STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 59–67, DOI 10.21307/stattrans-2020-030
Received – 31.01.2020; accepted – 30.06.2020

Rejoinder

Malay Ghosh[1]

I thank all the seven discussants for taking time to read the paper, and for their kind and valuable comments. In particular, they introduced some important current and potentially useful future topics of research, thus supplementing nicely the material covered in this article.

With the current exponential growth in the small area estimation (SAE) literature, I realized the near impossibility of writing a comprehensive review of the subject. Instead, I took the easier approach of tracing some of its early history, and bringing in only a few of the current day research topics, and that too reflecting my own familiarity and interest. I listed a number of uncovered topics in this paper, far outnumbering those that are covered. I am very glad to find that some of these topics are included in the discussion, in varied details.

I will reply to each discussant individually. Professor Molina and Dr. Newhouse have both discussed small area poverty indication, with some overlapping material. I will first discuss them jointly, and then individually on the distinct aspects of their discussion.

Gershunskaya

I thank Dr. Gershunskyaya for highlighting some of the potential problems that one may encounter in small area estimation. Yes, the assumption of known variances D_i, when indeed they are sample estimates, is a cause of concern. Joint modeling of $(y_i, \hat{D}_i)$, when possible, must be undertaken. Unfortunately, without the availability of microdata, especially for secondary users of surveys, modeling the $\hat{D}_i$ can be quite ad hoc, often resulting in very poor estimates. People in Federal Agencies, for example those in the BLS, US Census Bureau and others do have access to the microdata, which can facilitate their modeling. However, even then the issue may not always be completely resolved. I like the hierarchical Bayesian model of Dr. Gershunskaya, something similar to what I have used before. But I have always been concerned about the choice of hyperparameters. For example, in the inverse gamma hyperprior IG$(a_i, c_i\gamma)$, the choice of a_i and c_i can influence the inference considerably, and this demands sensitivity analysis. I wonder whether there is any real global justification of the choice $a_i = 2$ and $c_i = n_i^{-1}$ as proposed in Sugasawa et al. (2017). Added to this is modeling of the parameter γ, which enhances complexity.

[1]Department of Statistics, University of Florida, Gainesville, FL. USA. E-mail: ghoshm@stat.ufl.edu. ORCID: https://orcid.org/0000-0002-8776-7713

Following the same notations of Dr. Gershunskaya, another option may be to use a default half-Cauchy prior (Gelman, 2006) for $D_i^{1/2}$. This results in the prior $\pi(D_i) \propto D_i^{-1/2}(1+D_i)^{-1}$, the so-called "Horseshoe", which enjoys global popularity in these days. It may be noted though that the above prior is just a special case of a Type II beta prior $\pi(D_i) \propto D_i^{a-1}(1+D_i)^{-a-b}$ with $a=b=1/2$. In my own experience, even in the context of SAE research, the choice $a=b=1/2$ is not always the best choice. Other (a,b) combinations produce much better results.

I very much echo the sentiment of Dr. Gershunskaya that reliable estimates for thousands of small domains within a very narrow time frame is a real challenge for most Federal Agencies. With the present COVID-19 outbreak, the BLS is producing steady unemployment numbers for all the States in the US. In situations demanding a very urgent answer, I am quite in favor of a very pragmatic approach, for example, an empirical Bayes approach where one just uses estimates of the hyperparameters. Alternative frequentist approaches such as the jackknife and the bootstrap for mean squared error (MSE) estimation are equally welcome.

Dr. Gershunskaya has highlighted the importance of "external evaluation" of Current Employment Survey (CES) estimates, which I value as extremely important. However, is a six to nine month time lag on the availability of Quarterly Census of Employment and Wages (QCEW) seems a little too much for an ongoing survey like CES. Presumably, different QCEW data are used for production and evaluation. Otherwise, one is faced with the same old criticism of double use of the same data.

I agree wholeheartedly with Dr. Gershunskaya on the issue of robustness of models, and replacing the normal prior by mixtures of normals. In this article, I have mentioned the use of continuous "global-local shrinkage" priors which essentially attain the same goal and are easier for implementation.

Finaly, I thank Dr. Gershunskaya for bringing into our attention that the term "statistical engineering" was used by the late P.C. Mahalanobis, the founding father of statistics in India, back even in 1946 !

Han

I thank Dr. Han for her discussion of the current day research on probabilistic record linkage. While the theoretical framework of record linkage goes back to Fellegi and Sunter (1969), it seems that there was a long fallow period of research up until recent times. Indeed, in my opinion, research on record linkage has taken a giant leap in the last few years, mostly for catering to the needs of Federal Agencies, but its importance has been recognized by the industrial sectors as well.

While record linkage requires merging of two or more sources of data, often it is impossible to find a unique error-free identifier, for example, when there is an error in recording a person's Social Security Number. This necessitates the need for probabilistic record linkage.

While small area estimation seems to be a natural candidate for application of record linkage in merging survey and administrative data, research in this topic has taken off only very recently. I think that the major reason behind this is the formidable challenge of trustworthy implementation.

Let me elaborate this point a bit. It is universally recognized that small area estimators are model-based estimators. But as pointed out by Dr. Han, now one needs an integrated model based on three components: (1) a unit level SAE model, (2) a linkage error model and (3) a two-class mixture model on comparison vectors. Now, instead of model diagnostics for one single SAE model, one needs model diagnostics for all three models in order to have reliable SAE estimates. In my mind, this seems to be a formidable task. Nevertheless, I encourage Dr. Han and her advisor Partha Lahiri to pursue research in this very important area, and I am very hopeful that their joint venture will become a valuable resource for both researchers and practitioners.

I have some query regarding the assumptions (1)-(3) of Dr. Han. Can one always avoid duplicates in the source files ? Also, is the assumption $S_y \subset S_x$ always tenable ?

In summary, I thank Dr. Han again for her succinct discussion which will be a valuable source of information for the apparent two distinct groups of researchers, one on SAE and the other on record linkage.

Li

I congratulate Dr. Li for bringing in the very important issue of variable selection, a topic near and dear to me in these days. Variable selection is an essential ingredient of any model-based inference, and SAE is no exception.

Dr. Li has provided some very important information regarding necessary modifications of some of the standard criteria, such as the AIC, BIC, Mallows' C_p needed for variable selection in the SAE context. In my opinion though, AIC, BIC, C_p and their variants are more geared towards model diagnostics, and only indirectly towards variable selection. I admit that the two cannot necessarily be separated, but what I like in these days is a direct application of the LASSO (Least Absolute Shrinkage Selection Operator) which achieves simultaneously variable selection and estimation. This is achieved by getting some of the regression coefficients exactly equal to zero, which is extremely useful in the presence of sparsity. In some real life SAE examples that I have encountered, there is a host of independent variables. Rather than the classical forward and backward selection, LASSO and its variant such as LARS (Least Absolute Regrssion Shrinkage) can provide a very direct variable selection and estimation in one stroke.

For simpliicity of exposition, I restrict myself to linear regression models, although the application of LASSO can be extended to generalized linear models, Cox's proportional hazards models and others. For the familiar linear regression model given by $Y = X\beta + e$ notation. The LASSO estimator of β is given by

$$\hat{\beta}_{\text{LASSO}} = \text{argmin}_{\beta} \left[||Y - X\beta||^2 + \lambda \sum_j |\beta_j| \right],$$

where λ is the regularization or the penalty parameter. The choice of the penalty parameter can often become a thorny problem, and there are many proposals including an adaptive approach (Zou, 2006). It will be interesting to see an analog of LASSO in mixed effects models where there is a need for simultaneous selection of regression coefficients and random effects. Obviously, this is of direct relevance to small area estimation. The transformed model of Professor Li from random to fixed effects seems to facilitate the LASSO application in selecting the appropriate regression coefficients. I may add also that there is some recent work on the selection of random effects in the SAE context as discussed in the present paper. But the simultaneous selection problem can potentially be a valuable topic for future research.

I cannot resist the temptation of the well-known Bayesian interpretation of LASSO estimators. Interpreting the loss as the negative of the log-likelihood, and the regularization part as the prior, the LASSO estimator can be interpreted as the posterior mode of a normal likelihood with a double exponential prior. One interesting observation here is that the double exponential prior has tails heavier than that of the normal, but it is still exponential-tailed. Tang, Li and Ghosh (2018), pointed out that polynomial-tailed priors rectify certain deficiencies of exponential-tailed priors. Some of these priors were used in Tang, Ghosh, Ha and Sedransk (2018), as discussed in the present paper.

Molina and Newhouse

Both Professor Molina and Dr. Newhouse have presented very elegantly the current state of the art for estimation of small area poverty indicators. While Professor Molina has has provided a very up-to-date coverage of methodological advances in this area, Dr. Newhouse has focused very broadly on practical applications with examples, and finally a few pointers regarding possible alterations of the World Bank SAE methods with the advent of the so-called "big" data. As I mentioned at the beginning of this rejoinder, I will first present a few common things that I learnt from their discussion, and then reply separately to these two discussants.

One very interesting feature is that SAE of poverty indicators is based on unit level models, another good application of the classical model of Battese, Harter and Fuller (1988). Both discussants began their discussion mentioning the paper of Elbers, Lanjouw and Lanjouw (ELL, 2003), which in my mind, set the stage for further development. An important piece of information here is that while the SAE indicators both use survey and census data, they cannot be *linked* together at a household level due to data confidentiality. As described in details by Professor Molina, and also hinted at by Dr. Newhouse, ELL circumvented this problem by first fitting the survey data to estimate the model parameters, and then generating multiple censuses to estimate the SAE poverty indicators and their MSE by some sort of averaging of these censuses.

The second important aspect of this research is that unlike most SAE problems which involve estimation of totals, means or proportions, one needs to face nonlinear estimation in addressing the poverty indication problem. This poses further challenge. Variable transformation seems to be a way to justify approximate normality of transformed variables, and I will comment more on this while discussing Professor Molina.

Now I will respond individually to Professor Molina and Dr. Newhouse. Maintaining the alphabetical order throughout this rejoinder, I will first discuss Professor Molina and then Dr. Newhouse.

Molina

Professor Molina has pointed out the distinction of her 2010 joint paper with Dr. Rao with that of ELL. The Molina-Rao (MR) paper is an important contribution, which attracted attention of conventional small area researchers. I am not quite sure what Professor Molina means by "unconditional expectation" in ELL. What I understand though, and also essentially pointed out in Molina, that ELL is producing a synthetic estimator in contrast to an optimal composite estimator, namely the EBLUP as given in MR. This optimality is achieved by combining two sources of information, quite in conformity with the usual Bayesian paradigm, which combines a likelihood with a prior.

There are some important issues stemming out of the ELL and MR papers. One, which seems to have been addressed already in the 2019 paper of Dr. Molina, is how best one can utilize both survey and census data when they cannot be linked together. The second pertains to the question of variable transformation. The log transformation is often useful, especially since the moments of a log-normal distribution can easily be calculated via moment generating function of a normal distribution. While the log transfomation reduces skewness, resulting normality can sometimes be put to question. Professor Molina has mentioned the Box-Cox transformation, which is definitely useful. So are the skewed normal and generalized beta of the second kind. But what about a Bayesian nonparametric approach?

The Bayesian approach has a very distinct advantage of providing some direct measure of uncertainty associated with a point estimate via posterior variance. As recognized by Professor Molina, a hierarchical Bayesian approach avoids much of the implementation complexity, when compared to procedures such as the jackknife and bootstrap. But a Bayesian nonparametric approach seems equally applicable here. MR considered a general class of poverty measures given in Foster, Greer and Thorbecke (1984). These measures when simplified lead to estimation of either the distribution function or functionals of the distribution function. A Dirichlet process or its mixture with a normal or a heavy-tailed mixing distribution such as the double exponential can be used without much extra effort. This may be a potential topic of useful research.

Professor Molina has also pointed out that the revised World Bank approach of bootstrapped EB predictors can be severely biased. What about the double bootstrap of Hall and Maiti (2006)?

In summary, I thank Dr. Molina again for bringing in the salient features related to estimation of small area poverty indicators. There are potentials for further development, which I believe will take place in the next few years by Dr. Molina and her collaborators.

Newhouse

I thank Dr. Newhouse not only for bringing in the current World Bank practice of producing small area estimates of poverty indicators, but also for pointing out their global applications as well as some important directions for future research.

The World Bank produces small area estimates at a "subnational" level for 60 countries. Dr. Newhouse did not define subnational as its meaning inevitably varies from country to country. For me, it can be counties, census tracts, school districts, or sometimes even the states, depending on the problem at hand. What I admire though is the importance and relevance of this project from a global standpoint.

I agree with Dr. Newhouse about the need for separate models for urban and rural areas. In addition, in the US, variation between the states, for example, West Virginia and New York, also demands separate modeling. I do not think that this approach leads to reduction in efficiency. Rather, it has the potential to provide more meaningful measures of poverty indicators.

I agree wholeheartedly with Dr. Newhouse regarding the use of alternative sources of auxiliary data. But even there, one may often face the difficulty of proper linkage. Partha Lahiri and Ying Han are currently working quite extensively on probabilistic record linkage in the context of small area estimation. Some of their proposed methods may be helpful in other contexts as well.

"Big"" data offers a huge potential. Combining survey data with administrative data, whenever possible, is expected to provide better results than one that uses only one of these two sources of data. I may add that "non probability sampling" has started receiving attention as well because of the richness of administrative data. Whatever the source, model-based SAE is inevitable, and thus always has the potential danger of failing to provide the right answer. External evaluation of model-based procedures against some "gold standard" seems to be a necessity. This may not be feasible all the time. As an alternative, one may think of cross-validation.

Finally, I like to point out that a model may need to go through a thorough overhaul in the event of a natural or social catastrophe, as we are witnessing now in COVID-19, a "shock" in the general terminology of Dr. Newhouse. Many small area models, by necessity are spatial, temporal or spatio-temporal. Any prediction based on these models, assuming a smooth continuum, will be severely compromised with the occurrence of "shock" events even though some of the auxiliary variables may not be affected.

I thank Dr. Newhouse again for bringing in the current World Bank approach to the production of small area poverty indicators, and his insight into how to improve these estimates in the future.

Pfeffermann

I really appreciate all the valuable comments made by Dr. Pfeffermann in my original text, and they are all incorporated in the revision of this paper. Dr. Pfeffermann has years of both academic and administrative experience, and this is clearly reflected in his discussion. I will try point by point response to his comments, even though I really do not know proper answer to many of the issues that he has raised.

1. I agree with Dr. Pfeffermann that response rate, unless mandatory, is declining fast in most surveys. Further, the simplifying assumption of missing completely at random (MCAR) or missing at random (MAR) is often not very tenable. However, with not missing at random (NMAR) data, I do not see any alternative other than modeling the missingness. In the SAE context, this becomes an extra modeling in addition to the usual SAE modeling, and one requires validation of the integrated model. SAE models with a combination of survey and administrative data, can admit model diagnostics, or sometimes even external evaluation, for example with the nearest census data. Is there a simple way to validate the missingness model in this context? I simply do not know.
2. Again, I agree with Dr. Pfeffermann that present-day surveys offer the option of response via internet, telephone or direct face to face interview. In this cell phone era, I am not particularly fond of telephone interviews. A person living in Texas may have a California cell number. In an ideal situation, for example, a survey designed only for obtaining some basic non sensitive data, the response may not depend much on the mode used. But that is not the case for most surveys, and then the answer may indeed depend on the chosen mode as pointed out very appropriately by Dr. Pfeffermann. What I wonder though is that when there is modal variation in the basic response, is it even possible to quantify the modal difference in the data analysis?
3. Research on measurement errors in covariates for generalized linear models in the SAE context has not possibly started as yet, but it seems feasible. The approach that comes to mind is a hierarchical Bayes approach, both for functional and structural measurement error models.

4. Benchmarking for GLMM is possibly quite challenging from a theoretical point of view in a frequentist set up. It is not at all a problem in a Bayesian framework. Indeed, in Datta et al. (2011), as cited in the present paper, Bayesian benchmarking with squared error loss can be implemented knowing only the posterior mean vector and the posterior variance-covariance matrix.
5. The final point of Dr. Pfeffermann is extremely important as it opens up a new avenue of research. There is always a need for providing uncertainty measures associated with model-based estimates. As George Box once said: "All models are wrong, but some are useful". As a safeguard against potential model uncertainty, one option is to derive design-based MSE of model-based SAE estimators. This also has the potential for convincing conventional survey analysts that model-based SAE or even model-based survey sampling, in general, is not just an academic exercise. Research seems to have just started in this area. A paper that I have just become aware of, courtesy of Dr. Pfeffermann, and mentioned in the current version of the paper, is Pfeffermann and Ben-Hur (2018). Lahiri and Pramanik (2019) addressed the issue of average design-based estimator of design-based MSE, when the average is taken over similar small areas.

Rao

I very much appreciate the kind remarks of Professor Rao. It is needless to say that he is one of the pioneers who brought SAE in the forefront of not just survey statisticians, but for the statistics community at large. I have had the fortune of collaborating with him in a paper only once. But I have had the fortune of getting his advice on a number of occasions in my SAE research.

Regarding the points that he has raised, I agree virtually with all of them. Without a hierarchical Bayesian procedure, it is quite possible to get zero estimates of A, the random effect variance, by any of the standard methods, be it method of moments, ML or REML. Adjusted ML by Li and Lahiri (2010), and subsequent development by Yoshimori and Lahiri (2014), Molina et al. (2015) and Hirose and Lahiri (2018) are indeed very welcome as they rectify this deficiency.

The second point regarding external evaluation is also very useful. Census figures have often been used as "gold standard", used by many researchers including myself. Unfortunately, in many SAE examples, one does not have this opportunity of external validity. I do not have a real idea of an alternative approach with firm footing in this case, but think that cross validation may be an option.

Professor Rao has mentioned the need for design-based MSE computation of model-based SAE estimators. I have emphasized its relevance and importance, while discussing Dr. Pfeffermann. I reiterate that this topic will possibly be a fruitful research topic in the next few years.

I have not seen yet the review article of Jiming Jiang and Sunil Rao, but can appreciate their viewpoint. I have cherished the view for a long time that outliers should not necessarily be discarded for inferential purposes. Rather they can very well be a part of a model, typically a mixture model, which was advocated by Tukey many years ago.

I endorse also that it is high time to go beyond estimation of small area means. Estimation of small area poverty indicators where the World Bank people as well as Professors Rao and Molina have made significant contribution, has taken off the ground and research is pouring in this area. Another potential topic seems to be estimation of quantiles in general, since these parameters are less vulnerable to outliers.

Finally, I thank all the discussants once again for their thorough and informative discussion, supplementing very well the topics not covered in this paper. It is needless to say there is a plethora of other uncovered topics in my paper. We may need another review paper (not by myself) with discussion fairly soon to cover some of these other topics.

REFERENCES

FOSTER, J., GREER, J., THORBECKE, E., (1984). A class of decomposable poverty measures. *Econometrika*, 52, pp. 761–766.

GELMAN, A., (2006). Prior distributions for variance parameters in hierarchical models. (Comment on article by Browne and Draper). *Bayesian Analysis*, 1, pp. 515–553.

HALL, P., MAITI, T., (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, B*, 68, pp. 221–238.

HIROSE, M. Y., LAHIRI, P., (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics*, 46, pp. 1721–1741.

TONG, X., XU, X., GHOSH, M., GHOSH, P., (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80, pp. 215–246.

YOSHIMORI, M., LAHIRI, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriott small area model. *Journal of Multivariate Analysis*, 101, pp. 1418–1429.