STATISTICS IN TRANSITION new series, August 2020
Special Issue, pp. 31–35, DOI 10.21307/stattrans-2020-024
Submitted – 31.01.2020; paper accepted for published – 30.06.2020

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

Ying Han[1]

1. Introduction

I would like to thank Prof. Ghosh for his significant contributions to small area estimation, not only for his phenomenal research, but also for the talents that he cultivated and brought into this field. It is my great honor to be an invited discussant of Prof. Ghosh's paper "Small Area Estimation: Its Evolution in Five Decades".

In the paper, Prof. Ghosh presents a nice overview of the history and development of small area estimation. He clearly explains the reason why small area estimation techniques are important in providing accurate estimates for small regions or domains, illustrates the increasing importance of small area estimation through examples in different fields, introduces different small area estimates developed from area-level and unit-level models, etc. He traces back to the starting point of small area estimation, demonstrates its development, and shows us its bright future.

The basic idea of small area estimation is to increase the effective sample size by borrowing strengths from variable of interest from other related areas. This is primarily done by linking related small areas using auxiliary information related to the variable of interest. The auxiliary information often comes from administrative records. So, the availability of good administrative records is of great importance to small area estimation. As Prof. Ghosh said in the paper, "the eminent role of administrative records for small area estimation cannot but be underscored even today."

The unit-level small area estimation models require the joint observations on the variable of interest y and the auxiliary variables x for the sampled units in small areas. If administrative records are used, we need to know which administrative record represents the same population unit as one in the survey data. Consider the case where the data comes from two separate files: one survey data set containing the observations on y and an administrative data set containing the observations on x. If a unique and error-free identifier exists in both files, the two files can be linked without any errors and a merged dataset with joint observations on y and x is obtained. Under this data layout, a huge literature on small area estimation is available. We refer reader to Rao and Molina (2015), Jiang and Lahiri (2006), and Pfeffermann (2013).

[1]Gallup.Inc, USA. E-mail: ying_han@gallup.com. ORCID: https://orcid.org/0000-0003-0082-5654.

Most of the time, however, such identifier is not available in either the survey data set or the administrative data set. In this case, the administrative records can rarely be used for unit-level small area estimation model. This limits the application of small area estimation. Record linkage, a data integration technique, is a potential approach to link the files even when a unique and error-free identifier is not available. The application of record linkage extends the application of small area estimation to the case when administrative records cannot be linked to the survey data by using unique identifiers. This is one of the most emerging topics that was not covered in Prof. Ghosh overview paper. In this discussion, I would like to provide a brief description on this topic.

2. Probabilistic Record Linkage

Record linkage, or exact matching, is a technique to identify records for the same entity (e.g., person, household, etc.) that are from two or more files when a unique, error-free identifier (such as Social Security Number) is missing. The first theoretical framework for record linkage was developed by Fellegi and Sunter (1969). A linked dataset, created by record linkage, is of great interest to analysts interested in certain specialized multivariate analysis, which would be otherwise either impossible or difficult without advanced statistical expertise as variables are stored in different files.

However, the linked dataset is subject to linkage errors. If one simply ignores the linkage errors, analysis of the linked data could yield misleading results in a scientific study. Neter et al. (1965) demonstrated that a relatively small amount of linkage errors could lead to substantial bias in estimating a regression relationship. Therefore, the importance of accounting for linkage errors in statistical analysis cannot be overemphasized. In the past couple of decades, researchers have been focused on how to correct the bias caused by linkage errors when fitting linear regression model on linked data. Chambers (2009), Kim and Chambers (2012), Samart and Chambers (2014) tackled the problem from the second analyst point of view, assuming that they can only get access to the linked data and limited information is available about the linkage process. In contrast, Lahiri and Larsen (2005) solved the problem from the primary analyst point of view by taking advantage of the summary information generated during the record linkage process. But there is little literature on the how to apply small area estimation on the linked data generated through record linkage process.

The importance of integrating probabilistic record linkage in small area estimation was highlighted in the SAE International Statistical Institute Satellite Meeting held in Paris during July 10-12 , 2017. In his keynote address at the meeting, Professor Partha Lahiri introduce the concept of merging survey data with administrative records together through record linkage technique to obtain an enhanced dataset for small area estimation. It can cut down the cost in data collection by preventing the need to collect new survey data with all necessary information. Later, I worked with Professor Lahiri in proposing a unified way for performing small area estimation using data from multiple

files. A brief description of the methodology is provided in the next section. Readers interested in the details are referred to Lahiri (2017), Han (2018), and Han and Lahiri (2019).

3. Small area estimation within linked data

We are interested in predicting an area-specific parameter, which can be expressed as a function of fixed effects and random effects related to the conditional distribution of y given x. For simplicity, we restrict our research to the case where the observations on y and x come from two files, rather than more than two files (e.g., one survey dataset and multiple administrative data sets). Suppose the observations on y (x) are available for a sample S_y (S_x) and are recorded in file F_y (F_x). The matching status between any record in F_y and any record in F_x is unknown. We assume that (1) there is no duplicate in either F_y or F_x, (2) $S_y \subset S_x$, and (3) the records in both files can be partitioned into small areas without error.

We propose a general integrated model to propagate the uncertainty of the linkage process in the later estimation step under the assumption of data availability described above. The model is developed from a primary analyst point of view. The primary analyst can get access to the original two files, which contains both the separate observations on y and x and the values of matching fields (a set of variables for record linkage). The proposed model is built directly on the data values from the original two files (rather than on data in the linked dataset) and is based on the actual record linkage method that is used (rather than making a strong assumption on the linkage process afterwards). The general proposed integrated model includes three important components: a unit-level small area estimation model, a linkage error model, and a two-class mixture model on comparison vectors. The unit-level model is used to characterize the relationship between y and x in the target population. The linkage error model is used to characterize the randomness of the linkage process. It is developed by exploiting the relationship between X^* (the unobserved x values corresponding to the observed y values in F_y) and X (the observed x values in F_x). It is the key to the general integrated model, serving as a connector between the unit-level small area model and the record linkage model. The two-class mixture model is used to estimate the probability of a record pair being a match given the observed data and designate all record pairs into links and non-links.

Under the general integrated model, we provide a general methodology for obtaining an empirical best prediction (EBP) estimator of an area-specific mixed parameter. The unified jackknife resampling method proposed by Jiang et al. (2002) and its alternative proposed by Lohr and Rao (2009) can be used to estimate the mean squared error of the empirical best prediction estimator. The jackknife methods proposed by Jiang et al. (2002) and Lohr and Rao (2009) require closed-form expressions for the mean squared error (MSE) and conditional mean squared error (CMSE) of the best prediction estimator (BP), respectively. So, the choice of the jackknife methods depends on whether a

closed-form expression for MSE or CMSE is available.

Application of the general methodology is not limited to the mutual independence of measurements. It can be applied to measurements that are correlated within small areas but independent across small areas. Unit-level models such as general linear model with correlated sampling errors within small areas, general linear mixed model with nested errors can all be considered. To illustrate our general methodology, we consider the situation where the unit-level small area model of the general integrated model is set to be the general linear mixed model with block diagonal covariance structure. The Best Prediction (BP) estimator for the mixed parameter is derived under the general integrated model. The conditional mean squared error (CMSE) of its corresponding Best Prediction (BP) Estimator can be expressed in a closed form, making it possible to estimate its mean squared error using the jackknife method provided by Lohr and Rao (2009).

As a special example, we consider the estimation of small area means when a nested error linear model is used. We provide two methods for estimating the unknown parameters: the Maximum Likelihood (ML) method and the Pseudo Maximum Likelihood (PML) method. We also discuss the use of numerical algorithms in approximating the maximum likelihood estimates (MLE), including Newton-Raphson method and Fish scoring algorithm, and further propose a quasi-scoring algorithm in order to reduce the computational burden.

4. Summary

Due to the increasing demand of small area estimation in different fields and the accessibility of administrative records, it is of great interest for researchers and analysts to use probabilistic record linkage in extracting additional information from administrative records as additional auxiliary variable in unit-level small area models. It is an example of the more recent topics in small area estimation that are not covered by Prof. Ghosh in his overview paper. As Prof. Ghosh said, "the vastness of the area makes it near possible to cover each and every emerging topic". That means, small area estimation is still under its rapid development driven by its high demand, and it is a field full of vitality.

REFERENCES

CHAMBERS, R., (2009). Regression analysis of probability-linked data. *Statisphere*, 4.

FELLEGI, I., SUNTER, A., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, pp. 1183–1210.

HAN, Y., (2018). Statistical inference using data from multiple files combined through record linkage. PhD thesis, University of Maryland.

HAN, Y., LAHIRI, P., (2019). Statistical analysis with linked data. *Journal of the American Statistical Association*, 87, pp. S139–S157.

JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation. *Test*, 15(1), pp. 1–96.

JIANG, J., LAHIRI, P., WAN, S. W., (2002). A unified jackknife theory for empirical best prediction with m-estimation. *Annals of Statistics*, 30(6), pp. 1782–1810.

KIM, J., CHAMBERS, R., (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56(9), pp. 2756–2770.

LAHIRI, P., (2017). Small area estimation with linked data. Keynote address at the ISI Satellite Meeting on Small Area Estimation, Paris, France, July, pp. 10–12.

LAHIRI, P., LARSEN, M., (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469), 222–230.

LOHR, S. L., RAO, K., (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Boimetrika*, 96(2), pp. 457–468.

NETER, J., MAYNES, E., RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60(312), pp. 1005–1027.

PFEFFERMAN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28(1), pp. 40–68.

RAO, J., MOLINA, I., (2015). Small Area Estimation. Wiley, second edition.

SAMART, K., CHAMBERS, R., (2014). Linear regression with nested errors using probability-linked data. *Australian and New Zealand Journal of Statistics*, 56(1), pp. 27–46.