Small Area Estimation with Penalty for Specific Area Effects Selection

1st Novi Hidayat Pusponegoro [1], 2nd Anang Kurnia [2], 3rd Khairil Anwar Notodiputro[3], 4th Agus Mohamad Sholeh [4], 5th Erni Tri Astuti[5]
{novie@stis.ac.id[1], anangk@apps.ipb.ac.id [2], khairil@apps.ipb.ac.id [3], agusms@apps.ipb.ac.id [4], erni@stis.ac.id[5]}

Statistics Department, IPB University, Bogor, 16680, Indonesia [1,2,3,4], Statistics Department, STIS Polytechnic of Statistics, Jakarta, 13330, Indonesia [1,5]

Abstract. Small area estimation (SAE) techniques are now widely employed to produce parameter estimates for smaller domains where sample sizes cannot be used to deliver direct estimation. SAE as an indirect estimation method utilizes strength from other related small areas to improve the precision. In the 'Big data' era, database size and technology has developed rapidly. This leads to computational and statistical challenges since the availability of high data volume. Thus, the existing SAE methods can not longer handle the complexity of fixed effects or random effects in this data. The big data also provide large number areas as observation where not all them are small areas. This sparsity of random effects also brings out violation to its normal assumptions. Therefore, identifying the effective random effects is very important to ease the computational burden and to construct more interpretable models. This study presents a small area estimation method that is able to overcome the complexity of random effects with hard-ridge penalty. In this paper, simulations are delivered to demonstrate the performance of the methods and applied to estimate sub-district level mean of per capita income using the poverty survey data in Bangka Belitung Province at 2017.

Keywords: per capita income, sparsity of random effects, iterative algorithm.

1 Introduction

Survey is an effective tool to collect information of a certain population in a region or a country. The results are population characteristics estimates and used by the government to construct indicators for evaluating the development, making decisions or allocating funds. Sample size is an issue in survey estimation reliability, moreover if some of it needed to present by its subdomains. Thus, small area estimation can been used to overcome this problem by obtaining reliable indirect estimates. Small area refers to a small group or geographical area which is only a few samples is available from a given small area. In order to increase the estimate precision, it needs relevant auxiliary information such as data from other related small areas.

The Fay-Herriot model and the Nested Error Regression Model (NERM) are the typical models used for the small area estimation. Regarding the situation in which auxiliary information are available, NERM is the unit-level models entails information at the area and individual level. The area level-model known as Fay-Herriot model is more applicable in small area estimation since it does not require the confidential information of the observations

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290492

[1]. Then, the model represents the corresponding area-level auxiliary variables $\boldsymbol{X}_i$to the direct survey estimator of the ith small area $\boldsymbol{y}_i$can be written as:

$$y_i = \theta_i + e_i, \qquad \theta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i,\ i = 1, \dots, m \tag{1}$$

where m is the number of small areas, $\boldsymbol{\beta}$ is a $p \times 1$ unknown common vector of regression coefficients, and θ_i is a summary measure of the characteristic to be estimated for the ith small area. The sampling error $e_i \sim N(0, \sigma_{ei}^2)$and random effects $u_i \sim N(0, \sigma_u^2)$ are assumed to be independent. The random effects u_i denote the model error measuring the departure of θ_i from its linear regression on $\boldsymbol{X}_i, \boldsymbol{X}_i^T \boldsymbol{\beta}$. It also called specific small area effects since it present area-specific variations that cannot be explained by the auxiliary variables in the model. Thus, the good small area models have strong covariates that relate with $\theta_i$and expected that random area effects $u_i$are small. Regarding the sample size in survey areas, the random effects term may be negligible in the SAE model moreover if areas are similarly large. Moreover, a test for the presence of the random effects issuggested in the general linear model [2]. If the null hypothesis of no small-area effect is failed to reject, the random effects term is eliminated completely from the model. However, eradicating the random effects from the model may not be the best choice in terms of prediction accuracy.

Recently, the big data leads to confronting the high volume of public records in small area estimation. It also increases the sizes of databases tremendously and brings unexpected computational and statistical challenges. Large scale data such as one ‘modern data’ type has large observations or small areas. The existing SAE methods are not appropriate for modern data because they cannot handle the complexity of random effects. The number of small areas in fact is relatively small compared to the sample size n. The rate of convergence of the estimator is depend on the number of small areas. The rate of convergence of the estimator is faster when the number of small areas is smaller [2,3]. In other words, the speed of convergence in the SAE model is associated with the complexity of random effects. Therefore, reliable estimates can be obtained with selection the true small areas (true nonzero random effects) efficiently. Sparsity on u_i can be imposed by assigning zero for a large area i, whereas it preserve the nonzero value of u_i for a small area i. This sparsity of random effects brings heavy tails. As a result, the normality assumption of random effects is not retained any longer.

Therefore, this paper review the Small Area Estimation with Random Effects Selection (SARS) model in order to emphasized the selection methods under assumption the random effects sparsity which has been proposed by Lee [4] and apply the model to estimate sub-district level mean of per capita income using the poverty survey data in Bangka Belitung Province at 2017. This paper is organized into four sections. The first section presents the background, motivation and purpose of this research. Second section describes the material that is used in this study and the third section describes the Small Area Estimation with Random Effects selection Model. The empirical results obtained in the application and the main findings of the analysis and discusses the further possible researches are presented in section 4.

2 Methods

Consider estimator of the ith small area $\boldsymbol{y}_i$can be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{e}\,, \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{m}}$, $\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_2, \dots, \boldsymbol{X}_m)^T$is $m \times p$ explanatory variables, $\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{p}}$is a $p \times 1$ unknown vector of the regression coefficients, $\boldsymbol{u} \in \mathbb{R}^{\boldsymbol{m}}$and $\boldsymbol{e} \in \mathbb{R}^{\boldsymbol{m}}$ are vector of the random effects and vector of the random errors, respectively. Refers to linear mixed model, The Fay-Herriot model is a kind of specific models of LMM or two-level hierarchical model which can be written as:

$$\boldsymbol{y}|\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{X\beta} + \boldsymbol{u}, \boldsymbol{\sigma}_u^2). \tag{3}$$

Fay-Herriot model assumes that its random effects are normally distributed, but in can be violated such in case the point-mass at zero for large areas. As Fay-Herriot model in (1) where $\boldsymbol{\beta}$as fixed parameters and let $\boldsymbol{U}$ be a set of small areas, then $u_i \sim N(0, \sigma_u^2)$for $i \in \boldsymbol{U}$ and $u_i = 0$ for i.

In small area estimation, the method of estimating parameters are decided by $\boldsymbol{\sigma}_u^2$and $\boldsymbol{\beta}$ assumption. Assuming $\boldsymbol{\sigma}_u^2$and $\boldsymbol{\beta}$ are known, the best predictor (BP) or best linear unbiased predictor (BLUP) is derived by $E[\boldsymbol{u}|\boldsymbol{y}] = \boldsymbol{\partial}(\boldsymbol{y} - \boldsymbol{X\beta})$and $E[\boldsymbol{\theta}|\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{\partial})\boldsymbol{Xy} + \boldsymbol{\partial y}$where $\boldsymbol{V} = \boldsymbol{diag}(\sigma_u^2 + \sigma_{ei}^2)$and $\boldsymbol{\partial} = \boldsymbol{V}^{-\mathbf{1}}\boldsymbol{\sigma}_u^2$. Assuming $\boldsymbol{\sigma}_u^2 = \sigma_u^2 \boldsymbol{I}_n$is known, the estimates of $\boldsymbol{\beta}$is $\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$. The EBLUP of $\boldsymbol{\theta}$ is estimated by replacing $\boldsymbol{\beta}$ with its maximum likelihood estimator (MLE) and substituting $\boldsymbol{\sigma}_u^2$with its estimate by MLE, restricted maximum likelihood estimator (REML) using mean squared prediction error (MSPE) or mean squared error (MSE) approximation.

The best empirical estimator $\boldsymbol{\theta}$ does not depend on the model but on the weighting factor $\boldsymbol{\partial}$[5]. Whereas in estimating the best empirical linear parameter, the weighting factor $\boldsymbol{\partial}$ depends on the random effects variant $\boldsymbol{\sigma}_u^2$and σ_{ei}^2. So, it needs to be identified correctly to bring out the accurate estimation for the fixed effect of $\boldsymbol{\beta}$ and direct estimation $\boldsymbol{\theta}$. Somehow in the modern condition such as large-scale data, the normality assumptions in random effects $\boldsymbol{u}$ can be violated. So, estimators of the random effects variant through ML or REML are no longer appropriate.

If it is assumed that $\boldsymbol{U}$ is a small set of areas with $u_i \sim N(0, \sigma_u^2)$ for $i \in U$and $u_i = 0$ for $i \in U^C$ and $\boldsymbol{\sigma}_u^2$and σ_{ei}^2 are unknown,)Lee [4] developed a model Small Area Estimation with Random Effects Selection (SARS). Since sparsity in vector $\boldsymbol{u}$ is interpreted as the complexity of random effects, thus this model employ penalized regression in optimizing the function of the least squares difference for estimating the Fay-Herriot parameter. SARS model as a selective shrinkage estimator employs the penalized regression using the hard-ridge penalty. The SARS model uses multiple penalties for a fixed effects as well as for its random effects. With the multiple penalties the reference is a combination of penalties $\ell 0$ and $\ell 2$ which are referred to as Hard-Ridge Penalty [6]. In the SAE model, assuming that the random effects $\boldsymbol{u}$ are sparse, the hard-ridge penalty for select the random effects is stated as:

$$P_{02}(\boldsymbol{u}; \lambda_u, \eta_u) = \frac{\eta_u}{2}\|u\|^2 + \frac{\lambda^2{}_u}{2(1+\eta_u)}\|u\|_0 = \sum_{i=1}^{n}\left(\frac{\eta_u}{2}u_i^2 + \frac{\lambda^2{}_u}{2(1+\eta_u)}1_{u_{i\neq 0}}\right) \tag{4}$$

where λ_u is tuning parameter for hard penalty in order to optimized SARS prediction information criteria ($\lambda_u \geq 0$) and η_u tuning parameter for ridge penalty in order to select the random effects. So the objective functions in the small area estimation model with the random effects selection is:

$$\min_{\boldsymbol{u}\epsilon\mathbb{R}^{J(\boldsymbol{u})}} f(\boldsymbol{u};\beta,\lambda_u,\eta_u) \triangleq \frac{1}{2}\|y - Z\beta - u\|^2 + P_{02}(\boldsymbol{u};\lambda_u,\eta_u). \quad (5)$$

Under assumption the sparsity of $\boldsymbol{u}$, SARS prediction information criterion (PIC) is used as criteria to achieve the optimal prediction tuning parameter threshold in $\boldsymbol{u}$. SARS optimization problem is challenging due to the non-convex and non-smooth feature of the hard-ridge penalty. Then, an iterative technique can be employed to solve SARS problem. Lee [4] proposed an iterative selection-estimation (SE) algorithm to resolve the issue of multiple tuning parameters. The selection step is to screen small areas and significant coefficients by repeatedly applying quantile thresholds and the estimation step is associated with a ridge regression for estimating u with the support sets the true non-zero small area effects obtained from the selection step

To evaluate the performance of parameter estimators in the SAE model, the means absolute error (MAE) and median absolute error are employed in order to measuring the relative predictive accuracy of the SAE model with the formulation as follows:

$$MAE(\hat{\theta}_i) = m^{-1}\sum_{i=1}^{m}|\theta_i - \hat{\theta}_i| \quad (6)$$

$$Med.AE(\hat{\theta}_i) = m^{-1}\sum_{i=1}^{m}|\theta_i - \tilde{\theta}_i| \quad (7)$$

The more accurate prediction is achieved whenever the measurement above is smaller than the estimated model.

3 Materials

This paper illustrates the general results from simulation in order to verify parameter estimation performance of SARS model. The customized data are set with some different random effects variance components, sparsity levels and correlation between auxiliary. The random effects variance components in the simulation are set to be 0.3 and 1 since those value represent small and big value of correlation between model error. Simulation is used to emphasize that the model can be applied eventhough vector ***u***is sparse. Thus, it set the percentages of small area number are 10% and 50% to all analysis areas. The simulation in this study is also used to provide empirical evidence that model is a model that has considered the condition of the relationship between explanatory variables. For this reason, in this simulation the relationship between explanatory variables is also setup with values of 0, 0.5 and 0.8, each of which represents the condition of the uncorrelated, moderately correlated and highly correlated data, respectively.

This model also applied to Indonesian poverty data; it is used to estimate the mean of per capita income of each sub-district in Bangka Belitung Province at 2017. Since, one of substantial poverty indicator is regional mean of per capita income. In this paper, mean of per capita income is estimated from some auxiliary variables such as infrastructure, electricity and regional accessibility.

4 Results

Some simulated data are conducted with different random effects variance components, sparsity levels and correlation between auxiliary variables in order to emphasize the performance of SARS estimator. In the simulation, the sampling variance is set to be 1 and model error variance is $\sigma_u^2 = \{0.3,1\}$. Under the large-scale data condition, it set the number of small areas is 60 with the number of true nonzero small areas are 10% and 50%. The number of auxiliary variables is set to be 10 and the correlation between them are set up to be uncorrelated, 0.5 and 0.8.

The Selection-Estimation algorithm is applied in this paper to attain the SARS model parameter estimation. Since the study focus on the sparsity of $\boldsymbol{u}$, the simulation results are presented by the true nonzero proportion.

Table 1. MAE and Median Absolute Error of SARS model prediction value with 10% true non-zero random effects.

σ_u^2	Correlation	MAE	Med. AE
0.3	Uncorrelated	0.2574	15.4463
0.3	0.5	0.2574	15.4463
0.3	0.8	0.2574	15.4463
1	Uncorrelated	0.1154	6.9256
1	0.5	0.1754	10.5246
1	0.8	0.1776	10.6579

Table 1 and Table 2 present that the SARS model prediction values precision are not significantly different by the correlation between auxiliary variables but those differ slightly by the variance component.

Table 2.MAE and Median Absolute Error of SARS model prediction with 50% true non-zero random effects.

σ_u^2	Correlation	MAE	Med. AE
0.3	Uncorrelated	0.2160	12.9623
0.3	0.5	0.2160	12.9623
0.3	0.8	0.2160	12.9623
1	Uncorrelated	0.1849	11.0934
1	0.5	0.3380	20.2826
1	0.8	0.3483	20.8980

Thus, it gives strong evidence that estimating the regression coefficient and variable selection in linear mixed models is not only selecting fixed effects or random effects but also considering the random effects variance components.

In this section, SARS model is applied to sub-district means per capita income in Bangka Belitung province at 2017. The data consist of 140 sub-districts and 6 auxiliary variables. The auxiliary variables are total population, household total in each sub-district that have access to electricity from Indonesia Electricity Company, distance of the sub-district center to the municipality office, distance of the sub-district center to harbor, distance of the sub-district center hospital and distance means of the household resident to elementary schools. Those variables have linear correlation coefficients between two paired variables are between 0.3 and 0.7. Based on Figure 1, it can be presented the performance the SARS PIC as decision rule for finding the optimal tuning parameter that applied in real dataset. This model also have good performance since the means absolute error of the sub-district means per capita income in Bangka Belitung Province at 2017 is 0.5217 with number of small areas are 37 of 140 areas.

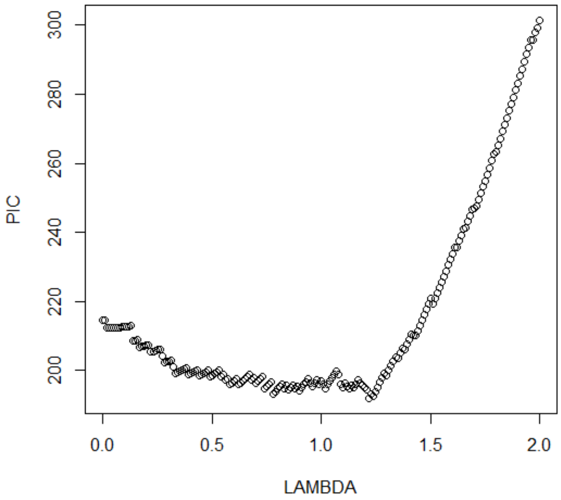

Fig. 1. SARS Predictive Information Criterion (PIC) Values against λ_u.

5 Conclutions

Summarizing the value of the regression coefficient and the selection of variables in the linear mixed model is not just a process of selecting a permanent or random effect but also taking into account the matrix components of the various random effects and error models.

Meanwhile, SARS model employs hard penalty to select the small areas over all areas. as one of the component hard-ridge penalty that is discrete and non-convex. And, according to "Big data" era and the current challenges, thus it is necessary to develop SAE methods that not only select the random effects but also shrinkage the value of their coefficients at once in order to get the parsimony.

References

[1]Rao J.N.K, Molina I.: Small Area Estimation. John Wiley & Sons, Inc., New Jersey (2015)

[2]Datta, G. S., Hall, P., and Mandal, A.: Model selection by testing for the presence of small-area effects, and application to area-level data. Journal of the American Statistical Association, 106(493). pp. 362-374 (2011)

[3]Datta, G. and Ghosh, M.: Small area shrinkage estimation. Statistical Science, 27(1). pp 95–114 (2012)

[4]Lee, J.: Small Area Estimation with Random Effects Selection. Dissertation, Florida University (2017)

[5]Jiang, J., Nguyen, T., and Rao, J. S.: Best predictive small area estimation. Journal of the American Statistical Association, 106(494). pp. 732-745 (2011)

[6]She, Y.: An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. Computational Statistics and Data Analysis, 56(10). pp. 2976-2990 (2012)