A nonparametric Bayesian prediction approach for modelling small data

12th International Conference on Bayesian Nonparametrics, University of Oxford, UK; June 24-28, 2019

Azizur Rahman - Charles Sturt University, Australia

ABSTRACT

In the 21st century's big data era, the small data snags still exist in many areas which causes modelling and evaluation are hard to make. Small dataset is insufficient to generate a reliable prediction especially in the small area estimation domain. This paper presents a Bayesian nonparametric prediction framework to models which are dealing with smaller data sets and/or give poor predictions. This approach uses a robust Gaussian model in weight-space notion and drive the prediction distribution of the future responses. Results revealed that the prediction distribution of a set of futures responses is conditional on a set of observed data and depends on the degree of the spline. It also provides an empirical illustration and demonstrated that the prediction outcomes depend on the realised responses only through the observations in design matrix and the sample residual sum of squares and products matrices with the Kronecker product.

INTRODUCTION

Most of the data scientists would agree that that all models are wrong, but some models are useful. Even in the Bayesian analysis there are problems where inference under the simplified model builds on the posterior distribution of parameters given the observed data can lead to inaccurate estimates and policy decisions [1], Although a number of authors have tried to establish conditions under which the selection method of unobserved/missing data for model-based inferences from the Bayesian, likelihood or sampling theory viewpoints [2-4], such a situation is especially common for models developed with small data in small area estimation (SAE) [5]. Nowadays indirect modelling approaches of SAE such as spatial microsimulation modelling (SMM) have received much attention [6]. Indirect SAE is the process of using statistical models and/or geographic models to link survey outcomes to a set of predictor variables known for small areas, in order to predict small area-level estimates [7]. However, building a spatial microsimulation model is very difficult for many reasons. The creation of reliable spatial microdata is still challenging due to a lack of mathematically sound reweighting algorithms [8,9]. Thus, a new reweighting tool for generating synthetic spatial micropopulation data at small area level is crucial [9], and this paper presents a Bayesian nonparametric prediction approach.

METHODOLOGY

This research uses an unfussy fundamental concept which is somewhat similar to the methods described in [2,3]. It is the newest development in SMM methodologies and particularly a very useful tool for dealing with small data. As for any area being sampled, a finite population usually has two parts - observed units called data and unobserved sampling units in the population (Fig. 1) [10].

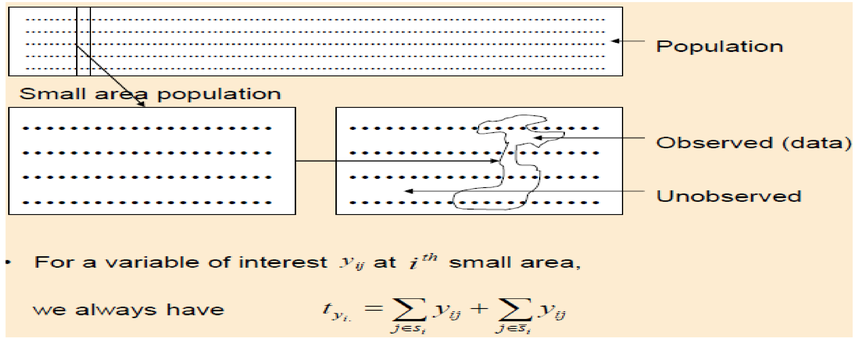

Fig. 1: A diagram of new system for generating spatial microdata.

The main challenge we found in this approach of microdata simulation is to establish the linkage of observed data to the unobserved sampling units in the small area population. Essentially it is a prediction problem, where a modeller tries to find a probability distribution of unobserved responses using the observed sample and the auxiliary data.

Adopting the notations from [2,5], a multivariate model for released data at the i[th] small area can be defined as $Y_i = X_i\beta + E_i$, where $E_i \sim T_{n_i,p}\left(0, I_{n_i \times n_i}, \Sigma_{p\times p}, \nu\right)$. Similarly, a respective model for unobserved units can be defined as $\bar{Y}_i = \bar{X}_i\beta + \bar{E}_i$ with $\bar{E}_i \sim T_{m_i,p}\left(0, I_{m_i \times m_i}, \Sigma_{p\times p}, \nu\right)$. Then the joint density function of observed sample units and unobserved units is given as

$$p(Y_i, \bar{Y}_i \mid \beta, \Sigma, G) \propto \left|\Sigma\right|^{-\frac{n_i+m_i}{2}} \left|I_p + \Sigma^{-1} Q\right|^{-\frac{1}{2}(\nu+p+n_i+m_i-1)} \quad (1)$$

where, $Q = (Y_i - X_i\beta)'(Y_i - X_i\beta) + (\bar{Y}_i - \bar{X}_i\beta)'(\bar{Y}_i - \bar{X}_i\beta)$.

Now using an appropriate set of nonparametric priors of p(G) for the infinite dimensional unknown distribution G with a non-informative joint distribution of $p(\beta, \Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$, the prediction distribution of unobserved units, given the observed units is obtained as

$$f(\bar{Y}_i \mid Y_i) = C(Y_i, H)\left|S_Y + [\bar{Y}_i' - \bar{X}_i\hat{\beta}]'H[\bar{Y}_i' - \bar{X}_i\hat{\beta}]\right|^{-\frac{N_i-k}{2}} \quad (2)$$

for $C(Y_i, H) = \dfrac{(\pi)^{-\frac{(N_i-n_i)p}{2}} \Gamma_p\left(\frac{n_i-k}{2}\right)\left|H\right|^{-\frac{p}{2}}}{\Gamma_p\left(\frac{N_i-k}{2}\right)\left|S_Y\right|^{\frac{n_i-k}{2}}}$, $S_Y = Y_i'[I - X_i(X_i'X_i)^{-1}X_i']Y_i$

$H = I - \bar{X}_i M^{-1} \bar{X}_i'$ and $M = X_i'X_i + \bar{X}_i'\bar{X}_i$.

Now using (2) by MCMC simulation method, we can obtain simulated copies of microdata for the entire population at the i[th] small area.

RESULTS AND DISCUSSION

It is observed that the prediction distribution in (2) is conditional on realised data only through the observations in design matrix and the sample residual sum of squares and products matrices with the Kronecker product, and it depends on the degree of the spline. An empirical illustration with EU data has also showed that the proposed reweighting technique is a robust tool for SMM in SAE (see, Fig. 2).

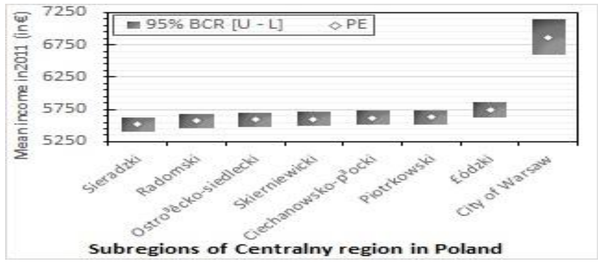

Fig. 2: Mean equivalised household income estimates of eights subregions in Poland, 2011.

Finally, the key feature of this new method is that it can simulate complete scenarios of the whole micro-population in a small area, which means it can produce more reliable small area estimates and their variance estimation. It is also able to create the Bayes credible region for spatial microsimulation models' estimates.

ACKNOWLEDGEMENTS

The travel costs were funded by a grant from the Faculty of Business, Justice and Behavioural Sciences - Charles Sturt University. Thanks are also due to Prof Murray Aitkin for the fruitful discussions concerning conceptual formulation.

REFERENCES

[1] Muller, P. & Mitra, R. 2013, Bayesian nonparametric inference - why and how. *Bayesian Analysis*, 8: 269-302.
[2] Zangeneh, S., Keener, R. & Little, R. 2011, Bayesian nonparametric estimation of finite population quantities in absence of design information on nonsampled units. Section on Survey Research Methods - JSM 2011: 3429-3440.
[3] Opsomer, et al. 2008, Non-parametric small area estimation using penalized spline regression. *JRSS-B*, 70: 265-286.
[4] Aitkin, M. 2010, Statistical Inference: an Integrated Bayesian/Likelihood Approach. NY: CRC.
[5] Rao, JNK. & Molina, I. 2015, Small area estimation, NY: Wiley.
[6] Rahman, A. & Harding, A. 2016, Small area estimation and microsimulation modelling. NY: CRC.
[7] Rahman, A. 2017, Small area housing stress estimation in Australia: Calculating confidence intervals for a spatial microsimulation model. *Communications in Statistics Part B: Simulation and Computation*, *46*: 7466-7484.
[8] Rahman, A., Harding, A., Tanton, R. and Liu, S. (2010), "Methodological issues in spatial microsimulation modelling for small area estimation", *The International Journal of Microsimulation* 3(2), pp. 3-22.
[9] Rahman, A., Harding, A., Tanton, R. & Liu, S. 2013, Simulating the characteristics of populations at the small area level: New validation techniques for a spatial microsimulation model in Australia. *Computational Statistics & Data Analysis*, 57: 149-165.
[10] Rahman, A. & Upadhyay, SK. 2015, A Bayesian reweighting technique for small area estimation. In U. Singh, A. Loganathan, S. K. Upadhyay, & D. K. Dey (Eds.), *Current trends in Bayesian methodology with applications* (1st ed., pp. 503-519). Florida: CRC.

Contact – Dr Azizur Rahman: azrahman@csu.edu.au

www.csu.edu.au