Bayesian Inference on Totals of Finite Populations based on Planned and Unplanned Domains*

Juan Carlos Martínez-Ovando[1], Sergio I. Olivares-Guzmán[2], Adriana Roldán-Rodríguez[3]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 24

[1] Banco de México, México D. F., México
juan.martinez@banxico.org.mx
[2] Banco de México, México D. F., México
solivares@banxico.org.mx
[3] Banco de México, México D. F., México
aroldan@banxico.org.mx

Abstract

In this paper, we develop an intuitive and flexible model-based framework to make inference on totals of finite populations. Statistical inference is based on the decomposition of the population total into sampled and unsampled parts, disaggregating both of them into planned domains, as well. Inference on the unsampled part is made using Bayesian nonparametric procedures within planned domains. An extension is developed to make inference on totals on unplanned domains, where simultaneous inference on the random composition of individuals across the partition induced by the unplanned domains is produced. As it is shown, both approaches produce consistent and coherent inferences.

Keywords: Finite population inference; Species-sampling models; Prior elicitation; Planned and unplanned domains; Convolution of probability distributions.

*The first author gratefully acknowledges a research stimulus from CONACYT (México). The views expressed in this article are those of the authors and do not necessarily reflect those of Banco de México.

1 Introduction

In this paper we address two inferential problems frequently appeared in finite population studies. In the first part, we develop a flexible model-based framework to inference on totals of finite populations. It is assumed that the characteristic to be observed in each individual is random and continuous. In our formulation, we make use of an intuitive decomposition of the total into two parts: sampled and unsampled. Prediction on the unsampled part is done using Bayesian nonparametric procedures within planned domains; see, *e.g.* [1]. Our development allows to make full inference on the population total through convolution type distributions. It is worth to notice that the derived point estimates resemble and encompass traditional stratified and post-stratified design-based estimators; see, *e.g.* [13].

In the second part, we derive an extension to the previous formulation, in order to make inference on disaggregations of the total induced by unplanned domains. This is an open problem in finite population inference; see, *e.g.* [7] and [9]. For that, we extend our scope of randomness in our formulation by considering the random composition of the partition associated with the unplanned domains. Inference then is made simultaneously on the partial totals and composition of the unplanned domains, through a nested disaggregation of the convolution distribution of the first formulation. Hence, inferences in both formulations are consistent and coherent with each other.

Let us introduce some notations and assumptions. Let $\mathcal{P}$ be the population, which is assumed to be divided into planned domains, $\{\mathcal{P}_{tj}\}$. Here, t stands for a time label and j stands for any other label of relevance. The combinations of t and j define the planned domains. We make emphasis on defining two dimensions on for the planned domains, as it is intended to produce time-series statistics on t. It is also assumed that the number of individuals in each planned domains, N_{tj}, is known.

Accordingly, the total of $\mathcal{P}_t$ can be decomposed as,

$$T_t = \sum_j T_{tj}, \tag{1}$$

where $T_{tj} = \sum_{l=1}^{N_{tj}} Y_{tjl}$, with Y_{tjl} being the characteristic of interest of the l-th individual in $\mathcal{P}_{tj}$. It is also assumed that Y_{tjl} is unknown and random. An additional assumption is that the T_{tj}'s are mutually independent.

Additionally, let $\mathcal{S}_{tj}$ stand for the sampled part of $\mathcal{P}_{tj}$, and let $\widetilde{\mathcal{S}}_{tj}$ be its associated unsampled part. Also, let $N_{tj}^{\mathcal{S}}$ and $N_{tj}^{\widetilde{\mathcal{S}}}$ be their corresponding compositions.

2 Species sampling models

The model-based framework we develop is flexible in that the structural assumptions related to the form of the distribution attained to the $\mathcal{P}_{tjl}$ is being relaxed. For that, we consider Bayesian nonparametric components; in particular, we consider a random distribution function in the class of *species-sampling models (SSM)*; see [10]. SSMs is a flexible class of countable random distribution functions that has received a lot of attention in the recent years; see, *e.g.* [2], [5] and [6]. In our context, we assume that the Y_{tjl}'s are conditionally i.i.d. given F_{tj}, assuming that each F_{tj} belongs to the class of SSMs. Among some of their most relevant properties, the marginalization property makes of SSMs the most appealing alternative of random probability measures in our context.

2.1 Marginalization property

The marginalization property of SSMs has been used extensively as a simulation devise in Bayesian nonparametric procedures. In a general setting, this property 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 vanishes once we incorporate relevant data. This property is highly relevant in our context, as we shall expose it below. See, [5] and [6].

2.2 Prior elicitation

However, the specification of the function parameter G_0 attained to SSMs is highly relevant in our formulation. In order to choose a sensible distribution, we have elicited it by means of comparing three alternative parametric distributions: Lognormal, Gamma and Weibull –other heavy-tailed distributions were considered as well–. Such a comparison is made in terms of the predictive behaviour of the parametric alternatives, in the spirit of [3] and [4]. Hence, the elicited distribution for each planned domain $\mathcal{P}_{tj}$ is the one that best describes the data $\{y_{(t-1)jl}\}$, which gathered on the previous period $(t-1)$.

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 of each planned domain. Accordingly, it is straightforward to derive the following estimates.

3.1 Bayesian estimate of totals

Let U_{tj} be the number of unique values of interest in $\mathcal{S}_{tj}$, *i.e.* $\{y^*_{tjk} : k = 1, \ldots, U_{tj}\}$. Then the posterior estimate of T_{tj} can be written as

$$\widehat{T}_{tj} \quad = \quad \sum_{g \in \mathcal{S}_{tj}} y_{tjg} + N^{\widetilde{\mathcal{S}}}_{tj} \cdot \left[\sum_{k=1}^{U_{tj}} \left(\rho_k(\boldsymbol{m}_{tj}) \cdot y^*_{tjk} \right) + \phi(\boldsymbol{m}_{tj}) \cdot \widehat{\mu_{tj0}} \right], \qquad (2)$$

with $\widehat{\mu_{tj0}} = \mathbb{E}_{\hat{G}_{tj}}\{Y_{tjl}|\widehat{\boldsymbol{\theta}}_{tj0}\}$ and $\widetilde{\mathcal{S}}$ being the complement of $\mathcal{S}$. Point estimates for aggregation of planned domains are naturally derived from (2). Here, (ρ_k) and ϕ define twoo functions, such that $\sum_{k=1}^{U_{tj}} \rho_k(\boldsymbol{m}_{tj}) + \phi(\boldsymbol{m}_{tj}) = 1$.

3.2 Full posterior inference

Moreover, our approach allows to full posterior inferences on T_{tj} through $N^{\widetilde{\mathcal{S}}}_{tj}$-fold convolution distribution induced by $\widehat{G}_{tj}$, the marginal predictive distribution of Y_{tjl}, *i.e.*

$$\mathbb{P}(T_{tj}|\mathcal{S}_{tj}) = \widehat{G}^{*N^{\widetilde{\mathcal{S}}}_{tj}}_{tj} \left(T^{\widetilde{\mathcal{S}}}_{tj} \right), \qquad (3)$$

which is shifted at the sampled part of the total, $T^{\mathcal{S}}_{tj} = \sum_{g \in \mathcal{S}_{tj}} y_{tjg}$, where $T^{\widetilde{\mathcal{S}}}_{tj} = \sum_{l \in \widetilde{\mathcal{S}}_{tj}} Y_{tjl}$. Inference on any aggregation of planned domains are produced through nested convolution procedures. Those predictive distributions are easily handled through simulation techniques; see, *e.g.* [12].

4 Totals on unplanned domains

As with many other survey sampling studies, the information collected in the questionnaire allows to ask interesting questions about further disaggregation of the totals to be estimated. In the design-based argot, those disaggregations are referred as unplanned domains.

A key consideration regarding unplanned domains is that no previous reference is known about their composition, and that composition is actually being regarded as random, see [9]. However, by using the decomposition of totals we worked before, it is possible to rewrite T_{tj} as the sum of partial totals on those unplanned domains, as

$$T_{tj} \quad = \quad T^{\mathcal{D}_1}_{tj} + \ldots + T^{\mathcal{D}_D}_{tj},$$

where $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_D\}$ define a partition of the population in D unplanned domains. Accordingly, the population total in domain $\mathcal{P}_{tj}$ can be decomposed and estimated by parts.

It is worth to notice that the relevant characteristic when working with unplanned domains, is the composition of the population $\mathcal{P}_{tj}$ between the unplanned domains, $\mathcal{D}$. Particularly, the composition over the unsampled portion, $\widetilde{\mathcal{S}}_{tj}$. That composition is being denoted by $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}} = (N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}, \ldots, N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D})$. Thus, inference on unplanned domains requires to extend our scope of uncertainty to include the unknown composition of the population in our formulation.

4.1 Prior on the composition of unplanned domains

We think of the composition vector $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ as a realization of a multinomial-Dirichlet distribution, with parameters $N_{tj}^{\mathcal{S}}$ (known) and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$, such that $\alpha_d > 0$, for $d = 1, \ldots, D$. The vector with the proportions between $\mathcal{D}$, denoted by $\boldsymbol{p}_{tj}^{\mathcal{D}}$, is treated as a latent variable. Hence, posterior inference and predictions are analytically produced using the conjugacy property of the multinomial-Dirichlet model.

As it has been described above, elicitation of the vector parameter $\boldsymbol{\alpha}$ is done using the information collected in the sample of the period $(t-1)$ for the samples sub-planned domain j.

4.2 Posterior estimates on unplanned domains

Posterior estimation of totals on unplanned domains is made in two steps. In the first step, point estimates for the composition vector $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ is produced as integer part of $N_{tmar}^{\mathcal{S}} \cdot \alpha_d + N_{tmar}^{\mathcal{S}\cap\mathcal{D}_d} / (\sum_{i=1}^{D} (\alpha_i + N_{tmar}^{\mathcal{S}\cap\mathcal{D}_i}))$. Consequently, posterior estimates of the composition of $\mathcal{P}_{tj}$ for each unplanned domain $\mathcal{D}_d$ is given by $\widehat{N_{tj}^{\mathcal{D}_d}} = N_{tj}^{\mathcal{S}\cap\mathcal{D}_d} + \widehat{N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}}$, where $N_{tj}^{\mathcal{S}\cap\mathcal{D}_d}$ is the composition of the $\mathcal{D}_d$ in the observed sample $\mathcal{S}_{tj}$.

Consistently with the above decomposition, point estimates of partial totals $T_{tj}^{\mathcal{D}_d}$'s, are being produced as

$$\widehat{T}_{tj}^{\mathcal{D}_d} = \sum_{g \in \mathcal{S}_{tj}\cap\mathcal{D}_d} y_{tjg} + \widehat{N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}} \cdot \left[\sum_{k=1}^{U_{tj}} \left(\rho_k(\boldsymbol{m}_{tj}) \cdot y_{tjk}^{*} \right) + \phi(\boldsymbol{m}_{tj}) \cdot \widehat{\mu_{tj0}} \right], \qquad (4)$$

with $\widehat{\mu_{tj0}}$ being defined as above.

4.3 Inference through a vector of convolutions

Full posterior inference requires to make inference on both vectors $\boldsymbol{T}_{tj}^{\mathcal{D}}$ and $\boldsymbol{N}_{tj}^{\mathcal{D}}$ through out their joint predictive distribution,

$$\mathbb{P}\{\boldsymbol{T}_{tj}^{\mathcal{D}}, \boldsymbol{N}_{tj}^{\mathcal{D}} | \mathcal{S}_{tj}\} \;=\; \mathbb{P}\{\boldsymbol{T}_{tj}^{\mathcal{D}} | \boldsymbol{N}_{tj}^{\mathcal{D}}, \mathcal{S}_{tj}\} \times \mathbb{P}\{\boldsymbol{N}_{tj}^{\mathcal{D}} | \mathcal{S}_{tj}\}. \qquad (5)$$

On the one hand, $\mathbb{P}\{\boldsymbol{N}_{tj}^{\mathcal{D}} | \mathcal{S}_{tj}\}$ is completely determined by the predictive distribution of the multinomial-Dirichlet conjugate model for the unsampled

part of the composition, $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$, described above. Thus, $\mathbb{P}\{\boldsymbol{N}_{tj}^{\mathcal{D}}|\mathcal{S}_{tj}\}$ is computed by shifting the predictive distribution of $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$ at to the sampled part of the composition, $\boldsymbol{N}_{tj}^{\mathcal{S}\cap\mathcal{D}}$.

On the other hand, $\mathbb{P}\{\boldsymbol{T}_{tj}^{\mathcal{D}}|\boldsymbol{N}_{tj}^{\mathcal{D}},\mathcal{S}_{tj}\}$, is being derived as a D-dimensional vector of $\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}$-fold convolutions,

$$\mathbb{P}\{\boldsymbol{T}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}}|\boldsymbol{N}_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}},\mathcal{S}_{tj}\}=\left(\widehat{G}^{*^{N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}}}(T_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_1}),\ldots,\widehat{G}^{*^{N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D}}}(T_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_D})\right), \quad (6)$$

shifted at the vector of sampled partial totals $\boldsymbol{T}_{tj}^{\mathcal{S}\cap\mathcal{D}}=(T_{tj}^{\mathcal{S}\cap\mathcal{D}_1},\ldots,T_{tj}^{\mathcal{S}\cap\mathcal{D}_D})$, and $T_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}=\sum_{l\in\widetilde{\mathcal{S}}_{tj}}^{N_{tj}^{\widetilde{\mathcal{S}}\cap\mathcal{D}_d}}Y_{tjl}$.

As before, the predictive distribution of any aggregation across sampled and unsampled parts is being defined through nested convolution procedures. Derived inferences are being consistently defined across the planned domains. Those distributions are easily recovered through simulation techniques.

5 Discussion

In this paper we develop an intuitive and appealing framework to make inference on totals of finite population based on individual continuous measurements. A key distinction of our framework with regards to traditional design-based alternatives, is that the characteristic of interest to be observed in each individual is assumed to be random. This assumption allows us to make full and interpretable posterior inferences on totals using convolution-type distributions. Predictive distributions are easily recovered via simulation techniques.

Another distinctive contribution of this paper consists in the formulation of a procedure to make inference on totals associated with unplanned domains, as well. In our formulation, uncertainty is spanned by considering the random composition associated with the unplanned domains. Thus, posterior inference is jointly made for the disaggregated totals and the intrinsic random composition attained to the unplanned domains, simultaneously. And such an inference is consistent with the aggregated inference based on planned domains solely. To the best of our knowledge, this is the first approach achieving this task.

A library in the `R` language [11] producing inferences on totals based for different specification of SSMs has been developed.

References

[1] D. A. Binder. **Non-Parametric Bayesian Models for Samples From Finite Populations**. *Journal of the Royal Statistical Society, Series B*; 1982; 44; pp. 388-393.

[2] S. Favaro, A. Lijoi, R. Mena and I. Prünster. **On Some Issues Related to Species Sampling Problems**. *Proceedings of the 7th Conference on Statistical Computing and Complex Systems - SCo 2011*, Padova; 2011; pp. 9pp (electronic).

[3] A. E. Gelfand, D. K. Dey and H. Chang. **Model Determination using Predictive Distributions with Implementation via Sampling-Based Methods**. *Bayesian Statistics 5, Oxford University Press*, Oxford; 1992; pp. 147-167.

[4] A. Gelman, X. L. Meng and H. Stern. **Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with Discussion)**. *Statistica Sinica*; 1996; 6; pp. 733-807.

[5] B. Hansen and J. Pitman. **Prediction Rules for Exchangeable Sequences Related to Species Sampling**. *Statistics & Probability Letters*; 2000; 46(3); pp. 251-256.

[6] J. Lee, F. A. Quintana, P. Müller and L. Trippa. **Defining Predictive Probability Functions for Species Sampling Models**. *Statistical Science*; 2012; to appear.

[7] R. Lehtonen and A. Veijanen. **Design-based Methods of Estimation for Domains and Small Areas**. *Handbook of Statistics Vol. 29B. Sample Surveys: Inference and Analysis* Elsevier, Amsterdam; 2009; pp. 219-249.

[8] A. Lijoi and I. Prünster. **Models Beyond the Dirichlet Process**. *Bayesian Nonparametrics, Cambridge University Press*, Cambridge; 2010; pp. 80-130.

[9] G. Meeden. **A noninformative Bayesian approach to domain estimation**. *J. Statistical Planning & Inference*; 2005; 129(1-2); pp. 85-92.

[10] J. Pitman. **Exchangeable and Partially Exchangeable Random Partitions**. *Probability Theory and Related Fields*, Hayward, CA; 1995; 102(2); pp. 145-158.

[11] R Core Team. **R: A Language and Environment for Statistical Computing, Reference Index**. *R Foundation for Statistical Computing, Reference Index version 2.15.2*; 2012; http://www.R-project.org.

[12] C. P. Robert and G. Casella. **Monte Carlo Statistical Methods**. *Springer*, New York; 2004.

[13] M. E. Thompson. **Theory of Sample Surveys**. *Monographs on Statistics and Applied Probability, Chapman & Hall*, London; 1997; 74.