STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 23–29, DOI 10.21307/stattrans-2020-023
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Julie Gershunskaya[1]

1. Introduction

I would like to begin by congratulating Professor Ghosh for his many contributions to small area estimation, both as an original researcher and effective communicator of complex ideas. The current paper provides a lucid overview of the history and developments in small area estimation (SAE) and offers a synopsis of some of the most recent innovations. As is well illustrated in the paper, the development of the field is driven by real-world demands and problems emerging in actual applications. Let us ponder on this practical side of the SAE methodology that, by offering a set of tools and concepts, provides an *engineering framework* for present day official statistics.

From the very beginning of large-scale sample surveys in the official statistics, there was the realization that the survey practice should be based on both theoretical developments and clear practical strategy. Morris Hansen (1987) applied the term "*total survey design*" to describe the fusion of theory and operational planning, a paradigm used from the early days of sampling surveys at the U.S. Bureau of Census. In a similar spirit, P. C. Mahalanobis (1946) characterized the whole complex of activities involved in the managing of large-scale sample surveys in the Indian Statistical Institute by calling it "*statistical engineering*".

Traditionally, a great deal of theory, experimentation, and practical considerations are focused on the design stage of sample surveys. Yet, no matter how well the survey is designed, there is a growing demand in extracting ever more information from already collected data. Even more, in many present day surveys, the required "unplanned" domains number in thousands. In such an environment, the production of small domain estimates becomes a substantial part of a large-scale enterprise. Developments in the SAE field address the demands by providing survey practitioners with necessary gear, whereas an applied statistician acts as *engineer* that employs a variety of available tools and creates an appropriate operational plan.

[1] U. S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212, USA.
E-mail: gershunskaya.julie@bls.gov. ORCID: https://orcid.org/0000-0002-0096-186X.

2. Model building considerations

To illustrate some aspects of the planning and model development for estimation in small domains, I will describe, in broad strokes, considerations involved in the model choice for the U.S. Bureau of Labor Statistics' Current Employment Statistics (CES) survey. The specific context that affects approaches to small domain modeling in CES includes:

- the tight production timeline, where estimates are produced monthly within only a few weeks after the data collection;
- the demand for estimates over a large number of small areas. Monthly estimates are published for about 10 thousands domains defined by intersections of detailed industry and geography. Of those, roughly 40 percent of domains have sufficient sample, so that direct sample-based estimates are deemed reliable for the use in publication; the other domains may have only a handful of sample units and require modeling;
- the dynamic and heterogeneous nature of the population of business establishments, a feature that could generally manifest itself – thus affecting the model fit – in two ways: 1. in the form of a frequent appearance of sample-influential observations or; 2. as irregularities in the signal for groups of domains.

Because of the above characteristics of the CES survey process, essential requirements for any model considered in CES are (i) computational scalability, (ii) flexibility of modeling assumptions, and (iii) robustness to model outliers. To demonstrate how the above aspects are taken into account, we examine three models.

Our baseline model M0 is the classical Fay-Herriot area level model. In the Bayesian formulation, using the notation of Professor Ghosh's paper, the *sampling model* for domain $i = 1\ldots, m$ is

$$y_i \mid \theta_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \tag{1}$$

and the *linking model* is

$$\theta_i \mid \mathbf{b} \overset{ind}{\sim} N\left(x_i^T \mathbf{b}, A\right). \tag{2}$$

The parsimonious structure and the ease of implementation of the FH model make it particularly appealing under the tight CES production schedule. The posterior mean in the form of the weighted average of direct sample based and synthetic estimators has clear intuitive interpretation, thus facilitating communication of the reasoning to a wider, less quantitatively oriented, community.

However, the dynamic nature of the population of business establishments affects the FH model fit and reduces the attractiveness of the model in two important respects:

1) On the one hand, sampling model (1) is not robust to extreme y_i values. Noisy direct estimates y_i could result from the appearance of influential observations in the sample data. In the ideal world, the additional variability induced by noisy sample data would be reflected in larger values of respective variances D_i's, that are assumed to be known. If that would be the case, larger D_i's would lessen the influence of noisy y_i's on the model fit. In practice, however, true variances are not known, and the usual method is to plug in values based on a generalized variance function (GVF). Such plug-in D_i's may not properly reflect the amount of noise in respective y_i's.
2) On the other hand, the linking model (2) normality assumption may fail, for example, when groups of domains form clusters or when some domains deviate from the linearity assumption $x_i^T\mathbf{b}$. This is especially likely to happen when a large number of domains is included in the same model.

In model M1, we address the concern regarding the non-robustness of sampling model (1). Here, sample-based estimates $\hat{D}_i$ of variances D_i are treated as data and modeled jointly with y_i's. The joint modeling approach was considered by Arora and Lahiri (1997), You and Chapman (2006), Dass et al. (2012), Liu et al. (2014), among others. Model M1 is related to the model proposed by Maiti et al. (2014) who used the EM algorithm for estimation of the model parameters within the empirical Bayes paradigm. The Bayesian extension of the model was developed by Sugasawa et al. (2017). Assume in domain i, $i = 1 \ldots, m$, the following model M1 holds for pair $\left(y_i, \hat{D}_i\right)$:

$$y_i \mid \theta_i, D_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \qquad \theta_i \mid \mathbf{b}, A \overset{ind}{\sim} N\left(x_i^T\mathbf{b}, A\right), \tag{3}$$

$$\hat{D}_i \mid D_i \overset{ind}{\sim} G\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2D_i}\right), \qquad D_i \mid \gamma \overset{ind}{\sim} IG\left(a_i, c_i\gamma\right), \tag{4}$$

where (3) is the usual FH model for the point estimate and (4) describes a companion model for observed variance $\hat{D}_i$ (here, direct sample-based estimates of variances are termed "observed variances" in the model input context); $G(\cdot)$ and $IG(\cdot)$ denote the gamma and inverse gamma distributions, respectively; γ is an unknown parameter; a_i and c_i are positive known constants, Sugasawa et al. (2017) suggested the choice of $a_i = 2$ and $c_i = n_i^{-1}$, n_i is the number of respondents in domain i.

Although model M1 mitigates the effect caused by noisy direct sample estimates, it still ignores the problem of possible deviations from the normality assumption in linking model (2). When there is a large number of domains, we can more fully explore the underlying structure and relax the assumption of linking model (2) by replacing the normality with a finite mixture of normal distributions. Model M2, proposed by Gershunskaya and Savitsky (2020), is given by (5) and (6):

$$y_i \mid \theta_i, D_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \qquad \theta_i \mid \boldsymbol{\pi}, \mathbf{b}_0, \mathbf{b}, A \overset{ind}{\sim} \sum\nolimits_{k=1}^{K} \pi_k N\left(b_{0k} + \tilde{\mathbf{x}}_i^T \mathbf{b}, A\right), \qquad (5)$$

$$\hat{D}_i \mid D_i \overset{ind}{\sim} G\left(\frac{sn_i}{2}, \frac{sn_i}{2D_i}\right), \qquad D_i \mid \boldsymbol{\gamma}, \boldsymbol{\pi} \overset{ind}{\sim} \sum\nolimits_{k=1}^{K} \pi_k IG\left(2, \exp\left(z_i^T \boldsymbol{\gamma}_k\right)\right). \qquad (6)$$

In this model, we assume the existence of K latent clusters having cluster-specific intercepts b_{0k}, $k = 1,\dots,K$, and common variance A; in addition, we relax the inverse gamma assumption of (4) by specifying a mixture of the inverse gamma distributions with the cluster-specific coefficient vectors $\boldsymbol{\gamma}_k$; z_i is a vector of covariates for the variance model for area i; s is a model parameter that regulates the shape and scale of the gamma distribution, it depends on the quality of variance estimates.

The Stan modeling language and the Variational Bayes algorithm within Stan proved to be effective in fitting the above models.

3. Model selection and evaluation plan

Due to the tight CES production schedule, a *production* model has to be chosen in advance, before a statistician obtains the actual data. Models for CES are pre-selected and pre-evaluated based on a comparison to historical employment series derived from the universe of data that is available from an administrative source, known as the Quarterly Census of Employment and Wages (QCEW) program. These data become available to BLS on a quarterly basis with the time lag of 6 to 9 months after the reference date and are considered a "gold standard" for CES. After an evaluation based on several years of data, that include periods of economic growths and downturns, the best model from a set of candidates would be accepted for the use in production.

Thus, the availability of a "gold standard" defines the CES strategy for the model development and evaluation. This approach differs from the usual model selection and checking methods used in statistics, yet it is common for government agencies.

4. Real-time analysis protocol

The quality of the production model is regularly re-assessed based on newly available data from QCEW. This kind of evaluation can be performed only post hoc, several months after the publication of CES estimates. While the "gold standard" based approach of model selection and evaluation works well overall and provides reassurance and the perception of objectivity of the chosen model, the following question remains: Suppose a particular model (say, model M2) is accepted for the production based on its historical performance; however, what if in a given month during the production such history-based best model would fit poorly for some of the domains? To diagnose possible problems in the real production time, analysts have to be equipped with formal tests and graphical tools allowing the efficient detection of potential problems, and with the guidelines for ways to proceed whenever problems arise.

One example of a tool for the routine diagnostics of outlying cases is given by the model-based domain screening procedure proposed by Gershunskaya and Savitsky (2020). The idea for this procedure is to flag the domains whose direct estimates y_i's have low probability of following the *posterior predictive distribution* obtained based on the model. The list of "suspect" domains is sent to analysts for checking; analysts review the list and decide if the reason for a given extreme direct estimate is one of the following: (i) the deficiency of the domain sample or (ii) a failure of modeling assumptions. In general, if the domain sample size is small, the outlyingness of the direct sample estimate would likely be attributed to the deficiency of the sample; in such a case, analysts would decide to rely on the model estimate for this domain. For domains with larger samples, the direct estimates may be deemed more reliable than the model-based estimates. In addition, to these general considerations, analysts would also have the ability to check the responses in the suspect domains to determine if there are any erroneous reports overlooked at the editing stage. Such reports would have to be corrected or removed from the sample. Analysts may also possess the knowledge of additional facts that may guide their decision, such as, information about the economic events not reflected in the modeling assumptions or, conversely, in the available sample.

5. Summary

The growing demand for estimates in "unplanned" domains instigated development of the SAE methods. Theoretical advances in SAE over past five decades, along with the proliferation of powerful computers and software, invited even more, ever increasing demand in estimates for small areas. Contemporary small area estimation becomes a *large-scale* undertaking. The present day *statistical engineers* require development of tools – as well as philosophy and guidelines – for the *quality control* in the *production environment* to help ensure estimates in small domains are reliable and impartial.

Acknowledgement

The views expressed here are those of the discussant and do not necessarily constitute the policies of the U.S. Bureau of Labor Statistics.

REFERENCES

ARORA, V., LAHIRI, P., (1997). On the superiority of the Bayesian methods over the BLUP in small area estimation problems. *Statistica Sinica* 7, pp. 1053–1063.

BUREAU OF LABOR STATISTICS, (2004). Employment, Hours, and Earnings from the Establishment Survey, BLS Handbook of Methods, Washington, DC: US Department of Labor.

DASS, S. C., MAITI, T., REN, H., SINHA, S., (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology,* 38, pp. 173–187.

GERSHUNSKAYA, J., SAVITSKY, T. D., (2020) Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, Vol. 8, Issue 2, pp. 181−205, https://doi.org/10.1093/jssam/smz004.

HANSEN, M. H., (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, Vol. 2, No. 2, pp. 180−190.

LIU, B., LAHIRI, P., KALTON, G., (2014). Hierarchical Bayes modelling of survey-weighted small area proportions. *Survey Methodology*, Vol. 40, No. 1, pp. 1−13.

MAHALANOBIS, P. C., (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, pp. 325−378.

MAITI, T., H. REN, A. SINHA, (2014). Prediction Error of Small Area Predictors Shrinking Both Means and Variances. *Scandinavian Journal of Statistics*, 41, pp. 775–790.

STAN DEVELOPMENT TEAM, (2017). Stan modeling Language User's Guide and Reference Manual, Version 2.17.0 [Computer Software Manual], available at http://mc-stan.org/. Accessed February 28, 2019.

SUGASAWA, S., TAMAE, H., KUBOKAWA, T., (2017). Bayesian Estimators for Small Area Models Shrinking Both Means and Variances. *Scandinavian Journal of Statistics*, 44, pp. 150–167.

YOU, Y., CHAPMAN, B., (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology,* 32, pp. 97–103.