Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2014-04

Predictive Inference on Finite Populations Segmented in Planned and Unplanned Domains

Juan Carlos Martínez-Ovando
Banco de México

Sergio I. Olivares-Guzmán
Banco de México

Adriana Roldán-Rodríguez
Banco de México

February 2014

Documento de Investigación
2014-04

Working Paper
2014-04

Predictive Inference on Finite Populations Segmented in Planned and Unplanned Domains*

Juan Carlos Martínez-Ovando†
Banco de México

Sergio I. Olivares-Guzmán‡
Banco de México

Adriana Roldán-Rodríguez§
Banco de México

Abstract: In this paper, we develop a new model-based method to inference on totals and averages of finite populations segmented in planned domains or strata. Within each stratum, we decompose the total as the sum of its sampled and unsampled parts, making inference on the unsampled part using Bayesian nonparametric methods. Additionally, we extend this method to make inference on totals of unplanned domains simultaneously modelling, within each stratum, the underlying uncertainty about the composition of the population and the totals across unplanned domains. Making inference on population averages is straightforward in both frameworks. To illustrate these methods, we develop a simulation exercise and evaluate the uncertainty surrounding the gender wage gap in Mexico.
Keywords: Survey methods, robustness, species-sampling models.
JEL Classification: C11, C14, C42, C81, C88, J31.

Resumen: En este trabajo, desarrollamos un nuevo método basado en modelos para hacer inferencia sobre totales y promedios de poblaciones finitas segmentadas en dominios planeados o estratos. Dentro de cada estrato, descomponemos el total como la suma de sus partes muestreadas y no muestreadas, haciendo inferencia sobre las partes no muestreadas usando métodos bayesianos no paramétricos. Adicionalmente, extendemos este método para hacer inferencia sobre totales de dominios no planeados modelando simultáneamente, dentro de cada estrato, la incertidumbre subyacente a la composición de la población y los totales entre los dominios no planeados. Hacer inferencia sobre promedios poblacionales es directo en ambos casos. Para ilustrar estos métodos, desarrollamos un ejercicio de simulación y evaluamos la incertidumbre en torno a la brecha salarial de género en México.
Palabras Clave: Métodos de encuestas, robustez, modelos de muestreo de especies.

*We would like to thank Alberto Padilla Terán, for sharing stimulating discussions at an early stage of this project, and to Adriana Martínez Guzman, for helping us to understand the content of the national survey on employment that we studied. We appreciate constructive comments on an earlier version of the paper from Brendan Murphy, as those received from seminar and conference participants at CIDE, CIMAT, IIMAS-UNAM, ITAM, IMATI-CNR and SESM. We also thank Angela Noufaily and Thomas W. Yee for making their source code available.

† Dirección General de Investigación Económica. Email: juan.martinez@banxico.org.mx.

‡ Dirección General de Investigación Económica. Email: solivares@banxico.org.mx.

§ Dirección General de Investigación Económica. Email: aroldan@banxico.org.mx.

1 Introduction

Most social and economic indicators concerning finite populations using survey data, and empirical studies derived from them, rely on design-based methods. See Thompson (1997). Those indicators typically require the estimation of totals or averages of the whole population, or sub-groups of the population, which are commonly computed using weight-based methods. Sample weights are built and used to make inference on the unsampled part of the population using the information collected in the survey sample data. This procedure relies only on the randomization process from which the survey sample data was collected and not on the range of possible outcomes that each unobserved individual characteristic may take. Sample weights are theoretically conceived as the inverse of the probability of selection of individuals in the sample, according to a well defined sampling design. Relevant examples of weight-based estimators are the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), the Hájek estimator (Hájek, 1971), and the post-stratified estimator (Holt and Smith, 1979).

However, in practice, sample weights are not necessarily built upon a well defined sampling design, *i.e.* data may be collected from a non or partially informative sampling scheme.[1] That is an implication of working with imperfect population frames, implementing corrections for missing data, or some other reasons. In those cases, weighting individual survey data may lack of a formal conceptual meaning.

In this paper we propose an intuitive model-based framework to make inference on totals and averages of a finite population, which also relies on some sort of weighting.

[1] An informative sampling scheme is that in which the data is collected using a selection scheme based on randomization in which the sampled units are associated with known unequal probabilities of being selected. In terms of modelling, the distribution or model of the outcome of interest is conditioned on the sampling process. The informativeness of the sample implies that this distribution differs from the population's model (Pfeffermann and Sverchkov, 2009). We understand by a partially informative sampling scheme as that where individuals are sampled with unequal probability of selection among strata or planned domains, but assuming that the units are uniformly sampled within them. In the latter case, the selection probabilities may not be known with entire precision.

However, in our formulation the sample weights are conceived differently and reflect the frequencies of individual outcomes in the sampled data, combined with some prior knowledge of the population.

Related model-based methods to inference on finite populations are explained in Särndal et al. (1992) and Chambers and Clark (2012). Here, we do not make use of additional information, as Särndal et al.'s approach does. Rather, we restrict ourselves to make use only of the information contained in the sampled data. Our main assumption is that unobserved individual outcomes are random and that the underlying distribution for such outcomes is Bayesian nonparametric, specifically in the class of species-sampling models (Pitman, 1996). Thus, structural assumptions are relaxed to the minimum possible.

Predictions are made accordingly to a preconceived segmentation of the population, assuming that the underlying composition of the population among that segmentation is known. Such a segmentation is referred in the survey data argot as stratification or partitioning based on planned domains. Additionally, we derive a framework to make inference on unplanned segmentations of the population, *i.e.* a segmentation of the population for which the underlying composition of the population is unknown. The latter is commonly referred as partitioning induced by unplanned domains. See Lehtonen and Veijanen (2009) for a revision of the current state-of-the-art concerning inference on planned and unplanned domains of finite populations. It is worth mentioning that traditional weight-based estimators get overwhelmed when dealing with estimation on unplanned domains; see *e.g.* Meeden (2005). Moreover, both frameworks are easily extended to make prediction on population averages as well, irrespective if the referred number of individuals used to compute the population average is known or unknown. In all cases, we show that the predictive distribution of the characteristic of interest is easily recovered through simulation methods.

1.1 Structure of the paper

The introduction is completed with notation used and relevant assumptions. Section 2 provides a review of species-sampling models and some of their most relevant properties.

Section 3 develops an integrated Bayesian nonparametric model-based framework to inference on totals of finite populations segmented in planned domains. Section 4 extends our formulation in order to make inference on the composition of the population across unplanned domains, and their associated sub-totals. Section 5 explains how our formulation can be adapted to make predictions on population averages. Section 6 develops a simulation study illustrating the inferential coverage of our method on the basis of samples that progressively cover up the whole population. Section 7 develops an evaluation of the uncertainty surrounding the gender wage gap in Mexico using data from a national survey on employment. Section 8 concludes with a brief discussion.

1.2 Notation and assumptions

The population of interest is denoted by $\mathcal{P}$, and the total number of individuals in $\mathcal{P}$ is denoted by N. It is assumed that the population is divided into J planned domains, $\{\mathcal{P}_j\}_{j=1}^{J}$, for which $N_j = \#\{\mathcal{P}_j\}$ is assumed known. Accordingly, the total of the population, T, can be decomposed as the sum of partial totals for planned domains,

$$T = \sum_{j=1}^{J} T_j, \tag{1.1}$$

where $T_j = \sum_{l=1}^{N_j} Y_{jl}$. Here, Y_{jl} stands for the observed characteristic in the l-th individual of the planned domain $\mathcal{P}_j$. That characteristic is assumed to be continuous and random for each individual, if not being observed. We also assume that the individuals are exchangeable with respect to the unknown underlying distribution for each $\mathcal{P}_j$. Notice that similar notions of symmetry are assumed as well on stratified and post-stratified design-based methods, but in those cases such an assumption regards only with the designed sampling-distribution. Here we assume that the sampling design is not informative or that it is partially informative, in the sense that it only takes into account that the data is being randomly collected according to the partition induced by the planned domains. Therefore, we place our attention to the relationship between individual measurements and their underlying distribution. Additionally, let $\mathcal{S}_j$ stand for the sampled part of $\mathcal{P}_j$, and let $\widetilde{\mathcal{S}}_j$ be its unsampled complement; *i.e.*, $\mathcal{S}_j \cup \widetilde{\mathcal{S}}_j = \mathcal{P}_j$ and $\mathcal{S}_j \cap \widetilde{\mathcal{S}}_j = \emptyset$. Also, let

$N_j^{\mathcal{S}} = \#\{\mathcal{S}_j\}$ and $N_j^{\widetilde{\mathcal{S}}} = \#\{\widetilde{\mathcal{S}}_j\}$. It is also assumed that individuals belonging to each $\mathcal{S}_j$ are uniformly sampled.

2 Species-sampling models

The model-based framework we develop is flexible in that structural assumptions related to the form of the underlying distribution for each planned domain are being relaxed to the minimum possible. The particular component that we take into account is taken from the class of *species-sampling models* (henceforth SSM); see Pitman (1996). SSMs define a flexible class of countable random distribution functions that has deserved a lot of attention among the Bayesian nonparametric community over the recent years. Some relevant properties of SSMs have been studied by Hansen and Pitman (2000) and Lee et al. (2013), among others. We relate SSM's to the finite population framework assuming that the Y_{jl}'s in $\mathcal{P}_j$ are conditionally i.i.d. given F_j, and that each F_j belongs to the class of SSMs. The most relevant property of SSMs in our context relates to its marginalization property, which we shall discuss below.

For the sake of exposition allow us for now to disregard the index j. It is said that a random distribution function, F, belongs to SSMs if it can be represented as,

$$F(\cdot) = \sum_{k=1}^{\infty} \rho_k \cdot \delta_{Z_k}(\cdot) + \left(1 - \sum_{k=1}^{\infty} \rho_k\right) \cdot G_0(\cdot), \tag{2.1}$$

where $(\rho_k)_{k=1}^{\infty}$ is a sequence of random positive variables, and $(Z_k)_{i=1}^{\infty}$ is a sequence of i.i.d. random variables with (absolutely continuous) distribution G_0. It is assumed that both F and G_0 share a common support. In our context, we have considered that this support is the positive real line. Additionally, SSMs assume that both random sequences, $(\rho_k)_{k=1}^{\infty}$ and $(Z_k)_{k=1}^{\infty}$, are stochastically independent. Particular cases of SSMs include the Dirichlet process (Ferguson, 1973), the normalized inverse-Gaussian process (Lijoi et al., 2005), and the geometric weight prior (Mena, 2012), among others. Basically, SSMs differ from each other in terms of the prior specification for the random weights $(\rho_k)_{k=1}^{\infty}$. A SSM is being completed with the specification of independent prior distributions on the sequences $(\rho_k)_{k=1}^{\infty}$ and $(Z_k)_{k=1}^{\infty}$, which typically satisfy that $\mathbb{E}\{F\} = G_0$.

2.1 Marginalization property

The marginalization property of SSMs has been extensively used as a simulation devise in Bayesian nonparametric procedures. In a general setting, the marginalization property basically expresses the predictive distribution of unknown individual characteristics as a weighted-sum of sampled measurements and a prior judgment of possible values that the unobserved characteristic can take. This property also guaranties that prediction becomes free of the infinite-dimensional object F when relevant information is being incorporated. Hence, all the uncertainty surrounding the auxiliary random variable F gets vanished once we incorporate relevant data. This property is very important in our context, as we shall expose it below. A general expression for the predictive distribution attained to SSM has been studied by Hansen and Pitman (2000) and Lee et al. (2013), among others, which basically rely on the Pólya urn model. We shall revisit this formula later on in the paper.

3 Totals on planned domains

It is crucial to observe that the total of the population can be decomposed as the sum of partial totals for each planned domain, and each of them can be treated independently. Let us notice that stratified and post-stratified weight-based methods take this decomposition into consideration, as well. Let us explore what happens at the interior of each planned domain $\mathcal{P}_j$.

Once the sampled units in $\mathcal{S}_j$ are being observed, *i.e.* the set $\boldsymbol{y}_j = \{y_{jl} : l = 1, \ldots, m_j\}$ is known, the partial total T_j is decomposed as the sum of sampled and unsampled parts,

$$T_j \;=\; \sum_{l \in \mathcal{S}_j} y_{jl} + \sum_{l \in \widetilde{\mathcal{S}}_j} Y_{jl}. \tag{3.1}$$

Inference is then to be made on the unobserved part of the sum, *i.e.* $\sum_{l \in \widetilde{\mathcal{S}}_j} Y_{jl}$. Chambers (1986) and Ghosh (2008) used a similar decomposition as we do here. Let us recall that the Y_{jl}'s are assumed to be exchangeable within the planned domain $\mathcal{P}_j$. Hence, under

the square-root loss function, the predictive estimate of T_j would be given by

$$\widehat{T}_j = \sum_{l\in\mathcal{S}_j} y_{jl} + \sum_{l\in\widetilde{\mathcal{S}}_j} \mathbb{E}\left\{Y_{jl}|\mathcal{S}_j\right\}, \tag{3.2}$$

where $\mathbb{E}\left\{Y_{jl}|\mathcal{S}_j\right\}$ is the individual predicted characteristic attained to the jth planned domain, given the information contained in the sample. This expectation is computed with respect to the predictive distribution induced by the SSM, which takes the form

$$\begin{aligned} \widehat{G}_j(\cdot) &= \mathbb{E}_{\mathbb{P}}\left\{F_j(\cdot)|y^*_{j1}, m_{j1}, \ldots, y^*_{jU_j}, m_{jU_j}\right\} \\ &= \sum_{k=1}^{U_j} \rho_k(\boldsymbol{m}_j)\cdot\delta_{y^*_{jk}}(\cdot) + \phi(\boldsymbol{m}_j)\cdot G_{j0}(\cdot), \end{aligned} \tag{3.3}$$

where U_j is the number of sample ties in $\mathcal{S}_j$, $\boldsymbol{y}^*_j = \{y^*_{jk} : k = 1,\ldots,U_j\}$, $\boldsymbol{m}_j = (m_{j1},\ldots,m_{jU_j})$, is the vector of sample frequencies attained to $\boldsymbol{y}^*_j$, *i.e.* $m_{jk} = \#\{y_{jl} \in \mathcal{S}_j : y_{jl} = y^*_{jk}\}$, for $k = 1,\ldots,U_j$, and G_{j0} is the distribution function that represents our prior judgment about the possible values that each Y_{jl} can take. The functions $\{\rho_k(\cdot)\}_{k=1}^{U_j}$ and $\phi(\cdot)$ are positive and satisfy that the sum, $\sum_{k=1}^{U_j}\rho_k(\boldsymbol{m}_j) + \phi(\boldsymbol{m}_j) = 1$, holds for any sample $\mathcal{S}_j$.

Let us point out that predictive point estimates of totals for any desired aggregation of planned domains are obtained as the sum of predicted estimates of the disaggregate planned domains involved. But, before getting into details about making predictions on the population total, let us introduce some thoughts about the meaning of the function $\phi(\boldsymbol{m}_j)$ in (3.3). This function represents one's strength of belief in G_{j0} with respect to what has been observed in the sample. In general, this function is defined in terms of the sample size, frequencies of the sample ties, and a set of parameters that represents one's priors belief in G_{j0}. In the finite population context, uncertainty on the range of possible individual outcomes that are still unobserved in the population is inversely related to the sample size. The limit case would be obtained when the sample fully covers up the whole population. In that limit case, there is no uncertainty about the possible individual measurements of the population, everything is known. The opposite limit case is obtained, of course, in the absence of sample data. There, all prior knowledge about the population

values is concentrated on G_{j0}. In order to reflect the above mentioned balance attained to finite population inference, we suggest to elicit the prior parameters related to $\phi(\cdot)$ (which intrinsically gives importance to G_{j0}) in an empirical Bayes framework (see, *e.g.* Casella, 1985). Thus, this parameter can be set as a function of the number of individuals in the population and the expected or planned sample size, in such a way to guaranty that $\phi(\boldsymbol{m}_j)$ would decrease monotonically toward zero as the sample size increases. By means of eliciting this parameter in this way we give a more intuitive and interpretable meaning to the inferential framework that we are proposing in the finite population context.

Now, setting $\phi(\boldsymbol{m}_j)$ up is strictly related with the amount of credible information that we might have concerning prior knowledge about the possible values that the unobserved outcomes may take, *i.e.* concerning G_{j0}. In the light of a credible opinion of a single or a group of experts, the procedure would require to set G_{j0} according to a prior elicitation scheme. That task is not necessarily simple, as it has been discussed in Gelfand et al. (1995) and French (1985) in detail. Alternatively, as it happens with recursing samples, we might have the chance of using data collected in previous samples to elicit a sensible choice for G_{j0} among a careful set of alternatives, which would not reflect any prior subjective opinion. This task is not an easy one either, as it lies right in the core of the problem of model comparison and selection (for parametric models, in this case). However, empirical methods as the one described in Gelman et al. (1996) can be of great use for this purpose. Now, in the absence of prior knowledge of experts of data from previous samples, the role of G_{j0} would be irrelevant for inference. In that case, we can set $\phi(\boldsymbol{m}_j)$ arbitrarily close to 0. In that case, our approach may resemble traditional weight-based methods; see Appendix A.

Despite of the above mentioned heuristic rules for setting up $\phi(\cdot)$ and G_{j0}, it should be acknowledged that setting both of these parameters up is actually case particular to the problem one would be aiming to solve. But that is the case for any frequentist or Bayesian inferential procedure in practice.

3.1 Bayesian estimate of totals

Under the square-root loss function, the predictive point estimate of T_j that we have sketched above can be written as

$$\widehat{T}_j = \sum_{g\in\mathcal{S}_j} y_{jg} + N_j^{\widetilde{\mathcal{S}}} \cdot \left[\sum_{k=1}^{U_j} \left(\rho_k(\boldsymbol{m}_j)\cdot y_{jk}^*\right) + \phi(\boldsymbol{m}_j)\cdot \widehat{\mu_{j0}}\right], \tag{3.4}$$

where $\widehat{\mu_{j0}} = \mathbb{E}_{G_{j0}}\{Y_{jl}|\boldsymbol{\theta}_{j0}\}$, and $\boldsymbol{\theta}_{j0}$ is the index parameter of G_{j0}.

A canonical example of (3.4) is given by a particular SSM model called the Dirichlet process (DP), see Ferguson (1973). The Dirichlet process has associated a prior distribution for $(\rho_k)_{k=1}^{\infty}$ that is characterized by a single positive parameter, $\alpha_{j,\mathrm{DP}}$. This parameter represents our prior degree of belief on G_{j0}; it also represents how disperse the random function F_j can be around G_{j0}. In this case, the updated functions associated with (3.3) are given by $\phi(\boldsymbol{m}_j) = \alpha_{j,\mathrm{DP}}/(\alpha_{j,\mathrm{DP}} + n_j)$ and $\rho_k(\boldsymbol{m}_j) = m_{jk}/(\alpha_{j,\mathrm{DP}} + n_j)$, for $k = 1,\ldots,U_j$, where $n_j = \sum_{k=1}^{U_j} m_{jk}$. Notice that the parameter $\alpha_{j,\mathrm{DP}}$ plays the role of tunning parameter to set one's prior strength of belief in G_{j0}. Following the reflection concerning $\phi(\cdot)$ we introduced above, the elicitation for the parameter $\alpha_{j,\mathrm{DP}}$ can be set as a function of N_j and the sample size, n_j, such that $\alpha_{j,\mathrm{DP}}(N_j, n_j) \to 0$ as $n_j \to N_j$. Hence, the predictive distribution (3.3) would converge to the censual distribution function of individual measurements, and the contribution of G_{j0} would get vanished.

Notice that traditional weight-based estimates of T_j correspond to the particular case of using the DP as the SSM for F_j; see Binder (1982) and Appendix A. Despite the resemblance of both approaches, our framework is more flexible in that the randomness involved in individual measurements is taken into consideration. Therefore, it is possible to produce robust predictive inferences for T_j. In the next section, we highlight that recovering the predictive distribution for T_j, and the one for the whole total T, is actually straightforward, by means of obtaining the predictive distribution of the corresponding partial sum via Monte Carlo simulation methods (Robert and Casella, 2004), and for the sub-totals of the planned domains involved.

3.2 Predictive inference through convolution

Our approach allows to produce full posterior inferences on T_j through a shifted $N_j^{\widetilde{\mathcal{S}}}$-fold convolution distribution induced by $\widehat{G}_j$, *i.e.*

$$\mathbb{P}(T_j \leq t|\mathcal{S}_j) = \widehat{G}_j^{*N_j^{\widetilde{\mathcal{S}}}}\left(T_j^{\widetilde{\mathcal{S}}} \leq t - T_j^{\mathcal{S}}\right), \tag{3.5}$$

where $T_j^{\widetilde{\mathcal{S}}} = \sum_{l\in\widetilde{\mathcal{S}}_j} Y_{jl}$ is the unsampled part of the total in $\mathcal{P}_j$, and $T_j^{\mathcal{S}} = \sum_{g\in\mathcal{S}_j} y_{jg}$ is the sampled part of the total, which plays the role of the shifting constant. Here, $\widehat{G}_j^{*N}$ stand for the N-fold convolution of $\widehat{G}_j$.

3.3 Simulation algorithm for planned domains

The predictive distribution (3.5) is analytically intractable, due in part that it is generated by a mixed-type distribution. However, one can handle it through simulation methods. Here we sketch a simple simulation procedure to draw an arbitrarily large collection of samples from the target distribution (3.5). The algorithm is designed for the Dirichlet process, but its adaptation to any other SSM is straightforward. The general steps of the algorithm are given as follows.

Step 0. Start by setting up the parameters of the model and the sampler, *i.e.* $\alpha_{j,\mathrm{DP}}$ and G_{j0}, for $j = 1, \ldots, J$.

Step 1. Generate $N_j^{\widetilde{\mathcal{S}}}$ i.i.d. random samples, $(y_{jl}^{(i)} : l = 1, \ldots, N_j^{\widetilde{\mathcal{S}}})$, from the predictive distribution $\widehat{G}_j$.

Step 2. Compute the unsampled part of the total, $T_j^{\widetilde{\mathcal{S}},(i)} = \sum_l y_{jl}^{(i)}$.

Step 3. Add up the sampled part of the total, $T_j^{\mathcal{S}}$.

Step 4. Compute $T_j^{(i)} = T_j^{\mathcal{S}} + T_j^{\widetilde{\mathcal{S}},(i)}$.

Steps 1 to 4 are repeated in each iteration of the algorithm. It is worth mentioning that any summary statistic and inferences of interest related to T_j can be produced using Monte Carlo methods using the simulated data produced with the steps described above. Similarly, simulated draws from the predictive distribution of any aggregate total of planned domains are produced as well by simply adding-up the simulated draws of the

involved planned domains at each iteration. We must warn the reader that despite the simplicity of this algorithm, its implementation may require heavy computations resources if the computation of the target population is relatively large, or if the sample size is also large.

4 Totals on unplanned domains

In this section we extend the formulation described in Section 3 in order to make inference on disaggregated totals associated with unplanned domains. Our formulation takes into consideration the randomness involved in the underlying composition induced by the unplanned domains, together with the compositional nature of the associated disaggregated sub-totals.

Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_D\}$ stand for the partition of the population induced by the categories of D unplanned domains. Such a partition is over imposed to that induced by the planned domains, $\{\mathcal{P}_j\}_{j=1}^{J}$. That is to say, each planned domain is being partitioned into D unplanned subsets $\{\mathcal{P}_j \cap \mathcal{D}_d : d = 1, \ldots, D\}$. Therefore, for inferential purposes, we can look at the unplanned domains of the whole population through the over imposition of the unplanned partition on each particular planned domain. As before, we shall focus our attention to what happens at the interior of a given planned domain. Aggregating the results turns out to be straightforward.

As we have discussed before, the key issue regarding the study of unplanned domains is that the composition of the population across them is unknown, *i.e.* the vector $(N_j^{\mathcal{D}_1}, \ldots, N_j^{\mathcal{D}_D})$, where $N_j^{\mathcal{D}_d} = \#(\mathcal{P}_j \cap \mathcal{D}_d)$, is intrinsically unknown. However, the number of individuals in the planned domain, N_j, it is know. Thus, the vector of composition should satisfy,

$$N_j = N_j^{\mathcal{D}_1} + \ldots + N_j^{\mathcal{D}_D}. \tag{4.1}$$

In a similar manner, the total T_j of the planned domain can also be decomposed in

the sum of partial totals associated with each unplanned domain,

$$T_j \quad = \quad T_j^{\mathcal{D}_1} + \ldots + T_j^{\mathcal{D}_D}. \tag{4.2}$$

However, in that case, T_j is intrinsically unknown, as well.

Once the sample is being observed, each element of the vector of composition of the population $\mathcal{P}_j$ satisfying (4.1) is decomposed as the sum of two components,

$$N_j^{\mathcal{D}_d} \quad = \quad N_j^{\mathcal{S}\cap\mathcal{D}_d} + N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d},$$

where

$$N_j^{\mathcal{S}\cap\mathcal{D}_d} = \#(\mathcal{S}_j \cap \mathcal{D}_d), \quad \text{and} \quad N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d} = \#(\widetilde{\mathcal{S}}_j \cap \mathcal{D}_d),$$

for $d = 1, \ldots, D$.

Similarly, each element of the sum (4.2) can be spread into the sum of sampled and unsampled parts,

$$T_j^{\mathcal{D}_d} \quad = \quad T_j^{\mathcal{S}\cap\mathcal{D}_d} + T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d},$$

where

$$T_j^{\mathcal{S}\cap\mathcal{D}_d} = \sum_{g\in\mathcal{S}\cap\mathcal{D}_d} y_{jg}, \quad \text{and} \quad T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d} = \sum_{l\in\widetilde{\mathcal{S}}\cap\mathcal{D}_d} Y_{jl},$$

for $d = 1, \ldots, D$.

In the above decompositions of the number of individuals and totals of the population, the sampled parts, $\{(T_j^{\mathcal{S}\cap\mathcal{D}_d}, N_j^{\mathcal{S}\cap\mathcal{D}_d})\}_{d=1}^D$, are completely known. The complementary unsampled parts, $\{(T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}, N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d})\}_{d=1}^D$, remain unknown, which inference is needed to be made. Notice that inference on the unsampled part of the composition of N_j across unplanned domains must satisfy,

$$N_j = \sum_{d=1}^{D} N_j^{\mathcal{S}\cap\mathcal{D}_d} + \sum_{d=1}^{D} N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}, \tag{4.3}$$

for any j.

Therefore, inference on totals of unplanned domains within $\mathcal{P}_j$ requires to extend the scope of uncertainty in order to include the underlying unknown composition of the population between unplanned domains. Accordingly, inference can be made in two steps. First, making inference on the underlying composition induced by the unplanned domains. And, second, making inference on the induced partial totals, given the inference on the underlying composition induced by the unsampled domains. In what follows we detail that inferential framework.

4.1 Prior on the composition of unplanned domains

We assume that the composition induced by unplanned domains in $\mathcal{P}_j$ is random. Given the sample, $\mathcal{S}_j$, only the unsampled part of that decomposition,

$$\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}} = (N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}, \ldots, N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D}), \tag{4.4}$$

remains unknown. Thus, it is reasonable to think of the vector $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ as a realization of a multinomial distribution, with

$$\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}|\boldsymbol{p}_j^{\mathcal{D}} \sim \text{Mult}(\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}|N_j^{\widetilde{\mathcal{S}}}, \boldsymbol{p}_j^{\mathcal{D}}), \tag{4.5}$$

where $N_j^{\widetilde{\mathcal{S}}} = N_j - N_j^{\mathcal{S}}$, is the number of unsampled individuals in $\widetilde{\mathcal{S}}_j$, which is known, and the vector of latent proportions of the composition across unplanned domains is $\boldsymbol{p}_j^{\mathcal{D}} = (p_j^{\mathcal{D}_1}, \ldots, p_j^{\mathcal{D}_{D-1}})$. That vector is defined on the $(D-1)$-dimensional simplex, with $p_j^{\mathcal{D}_D} = 1 - \sum_{d=1}^{D-1} p_j^{\mathcal{D}_d}$. Notice that each $p_j^{\mathcal{D}_d}$ can be interpreted as the probability that an individual in $\widetilde{\mathcal{S}}_j$ belongs in the unplanned domain $\mathcal{D}_d$, for $d = 1, \ldots, D$.

According to the paradigm we have been using, inference on (4.4) requires the specification of a probability distribution on the vector $\boldsymbol{p}_j^{\mathcal{D}}$. For that we use the natural conjugate $(D-1)$-dimensional Dirichlet prior, with some vector parameter $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \ldots, \alpha_{j,D})$, such that $\alpha_{j,d} > 0$, for $d = 1, \ldots, D$. That is to say,

$$\boldsymbol{p}_j^{\mathcal{D}} \sim \text{Dir}(\boldsymbol{p}_j^{\mathcal{D}}|\boldsymbol{\alpha}_j). \tag{4.6}$$

It is well known that (4.5) and (4.6) form a conjugate pair of distribution functions. See Gutiérrez-Peña and Smith (1997) for a comprehensive review on conjugate and exponential

families of distribution functions. Therefore, posterior inference and predictions on $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ are produced after updating (4.6) with data in the sample.

4.2 Predictive estimates on unplanned domains

We make predictive inference on domain totals and their composition across unplanned domains in two steps, without using additional information. However, we anticipate that introducing additional relevant information is do-able using multinomial-regression models, for example. We first propose to make inference on the composition attained to the unplanned domains, and then to make inference on the partial totals induced by the partition of the unplanned domains.

Under the multinomial-Dirichlet model, the posterior distribution of $\boldsymbol{p}_j^{\mathcal{D}}$ is also Dirichlet, with an updated vector parameter given by

$$\boldsymbol{\alpha}_j(\mathcal{S}_j) = (\alpha_{j,1} + N_j^{\mathcal{S}\cap\mathcal{D}_1}, \ldots, \alpha_{j,D} + N_j^{\mathcal{S}\cap\mathcal{D}_D}). \tag{4.7}$$

Therefore, the predictive estimate of the vector of compositions for the unplanned domains in $\mathcal{P}_j$ is being defined as the integer part of the D-dimensional vector with entries given by,

$$\widehat{N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}} = N_j^{\widetilde{\mathcal{S}}} \cdot \frac{\alpha_{j,d} + N_j^{\mathcal{S}\cap\mathcal{D}_d}}{\sum_{i=1}^{D}(\alpha_{j,i} + N_j^{\mathcal{S}\cap\mathcal{D}_i})}, \tag{4.8}$$

for $d = 1, \ldots, (D-1)$, and

$$\widehat{N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D}} = N_j^{\widetilde{\mathcal{S}}} - \sum_{d=1}^{D-1} \widehat{N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}}, \tag{4.9}$$

where $N_j^{\widetilde{\mathcal{S}}}$ is known. For the above estimates, we take compute their integer part.

Notice that the prior specification (4.5) and (4.6) complement the prior specification on the total T_j, for each planned domain $\mathcal{P}_j$, that we have discussed in Section 2. Just as it happens with T_j, our uncertainty about $\boldsymbol{N}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ vanishes as long as the sample size increases, and all relevant information concentrates in the sampled frequencies. Following the ideas we discussed in Section 3, concerning the prior specification of the SMM, we

suggest to define each vector parameter of the Dirichlet distribution as a function of the planned sample size, n_j, such that $\boldsymbol{\alpha}_j(n_j) \to \mathbf{0}$ as $n_j \to N_j$. In this way, $\boldsymbol{\alpha}_j(\mathcal{S}_j)$ shall converge to the population composition of $\mathcal{P}_j$ across $\mathcal{D}$.

Now, similarly to the derivations exposed in section 3, predictive estimates for each partial total associated with a given unplanned domain, $\mathcal{D}_d$, is written as

$$\widehat{T_j^{\mathcal{D}_d}} = \sum_{g \in \mathcal{S}_j \cap \mathcal{D}_d} y_{jg} + \widehat{N_j^{\tilde{\mathcal{S}} \cap \mathcal{D}_d}} \cdot \left[\sum_{k=1}^{U_j} \rho_k(\boldsymbol{m}_j) y_{jk}^* + \phi(\boldsymbol{m}_j) \widehat{\mu_{j0}} \right], \tag{4.10}$$

with $\widehat{\mu_{j0}}$ given as above. Consequently, the posterior estimate of the number of individuals in $\mathcal{P}_j$ that belong to the unplanned domain $\mathcal{D}_d$ is given by the sum of sampled part of the composition and the predicted estimate of the unsampled part,

$$\widehat{N_j^{\mathcal{D}_d}} = N_j^{\mathcal{S} \cap \mathcal{D}_d} + \widehat{N_j^{\tilde{\mathcal{S}} \cap \mathcal{D}_d}}, \tag{4.11}$$

for any d.

As before, it is also possible to produce more general inferences on the unplanned domains totals via the computation of the joint predictive distribution of the composition of the population and totals.

4.3 Inference through a vector of convolutions

Posterior estimates for $\boldsymbol{T}_j^{\mathcal{D}}$ and $\boldsymbol{N}_j^{\mathcal{D}}$, given in (4.10) and (4.11), respectively, are derived from a conditional dependence structure between the blocks of variables of totals and composition on unplanned domains. Accordingly, posterior inference about this quantities is summarized in terms of the predictive distribution,

$$\mathbb{P}\{\boldsymbol{T}_j^{\mathcal{D}}, \boldsymbol{N}_j^{\mathcal{D}} | \mathcal{S}_j\} = \mathbb{P}\{\boldsymbol{T}_j^{\mathcal{D}} | \boldsymbol{N}_j^{\mathcal{D}}, \mathcal{S}_j\} \times \mathbb{P}\{\boldsymbol{N}_j^{\mathcal{D}} | \mathcal{S}_j\}. \tag{4.12}$$

Through out this joint distribution, it is possible to make inference on the composition and totals attained to the unplanned domains $\mathcal{D}$ within each planned domain $\mathcal{P}_j$, simultaneously.

In the above representation, $\mathbb{P}\{\boldsymbol{N}_j^{\mathcal{D}} | \mathcal{S}_j\}$ is completely determined by the predictive distribution of the multinomial-Dirichlet conjugate component for the unsampled part of

the vector of compositions, $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$, *i.e.*

$$\mathbb{P}\{\boldsymbol{N}_j^{\mathcal{D}} \leq \boldsymbol{n}|\mathcal{S}_j\} = \int \text{Mult}(\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}} \leq \boldsymbol{n} - \boldsymbol{N}_j^{\mathcal{S}\cap\mathcal{D}}|N_j^{\widetilde{\mathcal{S}}}, \boldsymbol{p}_j^{\mathcal{D}})\,\text{Dir}(\text{d}\boldsymbol{p}_j^{\mathcal{D}}|\boldsymbol{\alpha}_j(\mathcal{S}_j)), \tag{4.13}$$

where $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ is the vector of compositions of $\widetilde{\mathcal{S}}_j$ across the unplanned domains, and

$$\boldsymbol{N}_j^{\mathcal{S}\cap\mathcal{D}} = (N_j^{\mathcal{S}\cap\mathcal{D}_1}, \ldots, N_j^{\mathcal{S}\cap\mathcal{D}_D}),$$

is the vector of compositions of $\mathcal{S}_j$ across the D-unplanned domains. This vector plays the role of a shifting constant.

Similarly, the conditional predictive distribution of the vector with partial totals for the unplanned domains, $\mathbb{P}\{\boldsymbol{T}_j^{\mathcal{D}}|\boldsymbol{N}_j^{\mathcal{D}}, \mathcal{S}_j\}$, is being completely determined by the predictive distribution of the vector of the unsampled parts associated with the unplanned domains, $\boldsymbol{T}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$, which is defined as a $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$-fold D-dimensional vector of convolutions, *i.e.*

$$\begin{aligned}\mathbb{P}\{\boldsymbol{T}_j^{\mathcal{D}} \leq \boldsymbol{t}|\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}}, \mathcal{S}_j\} = \\ = \Big(\widehat{G}^{*N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}}\left(T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1} \leq t_1 - T_j^{\mathcal{S}\cap\mathcal{D}_1}\right), \\ \ldots, \\ , \widehat{G}^{*N_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D}}\left(T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D} \leq t_D - T_j^{\mathcal{S}\cap\mathcal{D}_D}\right),\end{aligned} \tag{4.14}$$

where $T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}$ and $T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}$ are given as above, for $d = 1, \ldots, D$. Notice that the vector of the sampled composition, $(T_j^{\mathcal{S}\cap\mathcal{D}_1}, \ldots, T_j^{\mathcal{S}\cap\mathcal{D}_D})$, plays the role of the shifting vector constant for the predictive distribution of the composition of the whole planned domain.

As before, it is quite difficult to derive an analytic expression for the joint predictive distribution $\mathbb{P}\{\boldsymbol{T}_j^{\mathcal{D}}, \boldsymbol{N}_j^{\mathcal{D}}|\mathcal{S}_j\}$ in a closed form. However, we can circumvent that problem through simulation methods. In the next subsection we sketch an algorithm to drawing samples from (4.14) and, therefore, from the predictive distribution of composition and totals of unplanned domains over imposed to any aggregation of planned domains.

4.4 Simulation algorithm for planned and unplanned domains

Here we sketch the simulation steps designed to draw samples from the joint predictive distribution (4.14). The algorithm is designed for the Dirichlet process, but its adaptation

to any other SSM is straightforward. The basic steps are given as follows.

Step 0. Start by setting up the parameters of the model and the sampler, *i.e.* $\alpha_{j,DP}$, G_{j0} and $\boldsymbol{\alpha}_j$, for $j = 1, \ldots, J$.

Step 1. Generate a D-dimensional vector, $\boldsymbol{p}_j^{\mathcal{D},(i)}$, from the updated Dirichlet distribution with parameter $\boldsymbol{\alpha}_j(\mathcal{S}_j)$.

Step 2. Generate a random sample of the vector of composition across unplanned domains, $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D},(i)}$, from the multinomial distribution with updated parameters $N_j^{\widetilde{\mathcal{S}}}$ and $\boldsymbol{p}_j^{\mathcal{D},(i)}$.

Step 3. Compute a sample of the composition vector across unplanned domains, where

$$\boldsymbol{N}_j^{\mathcal{D},(i)} = \boldsymbol{N}_j^{\mathcal{S}\cap\mathcal{D}} + \boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D},(i)}$$

for $j = 1, \ldots, J$

Step 4. Generate a collection of $N_j^{\widetilde{\mathcal{S}}}$ individual characteristics, $(y_{jl}^{(i)} : l = 1, \ldots, N_j^{\widetilde{\mathcal{S}}})$, from the predictive distribution $\widehat{G}_j$.

Step 5. Distribute the collection of the simulated individual characteristics, $(y_{jl}^{(i)})$, across the D unplanned domains, accordingly to the vector of compositions $\boldsymbol{N}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D},(i)}$.

Step 6. Generate the vector of partial totals across unplanned domains, as $\boldsymbol{T}^{\mathcal{D},(i)} = \boldsymbol{T}^{\mathcal{S}\cap\mathcal{D}} + \boldsymbol{T}_j^{\widetilde{\mathcal{S}}\cap\mathcal{D},(i)}$. Note that $T_{tmar}^{\mathcal{S}\cap\mathcal{D}_d}$ and $T_j^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}$ are given as before.

As before, steps 1 to 6 are repeated in each iteration of the sampler. Notice that all the simulation steps involved there are relatively simple to implement. However, slight complexities may show-up when considering different types of SSMs. It is worth mentioning that the algorithm may demand heavy computational resources either when dealing with either relatively large samples or when the composition of the target population is relative large.

5 Population averages

Most finite population studies also concern with average characteristics of the population, *i.e.* $\eta = T/N$, where T represents the population total and N represents the number

of individuals in the target population. Weight-based estimates of population averages are expressed as the ratio $\eta = \sum_{g\in\mathcal{S}} w_g y_g / \sum_{g\in\mathcal{S}} w_g$, where $(w_g)_{g\in\mathcal{S}}$ corresponds to the sample weights. Well controlled design-based weights are built in such a way that $N = \sum_{g\in\mathcal{S}} w_g$ is fixed and known. However, when the target population makes reference to unplanned domains, the sum $\sum_{g\in\mathcal{S}} w_g$ would merely be an estimation of the composition of the target population, and the traditional estimator of μ would be written as a ratio estimator $\widehat{\mu} = \widehat{T}/\widehat{N}$; see Hájek (1971). Following this idea, when dealing with unplanned domains, a proper assessment of a ratio estimator for averages on unplanned domains would require a joint assessment of the estimated compositional vector of the population, or its marginal distribution. Recall that this problem remains active in the literature; see, *e.g.* Aronow and Lee (2013). For the best of our knowledge, a proper assessment of ratio estimators is still difficult to perform using traditional weight-based methods. The later problem is another reason why weight-based estimators get overwhelmed when dealing with unplanned domains. The methodology proposed in this paper allows to make predictive distributional inference on η, disregarding if we are dealing with planned or unplanned domains. For doing so, it is convenient to take into account the cases of whether N is known or unknown. Thus, we can produce inferences on η beyond point estimates. Refer to Ghosh and Meeden (1997) and Hammer et al. (2001) for related approaches to the one presented in this paper.

5.1 Case N known

For the sake of simplicity, allow us to describe the inferential procedure an the interior of a given planned domain. Let T_j be defined as before, and let N_j be the total number of individuals in $\mathcal{P}_j$. The average mean for j is then given by $\eta_j = T_j/N_j$. Hence, the predictive distribution for the population average on the planned domain j would be given in terms of the convolution distribution (3.5), as

$$\mathbb{P}(\eta_j \leq t|\mathcal{S}_j) = \widehat{G}_j^{*N_j^{\widetilde{S}}}\left(T_j^{\widetilde{S}} \leq N_j \cdot t - T_j^{\mathcal{S}}\right) \tag{5.1}$$

where N_j corresponds to a constant term, with the remaining components given as above.

Therefore, full posterior inferences on η_j are obtained from (5.1). As sketched in Subsection 3.3, those inferences can rely on simulation results. Draws from (5.1) are computed from draws from (3.5) by means of transforming the draws $(T_j^{(i)})_{i\geq 1}$ into $(\eta_j^{(i)})_{i\geq 1}$, with $\eta_j^{(i)} = T_j^{(i)}/N_j$. Thus, complementary to Algorithm 3.3, it would only be required to add the follow step:

Step 5. Compute $\eta_j^{(i)} = T_j^{(i)}/N_j$, for any $i \geq 1$.

5.2 Case N unknown

The case N unknown we are interested here is the one derived from the interest of computing population averages of unplanned domains. Let us focus in what happens within a planned domain $\mathcal{P}_j$. As described above, the computation of population averages across unplanned domains in $\mathcal{P}_j$ must be computed as a vector of averages, $\boldsymbol{\eta}_j^{\mathcal{D}} = (\eta_j^{\mathcal{D}_1}, \ldots, \eta_j^{\mathcal{D}_D})$, where $\eta_j^{\mathcal{D}_d} = T_j^{\mathcal{D}_d}/N_j^{\mathcal{D}_d}$ for any d. The distribution associated with $\boldsymbol{\eta}_j$ is induced, by transformation, from (4.14). We can deal with that distribution via simulation, as well. That would require to transform the simulated outcomes from the joint distribution (4.14) into $(\boldsymbol{\eta}_j^{\mathcal{D},(i)})_{i\geq 1}$. That can be achieved by means of adding up the follow step into the Algorithm 4.4:

Step 7. Compute $\eta_j^{\mathcal{D}_d,(i)} = T_j^{\mathcal{D}_d,(i)}/N_j^{\mathcal{D}_d,(i)}$, for $d = 1, \ldots, D$ and any $i \geq 1$.

Similarly, draws from the distribution of averages over unplanned domains derived from aggregations of planned domains are obtained by means of aggregating the draws produced with the algorithm sketched in Sub-section 4.4 in each iteration, accordingly. For example, if the aim is to produce the vector of averages for the whole population across unplanned domains, it would be necessary to transform the draws for each iteration, *i.e.* $(\boldsymbol{T}_j^{\mathcal{D},(i)}, \boldsymbol{N}_j^{\mathcal{D},(i)})_{j=1}^J$, into the vector $\boldsymbol{\eta}^{\mathcal{D},(i)}$. That can also be done by means of including the additional step into the Algorithm 4.4:

Step 8. Compute $\eta^{\mathcal{D}_d,(i)} = \sum_{j=1}^J T^{\mathcal{D}_d,(i)} / \sum_{j=1}^J N^{\mathcal{D}_d,(i)}$, for any $i \geq 1$.

It is worth mentioning that computing steps 5, 7 and 8 that we described in this Section do not require to run Algorithms 3.3 and 4.4 over and over, it only require to transform the outcomes produced from previous runs of those algorithms.

6 Simulation study

For the simulation study, we have generated a fictitious population of 2 thousand individuals grouped in two planned domains. Domain A is formed of 800 individuals, and domain B is composed of 1.2 thousand individuals. Each planned domain of the population has been split, as well, into the three unplanned domains. The composition of the population across planned and unplanned domains is summarized in Table 1. Individual measurements were simulated using the log-normal and Weibull distributions with different parameterizations for each combination of planned and unplanned domains. Table 2 summarizes the distributions from which the data were generated.

Table 1: Composition of the simulated population across planned and unplanned domains.

Planned	Unplanned			Total
	Domain I	Domain II	Domain III	
Domain A	200	400	200	800
Domain B	780	400	20	1,200
Total	980	800	220	2,000

In Figures 1(b) and 1(c) we display the outcomes for the simulated population, which we take it as the population of reference in the simulation study. Sub-figure 1(a) displays the distribution of the aggregate population, whereas that Sub-figures 1(b) and 1(c) display the disaggregate distributions for the planned domains A and B, respectively. The associate disaggregation for the three unplanned domains is being displayed in Figures 2(a), 2(b) and 2(c). Individual simulated measurements of the reference data were also drawn from the distributions summarized in Table 2. Notice that some of this distributions exhibit slight heavy tails, which resemble the behavior of some actual measurements in real life problems.

The simulation exercise reported in this Section, consists in extracting a random sample of five percent of the population, in the first instance. Taking that sample as a reference, we gradually cover up the whole population by adding up a random sample of the same

Table 2: Distributional origins of individual measurements for the simulated population.

Planned	Unplanned					
	Domain I		Domain II		Domain III	
Domain A	μ	σ	μ	σ	λ	ϕ
	6.198	0.840	6.801	0.701	1.850	1646.4
Domain B	μ	σ	μ	σ	μ	σ
	6.152	0.830	6.670	0.614	6.987	0.698

NOTE: All measurements were generated from log-normal distributions, with exception of domain A-III, which were generated from a Weibull distribution. The parameters μ and σ correspond to the mean and standard deviation of the log-normal distribution, whereas λ and ϕ correspond to the shape and scale parameters, respectively, of the Weibull distribution.

size extracted from the rest of the unsampled population at each time. In this way, we are able to produce 20 random samples, which gradually cover up the whole population. Now, on each one of those samples we implement our proposed methodology, using the Dirichlet process as the particular SSM. In each implementation, we elicit the parameter $\alpha_{j,\mathrm{DP}}$ in terms of the sample size, as a linear decreasing function according to the ideas discussed in Section 3. We follow the same idea for eliciting the three-dimensional vector parameters $\boldsymbol{\alpha}_j$ for the multinomial-Dirichlet component. As reference data for eliciting the two baseline distribution functions, G_{j0}'s, for the two planned domains, we use an additional simulated population of 250 individual measurements distributed across planned and unplanned domains consistently with the distribution in Table 1.

Prior elicitation of each G_{j0} is based on a predictive model comparison and selection among four alternative parametric distributions: Weibull, log-normal, gamma and inverse-Gaussian distribution.[2] We also explored the generalized extreme value distribution and the generalized Pareto distribution, because some data groups are extreme values validated data Chambers (1986), but we could not found a suitable fit to the data in the two cases. Additionally, the generalized gamma distribution was tested as an alternative. We found

[2]Some of these distributions were fitted using algorithms related to Yee (2010).

suitable the generalized gamma distribution because the distribution encompasses cases such as the gamma, Weibull and exponential distributions, which includes the log-normal as a limit case. The challenge involving the use of the generalized gamma distribution is the estimation of the shape parameters. We resort to the application of a maximization iterative algorithm to estimate the parameters of the distribution, proposed by Noufaily and Jones (2013). However, our database did not allow to identify a proper optimization of the likelihood. In the end, we preferred to use only functions of the following distributions: Weibull, log-normal and gamma distributions. Finally, we decided to consider the inverse-Gaussian distribution, because it is used in finance and insurance, and the distributional families included in the simulation exercise would be more exhaustive to model data with extreme value.

The results are computed with 30,000 Monte Carlo simulations for each sample size. Figure 3 shows trends and predictions of the population totals for aggregate and disaggregate planned domains, as a function of the sample size. Figure 3(a) displays the trajectories for the predictive estimates of the aggregate total, whereas that figures 3(b) and 3(c) do the same for the totals attained to the planned domains A and B, respectively. These figures display the trajectories of traditional weight-based estimates (the dotted-lines), naïve estimates (dotted and solid lines) and actual totals (solid constant lines).[3] As we can observe, the behavior of traditional estimates and predictive means of our proposed methodology is similar. However, it is bluffing the apparent erratic behavior of traditional weight-based estimator in relatively small sample sizes. Our conjecture for that result is that weight-based estimators are extremely sensible to extreme values. Planned domain B does not exhibit extreme measurements as planned domains A does. Thus, its estimates in small sample sizes are less erratic.

This simulation study exhibits the erratic behavior that traditional weight-based estimators may have in the presence of representative extreme measurements or outliers (see, Chambers, 1986), particularly for relatively small sample sizes. That is not a surprising result, as it has been widely documented in the literature that weight-based estimators

[3] See Appendix B for further references.

are highly sensitive to those cases. See, for instance, Chambers (1986) and Ghosh (2008). It is worth to noting that recent approaches to handle with extreme measurements resort on model-based methods to either calibrate the sample weights or to make inference directly with the model. See Beaumont and Rivest (2009) and Beaumont et al. (2013), for example. In that respect, our methodology gives some insights to the problem and handles with extreme measurements within an integrated inferential framework. But, we shall point out that informative and relevant reference information is needed to properly elicit the baseline parameters of the SSM required in our formulation.

On the other hand, Figure 4 displays predictive trajectories for the composition and partial totals associated with the three unplanned domains. Figures 4(a), 4(c) and 4(e) make reference to the predictive composition of the population across the three unplanned domains in terms of the random samples. As we can observe, the evolution of the composition of the three unplanned domains evolves consistently. Figures 4(b), 4(d) and 4(f) display the trajectories of the predictive distributions for the partial totals of each one of the three unplanned domains. As it was expected, those trajectories display an erratic behavior. The trajectories of traditional weight-based estimators are displayed in dotted lines. Differently to what happened with planned domains, in unplanned domains we observe a discrepant evolution of the estimators, particularly in domains I and II. However, in domain III both methods produce consistent results for any sample size. We believe that what we observe in these trajectories is a consequence of the way that the composition and the extreme measurements in the sample combine. We acknowledge that the composition of the population across unplanned domains plays an important role, and even more when the sub-populations have extreme measurements, as it happens with domains I and II. Even thought, when the subpopulation measurements range more uniformly in a narrower interval, as in domain III, both approaches converge quickly consistently to the actual measurement.

7 Uncertainty surrounding the gender wage gap in Mexico

In this section, we illustrate the usefulness of our method for the study of the uncertainty surrounding the OECD definition of wage differentials by gender in Mexico.[4] In doing that, we use data from the *Encuesta Nacional de Ocupación y Empleo* (ENOE, for its abbreviation in Spanish), the largest national household survey in Mexico. It is worth mentioning that our illustration does not attempt to elaborate an exhaustive analysis on the possible explanatory factors that may cause such differences, as Meza González (2001) and Popli (2013), for example, do. Such an aim would require a more elaborate analysis based on matched micro-data. Instead, we develop our illustration as a descriptive exercise in which we give some insights about the uncertainty surrounding this statistic, and the way the proposed methodology can be applied to a real life problem. For that, we adopt a predictive distributional approach, which goes in line with related work developed by DiNardo et al. (1996).[5]

The ENOE is nowadays the largest national household survey in Mexico. This survey was preceded by a national survey on urban labor. The agency responsible for administering and producing results for this survey is the *Instituto Nacional de Estadística y Geografía* (INEGI, for its abbreviation in Spanish). The ENOE uses stratified and cluster sampling schemes in two stages to make the survey data representative nationally and at some other levels of disaggregation, such as rural and urban areas, but it is not representative of wages by gender. The ENOE is collected in quarterly basis since 2005. Each quarterly edition of the survey comprises over 120 thousand households measurements. The data is intended to make reference to a target population of approximately 46 millions of individual employees. For the sake of illustration, we restrict this exercise

[4] We make reference to the official definition of unadjusted gender wage gap used by the OECD, which summarizes the difference between male and female earnings relative to male earnings. This statistic is computed using median or average gross hourly earnings of full time employees (see OECD, 2012).

[5] By distributional approach, we mean that the aim is at reproducing the underlying distribution of this statistic. Thus, relevant aspects of the uncertainty surrounding this statistics can be assessed.

to analyze data corresponding to the third quarter of 2011.[6] Micro-data obtained from the ENOE are accompanied with sample weights, which are defined consistently with the survey design. See INEGI (2007) for further details.

Official computations on wage statistics in Mexico produced by INEGI make reference to the employed population, 15 years old and older, whose monetary earnings were strictly positive. The computation of official statistics on wages by gender typically uses weight-based estimators. Since gender groups are not considered during the sample design, they are treated as unplanned domains. Official statistics on earnings are computed as medians or averages; in our illustration we make reference to average earnings. Here, we illustrate the computations that can be produced using the proposed methodology, which allows us to evaluate the predictive distribution underlying the gender wage gap. According to the OECD definition, the gender wage gap based on average earnings is computed as

$$\text{wage.gap}_t = 100 \times \frac{\overline{\text{wage}}_t^{\texttt{male}} - \overline{\text{wage}}_t^{\texttt{female}}}{\overline{\text{wage}}_t^{\texttt{male}}}, \tag{7.1}$$

where $\overline{\text{wage}}_t^{\texttt{male}}$ and $\overline{\text{wage}}_t^{\texttt{female}}$ denote the population average wage per hour for male and female at quarter t, respectively. Following the proposed methodology, the distribution underlying the gender wage gap can be reproduced from Monte Carlo samples generated for $\overline{\text{wage}}_t^{\texttt{male}}$ and $\overline{\text{wage}}_t^{\texttt{female}}$, which are obtained according to the algorithm described in Section 4 and the extensions presented in Section 5.

Figure 5 summarizes the uncertainty surrounding the gender wage gap in Mexico through its predictive distribution. There it can be seen that the gender wage gap in Mexico is likely to be concentrated between -0.5 and 2.8 percent, with respect to the male average earnings per hour for the period of analysis, with approximately 0.95 of probability mass. It is worth to noting that assessing the dispersion of the gender wage gap using traditional weight-based methods is quite challenging, due that this statistic is expressed as a double ratio estimator for complex survey samplings (see, Meng, 1993,

[6] The results presented here were produced using data published before the recent adjustment for population predictions, and their retrospective adjustments, that INEGI carried out in the second quarter of 2013.

and the references therein). To clarify, weight-based methods and the proposed methodology take into account different sources of uncertainty, as it has been described in the core of this document. Accordingly, the way in which both methods (weight-based and proposed) assess the uncertainty surrounding the type of statistic under study is different, but complementary.[7] However, with the aim of having a benchmark for comparison with traditional methods of estimation, it can be pointed out that the estimated gender wage gap using the weight-based methods (which is approximately equal to 2.3 percent) is contained in the region of highest concentration of the predictive distribution derived from the proposed method. This result illustrates that the proposed methodology may provide a complementary view at the uncertainty surrounding this type of statistics for finite populations.

8 Discussion

In this paper we develop an intuitive and simple framework to inference on totals and averages of finite populations, assuming that the population is segmented in planned domains. We take into consideration that unobserved individual characteristics are continuous and random, and that the sampling scheme from which the outcomes are obtained is non or partially informative. Assuming that unobserved individual measurements are random allows for the possibility of providing a more robust reading of the uncertainty surrounding the inferential process. The predictive distribution of the characteristic of interest is recovered as a convolution-type distribution, which is evaluated using Monte Carlo methods. That is a key distinction of our framework with regard to traditional weight-based alternatives. It is worth mentioning that the structural assumptions on the model used to produce predictions are being relaxed to the minimum possible by means of incorporating a Bayesian nonparametric component.

Another contribution of this paper consists in the formulation of a procedure to make

[7] Traditional weight-based methods focus on estimating variances of estimators (mean, median, etc.). The proposed method, on the contrary, aims at assessing the uncertainty of the population variable (in this case, the gender wage gap). Thus, these methods are complementary.

inference on totals segmented in unplanned domains. In our formulation, the notion of uncertainty is spanned by means of incorporating the random composition of the population induced by the unplanned domains, together with their underlying sub-totals. Therefore, predictive inferences are simultaneously obtained on the composition of the population and on sub-totals across unplanned domains, and they are consistent with the inferences produced for planned domains, as well. To the best of our knowledge, this is the first approach in the literature achieving this aim.

Our framework also allows for the incorporation of relevant additional prior information, in the form of G_{j0}, which reflect knowledge from reference data or from subjective or expert opinions concerning the underlying distribution attained to each planned domain. In the absence of any type of prior information, the contribution of G_{j0} may be forced to be vanished away by means of setting $\phi(n_j)$ close enough to 0. In that case, no prior opinion would be taken into consideration for making prediction. Of course, in practice, we must look for sensible ways of incorporating prior knowledge when setting G_{j0}, if such prior information exists. It is our belief that there is no standard guideline for doing that, it is mostly case-specific. The incorporation of prior information also brings up new elements to control for the presence of extreme representative measurements by means of controlling the weights associated with the sample frequencies, and by means of balancing prior and sample information through $\phi(\cdot)$, the tuning parameter of the species-sampling model.

References

Aronow, P. M. and Lee, D. K. K. (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240.

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, To appear.:doi: 10.1093/biomet/ast010.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In Rao,

C. R. and Pfefferman, D., editors, *Sample Surveys: Design, Methods and Applications, Vol. 29A*, pages 247–279. Elsevier B.V., Amsterdam.

Binder, D. A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B (Statistical Methodololy)*, 44:388–393.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069.

Chambers, R. L. and Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling Applications*. Oxford University Press, Oxford.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

French, S. (1985). Group consensus probability distributions: A critical survey. In Bernaro, J. M., DeGrott, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 183–202. Elseiver, Amsterdam.

Gelfand, A. E., Mallick, B. K., and Dey, D. K. (1995). Modeling expert opinion arising as a partial probabilistic specification. *Journal of the American Statistical Association*, 90(430):598–604.

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6:733–807.

Ghosh, M. (2008). Robust estimation in finite population. In *Beyond Parametrics in*

Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, volume 1, pages 116–122. Institute of Mathematical Statistics, Cambridge.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Gutiérrez-Peña, E. and Smith, A. F. M. (1997). Exponential and Bayes in conjugate families: Review and extensions (with discussion). *Test*, 6(1):1–90.

Hájek, J. (1971). Comment on: "An essay on the logical foundations of survey sampling, part one". In Godambe, V. P. and Sprott, D. A., editors, *The Foundations of Survey Sampling*, page 246. Holt, Rinehart & Winston, Toronto.

Hammer, M. S., Seaman, J. W., and Young, D. M. (2001). Bayesian methods in finite population sampling. In *Proceeedings of the Anual Meeeting of the American Statistical Association*. American Statistical Association.

Hansen, B. and Pitman, J. (2000). Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters*, 46(3):251–256.

Holt, D. and Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142(1):33–46.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

INEGI (2007). *Encuesta Nacional de Ocupación y Empleo (ENOE) - Métodos y procedimientos*. Instituto Nacional de Estadística y Geografía (INEGI), Aguascalientes, México.

Lee, J., Quintana, F. A., Müller, P., and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statistical Science*, To appear.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In Pfeffermann, D. and Rao, C. R., editors, *Handbook of Statistics.*

Sample Surveys: Inference and Analysis, Vol. 29B, pages 219–249. Elsevier B.V., Amsterdam.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Meeden, G. (2005). A noninformative Bayesian approach to domain estimation. *Journal of Statistical Planning & Inference*, 129(1-2):85–92.

Mena, R. H. (2012). Geometric weight priors and their application in Bayesian nonparametrics. In Damien, P., Dellaportas, P., Polson, N. G., and Stephen, D. A., editors, *Bayesian Theory and Applications*. Oxford University Press, Oxford.

Meng, X.-L. (1993). On the absolute bias ratio of ratio estimators. *Statistics & Probability Letters*, 18(5):345–348.

Meza González, L. (2001). Wage inequality and the gender wage gap in Mexico. *Economía Mexicana*, X(2):291–323.

Noufaily, A. and Jones, M. C. (2013). On maximisation of the likelihood for the generalised gamma distribution. *Computational Statistics*, 28(2):505–517.

OECD (2012). LMF1.5: Gender pay gaps for full-time workers and earnings differentials by educational attainment. Social Policy Division - Directorate of Employment, Labour and Social Affairs.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In Rao, J. K. N., editor, *Sample Surveys: Inference and Analysis, Vol. 29B*, chapter 39, pages 455–487. Elsevier B.V., Amsterdam.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T. S., Shapley, L. S., and MacQueen, J. B., editors, *Statistics, Probability and Game Theory*, pages 245–267. Institute of Mathematics and Statistics, Hayward, CA.

Popli, G. K. (2013). Gender wage differentials in Mexico: A distributional approach. *Journal of the Royal Statistical Society, Series A*, 176(2):295–319.

R Core Team (2012). R*: A Language and Environment for Statistical Computing, Reference Index*. R Foundation for Statistical Computing, Vienna, Austria, reference index version 2.15.2 edition.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Thompson, M. E. (1997). *Theory of Sample Surveys*. Number 74 in Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Yee, T. W. (2010). The vgam package for categorical data analysis. *Journal of Statistical Software*, 32(10):1–34.

A Connection with traditional weight-based estimators

Our predictive estimates for totals on planned domains encompass traditional weight-based estimators. For instance, it is straightforward to see that the predictive point estimate of the total (3.4), based on the Dirichlet process, encompasses the Horvitz-Thompson estimator as a particular case. Following the notation we used above, the Horvitz-Thompson estimator for T_j can be rewritten in terms of the collection of sampled ties and frequencies, as

$$\begin{aligned} \widehat{T_j^{\mathrm{HT}}} &= \sum_{g \in \mathcal{S}_j} \left(\frac{N_j}{n_j} \right) y_{jg} \\ &= \sum_{k=1}^{U_j} \left(\frac{N_j}{n_j} \right) m_{jk} \cdot y^*_{jk}, \end{aligned} \tag{A.1}$$

where $n_j = N_j^{\mathcal{S}}$, with U_j and $\{(m_{jk}, y^*_{jk}) : k = 1, \ldots, U_j\}$ are given as in the document.

Additionally, by taking the parameter $\alpha_{j,\mathrm{DP}}$ arbitrarily close to 0 it follows that for any sample the function $\phi(\boldsymbol{m}_j)$ would be arbitrarily close to 0, as well. Hence, in the limit, our predictive point estimate for T_j can be written as

$$\begin{aligned}
\widehat{T}_j &= \sum_{g\in\mathcal{S}_j} y_{jg} + N_j^{\widetilde{\mathcal{S}}} \cdot \sum_{k=1}^{U_j} \rho_k(\boldsymbol{m}_j)\cdot y_{jk}^* \\
&= \sum_{k=1}^{U_j} \left(1 + \frac{N_j^{\widetilde{\mathcal{S}}}}{n_j}\right) m_{jk}\cdot y_{jk}^* \\
&= \sum_{k=1}^{U_j} \left(1 + \frac{N_j - n_j}{n_j}\right) m_{jk}\cdot y_{jk}^*, \qquad \text{(A.2)}
\end{aligned}$$

from which, it can be seen that (A.1) is a particular case. Note that this result does not necessarily hold for other specifications of species-sampling models, apart of the aforementioned Dirichlet process.

B Review of traditional estimators for replicate samples

The simulation study reported in Section 6 is build upon a progression of 20 samples, chosen randomly without replacement from the referred population, which cover up the population in blocks of five percent of the population. Each sample in that progression is denoted by $\mathcal{S}^{[h]}$, for $h = 1,\ldots,20$, such that their sample size is $n^{\mathcal{S}^{[h]}} = N\cdot(h/20)$. Let us recall that each sample is balanced across planned domains, *i.e.* $n_j^{\mathcal{S}^{[h]}} = N_j\cdot(h/20)$ for each jth planned domain. However, they are not necessarily balanced across unplanned domains, as it happens in practice.

Concerning weight-based estimators, each outcome observed in the sample $\mathcal{S}^{[h]}$ is being assigned a sample weight given by $w_{jl}^{[h]} = (n_j^{\mathcal{S}^{[h]}}/N_j)^{-1}$, for $l = 1,\ldots,n_j^{\mathcal{S}^{[h]}}$, according to whether the ith observed outcome belongs to the jth planned domain. Thus, the weight-based estimate of the total for each jth planned domain is given by $\widehat{T}_j^{[h]} = \sum_{l=1}^{n_j^{\mathcal{S}^{[h]}}} w_{jl}^{[h]}\cdot y_{jl}$. Additionally, the weight-based estimate of the total for unplanned domains, within planned domains, is given by $\widehat{T}_j^{[h],\mathcal{D}_d} = \sum_{l=1}^{n_j^{\mathcal{S}^{[h]}}} w_{jl}^{[h]}\cdot y_{jl}\cdot I(l\in\mathcal{D}_d)$, for $d = 1,\ldots,D$.[8]

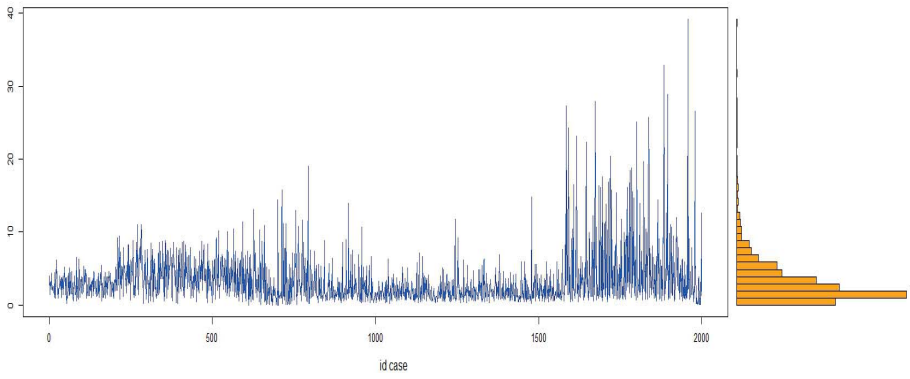
[8]Here, $I(A)$ stands for the indicator function which is equal to 1 is the condition A holds, and 0

Concerning naïve estimators, they are computed as the product of the population size and the (unweighted) sample mean, *i.e.* $\widehat{T}_j^{[h]} = N_j \cdot \bar{y}_j$, where $\bar{y}_j = (1/n_j^{\mathcal{S}^{[h]}}) \cdot \sum_{l=1}^{n_j^{\mathcal{S}^{[h]}}} y_{jl}$. It is trivial to see that across planned domains weight-based and naïve estimators are the same. However, discrepancies are found when computing the estimates for unplanned domains. In that case, the naïve estimate for the total across unplanned domains is given by $\widehat{T}_j^{[h],\mathcal{D}_d} = \widehat{N}_j \cdot \bar{y}_j$, where $\widehat{N}_j^{\mathcal{D}_d} = N_j \cdot \sum_{l=1}^{n_j^{\mathcal{S}^{[h]}}} I(l \in \mathcal{D}_d)/n_j^{\mathcal{S}^{[h]}}$ is the naïve estimate of the composition of the unplanned domain $\mathcal{D}_d$ within the jth planned domain, for $d = 1, \ldots, D$.

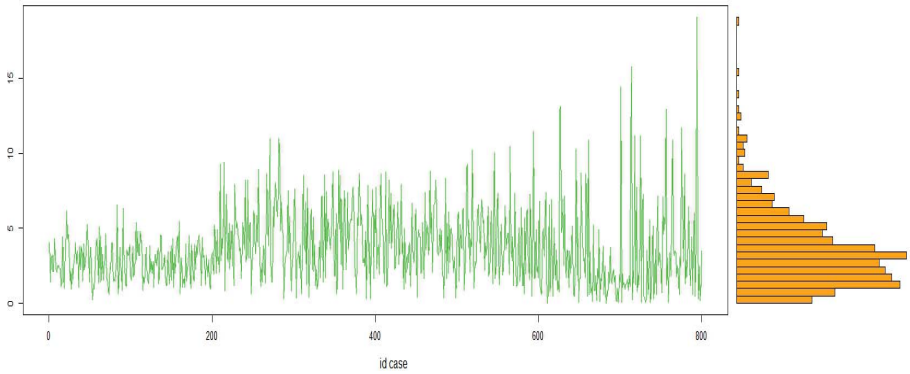
C Additional material

`predfinitepop`. This is a package written in `R` (R Core Team, 2012) which implements the methodology introduced in this paper and replicates the simulation results presented in Section 6. The package is available from the authors upon request.

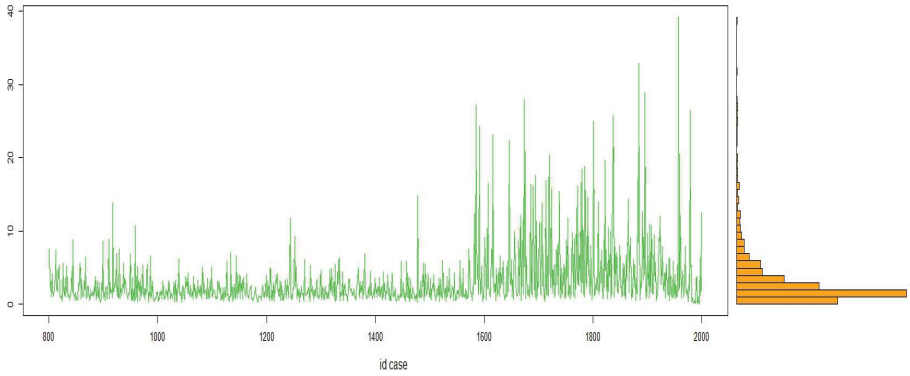
otherwise.



(a) Whole population

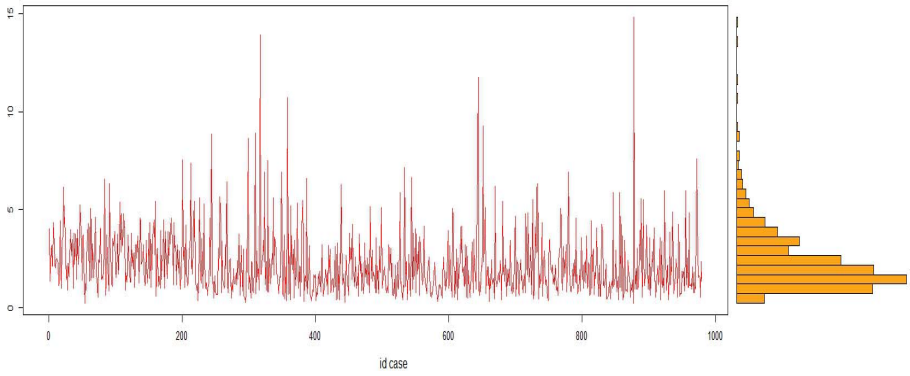


(b) Planned domain A

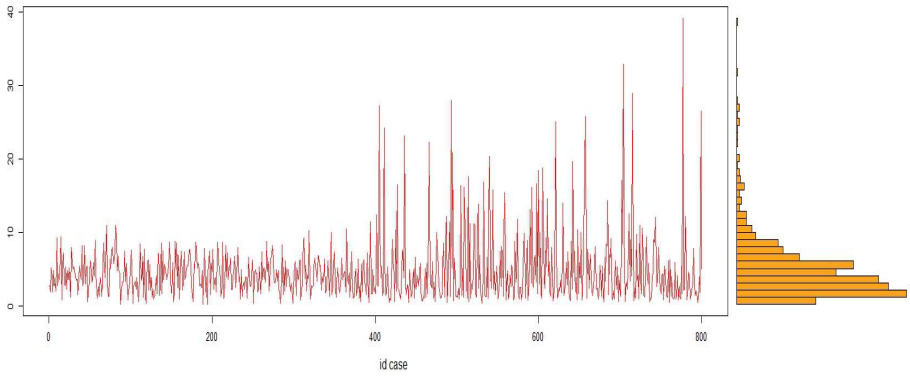


(c) Planned domain B

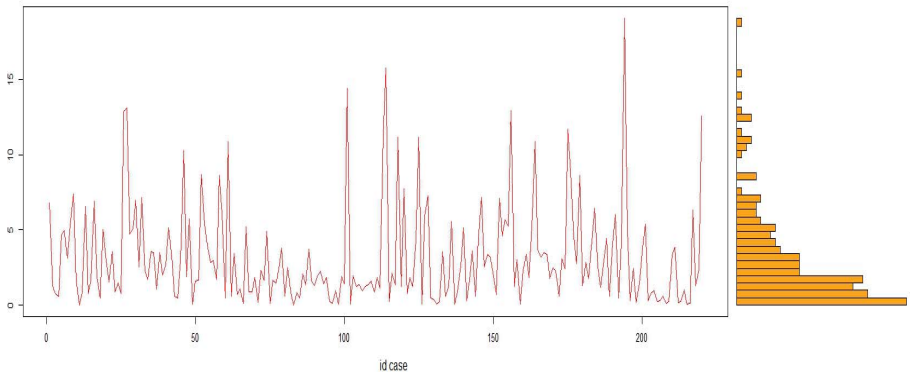
Figure 1: Distribution of individual measurements for the whole simulated population and the two planned domains.



(a) Unplanned domain I

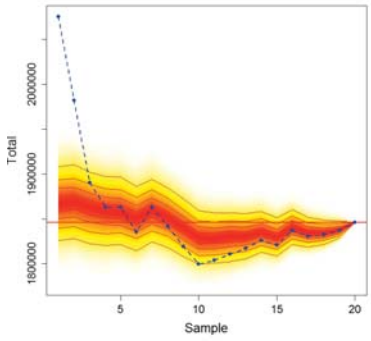


(b) Unplanned domain II

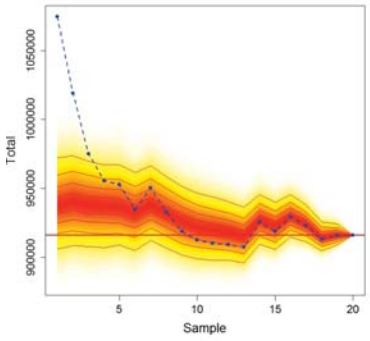


(c) Unplanned domain III

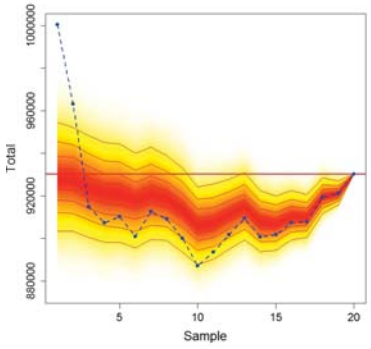
Figure 2: Distribution of individual measurements for the three unplanned domains.



(a) Whole population

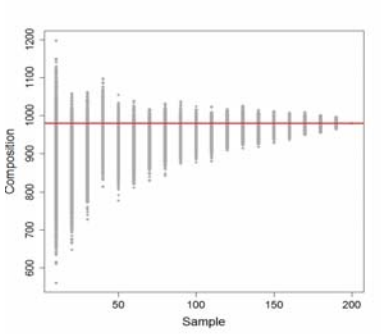


(b) Domain A

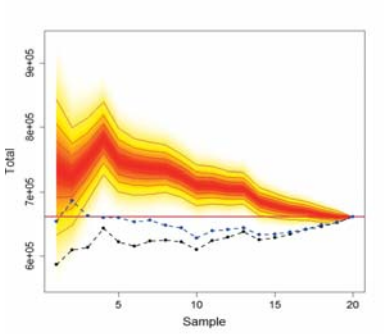


(c) Domain B

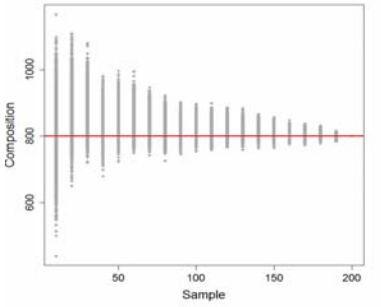
Figure 3: Predictive distributions for totals in terms of the twenty progressive samples. Constant solid lines represent actual totals and dotted lines represent the trajectories of traditional weight-based and naïve estimates (dotted lines).



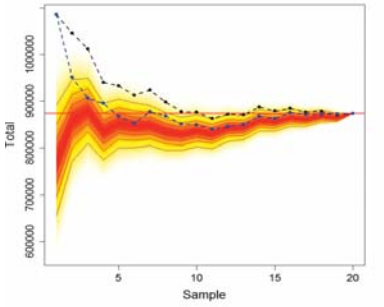
(a) Domain I - Composition



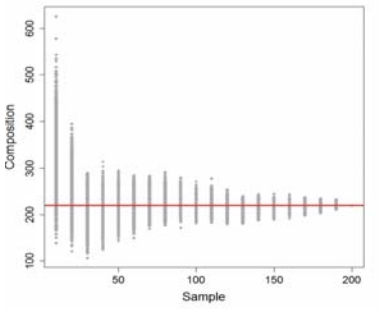
(b) Domain I - Total



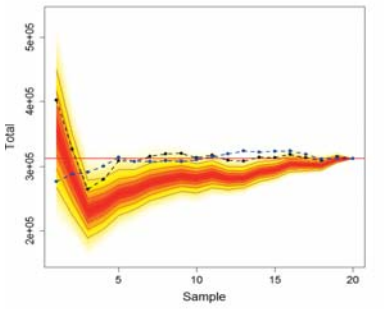
(c) Domain II - Composition



(d) Domain II - Total



(e) Domain III - Composition



(f) Domain III - Total

Figure 4: Predictive distributions for the composition and partial totals of the three unplanned domains in terms of the twenty progressive samples. Constant solid lines represent actual totals, dotted lines represent traditional weight-based, and dotted and solid lines represent naïve estimates.

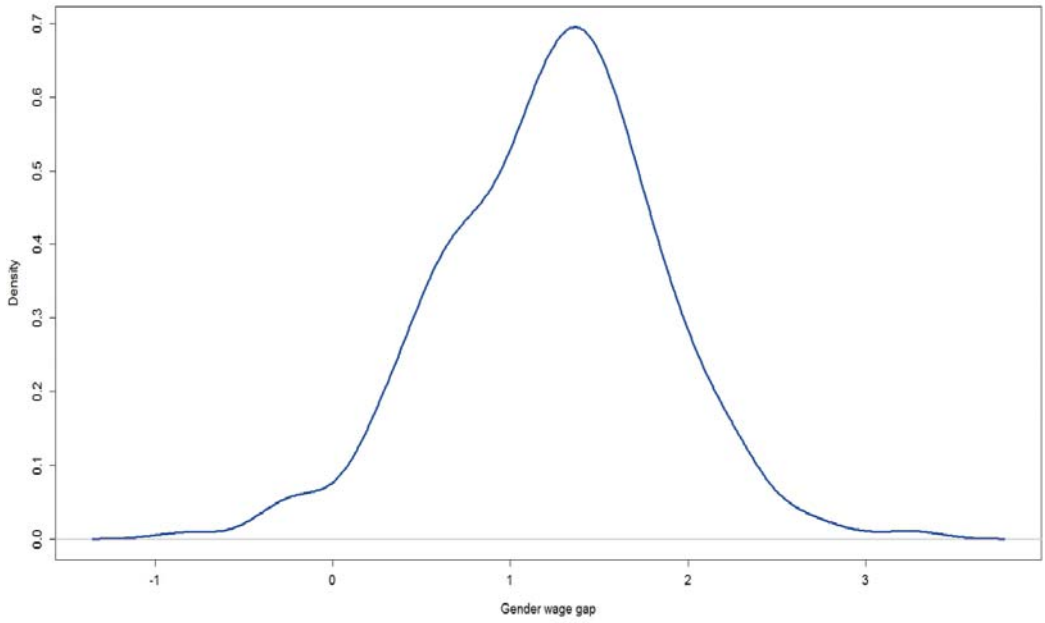


Figure 5: Predictive distribution for the gender wage gap in Mexico.