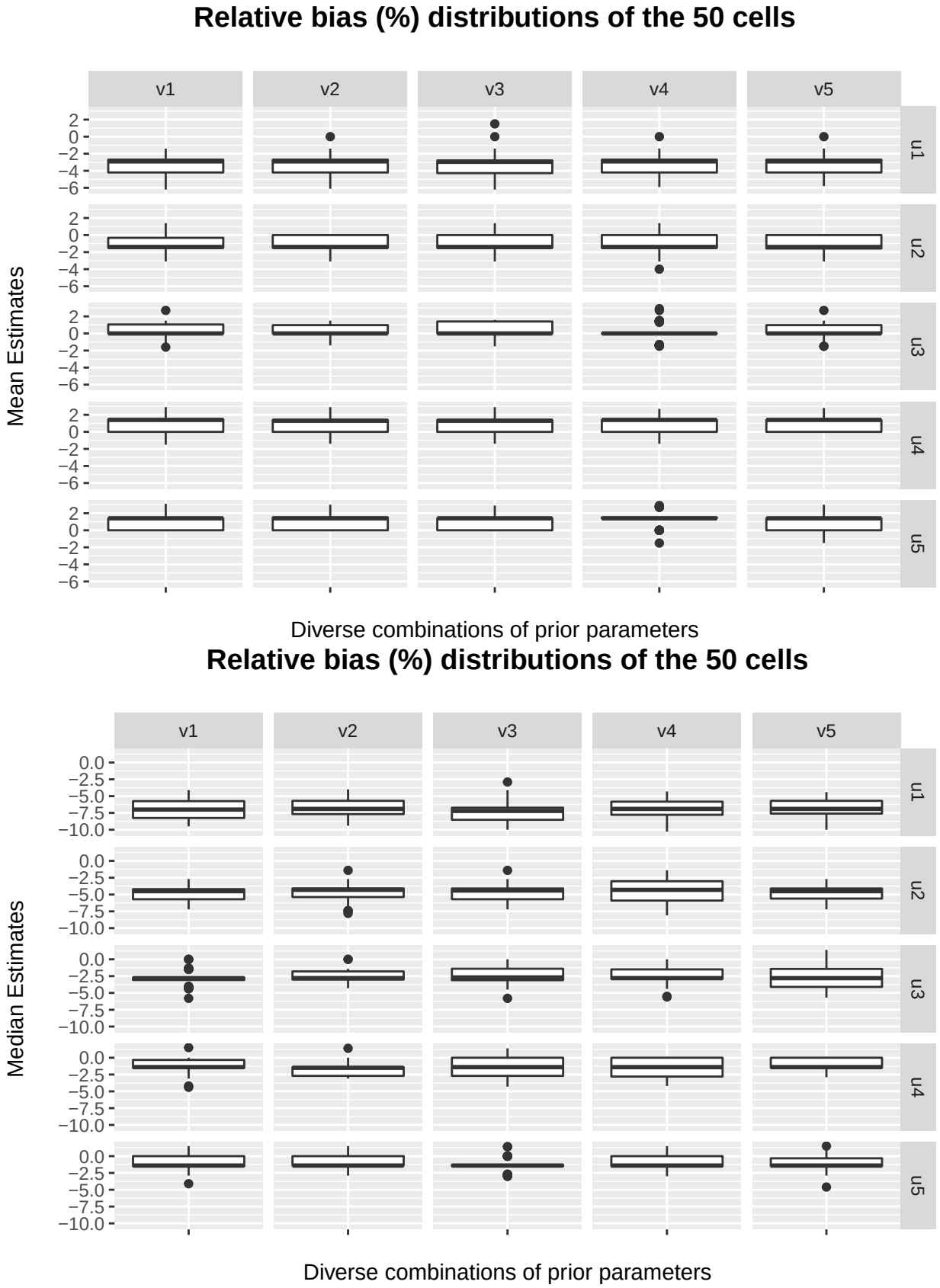

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
Mean Estimates
Diverse combinations of prior parameters

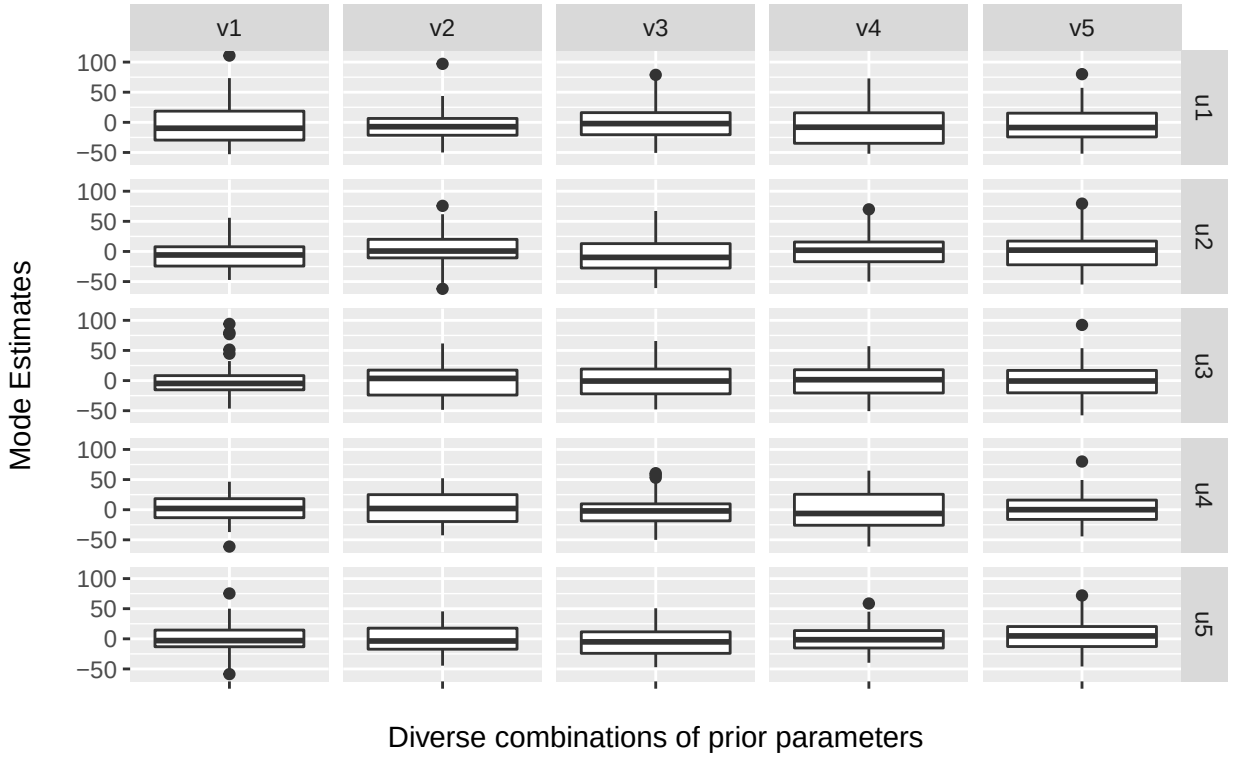

Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
Median Estimates
Diverse combinations of prior parameters



6.3.6 $f_1 \simeq \mathbf{Triang}(u_m, u_M, u^*)$, $f_2 \simeq \Gamma(a+1, \frac{N^{\mathbf{Reg}}}{a})$

Finally we combine a triangular distribution for f_1 and a gamma distribution for f_2.


Relative bias (%) distributions of the 50 cells
v1
v2
v3
v4
v5
u1
u2
u3
u4
u5
Mean Estimates
2
0
−2
−4
−6
Diverse combinations of prior parameters
Relative bias (%) distributions of the 50 cells
Median Estimates
0.0
−2.5
−5.0
−7.5
−10.0
Diverse combinations of prior parameters



6.4 Simulations with aggregated mobile phone data and the Spanish population register

For more realistic values (more than 3-digit figures) we have detected some needs of improvement in the implementation of the Kummer function computation and of the optimization algorithm. We will carry these simulations in a separate document as soon as these improvements have been completed.

For the time being we will illustrate in a complementary document the IT side of the preceding computations so that these tools can be used for any user upon his/her own data.

6.5 Next steps

This document illustrates our proposal to infer population counts $N_{i,0}$ over a population partition into cells i at a given time instant t_0 from the input values $(N_{i,0}^{\text{Reg}}, N_{i,0}^{\text{MNO}})$ of population counts in a population register or a similar auxiliary data source and detected by a cellular telecommunication network.

The next step is to introduce the time dimension by considering how these population counts evolve. This work is already ongoing and is based on the main working hypothesis that the displacement of individuals among territorial cells is independent of the MNO of each subscriber. Taking the matrix $N^{\text{MNO}}(t_0, t_1) = [N_{ij}^{\text{MNO}}(t_0, t_1)]_{1 \le i,j \le I}$ of number of displaced subscribers from cell i to cell j in the time interval $[t_0, t_1]$ and the registered population counts $N_{i,0}^{\text{Reg}}$ as input data we propose the next hierarchical model:

$$N_i(t_1) = \lfloor \sum_{j=1}^{I} p_{j\to i}(t_0,t_1)N_j(t_0) - \sum_{\substack{j=1\\ j\neq i}}^{I} p_{i\to j}(t_0,t_1)N_i(t_0) \rfloor, \quad i = 1,\ldots,I \tag{20}$$

$$\begin{aligned}
\mathbf{p}_{i\cdot}(t_0,t_1) &\simeq \text{Dirichlet}(\alpha_{i\cdot}(t_0,t_1)), \quad i = 1,\ldots,I \\
\alpha_{i\cdot} &\simeq \frac{\prod_{j=1}^{I-1} f_1\left(\frac{\alpha_{ij}(t_0,t_1)}{\alpha_i^{(0)}(t_0,t_1)}; \frac{N_{ij}^{\text{MNO}}(t_0,t_1)}{N_i^{\text{MNO}}(t_0)}\right)\cdot f_2\,()}{(\alpha_i^{(0)}(t_0,t_1))^{I-1}}, \quad i = 1,\ldots,I \\
N_i^{\text{MNO}}(t_0) &\simeq \text{Bin}\left(N_i(t_0), p_i(t_0)\right), \qquad N_i^{\text{MNO}}(t_0) \perp N_j^{\text{MNO}}(t_0), \quad i \neq j = 1,\ldots,I \\
N_i(t_0) &\simeq \text{Po}\left(\lambda_i(t_0)\right), \qquad N_i(t_0) \perp N_j(t_0), \quad i \neq j = 1,\ldots,I \\
p_i(t_0) &\simeq \text{Beta}\left(\alpha_i(t_0), \beta_i(t_0)\right), \qquad p_i(t_0) \perp p_j(t_0) \quad i \neq j = 1,\ldots,I \\
(\alpha_i(t_0), \beta_i(t_0)) &\simeq \frac{f_1(\frac{\alpha_i(t_0)}{\alpha_i(t_0)+\beta_i(t_0)}; \mathbf{N}^{\text{REG}}(t_0)) \cdot f_2(\alpha_i(t_0)+\beta_i(t_0); \mathbf{N}^{\text{REG}}(t_0))}{\alpha_i(t_0)+\beta_i}(t_0), \qquad (\alpha_i(t_0),\beta_i(t_0)) \perp (\alpha_j(t_0),\beta_j(t_0)), \\
\lambda_i(t_0) &\simeq f_3(\lambda_i(t_0); N_i^{\text{REG}}(t_0)) \quad (\lambda_i(t_0) > 0, \lambda_i(t_0) \perp \lambda_j)(t_0), \quad i = 1,\ldots,I,
\end{aligned}$$

where $\alpha_i^{(0)}(t_0,t_1) = \sum_{j=1}^{I} \alpha_{ij}(t_0,t_1)$. The posterior probability distribution to compute is

$$\mathbb{P}\left(N_i(t_1)\middle|N^{\text{MNO}}(t_0,t_1), N^{\text{Reg}}(t_0)\right) = \sum_{n=0}^{\infty} \mathbb{P}\left(N_i(t_1)\middle|N_i(t_0)=n, N^{\text{MNO}}(t_0,t_1), N^{\text{Reg}}(t_0)\right) \mathbb{P}\left(N_i(t_0)=n\middle|N^{\text{MNO}}(t_0,t_1), N^{\text{Reg}}(t_0)\right).$$

This will be undertaken in a separate document. Notice that $\mathbb{P}\left(N_i(t_0)=n\middle|N^{\text{MNO}}(t_0,t_1), N^{\text{Reg}}(t_0)\right)$ is the result of the preceding proposal and $\mathbb{P}\left(N_i(t_1)\middle|N_i(t_0)=n, N^{\text{MNO}}(t_0,t_1), N^{\text{Reg}}(t_0)\right)$ can be computed using simulations.

7 Appendix A: Computation of $J_{n+m,p}(N)$

The integral $J_{n+m,p}(N)$ can be computed using the residue theorem (see e.g. Brown and Churchill (2003)). Applying this theorem to $g(z) = f_N(z) \cdot \frac{z^p}{\prod_{k=1}^{n+m-1}(z+k)} \cdot \log(z)$ in the closed path around the origin composed by a straight path γ_1 along and above the positive real axis (from $+\epsilon$ to $+R$), a counterclockwise circular path γ_R at radius R, a straight path γ_2 along and below the positive real axis (from $+R$ to $+\epsilon$) and a clockwise circular path γ_ϵ at radius ϵ. We place a branch cut at the positive real axis. Then it is easy to prove (via Jordan's lemma) that $\int_{\gamma_R} g(z)dz \to 0$ and $\int_{\gamma_\epsilon} g(z)dz$ when $R \to \infty$ and $\epsilon \to 0$, while

$$\int_{\gamma_1} g(z)dz \to \int_0^\infty dx\ f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} \log(x), \tag{21}$$

$$\int_{\gamma_2} g(z)dz \to -\int_0^\infty dx\ f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} \left(\log(x) + 2\pi i\right). \tag{22}$$

The poles of $g(z)$ are simple and located at $z = -k$, $k = 1, \ldots, n+m-1$ and the residues can be computed easily:

$$\begin{aligned}
\text{Res}(g, -k) &= f_N(-k) \cdot \frac{(-k)^p}{\prod_{\substack{i=1 \\ i\neq k}}^{n+m-1}(i-k)} \log\left(ke^{i\pi}\right) \\
&= f_N(-k) \frac{(-1)^p k^p}{(-1)^{k-1}(k-1)!(n+m-1-k)!} \left(\log(k) + i\pi\right) \\
&= f_N(-k) \cdot \frac{(-1)^{p-k} k^{p+1}}{(n+m-1)!} \binom{n+m-1}{k} \left(\log(k) + i\pi\right)
\end{aligned} \tag{23}$$

Substituting on the residue theorem and focusing on the imaginary part of the expressions we have

$$-2\pi i \int_0^\infty dx \quad f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} = 2\pi i \sum_{k=1}^{n+m-1} f(-k) \cdot \frac{(-1)^{p-k} k^{p+1}}{(n+m-1)!} \binom{n+m-1}{k} \left(\log(k) + i\pi\right), \tag{24}$$

thus arriving at

$$J_{n+m,p}(N) = \int_0^\infty dx \quad f_N(x) \frac{x^p}{\prod_{k=1}^{n+m-1}(x+k)} = \frac{1}{(n+m-1)!} \sum_{k=1}^{n+m-1} (-1)^{p-k-1} \binom{n+m-1}{k} f(-k) \cdot k^{p+1} \log(k). \tag{25}$$

Setting $n = N_i^{\text{MNO}}$, $m = n_i - N_i^{\text{MNO}}$ and $N = N_i^{\text{REG}}$, we have

$$J_{n_i,p_i}(N_i^{\text{REG}}) = \frac{1}{(n_i-1)!} \sum_{k_i=1}^{n_i-1} (-1)^{p_i-k_i-1} \binom{n_i-1}{k_i} f_{N_i^{\text{REG}}}(-k_i) \cdot k_i^{p_i+1} \log(k_i). \tag{26}$$

8 Appendix B: computation of $a_{N,n-N}(m)$

Let us consider the following attempt to compute the numerical factors

$$a_{N,n-N}(p) = \sum_{q=0}^{p} q! \begin{bmatrix} n \\ q \end{bmatrix} (p-q)! \begin{bmatrix} n-N \\ p-q \end{bmatrix}.$$

Let us consider the generating function $G_k(z) = \frac{(\log(1+z))^k}{k!} = \sum_{n=0}^{\infty}(-1)^{n-k} \begin{bmatrix} n \\ k \end{bmatrix} \frac{z^n}{n!}$. Notice that $G_k^{(n)}(0) = (-1)^{n-k} \begin{bmatrix} N \\ s \end{bmatrix}$.

We apply Bruno di Faà's formula (Johnson (2002)) to $k!G_k^{(n)}(z) = (\log(1+z))^k$:

$$\frac{d^n}{dz^n} f(g(z)) = \sum_{m=1}^{n} \sum_{\substack{k_1,\ldots,k_n=0 \\ k_1+2\cdot k_2+\cdots+n\cdot k_n=n}}^{m} \frac{n!}{k_1!\ldots k_n!} f^{(m)}(g(z)) \left(\frac{g^{(1)}(z)}{1!}\right)^{k_1} \cdots \left(\frac{g^{(n)}(z)}{n!}\right)^{k_n}.$$

Now, in our case we have $f(z) = z^k$ and $g(z) = \log(1+z)$. Then we can write

$$\begin{aligned} k! \begin{bmatrix} n \\ k \end{bmatrix} = (-1)^{n-k} k! G_k^{(n)}(0) &= (-1)^{n-k} \sum_{m=1}^{k} \sum_{\substack{k_1,\ldots,k_n=0 \\ k_1+2\cdot k_2+\cdots+n\cdot k_n=n}}^{m} \frac{n!}{k_1!\cdots k_n!} \frac{k!}{(k-m)!} \delta_{km} \left(\frac{(-1)^{1-1}}{1}\right)^{k_1} \cdots \left(\frac{(-1)^{n-1}}{n}\right)^{k_n} \\ &= (-1)^{n-k} \cdot n! \cdot k! \sum_{\substack{k_1,\ldots,k_n=0 \\ k_1+2\cdot k_2+\cdots+n\cdot k_n=n}}^{k} \frac{1}{k_1!\ldots k_n!} \left(\frac{1}{1}\right)^{k_1} \cdots \left(\frac{(-1)^{n-1}}{n}\right)^{k_n}. \quad (27) \end{aligned}$$

We have found no way to further simplify this expression. Notice also that for large n and k as in our case, it can hardly be implemented in a computer routine (indeed, Stirling numbers are often computed using recurrence relations in $O(n^2)$ steps). So far, no further simplification has been found.

References

Brown, J.W., and R.V. Churchill. 2003. "Complex Variables and Applications (7th Ed.)."

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

Devroye, L. 1986. "Non-Uniform Random Variable Generation." Springer.

ESSnet on Big Data WP5. 2017. "Current Status of Access to Mobile Phone Data in the ESS."

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. "Bayesian Data Analysis (3rd Ed)." CRC Press.

Graham, R.L., D.E. Knuth, and O. Patashnik. 1996. "Concrete Mathematics (2nd Ed.)." Addison-Wesley.

Johnson, W.P. 2002. "The Curious History of Faà Di Bruno's Formula." *American Mathematical Monthly 109, 217-234.*

Manly, B.F.J., and J.A. Navarro Alberto, eds. 2015. "Introduction to Ecological Sampling." CRC Press.

Royle, J.A., and R.M. Dorazio. 2008. "Hierarchical Modeling and Inference in Ecology." Academic Press.