Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 8

2-26-2020

Small Area Estimation on Zero-Inflated Data Using Frequentist and Bayesian Approach

Kusman Sadik
Bogor Agricultural University, kusmansadik@gmail.com

Rahma Anisa
Bogor Agricultural University

Euis Aqmaliyah
Bogor Agricultural University

Recommended Citation

Sadik, K., Anisa, R., & Aqmaliyah, E. (2019). Small area estimation on zero-inflated data using frequentist and Bayesian approach. Journal of Modern Applied Statistical Methods, 18(1), eP2677. doi: 10.22237/jmasm/1582727606

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2677.
doi: 10.22237/jmasm/1582727606

בס"ד

ISSN 1538 − 9472

Small Area Estimation on Zero-Inflated Data Using Frequentist and Bayesian Approach

Kusman Sadik
Bogor Agricultural University
Bogor, Indonesia

Rahma Anisa
Bogor Agricultural University
Bogor, Indonesia

Euis Aqmaliyah
Bogor Agricultural University
Bogor, Indonesia

The most commonly used method of small area estimation (SAE) is the empirical best linear unbiased prediction method based on a linear mixed model. However, it is not appropriate in the case of the zero-inflated target variable with a mixture of zeros and continuously distributed positive values. Therefore, various model-based SAE methods for zero-inflated data are developed, such as the Frequentist approach and the Bayesian approach. Both approaches are compared with the survey regression (SR) method which ignores the presence of zero-inflation in the data. The results show that the two SAE approaches for zero-inflated data are capable to yield more accurate area mean estimates than the SR method.

Keywords: Bayesian, frequentist, Markov chain Monte Carlo, small area estimation, zero-inflated data

Introduction

A small area is a subset of a population which has a small sample size with a variable of concern (Rao & Molina, 2015). That small area may be a geographic area or socio-demographic group. Nowadays, the demand for small areas statistic is increasing because that statistic is needed as regional planning material in the small area. However, very few data are available in a small area because these data are collected from a national survey. Moreover, there is a possibility that the data in a small area are not available if that small area is not represented in the national survey.

With small sample sizes, a direct estimation of small area estimation (SAE), which based on the sampling design (design-based), will yield a low precision

doi: 10.22237/jmasm/1582727606 | Accepted: October 30, 2017; Published: February 26, 2020.
Correspondence: Kusman Sadik, kusmansadik@gmail.com

estimator (Hanike et al., 2016). Meanwhile, increasing of the sample size can increase the cost, time, and labor of the national survey. SAE with an indirect estimation which is based on a model (model-based), by utilizing data from the national survey and the addition of auxiliary variables, is an alternative to that problem. Those auxiliary variables may be other variables that are related to the variable of concern (Suhartini et al., 2016; Asfar & Sadik, 2016). The variable of concern can be called the target variable.

The most commonly used indirect method of small area estimation is the empirical best linear unbiased prediction method (EBLUP), which is based on a linear mixed model (LMM) with normality assumption on the target variable (Rao & Molina, 2015). However, this model will be not appropriate if the target variable is a zero-inflated variable. The zero-inflated variable is a variable that follows the semi-continuous distribution with a mixture of zeros and continuously distributed positive values (Krieg et al., 2016). In many surveys, such as business, income, expenditure, agriculture, and ecology surveys, the observed target variables are often zero-inflated variables. For example, the expenditure of households to buy furniture in the past month, literacy proficiency of the community in an interior, and the level of consumption of illicit drugs are variables where observed values are zeros or positives.

Zero-inflation in the data can make this data tend to be skewed so that normality assumption cannot be fulfilled. Chandra and Chambers (2011b) and Karlberg (2014) explained SAE for skewed data because the existence of zero-inflation can use a mixture model, that is a mixture of log-log normal model and logistic model. Meanwhile, Chandra and Chambers (2011a) proposed three mixture model methods and compared them with the EBLUP method. However, Chandra and Sud (2012) and Pfeffermann et al. (2008) developed estimators for zero-inflated data using the frequentist approach and the Bayesian approach, respectively. Both approaches are based on two models, they are a linear mixed model (LMM) for the nonzero values of target variable and a generalized linear mixed model (GLMM) for the probability of nonzero values of target variable. Krieg et al. (2016) used both approaches and compared them with the survey regression (SR) method and the EBLUP method. The SR and the EBLUP methods ignore zero-inflation in the data. The SR method adopts design-based estimation.

The aim of this study is to review the use of the SR method and the two SAE approaches for zero-inflated data with a simulation. Four data sets were created with different proportions of zero values of each area. Then, samples were taken from each data set with various sample sizes. That sampling was repeated with various sample sizes. The objectives were to compare the frequentist with the

Bayesian approach in estimation of a small area mean with a zero-inflated target variable and to compare both approaches with the SR method, and to determine the method that yields estimator with high accuracy based on the relative root mean squared error (RRMSE).

Methodology

Generally, a population mean for area j is

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N_j} y_{ij} ,$$

where y_{ij} is the target variable for unit i in area j and N_j is the population size in area j. For all areas, $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$. Area means using direct estimation can be calculated based on information from the sample and depend on design sampling. If the samples are drawn using simple random sampling without replacement, the area mean can be estimated by

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} ,$$

where n_j is the sample size in area j (Scheaffer et al., 2006).

Survey Regression

Survey Regression (SR) is a design-based model-assisted estimator because this method adopts design-based estimation but using the auxiliary variables. Suppose for unit i in area j, p auxiliary variables $\mathbf{x}_{ij} = [1\ x_{1ij}\ \ldots\ x_{pij}]$t for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$. According to Park (2002), the area mean estimate for area j can be calculated by

$$\bar{y}_{\mathrm{SR},j} = \bar{y}_j + \left(\boldsymbol{\mu}_{x,j} - \bar{\mathbf{x}}_j\right)^{\mathrm{t}} \hat{\boldsymbol{\beta}} , \qquad (1)$$

with $\bar{y}_j$ defined as above and

$$\bar{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{ij} .$$

Here,

$$\boldsymbol{\mu}_{x,j} = \frac{1}{N_j}\sum_{i=1}^{N_{ij}}\mathbf{x}_{ij}$$

is a population mean vector of the auxiliary variables in area j and $\hat{\boldsymbol{\beta}}$ can be calculated by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y}$ based on the sample information.

Frequentist Approach

Estimation with the frequentist approach is based on the sample information and assumes the parameter as a fixed component. The first model, that is LMM, describes the distribution of nonzero values target variable.

$$y_{ij}^{*} = \mathbf{x}_{nz,ij}^{\mathrm{t}}\boldsymbol{\beta}_{nz} + \vartheta_{nz,j} + e_{ij} \tag{2}$$

for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$, where $\vartheta_{nz,j}$ is a random effect in area j that follows the normal distribution $\mathrm{N}\left(0,\sigma_{r,nz}^2\right)$ and e_{ij} is a unit-level error that follows the normal distribution $\mathrm{N}\left(0,\sigma_{e,nz}^2\right)$. Meanwhile, the second model describes the probability $p_{ij} = \mathrm{P}(y_{ij} \neq 0)$,

$$\mathrm{logit}\left(p_{ij}\right) = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mathbf{x}_{z,ij}^{\mathrm{t}}\boldsymbol{\beta}_z + \vartheta_{z,j} \tag{3}$$

or

$$p_{ij} = \frac{\exp\left(\mathbf{x}_{z,ij}^{\mathrm{t}}\boldsymbol{\beta}_z + \vartheta_{z,j}\right)}{1+\exp\left(\mathbf{x}_{z,ij}^{\mathrm{t}}\boldsymbol{\beta}_z + \vartheta_{z,j}\right)} \tag{4}$$

for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$, where $\vartheta_{z,j}$ is a random effect in area j that follows the normal distribution $\mathrm{N}\left(0, \sigma_{r,z}^2\right)$.

Model (2) is estimated by the nonzero part of the sample using restricted maximum likelihood (REML), whereas model (3) is estimated by the complete sample using maximum likelihood estimation (MLE; Bates, 2010). Therefore, y_{ij} and p_{ij} are estimated by

$$\hat{y}_{ij}^{*} = \mathbf{x}_{nz,ij}^{\mathrm{t}} \hat{\boldsymbol{\beta}}_{nz} + \hat{\vartheta}_{nz,j} \tag{5}$$

and

$$\hat{p}_{ij} = \frac{\exp\left(\mathbf{x}_{z,ij}^{\mathrm{t}} \hat{\boldsymbol{\beta}}_{z} + \hat{\vartheta}_{z,j}\right)}{1 - \exp\left(\mathbf{x}_{z,ij}^{\mathrm{t}} \hat{\boldsymbol{\beta}}_{z} + \hat{\vartheta}_{z,j}\right)}. \tag{6}$$

In this approach, the estimate for y_{ij} is $\hat{y}_{ij} = \hat{y}_{ij}^{*} \hat{p}_{ij}$ so that the estimate for area mean is

$$\bar{y}_{\mathrm{F},j} = \frac{1}{N_j} \sum_{i=1}^{N_j} \hat{y}_{ij}^{*} \hat{p}_{ij} \tag{7}$$

for $j = 1, 2,\ldots, m$.

Bayesian Approach

In parameter estimation, the Bayesian approach assumes a parameter is a random variable (Rao & D'Cunha, 2016). This approach using prior information regarding the parameter will be estimated. This prior information is called the prior distribution. Then, the sample is drawn from the population and the prior distribution is updated with the sample information so that it will be a distribution that is called posterior distribution (Casella & Berger, 2002). Some types of well-known prior distributions are, according to Gelman et al. (2014), informative prior that consisting of conjugate and nonconjugate prior, noninformative prior that consisting of proper and improper prior, and weakly informative prior.

With the Bayesian technique, models (2) and (3) can be estimated using Markov chain Monte Carlo (MCMC) simulation with a Gibbs sampling algorithm

(Jacyna & Rosen, 2016). As in the frequentist approach, model (2) is estimated by the nonzero part of the sample whereas model (3) is estimated by the complete sample. The parameter estimates obtained with REML and MLE methods in the frequentist approach can be used as the starting value of the parameter in both models. In this simulation, the length of the chains that will be built are $R = 100000$, but the first part $b = 10000$ of the chains as burn-in aren't used because they are biased. Then, those chains can be made thinner by only retaining the generated values every 90^{th} chain. This number is called a thinning interval. Therefore, $r = 1000$ iterations will be used for further analysis.

R has to be chosen sufficiently large so that the chain can converge. One of the methods that can be used for convergence inspection is a trace plot. However, this method is a graphic or explorative method that tends to be subjective. Therefore, the convergence inspection can be performed formally by hypothesis testing that is called the Geweke test (Sahlin, 2011).

Determine the prior distribution for all parameters that will be estimated. The prior distribution is according to Krieg et al. (2016) and Gelman et al. (2014), which is a weakly informative prior for regression parameters and random effect variance parameters in both models. Normal prior distribution with zero mean and variance equal to 1×10^8 for regression parameters and parameter expansion inverse chi-square prior distribution for random effect variance parameters that are implied to be a half-Cauchy prior distribution for random effect standard deviation parameters. Meanwhile, the prior distribution used for the residual variance parameter in the first model is noninformative prior, that is $\mathrm{p}\left(\sigma^2_{e,ns}\right) \propto 1/\sigma^2_{e,nz}$. This prior can be obtained with Jeffrey method from the normal distribution.

Based on MCMC simulation, the estimates for both models for unit i in area j and iteration ρ with $\rho = 1, 2,\ldots, r$ are

$$\hat{y}^*_{ij,\rho} = \mathbf{x}^{\mathrm{t}}_{nz,ij}\hat{\boldsymbol{\beta}}_{nz,\rho} + \hat{\vartheta}_{nz,j,\rho} \tag{8}$$

and

$$\hat{p}_{ij,\rho} = \frac{\exp\left(\mathbf{x}^{\mathrm{t}}_{z,ij}\hat{\boldsymbol{\beta}}_{z,\rho} + \hat{\vartheta}_{z,j,\rho}\right)}{1-\exp\left(\mathbf{x}^{\mathrm{t}}_{z,ij}\hat{\boldsymbol{\beta}}_{z,\rho} + \hat{\vartheta}_{z,j,\rho}\right)}. \tag{9}$$

Based on both model estimates, the estimate for area mean is

$$\bar{y}_{\mathrm{B},j} = \frac{1}{r}\sum_{\rho=1}^{r} Y^*_{j,\rho}\,, \tag{10}$$

with

$$Y^*_{j,\rho} = \frac{1}{N_j}\sum_{i=1}^{N_j} \hat{y}^*_{ij,\rho}\delta^*_{ij,\rho}\,, \tag{11}$$

where $\delta^*_{ij,\rho} \sim \mathrm{Be}\left(\hat{p}_{ij,\rho}\right)$ for $j = 1, 2, \ldots, m$.

Evaluation Measures of the Estimators

To evaluate the performance of the estimators consider the accuracy of how close the estimator is to the true value (Walther & Moore, 2005). One of the measures that can be used to measure the estimator accuracy is the relative root mean squared error (RRMSE) calculated by

$$RRMSE_j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\bar{y}_{j,k} - \mu_j}{\mu_j}\right)^2} \times 100\% \tag{12}$$

for $j = 1, 2, \ldots, m$, where $\bar{y}_{j,k}$ = area mean estimate that is yielded by the used method in repetition k and K is the number of repetition or number of sampling. The method that yields the area mean estimator with the highest accuracy is the method that is capable of yielding the lowest RRMSE.

The relative bias can also be used to evaluate the performance of estimators, calculated by

$$RB_j = \left|\frac{\frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_{j,k} - \mu_j\right)}{\mu_j}\right| \times 100\% \tag{13}$$

for $j = 1, 2, \ldots, m$.

Simulation Study

The finite population data is generated under the model, via R software. The data consists of two variables: the zero-inflated target variable and an auxiliary variable. The population size is N = 1000 units with the total of m = 20 areas. The area population sizes are in the range 43 to 59. Simulation scenarios were used with different proportion of zero values of each area, that is 0.35, 0.50, 0.75, and 0.90.

The steps of the model-based simulation are as follows:

1) The auxiliary variables x_{1ij} for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$ were generated from the uniform distribution U(1, 7) for each area.
2) The population values y_{ij} for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$ were generated via model (2) with regression parameters $\boldsymbol{\beta}_{nz} = [10\ 1]^t$, area random effects $\vartheta_{nz,j}$ were independently generated from the normal distribution N(0, 22), and unit level errors e_{ij} were independently generated from the normal distribution N(0, 1).
3) The probability of nonzero values p_{ij} for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$ were generated via model (4) with the same regression parameters, and area random effects $\vartheta_{z,j}$ were independently generated from the normal distribution N(0, 1).
4) Define a new variable u_{ij} for $i = 1, 2,\ldots, N_j$ and $j = 1, 2,\ldots, m$ generated from the uniform distribution U(0, 1/P) for each area, where P is the proportion of nonzero values of each area so that P = 0.65, 0.50, 0.25, and 0.10.
5) Set $\delta_{ij} = 1$ if $u_{ij} \leq p_{ij}$ and $\delta_{ij} = 0$ if $u_{ij} > p_{ij}$ so that $\delta_{ij} \sim \text{Be}(p_{ij})$.
6) The zero-inflated target variable can be obtained from $y_{ij} = y_{ij}^{*}\delta_{ij}$.

A random sample of size n = 300 was drawn repeatedly with K = 200 repetitions from every finite population using simple random sampling without replacement. Then, these samplings were repeated with a smaller sample size n = 200. All those samplings with various sample sizes were also repeated with higher repetitions, K = 500 and K = 1000.

Results

Simulation Data

The existence of zero-inflation in the target variable data can affect the shape of the data distribution. Show in Figure 1 are the histograms from all populations to show

the shape of the data distributions. Zero values in the target variable data make the shape of the data distributions not symmetric so that normality assumption in the data cannot be fulfilled. If zero is not in the data, the shape of the data distributions tends to be symmetric. The histograms indicate if the nonzero values are in the range 5 to 20.

Estimation Parameters on LMM and GLMM

The estimated parameters on LMM are intercept ($\boldsymbol{\beta}_{0,nz}$), regression coefficient ($\boldsymbol{\beta}_{1,nz}$), random effect variance $\left(\sigma^2_{r,nz}\right)$, and residual variance $\left(\sigma^2_{e,nz}\right)$, whereas the estimated parameters on GLMM are intercept ($\boldsymbol{\beta}_{0,z}$), regression coefficient ($\boldsymbol{\beta}_{1,z}$), and random effect variance $\left(\sigma^2_{r,z}\right)$. In the frequentist and Bayesian approaches, the average of every parameter estimate on LMM from all numbers of samplings have values close to the simulated values. However, this result did not happen for the average of every parameter estimate on GLMM.

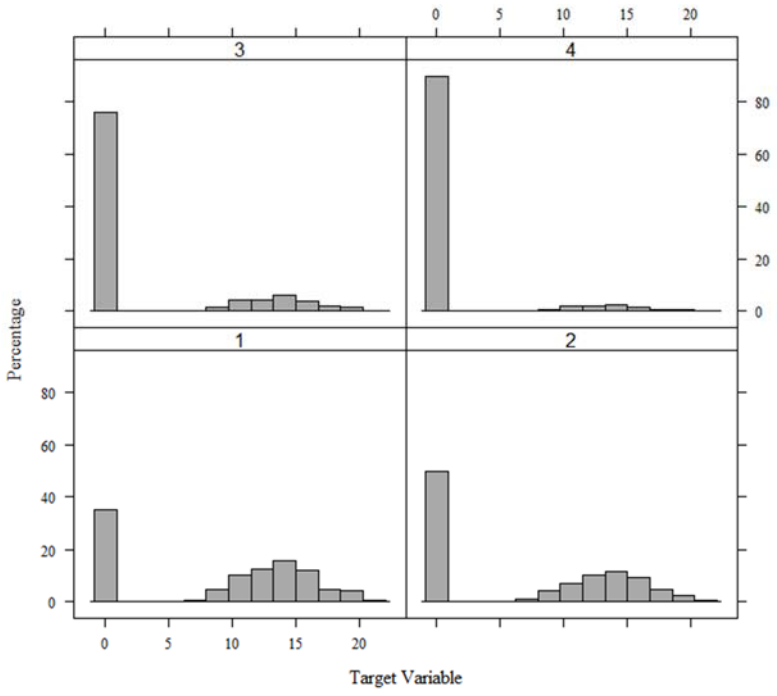

Figure 1. Histogram of the target variable

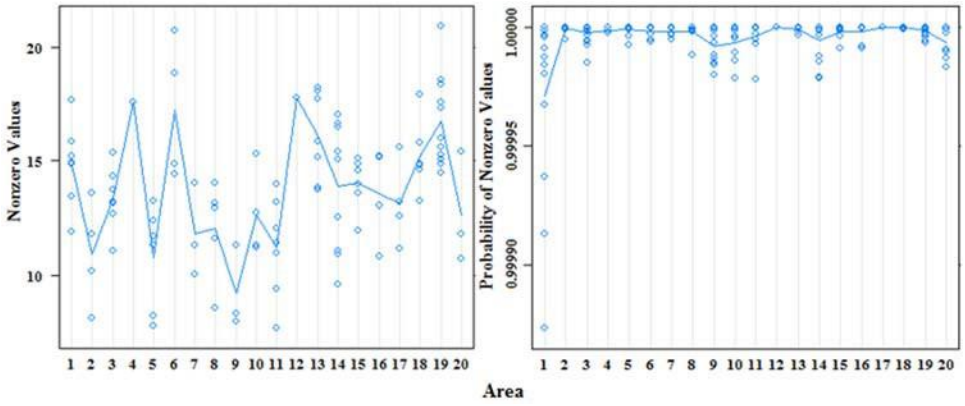

Figure 2. The variability of inter-area response on LMM (left) and GLMM (right)

The estimation of the model parameters using the frequentist approach often yields the estimate that the random effect variance parameters on GLMM are equal to zero. Although the estimate of the random effect variance parameter is equal to zero, it does not mean that there is no inter-area variability; rather this variability is relatively small compared with the inter-unit variability. As in the illustration, in the first repetition from 200 repetitions, the estimation of both models based on the drawn sample of size 200 from the population with a proportion of zero values equal to 0.50 for each area yields the estimate of random effect variance parameter on GLMM is equal to zero but the estimate of random effect variance parameter on LMM is not zero. This result is caused by the difference in the variability of inter-area response in both models, as can be seen in Figure 2.

The response on LMM is the nonzero values of the target variable whereas the response on GLMM is the probability of nonzero values of the target variable. The line in the graphic connects the averages of the response from one area to another area. The variability of inter-area response is described by the movement of the averages of the response from one area to another area. Based on Figure 2, the averages of response from one area to another area on LMM are more fluctuating than the averages of the response from one area to another area on GLMM. Therefore, the variability of inter-area response on LMM is higher than the variability of inter-area response on GLMM. Besides that, on GLMM, the averages of the response from one area to another area tend to be constant so that there is a possibility the estimation yields the estimate of random effect variance parameter is equal to zero.

In the Bayesian approach, that case can be handled by using parameter expansion inverse chi-square prior distribution for random effect variance

parameters in both models that implied be a half-Cauchy prior distribution for random effect standard deviation parameters. Therefore, the average of the estimates of the random effect variance parameter on GLMM by the Bayesian approach is higher than the average of the estimates of the random effect variance parameter on GLMM by the frequentist approach. This parameter expansion is also useful for speeding up the Markov chain convergence on the Gibbs sampling algorithm.

The Inspection of Markov Chain Convergence

In the Bayesian approach, LMM and GLMM are estimated using Markov chain Monte Carlo (MCMC) simulation with a Gibbs sampling algorithm. In this section, the inspection of Markov chain convergence that will be discussed is just from one repetition. This inspection can be performed using the exploration method by seeing the trace plot. Figure 3 shows the trace plot for all parameter estimates on LMM.

"x1" is the estimate of the regression coefficient parameter, "area" is the estimate of the random effect variance parameter, and "units" is the estimate of the residual variance parameter. The trace plot for all the estimates of parameters on LMM tend to be constant or stationary. These situations show that the burn-in process has been completed. Therefore, the Markov chain estimate for all parameters on LMM have converged.

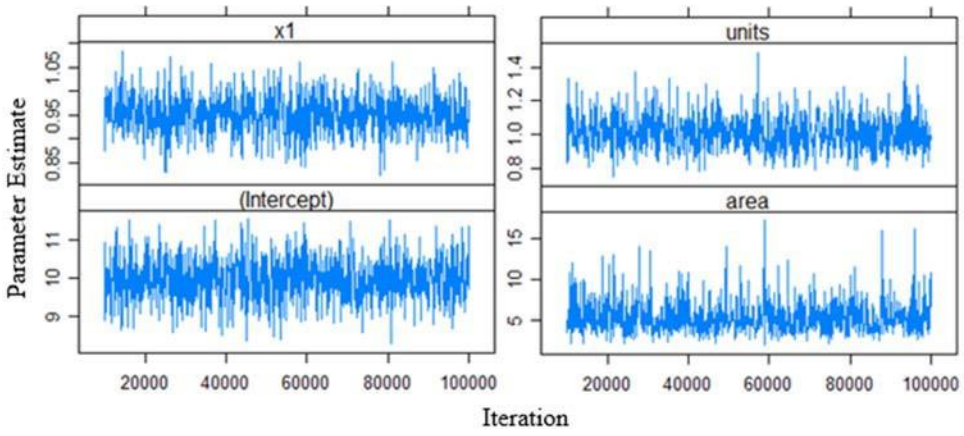

Figure 3. Trace plot for all parameter estimates on LMM

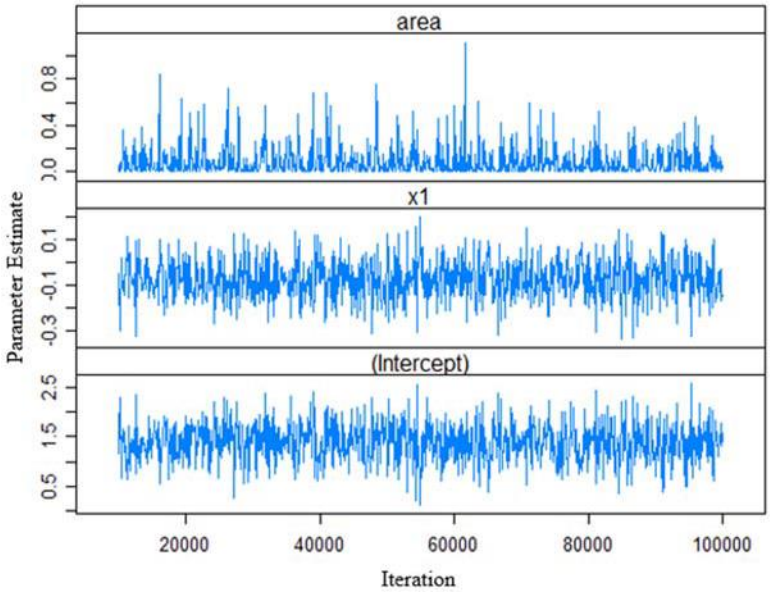

Figure 4. Trace plot for all parameters on GLMM

Table 1. Convergence inspection with the Geweke test

	Parameter	*Z*	*p*-value
LMM	$\boldsymbol{\beta}_{0,nz}$	−0.8602	0.3897
	$\boldsymbol{\beta}_{1,nz}$	1.8892	0.0589
	$\sigma^2_{r,nz}$	0.8452	0.3980
	$\sigma^2_{e,nz}$	1.1996	0.2303
GLMM	$\boldsymbol{\beta}_{0,z}$	−0.1479	0.8825
	$\boldsymbol{\beta}_{1,z}$	−0.1538	0.8778
	$\sigma^2_{r,z}$	0.2900	0.7718

The trace plot for all parameter estimates on GLMM can be seen in Figure 4. From these trace plot, the Markov chain estimate for all parameters on GLMM have converged, shown by the trace plot from the three parameters on GLMM that tend to be constant or stationary.

The inspection of Markov chain convergence also can be performed using the Geweke test. The results from this test are presented in Table 1. With α = 5%, the

absolute value of the Z statistic for all parameters in both models is not greater than $Z_{\alpha/2} = Z_{0.025} = 1.96$. This means p-values that are yielded for all parameters in both models are not smaller than α so that the decision of hypothesis testing is to not reject $H0$, with $H0$ stating that the Markov chain has converged. It can be concluded that the Markov chains in the MCMC simulations to estimate parameters on LMM and GLMM have converged with $\alpha = 5\%$.

Area Mean Estimation

Direct estimation which applies design-based estimation cannot estimate the area mean if there are no samples in that area. In the case of the zero-inflated target variable, the area mean estimate with direct estimation can be equal to zero if all drawn samples are zeroes. They are caused by the direct estimation, which only use sample information to estimate the area mean. The SR method applies design-based estimation but is model-assisted because it uses the auxiliary variable. Therefore, the SR method cannot be applied if there are no samples in that area.

With 200 repetitions, there is no area that has zero sample size so that area means can be estimated using the SR method and are compared with the two SAE approaches which take zero-inflation in the data. However, with 500 repetitions area 11 has zero sample size. This happened to the population with the proportion of zero values of each area equal to 0.75 based on the sample of size 200. The same thing happened with 1000 repetitions. There is one area that has zero sample size. That area is area 15 on the population with the proportion of zero values of each area equal to 0.90 based on the sample of size 200.

Consider next the averages of the mean estimates of each area over 200 repetitions based on samples of size 300 and 200. These averages are shown in Figures 5 and 6. Based on these figures, the averages of the area mean estimates decrease as the proportion of zero values of each area increases. The inter-area variability of the average of the mean estimate decreases as the proportion of zero values of each area increases. This can be shown by the movement of the averages of the mean estimate from one area to another area. On the population with the proportion of zero values of each area equal to 0.35, the averages of the mean estimate from one area to another area tend to be fluctuating. The averages of the mean estimate from one area to another area are more constant as the proportion of zero values of each area increases.

The averages of the mean estimate of each area that are yielded by the frequentist and the Bayesian approach have the same pattern with little differences between the two averages. However, the averages of the mean estimates of each

area that are yielded by the SR method are different from the averages of the mean estimates of each area that are yielded by the two SAE approaches. For the population with the proportion of zero values of each area equal to 0.50, the averages of the mean estimates of each area that are yielded by the three methods are almost similar.

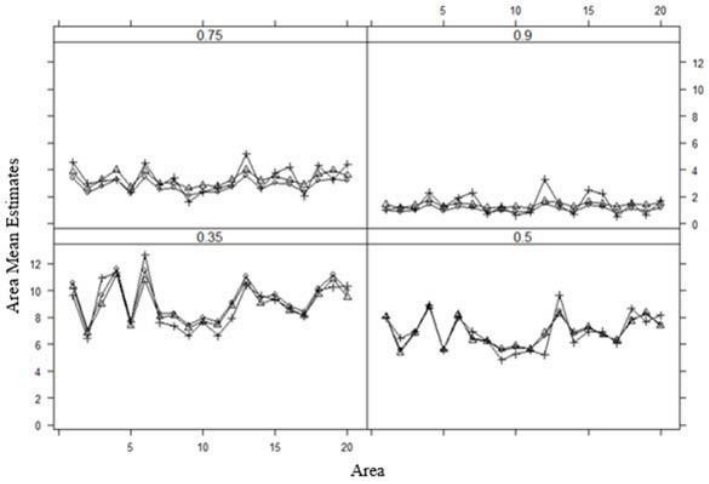

Figure 5. The average of the mean estimate of each area over 200 repetitions based on the sample of size 300; (+) SR, (∆) frequentist, and (o) Bayesian

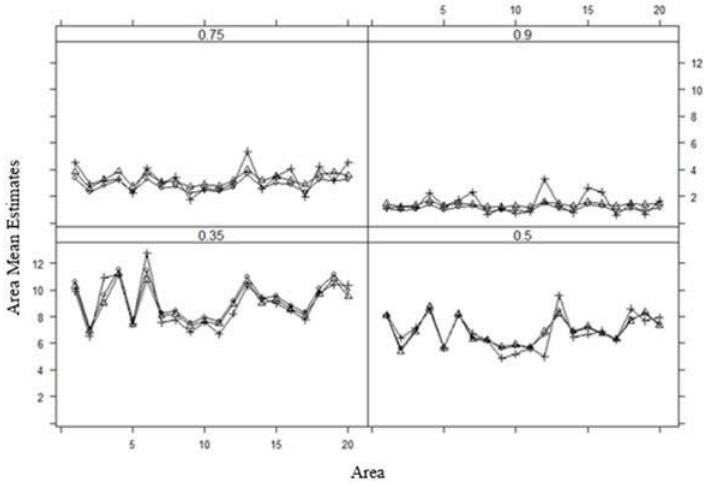

Figure 6. The average of the mean estimate of each area over 200 repetitions based on the sample of size 200; (+) SR, (∆) Frequentist, and (o) Bayesian

The averages of the mean estimates of each area over 500 and 1000 repetitions have the same results as with the averages of the mean estimates of each area over 200 repetitions and are not shown. However, in the SR method, the average of the mean estimate of area 11 over 500 repetitions on the population with the proportion of zero values of each area equal to 0.75 based on the sample of size 200 cannot be calculated. The same is true of the average of the mean estimate of area 15 over 1000 repetitions on the population with the proportion of zero values of each area equal to 0.90 based on the sample of size 200.

Comparison of Accuracy Measure of Area Mean Estimator

The accuracy of the estimators that are yielded by the SR method, the frequentist approach, and the Bayesian approach can be measured using RRMSE. Shown in Table 2 are the averages of RRMSE over 20 areas for all methods in all cases. From 500 repetitions on the population with the proportion of zero values of each area equal to 0.75 based on the sample of size 200 and from 1000 repetitions on the population with the proportion of zero values of each area equal to 0.90 based on the sample of size 200, the average of RRMSE that is yielded by the SR method over 20 areas are not available. This is caused by the existence one area that has zero sample size in every case so that area means cannot be estimated. Therefore, RRMSE of the areas and the averages of RRMSE over 20 areas cannot be calculated.

Based on Table 2, using all methods for all numbers of sampling, the averages of RRMSE increased as the proportion of zero values of each area increased. The averages of RRMSE increased as sample size decreased. In many cases, the average of RRMSE that is yielded by the SR method decreases as the number of sampling increases whereas the averages of RRMSE that yielded by both approaches increase as the number of sampling increases.

In all cases, the average of RRMSE that is yielded by the SR method is higher than the averages of RRMSE that are yielded by both approaches for zero-inflated data. The differences of the average of RRMSE that is yielded by the SR method with the averages of RRMSE that are yielded by the frequentist and the Bayesian approach are very large; the average of RRMSE that is yielded by the SR method is around two times higher than the averages of RRMSE that are yielded by the frequentist and the Bayesian approach.

Because the SR method applies design-based estimation, like in direct estimation, the high average of RRMSE is an indirect effect from the small sample size. The small sample size affects the variance of the area mean estimator directly.

That variance will be high. This means the area mean estimator that is yielded by the SR method has low precision. According to Walther and Moore (2005), precision is a variability measure of the estimator that measures how close an estimator is to the average of the estimator from repeated estimation.

If the frequentist approach and the Bayesian approach are compared, the average of RRMSE that is yielded by the frequentist approach is lower than the average of RRMSE that is yielded by the Bayesian approach on the populations with the proportion of zero values of each area not greater than 0.50. However, when the populations with the proportion of zero values of each area is greater than 0.50, the average of RRMSE yielded by the Bayesian approach tends to be lower than the average of RRMSE yielded by the frequentist approach.

Table 2. The average of RRMSE (%) of area mean estimates

Repetitions	Sample Size	Proportion of Zero Values	Method		
			SR	Frequentist	Bayesian
200	300	0.35	17.24433	8.29568	9.79165
		0.50	23.52507	11.06313	11.57329
		0.75	41.87518	22.68900	22.40839
		0.90	73.00216	48.92691	38.40262
	200	0.35	23.06639	9.64062	11.37498
		0.50	30.92323	12.45302	13.54023
		0.75	56.01238	25.62869	26.24839
		0.90	95.09355	54.04865	43.73995
500	300	0.35	17.14662	8.46099	10.02558
		0.50	22.85136	11.14344	11.60534
		0.75	42.01697	22.81530	22.71915
		0.90	71.69520	49.81725	39.35387
	200	0.35	23.02844	9.62962	11.33199
		0.50	31.08346	12.67838	13.88093
		0.75	-	25.49146	25.82837
		0.90	97.98052	54.51073	44.85474
1000	300	0.35	17.10552	8.58378	10.15960
		0.50	23.24979	11.03258	11.50636
		0.75	41.84773	22.88786	22.67891
		0.90	72.72560	49.24736	39.31888
	200	0.35	23.00702	9.66277	11.38763
		0.50	30.84812	12.53227	13.67035
		0.75	55.64881	25.45492	25.13955
		0.90	-	53.90447	44.43489

Another evaluation measure of area mean estimators is relative bias. According to Walther and Moore (2005), a good estimator is unbiased or has a low bias. The averages of relative bias over 20 areas that are yielded by the SR method and the two SAE approaches in all cases can be seen in Table 3. Like in the averages of RRMSE, the averages of relative bias also are not available for the population with the proportion of zero values of each area equal to 0.75 based on the sample of size 200 in 500 repetitions and for the population with the proportion of zero values of each area equal to 0.90 based on the sample of size 200 in 1000 repetitions. This situation is caused by the existence of areas that have zero sample size in each case so that area means cannot be estimated. Therefore, relative bias for these areas cannot be calculated and the averages of relative bias over 20 areas also cannot be calculated.

Table 3. The average of relative bias (%) of area mean estimates

Repetitions	Sample Size	Proportion of Zero Values	Method		
			SR	Frequentist	Bayesian
200	300	0.35	1.22435	6.29250	7.35980
		0.50	0.89763	7.83122	6.50540
		0.75	2.35963	18.36507	16.33679
		0.90	4.58776	39.32193	28.97403
	200	0.35	1.40807	6.45899	7.29269
		0.50	1.58011	8.03656	6.93242
		0.75	2.85852	18.60992	16.48173
		0.90	4.23996	40.57726	30.13948
500	300	0.35	0.54063	6.35326	7.50995
		0.50	0.62065	7.94392	6.65642
		0.75	1.73407	18.37823	16.35457
		0.90	1.59978	40.08095	29.99637
	200	0.35	0.80230	6.22522	7.04667
		0.50	1.01084	7.78694	6.50939
		0.75	-	18.60782	15.86594
		0.90	1.91971	40.06565	29.40904
1000	300	0.35	0.44347	6.42496	7.60598
		0.50	0.58945	7.99401	6.75530
		0.75	1.14593	18.38040	16.31381
		0.90	1.46579	39.28203	29.74171
	200	0.35	0.53514	6.33230	7.19938
		0.50	0.85549	7.78478	6.44458
		0.75	1.46986	19.18528	15.94550
		0.90	-	39.72740	29.33292

By using the SR method, in almost all cases, the averages of relative bias increase as sample size decreases. As in Ramachandran and Tsokos (2009), the bias will be near zero as the sample size increases. For all methods, the averages of relative bias have a trend to increase as the proportion of zero values of each area increases.

In almost all cases, the average of relative bias yielded by the SR method is lower than the averages of relative bias yielded by the frequentist and the Bayesian approaches. This is caused by the SR method that applies design-based estimation but is model-assisted. The SR method is generally approximately direct estimation that yields an unbiased estimator. In direct estimation, the sample mean from a sample drawn using simple random sampling is a sum of all samples divided by the sample size. Based on Ramachandran and Tsokos (2009), the sample mean is always an unbiased estimator for the population mean. In Krieg et al. (2016), the SR method yields an unbiased area mean estimator whereas the EBLUP method and both SAE approaches for zero-inflated data yield biased area mean estimators. In this research, area mean estimators that are yielded by the frequentist and the Bayesian approaches are biased estimators. This is shown by the average of their relative biases are large enough; they even reach about 40% for populations with the proportion of zero values of each area equal to 0.90. This issue can be caused by few possibilities, such as simulation procedure, the equation used to calculate area mean estimates in both approaches, or the prior distributions used in the Bayesian approach.

Although the two SAE approaches for zero-inflated data yield higher averages of relative bias than the SR method, both approaches are capable of yielding lower averages of RRMSE than the SR method. This is caused by the variance of the area mean estimators from the frequentist and the Bayesian approaches being lower than the variance of the area mean estimator from the SR method. Therefore, the area mean estimators yielded by both approaches have a high precision. Bias and variance of the area mean estimator are RMSE components. In the frequentist and the Bayesian approaches, the low variances are capable of defeating the high relative biases so that the RRMSE from those approaches are lower than the RRMSE from the SR method.

From the two SAE approaches, the Bayesian approach is capable of yielding a lower average of relative bias than the frequentist approach. However, for the population with the proportion of zero values of each area equal to 0.35, the average of relative bias yielded by the frequentist approach is lower than the average of relative bias yielded by the Bayesian approach.

Conclusion

In the zero-inflated case, the target variable is a mixture of zero values and positive values. Area means estimation using the frequentist and the Bayesian approach figure out the existence of zero-inflation in the data whereas the SR method ignores it. Besides that, the SR method is based on design-based estimation. Through simulation, with various proportions of zero values of each area and various sample sizes including various numbers of sampling, the results obtained are that the accuracy of the area mean estimators yielded by the three methods decrease as the proportion of zero values of each area increases and as sample size decreases. For a substantial case, the accuracy of the area mean estimator yielded by the SR method increases as number of sampling increases. However, the accuracy of the area mean estimators yielded by the two SAE approaches for zero-inflated data decrease as number of sampling increases.

The SR method yielded the lowest bias of area mean estimator whereas the two SAE approaches for zero-inflated data yield high bias of area mean estimators. Generally, the average of relative bias yielded by the Bayesian approach is lower than the average of relative bias yielded by the frequentist approach. However, the two SAE approaches are capable of yielding higher accuracy of the area mean estimates than the SR method. On the populations with the proportion of zero values less than 0.50, the frequentist approach is more accurate than the Bayesian approach. However, the Bayesian approach tends to be more accurate than the frequentist approach in the populations with the proportion of zero values of each area greater than 0.50.

References

Asfar, A. K., & Sadik, K. (2016). Optimum spatial weighted in small area estimation. *Global Journal of Pure and Applied Mathematics, 12*(5), 3977-3989. Retrieved from https://www.ripublication.com/gjpam16/gjpamv12n5_10.pdf

Bates, D. M. (2010). *Lme4: Mixed-effects modeling with R*. Madison, WI: Springer.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). New York, NY: Duxbury Press.

Chandra, H., & Chambers, R. L. (2011a). Small area estimation for semicontinuous data. *Biometrical Journal, 58*(2), 303-319. doi: 10.1002/bimj.201300233

Chandra, H., & Chambers, R. (2011b). Small area estimation for skewed data in presence of zeros. *Calcutta Statistical Association Bulletin, 63*(1-4), 249-252. doi: 10.1177/0008068320110113

Chandra, H., & Sud, U. C. (2012). Small area estimation for zero-inflated data. *Communication in Statistics – Simulation and Computation, 41*(5), 632-643. doi: 10.1080/03610918.2011.598991

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). London, UK: CRC Press.

Hanike, Y., Sadik, K., & Kurnia, A. (2016). Post-stratification sampling in small area estimation (SAE) model for unemployment rate estimation by Bayes approach. *AIP Conference Proceedings, 1707*(1), 080019. doi: 10.1063/1.4940876

Jacyna, G. M., & Rosen, S. L. (2016). Developing Bayesian-based confidence bounds for non-identically distributed observations using the Lyapunov condition. *Journal of Modern Applied Statistical Methods, 15*(2), 536-562. doi: 10.22237/jmasm/1478003460

Karlberg, F. (2014). Small area estimation for skewed data in the presence of zeroes. *Statistics in Transition New Series, 16*(4), 541-562. doi: 10.21307/stattrans-2015-032

Krieg, S., Boonstra, H. J., & Smeets, M. (2016). Small-area estimation with zero-inflated data – A simulation study. *Journal of Official Statistics, 32*(4), 963-986. doi: 10.1515/jos-2016-0051

Park, M. (2002). *Regression estimation of the mean in survey sampling* (Unpublished doctoral dissertation). Iowa State University, Ames, IA. Retrieved from https://lib.dr.iastate.edu/rtd/1020/

Pfeffermann, D., Terryn, B., & Moura, F. A. S. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, *34*(2), 235-249. Retrieved from https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10764-eng.pdf

Ramachandran, K. M., & Tsokos, C. P. (2009). *Mathematical statistics with applications*. Boston, MA: Academic Press.

Rao, K. A., & D'Cunha, J. G. (2016). Bayesian inference for median of the lognormal distribution. *Journal of Modern Applied Statistical Methods, 15*(2), 526-535. doi: 10.22237/jmasm/1478003400

Rao, J. N. K., & Molina, I. (2015). Small area estimation (2nd ed.). New York, NY: John Wiley & Sons, Inc. doi: 10.1002/9781118735855

Sahlin, K. (2011). *Estimating convergence of Markov chain Monte Carlo simulations* (Unpublished Master's thesis). Stockholm University, Stockholm, Sweden. Retrieved from https://www2.math.su.se/matstat/reports/master/2011/rep2/report.pdf

Scheaffer, R. L., Mendenhall, W., Ott, L. R., & Gerow, K. G. (2006). *Elementary survey sampling* (6th ed.). Boston, MA: Duxbury Press.

Suhartini, T., Sadik, K., & Indahwati. (2016). Small area estimation (SAE) model: Case study of poverty in West Java Province. *AIP Conference Proceedings, 1707*(1), 080016. doi: 10.1063/1.4940873

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography, 28*(6), 815-829. doi: 10.1111/j.2005.0906-7590.04112.x