STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 40–44, DOI 10.21307/stattrans-2020-026
Received – 31.01.2020; accepted – 30.06.2020

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

Isabel Molina[1]

Extending on poverty mapping methods

The paper gives a nice overview of small area estimation, putting emphasis on important applications that have led to notable methodological contributions to the field. I would like to extend further on one of the important applications of unit level models that is mentioned in the paper, which is the estimation of poverty or inequality indicators in small areas. The characteristic of this application that makes it particular is that many of these indicators are defined as much more complex functions of the values of the target variable in the area units than simple means or totals.

The traditional method used by the Word Bank, due to Elbers, Lanjouw and Lanjouw (2003 – ELL), was designed to estimate general small area indicators (and perhaps several of them together), defined in terms of a welfare measure for the area units (i.e. households) with a single unit level model for the welfare variable. The model is traditionally a nested error model similar to that of Battesse et al. (1988), for the log of the welfare variable in the population units. This model is fit to the survey data, and the resulting model parameter estimates are then used to generate multiple censuses based on census auxiliary information. With each census, indicators are calculated for each area, and averages across the censuses are taken as ELL estimators. Similarly, variances across the indicators from the different censuses are taken as ELL noise measures of the estimators.

When estimating simple area means with a model for the welfare variable without transformation, the final averaging makes the area effect vanish (it has zero expectation), making ELL estimators essentially synthetic. In fact, ELL method seems to be inspired by the literature on multiple imputation rather than by the small area estimation literature.

[1] Department of Statistics, Universidad Carlos III de Madrid, Spain. E-mail: isabel.molina@uc3m.es. ORCID: https://orcid.org/0000-0002-4424-9540.

Molina and Rao (2010 - MR) proposed to consider empirical best/Bayes (EB) estimators of general small area indicators based on a similar nested error model as in ELL method. The only difference in the model was that, in the traditional applications of ELL method, the random effects were for the clusters of the sampling design (i.e. primary sampling units), which are generally nested in the small areas of interest (e.g., census tracks). In the EB procedure by MR, as in typical small area applications with unit level models, the random effects in the nested error model are for the areas of interest. Considering the clusters as the small areas of interest for more fair comparisons, MR showed substantial gains of EB estimators with respect to ELL ones in a (limited) simulation experiment. In fact, EB estimators are optimal in the sense of minimizing the mean squared error (MSE) under the assumed model and hence cannot be worse than ELL estimators under the same model assumptions. The main reason for the large gains in efficiency is that the EB estimator is theoretically (i.e., under completely known model) defined as the conditional expectation of the indicator given the survey welfares, whereas ELL estimator is theoretically defined as the unconditional expectation which does not make use of the precious information on the actual welfare variable, coming from the survey.

The MSE of the EB estimators in MR (2010) was estimated using the parametric bootstrap approach for finite populations of González-Manteiga et al. (2008), which can be computationally very intensive for large populations and very complex indicators. Molina, Nandram and Rao (2014) proposed a hierarchical Bayes (HB) alternative that avoids performing a bootstrap procedure for MSE estimation, since posterior variances are obtained directly from the predictive distribution of the indicators of interest. They use a reparameterization of the nested error model in terms of the intraclass correlation coefficient, which allows to draw directly from the posterior using the chain rule of probability, avoiding MCMC methods.

Ferretti and Molina (2011) introduced a fast EB approach for the case when the target area parameter is computationally very complex, such as when the indicators are based on pairwise comparisons or sorting area elements, or when the population is too large. Faster HB approaches can be implemented similarly.

Marhuenda et al. (2017) extended the EB procedure for estimation of general parameters to the twofold nested error model with area and (nested) subarea effects, considered in Stukel and Rao (1999) for the case of linear parameters. They obtained clear losses in efficiency when the random effects are specified for the subareas (e.g. clusters) but estimation is desired for areas, except for the case when the areas of interest are not sampled. In this case, they recommend the inclusion of both area and subarea random effects.

Another subtle difference between the traditional ELL approach and the EB method of MR lies in the fact that the original EB method requires to link the survey and census units, because the expectation defining the EB estimator is with respect to the distribution of the non-sample welfares given the sample ones. The Census EB estimator (Molina, 2019) is a slight variation of the original EB estimator based on the nested error model, which does not require linking the survey and census data sets, similarly as ELL procedure. Molina (2019) presents a slight variation of the parametric bootstrap procedure of González-Mateiga et al. (2008) for estimation of the MSE of the Census EB estimator that avoids linking the survey and census data sets.

The Word Bank revised their methodology in 2014 introducing a new bootstrap procedure intended to obtain EB predictors according to Van der Weide (2014), but this procedure is not leading to the original EB (or Census EB) predictors. They also incorporated heteroscedasticity and survey weights, to account for complex sampling designs. They include the survey weights in the estimates of the regression coefficients and variance components according to Huang and Hidiroglou (2003), and also in the predicted area effects following You and Rao (2002). Recently, Corral, Molina and Nguyen (2020) show that the resulting bootstrap procedure leads to substantially biased small area estimators. They also show that MSEs are not correctly estimated with this approach. This has lead to a very recent revision of the World Bank methodology and software, incorporating now the original Census EB estimators and the parametric bootstrap procedure of González-Manteiga et al. (2008), adapted for the case when the survey and census data cannot be linked. The new estimators account for heteroscedasticity and include also survey weights in the model parameter estimators and in the predicted area effects similarly as in Van der Weide (2014). The implemented estimators are the Census versions of the pseudo EB estimators of Guarrama, Molina and Rao (2018) designed to reduce the bias due to complex sampling designs, accounting for heteroscedasticity and using estimates of the variance components that include the survey weights as well.

In small area estimation of welfare-related indicators, another important issue is the transformation taken to the welfare variable in the model. Since welfare variables are most often severely right-skewed and may show heteroscedasticity, log transformation is customarily taken in the nested error model. For the special parameters of area means of the original variables, Molina and Martín (2018) studied the analytical EB predictors under the model with log transformation and obtained second-order correct MSE estimators.

In fact, the EB method of MR for the estimation of general indicators requires normality of area effects and unit level errors, so care should be taken with the transformation taken in order to achieve at least approximate normality. Popular

families of transformations are the power or Box-Cox families. The appropriate member of these families may be selected beyond log in the implemented function for EB method ebBHF() from the R package sae (Molina and Marhuenda, 2015). In fact, in the presence of very small values of the welfare variable, the log transformation shifts these small values towards minus infinity, which may produce now a thin yet long tail in the distribution. A simple way of avoiding such effect is just adding a shift to the welfare variable before taking log. A drawback is that selection of this shift, as well as selection of the Box-Cox or power transformation, needs to be based on the actual survey data. A different approach is to consider a skewed distribution for welfare. Diallo and Rao (2018) extended the EB procedure to the skew normal distribution and Graf, Martín and Molina (2019) considered the EB procedure under a generalized beta of the second kind (GB2). This distribution contains four parameters, one for each tail, offering a more flexible framework for modeling skewed data of different shapes.

Acknowledgement

This work was supported by the Spanish grants MTM2015-69638-R and MTM2015-64842-P from Ministerio de Economía y Competitividad.

REFERENCES

BATTESE, G. E., HARTER, R. M., FULLER, W. A., (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association*, 83, pp. 28–36.

CORRAL, P., MOLINA, I., NGUYEN, M., (2020). Pull your small area estimates up by the bootstraps. World Bank Policy Research Working Paper 9256.

DIALLO, M., RAO, J., (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, pp. 1092–1116.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, pp. 355–364.

FERRETTI, C., MOLINA, I., (2012). Fast EB Method for Estimating Complex Poverty Indicators in Large Populations. *Journal of the Indian Society of Agricultural Statistics*, 66, pp. 105–120.

GONZÁLEZ -MANTEIGA, W., LOMBARDÍA,M.J.,MOLINA,I.,MORALES,D.,SANTAMARÍA,L., (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78, pp. 443–462.

GRAF, M., MARÍN, J. M., MOLINA, I., (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 28, pp. 565–597.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis*, 121, pp. 20–40.

HUANG, R., HIDIROGLOU, M., (2003). Design consistent estimators for a mixed linear model on survey data. Proceedings of the Survey Research Methods Section, American Statistical Association (2003), pp. 1897–1904.

MARHUENDA, Y., MOLINA, I., MORALES, D., RAO, J. N. K., (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, 180, pp. 1111–1136.

MOLINA, I., (2019). Desagregación de datos enencuestas de hogares: metodologías de estimación en áreas pequeñas. Seriesdela Comisión Económicapara América Latinayel Caribe (CEPAL), Estudios Estadísticos LC/TS.2018/82/Rev.1, CEPAL.

MOLINA, I., MARHUENDA, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7, pp. 81–98.

MOLINA, I., NANDRAM, B., RAO, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8, pp. 852–885.

MOLINA, I., RAO, J. N. K., (2010). Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics*, 38, pp. 369–385.

STUKEL, D., RAO, J. N. K., (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, 78, pp. 131–147.

VAN DER WEIDE, R., (2014). Gls estimation and empirical Bayes prediction for linear mixed models with heteroskedasticity and sampling weights: a background study for the POVMAP project. World Bank Policy Research Working Paper 7028.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, pp. 431–439.