Mathematics and Statistics 11(3): 579-585, 2023
DOI: 10.13189/ms.2023.110315 http://www.hrpub.org

Small Area Estimation of Illiteracy Rates based on Beta-Binomial Model using Hierarchical Likelihood Approach

Etis Sunandi[1,2], Khairil Anwar Notodiputro[1,*], Indahwati[1], Agus M Soleh[1]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Indonesia

Received December 31, 2022; Revised April 20, 2023; Accepted May 9, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] Etis Sunandi, Khairil Anwar Notodiputro, Indahwati, Agus M Soleh , "Small Area Estimation of Illiteracy Rates based on Beta-Binomial Model using Hierarchical Likelihood Approach," Mathematics and Statistics, Vol. 11, No. 3, pp. 579 - 585, 2023. DOI: 10.13189/ms.2023.110315.*

(b): *Etis Sunandi, Khairil Anwar Notodiputro, Indahwati, Agus M Soleh (2023). Small Area Estimation of Illiteracy Rates based on Beta-Binomial Model using Hierarchical Likelihood Approach. Mathematics and Statistics, 11(3), 579 - 585. DOI: 10.13189/ms.2023.110315.*

Abstract Small Area Estimation (SAE) is a statistical method used to estimate parameters in sub-populations with small samples. This study aims to develop a Beta-Binomial model on SAE with a Hierarchical Likelihood (HL) approach. The model built is called the SAE-BB-HL model. This research begins by deriving a formula for estimating model parameters analytically. A good fit is calculated with the Mean Square Error of Prediction (MSEP) and bias. This study used simulation data and data from the National Socio-Economic Survey (SUSENAS) and Village Potential (PODES) of Bengkulu Province for 2021 collected by Statistics Indonesia (BPS). The simulation study aims to evaluate the SAE-BB-HL model. Simultaneously, the application study aims to predict the illiteracy rate per sub-district in Bengkulu Province. The simulation study results show that the parameter estimates of random area distribution are very close to the actual parameters. It also reveals that the bias and MSEP estimates of the proportion of HL are lower than the direct estimates. In addition, the results of this study show that the SAE-BB-HL model can improve the accuracy and precision of proportion estimation. Applying the SAE-BB_HL model to real data shows that the predictive value of the illiteracy rate tends to be higher when compared to the direct estimator.

Keywords Binary Response, Mean Square Error, Overdispersion, Small Sample

1. Introduction

Small Area Estimation (SAE) is a statistical method for estimating parameters in a subpopulation based on a small number of samples to provide estimates with adequate precision [1]. Small area models are classified into two types based on the type of data available. The first is the area-level model. This model connects the direct estimator and the auxiliary area variable. The second model is a unit-level model that connects the unit values of the estimator directly to the unit level of auxiliary variables with known area mean values and certain area auxiliary variables. The SAE model can be applied to both continuous and discrete response variables such as binary. Research using binary data is prone to overdispersion. The Beta-Binomial distribution can be used to overcome the problem of overdispersion in binary data [2]. Beta-Binomial model development on small area estimation has been carried out. The model parameters are estimated through the Bayesian approach.

According to [3], parameter estimation in a mixed model with a Beta-Binomial hierarchy can be done through Hierarchical Likelihood (HL). This method is claimed to be better than the Bayes approach analytically. The HL method can reduce the bias of estimating binary data

parameters, which is a problem in the Generalized Linear Mixed Model (GLMM).

Based on the description above, this paper aims to develop a Beta-Binomial model for small area estimation. Parameter estimation is carried out using the Hierarchical Likelihood approach. The developed model is called the SAE-BB-HL model. Evaluation of the model used simulation data. Then, this model was applied to predict the proportion of illiteracy at the sub-district level in Bengkulu Province, Indonesia.

2. Literature Review

2.1. Small Area Estimation (SAE)

A small area is a subset of the population whose sample size is small with a variable of concern. Parameter estimation in a small area can be done by direct or indirect estimation. Direct estimation is an estimation based on sample data from the area. The result of direct estimation in a small area is an unbiased estimator but has a significant variance. This is because the estimated parameters are obtained from small samples. Therefore, the indirect estimation method is more appropriate for utilizing the strength of the surrounding area and data sources inside or outside the area whose statistics are to be obtained [4]. A small area estimation model has been developed by several researchers [5], [6].

Based on the basic model, the small area model consists of two types: based on the area level and the unit level. The area-level-based model is a General Mixed Linear Model, which Fay-Herriot introduced. This model is based on supporting data that only exists for a specific area level. Let $z_i = (z_{1i}, z_{2i}, \dots, z_{pi})'$and the parameters to be assumed are θ_i, assumed to have a linear relationship with z_i. The supporting data is used to build the following model:

$$\boldsymbol{\theta_i = z'_i \beta_z + v_i}, \text{ for } \boldsymbol{i = 1,\dots,m},\ \boldsymbol{y_i = \theta_i + \varepsilon_i} \quad (1)$$

Where $v_i \sim N(0, \sigma_v^2)$ as a random effect the *ith* area error is assumed to be normally distributed. Meanwhile, θ_i, it can be known by assuming that the estimator is direct y_i *w*ith sampling error $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and σ_ε^2 is known [1].

The model that can be used for binary response variable data is a two-level model, namely Beta-Binomial. A small area estimation model for Binomial response has been developed by several researchers. The Beta-Binomial Model can be written as follows [1]:

$$\begin{aligned} &(i)\ \ y_i|p_i \sim iid\ \ Binomial(n_i, p_i), \\ &(ii)\ \ p_i|\alpha,\gamma \sim iid\ \ Beta(\alpha,\gamma\ ), \\ &\alpha > 0, \gamma > 0, \\ &(iii)\ \ \alpha,\gamma \text{ are mutually independent} \end{aligned} \quad (2)$$

2.2. Hierarchical Likelihood (HL) Method

Generalized Linear Mixed Model (GLMM) is defined as $E(y|v) = X\beta + Zv$, where X is a covariate matrix of fixed parameter, β is a vector fixed parameter, Z is a covariate matrix of random parameter, and v is a vector random effect. GLMM is extended to HGLM by accommodating a non-normal distribution for random effects [7]. HGLM uses HL to avoid complex integration. The HL formula can be written as follows:

$$\boldsymbol{h = logl_1(\beta, \phi; y|v) + logl_2(\lambda; v)} \quad (3)$$

Where $l_1(\beta, \phi; y|v)$ and $l_2(\lambda; v)$ are conditional probability functions $y|v$ and probability functions v. Vector β is a canonical parameter, (ϕ, λ) denotes the dispersion parameter, and λ is the distribution parameter v. Let $v(.)$ be the corresponding link function that defines HL such that $v = v(u)$. In the HL form, the selection of the random effect scale is essential in that the scale change requires Jacobian adjustment. Suppose that $v = v(u)$ is used to show the scale in which the random effect of v is assumed to be linear, and the predictor is linear. Using the score function of HL, the parameters β and v are estimated as follows [8]:

$$\boldsymbol{\frac{\partial h}{\partial \beta} = 0, \frac{\partial h}{\partial v} = 0} \quad (4)$$

Furthermore, the estimated dispersion parameter $\hat{\tau} = (\hat{\phi}, \hat{\sigma}_v^2)$ is the maximization solution of the APHL function. APHL, $p_{v,\beta_z}\ (h)$, which is defined as follows:

$$\boldsymbol{p_{\beta,v}(h) = \left(h + \frac{1}{2}\log(2\pi H^{-1})\right)\Big|_{\beta=\hat{\beta}, v=\hat{v}}} \quad (5)$$

Where $\boldsymbol{H = \begin{bmatrix} X'W_1X & X'W_1Z \\ Z'W_1X & Z'W_1Z + W_2 \end{bmatrix}}$, $\boldsymbol{W_1 = \left(\frac{\partial \mu}{\partial \eta}\right)^2 (\phi V(\mu))^{-1}}$, and $\boldsymbol{W_2 = -\frac{\partial^2 h_1}{\partial v^2}}$.

3. Methodology

This research uses simulation and real data. In the simulation study, the response variable data was generated $y_i \sim BB(n_i, p_i, \phi)$, $i = 1,2,\dots,m = 50$. Initial values for the covariates and parameters considered based on previous research [9]–[12].

This study applied the SAE model with a variance for the random effect of $\sigma_v^2 = 2$. The value follows the research of [13]. Meanwhile, the initial value of the influence parameter is fixed for as many as three variables $(p = 3)$ with the initial value: $\beta_Z \epsilon (1,1,1)$ and $\beta_0 = 0$. The number of household samples per census block in area-i, $n_{i(k)}$, is 2%. The simulation has three main steps. There are population generation, sampling, and analysis. The simulation is repeated $B = 100$ times and then it is carried out

$$Bias\ \ (\hat{p}_i^{HL}) = p_i\ - E(\hat{p}_i^{HL})\ \text{ with } E(\hat{p}_i^{HL}) = \frac{\sum_{b=1}^{B} \hat{p}_i^{HL(b)}}{B}$$

$$MSEP(\hat{p}_i^{HL}) = \frac{1}{B}\sum_{b=1}^{B}\left(p_i - E(\hat{p}_i^{HL})\right)^2$$

The actual data comes from Statistics of Bengkulu Province, The National Socio-economic Survey (SUSENAS), and Potential Village (PODES) 2021. This study made predictions on the proportion of illiteracy at the sub-district level in Bengkulu Province in 2021. The variables used can be seen in Table 1.

Table 1. The Research Variables

	Variable	Code	Source
Response	The proportion of illiteracy	BBH	SUSENAS 2021
Auxiliary Variables	Average recipient letter description no able (SKTM)	Z1	PODES 2021
	The number of malnutrition	Z2	PODES 2021
	The number of facility health	Z3	PODES 2021
	The number of activities eradicating illiteracy	Z4	PODES 2021
	The number of facility education	Z5	PODES 2021

4. Result and Discussions

4.1 Development Model

This research begins with developing a model called the SAE-BB-HL model. The models used as the basis for the development of the SAE-BB-HL model are the Fay-Herriot Area Level SAE model in (1) and the Beta-Binomial Model (2) [14]. The SAE-BB-HL model is a development of the compilation of the Fay-Herriot model and the Beta-Binomial Model, which is defined as follows:

$$\begin{aligned} &y_i|p_i \sim Binomial(n_i, p_i),\ i = 1, \dots, m \\ &p_i|v_i \sim Beta(\alpha, \gamma),\ \alpha > 0, \gamma > 0 \\ &\log\left(\frac{p_i}{1-p_i}\right) = \xi_i = \boldsymbol{z_i'\beta_z} + v_i \end{aligned} \tag{6}$$

Where y_i is the response variable of the i-th area, the parameter p_i is the proportion of the i-area. At the same time, v_i is an i-area random effect assuming $v_i \sim N(0, \sigma_v^2)$. The covariate z_i of size $p \times 1$ is a constant covariate. The parameter $\boldsymbol{\beta}_z$ is a fixed effect parameter of size $p \times 1$. Then v_i, and $y_i|p_i$ are assumed to be independent. Parameters γ are assumed to be fixed.

The random response variable $\boldsymbol{y} = (y_1, \dots, y_m)$ has a Beta-Binomial distribution that can be written as follows:

$$\begin{aligned} &\boldsymbol{f(y|\phi, \beta_z, v)} \\ &= \prod_{i=1}^{m}\left\{\begin{array}{l} \binom{\boldsymbol{n_i}}{\boldsymbol{y_i}}\left(\frac{\boldsymbol{1}}{\boldsymbol{\phi}}\right)^{\boldsymbol{y_i}}\prod_{\boldsymbol{k=0}}^{\boldsymbol{y_i-1}}[\boldsymbol{p_i + k\phi}]\left(\frac{\boldsymbol{1}}{\boldsymbol{\phi}}\right)^{\boldsymbol{n_i-y_i}} \\ \prod_{\boldsymbol{k=0}}^{\boldsymbol{n_i-y_i-1}}[\boldsymbol{1-p_i+k\phi}]\left(\frac{\boldsymbol{1}}{\boldsymbol{\phi}}\right)^{-\boldsymbol{n_i}} \\ \prod_{\boldsymbol{k=0}}^{\boldsymbol{n_i-1}}[\boldsymbol{1+k\phi}]^{-\boldsymbol{1}} \end{array}\right\} \end{aligned} \tag{7}$$

Where $\boldsymbol{\beta}_z$ implicit in p_i with equation $p_i = \frac{\exp(z_i'\boldsymbol{\beta}_z + v_i)}{1+\exp(z_i'\boldsymbol{\beta}_z + v_i)}$

The log-likelihood function according to Equation (7) is as follows:

$$\begin{aligned} &l(\boldsymbol{\beta}_z, \phi|\boldsymbol{v}, \boldsymbol{y}) \approx \log f(\boldsymbol{y}|\phi, \boldsymbol{\beta}_z, \boldsymbol{v}) \\ &= \sum_{i=1}^{m}\left[\begin{array}{c} \log\left\{\binom{n_i}{y_i}\right\} \\ +\sum_{k=0}^{y_i-1}\log(p_i + k\phi) \\ +\sum_{k=0}^{m_i-y_i-1}\log(1-p_i+k\phi) \\ -\sum_{k=0}^{m_i-1}\log(1+k\phi) \end{array}\right] \end{aligned} \tag{8}$$

The random effect area $\boldsymbol{v} = (v_1, \dots, v_m)$ has a normal distribution with a vector zero expected value and covariance $\sigma_v^2\boldsymbol{I}$. The PDF is as follows:

$$\boldsymbol{f(v|\sigma_v^2)} = \frac{\boldsymbol{1}}{\sqrt{\boldsymbol{2\pi\sigma_v^2}}}\boldsymbol{e}^{-\frac{\boldsymbol{1}}{\boldsymbol{2\sigma_v^2}}\boldsymbol{v'v}} \tag{9}$$

The log-likelihood function according to (9) is as follows:

$$\boldsymbol{l(\sigma_v^2|v)} \approx \boldsymbol{\log f(v|\sigma_v^2) = \log\left([2\pi\sigma_v^2]^{-\frac{1}{2}}\right) - \frac{v'v}{2\sigma_v^2}} \tag{10}$$

Using (3), (8), and (9), the HL function can be written as follows [15], [16]:

$$\begin{aligned} &h = \ell(\boldsymbol{\beta}_z, \phi, \sigma_v^2|\ \boldsymbol{y}, \boldsymbol{v}) \\ &= log\, f(\boldsymbol{y}|\phi, \boldsymbol{\beta}_z) + log\, f(\boldsymbol{v}|\sigma_v^2) \\ &= \sum_{i=1}^{m}\left[\begin{array}{c} \log\left\{\binom{n_i}{y_i}\right\} \\ +\sum_{k=0}^{y_i-1}\log(p_i + k\phi) \\ +\sum_{k=0}^{n_i-y_i-1}\log(1-p_i+k\phi) \\ -\sum_{k=0}^{n_i-1}\log(1+k\phi) \end{array}\right] \\ &+\log\left([2\pi\sigma_v^2]^{-\frac{1}{2}}\right) - \frac{\boldsymbol{v'v}}{2\sigma_v^2} \end{aligned} \tag{11}$$

Estimating the fixed parameter $\boldsymbol{\beta_z}$ and the random variable $\boldsymbol{v}$ SAE-BB-HL model is done by maximizing the HL function, which is to find the first derivative h with respect to $\boldsymbol{\beta_z}$ and $\boldsymbol{v}$ by solving $\partial h/\partial \boldsymbol{\beta_z} = 0$ and $\partial h/\partial \boldsymbol{v} = 0$. However, the derivative does not have a closed form, so it will be challenging to estimate manually. Therefore, getting the estimated parameters $\boldsymbol{\beta_z}$ and $\boldsymbol{v}$ is done using the Delta iteration method.

Based on (11), let $h = h_0 + h_1$

$$h_0 = \sum_{i=1}^{m}\left[\log\left\{\binom{n_i}{y_i}\right\} + \sum_{k=0}^{y_i-1}\log(p_i + k\phi) + \sum_{k=0}^{n_i-y_i-1}\log(1-p_i+k\phi) - \sum_{k=0}^{n_i-1}\log(1+k\phi)\right]$$

And $h_1 = \log\left([2\pi\sigma_v^2]^{-\frac{1}{2}}\right) - \frac{\boldsymbol{v'v}}{2\sigma_v^2}$

From the results of the maximization of the HL function, the score function of the fixed parameters and random effect is obtained as follows:

$$\boldsymbol{u'}_{\boldsymbol{\beta_z}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta_z}} = \frac{\partial \boldsymbol{h_0}}{\partial \boldsymbol{\beta_z}} + \frac{\partial \boldsymbol{h_1}}{\partial \boldsymbol{\beta_z}} = \boldsymbol{\zeta' SZ} \quad (12)$$

$$\boldsymbol{u'}_{\boldsymbol{v}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{v}} = \frac{\partial \boldsymbol{h_0}}{\partial \boldsymbol{v}} + \frac{\partial \boldsymbol{h_1}}{\partial \boldsymbol{v}} = \boldsymbol{\zeta' S} - \frac{\boldsymbol{v'}}{\boldsymbol{\sigma_v^2}} \quad (13)$$

Where $\boldsymbol{\zeta'} = \frac{\partial h_0}{\partial \boldsymbol{p}} = \left(\frac{\partial h_0}{\partial p_1}, \frac{\partial h_0}{\partial p_2}, \ldots, \frac{\partial h_0}{\partial p_m}\right)$, $\boldsymbol{S} = diag[p_i(1-p_i)]$, and $\boldsymbol{Z}$ is a $m \times p$ covariate matrix.

The Delta algorithm defines the Hessian matrix of the model as an approximation procedure of the second derivatives of the log-likelihood. The formula is as follows:

$$\boldsymbol{H} = \begin{bmatrix} E\left\{-\frac{\partial^2 h}{\partial \boldsymbol{\beta_z'}\partial \boldsymbol{\beta_z}}\right\} & E\left\{-\frac{\partial^2 h}{\partial \boldsymbol{\beta_z'}\partial \boldsymbol{v}}\right\} \\ E\left\{-\frac{\partial^2 h}{\partial \boldsymbol{v'}\partial \boldsymbol{\beta_z}}\right\} & E\left\{-\frac{\partial^2 h}{\partial \boldsymbol{v'}\partial \boldsymbol{v}}\right\} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z'SUSZ} & \boldsymbol{Z'SUS} \\ \boldsymbol{SUSZ} & \boldsymbol{WS + D^{-1}} \end{bmatrix} \quad (14)$$

Based on (12), (13), and (14), the estimator (β_z, v) is obtained through iteration as follows

$$\begin{pmatrix} \boldsymbol{\beta_z} \\ \boldsymbol{v} \end{pmatrix}^{(\boldsymbol{r+1})} = \begin{pmatrix} \boldsymbol{\beta_z} \\ \boldsymbol{v} \end{pmatrix}^{(\boldsymbol{r})} + \begin{bmatrix} \boldsymbol{Z'SUSZ} & \boldsymbol{Z'SUS} \\ \boldsymbol{SUSZ} & \boldsymbol{WS + D^{-1}} \end{bmatrix}^{-\boldsymbol{1}} \begin{pmatrix} \boldsymbol{u_{\beta_z}} \\ \boldsymbol{u_v} \end{pmatrix} \quad (15)$$

Estimation of the dispersion parameter, σ_v^2, in the SAE-BB-HL model is carried out by maximizing the APHL function in (5) by solving $\partial p_{\beta,v}(h) \ /\partial \sigma_v^2 = 0$. Estimating this parameter is carried out assuming that β and v are constant. Using iteratively the adjusted profile estimator of $\boldsymbol{\sigma_v^2}$ is obtained

$$\boldsymbol{\hat{\sigma}_v^2} = \frac{\boldsymbol{1}}{\boldsymbol{m}}(\boldsymbol{\hat{v}'\hat{v} + G}) \quad (16)$$

Where

$G = \boldsymbol{tr}([\boldsymbol{WS + D^{-1} - SUSZ(Z'SUSZ)^{-1}Z'SUS}]^{-\boldsymbol{1}})$.

Overdispersion parameter estimation is done by using the moment method. This method is modified from the Klainman method which has been developed by [1].

$$\hat{\phi} = \left(1 + \frac{(n-1)}{1+\hat{\alpha}+\hat{\gamma}}\right) \quad (17)$$

Where $\hat{\alpha} = \hat{p}\left[\frac{\hat{p}(1-\hat{p})[n-\sum_{i=1}^{m} n_i^2-(m-1)]}{n s_p^2 - \hat{p}(1-\hat{p})(m-1)} - 1\right]$, $\hat{\gamma} = \hat{p}\left[\frac{\hat{p}(1-\hat{p})[n-\sum_{i=1}^{m} n_i^2-(m-1)]}{n s_p^2 - \hat{p}(1-\hat{p})(m-1)} - 1\right]\left[\frac{1}{\hat{p}} - 1\right]$,

$\hat{p} = \sum_{i=1}^{m} \frac{n_i}{n}\hat{p}_i^{DE}$ with $\hat{p}_i^{DE}$ *is direct estimator*, $n = \sum_{i=1}^{m} n_i$, and $s_p = \sum_{i=1}^{m} \frac{n_i}{n}(\hat{p}_i^{DE} - \hat{p})^2$

Prediction of the proportion of the SAE-BB-HL model is made through the formula

$$\boldsymbol{\hat{p}_i^{HL}} = \frac{\boldsymbol{\exp(z_i'\hat{\beta}_z + \hat{v}_i)}}{\boldsymbol{1 + \exp(z_i'\hat{\beta}_z + \hat{v}_i)}} \quad (18)$$

4.2. Simulation Results

We conducted a simulation study to evaluate the performance of the SAE-BB-HL model. The simulation was carried out with 100 replications. The objectives of this simulation study were threefold: (a) to analyze the bias of the parameter estimators, (b) to analyze the prediction bias of proportions and MSEP, and (c) to compare the HL estimate ($\hat{p}_i^{HL}$) and the direct estimate ($\hat{p}_i^{DE}$). This SAE-BB-HL modeling uses the R programming language.

The results of the estimation of fixed parameters of SAE-BB-HL model through HL are summarized in Table 2. Based on the table, it can be seen that $\boldsymbol{\hat{\beta}_Z} = (-7.2, -0.01, 1.1, -0.01)$. It is known that the initial value of the fixed effect parameter is $\boldsymbol{\beta_Z} = (0,1,1,1)$. It appears that $\hat{\beta}_Z$ is biased. At the same time, the dispersion parameter on the random variable v is $\hat{\sigma}_v^2 = 2.05$. When compared with the initial value of the dispersion parameter set, namely $\sigma_v^2 = 2$, it can be said that the assumption is almost unbiased. The overdispersion parameter estimator is $\hat{\phi} = 1$. It can be seen that the estimation of the overdispersion parameter using the moment method is biased. This study also carried out the proportion estimation in a small area. We compared two estimators: the direct estimate ($\hat{p}_i^{DE}$) and the SAE-BB-HL estimate ($\hat{p}_i^{HL}$). These results show that the proportion estimator ($\hat{p}_i^{HL}$) is almost the same as the parameter. The estimator of the proportion ($\hat{p}_i^{HL}$) from several areas has the same tendency. In other words, the SAE-BB-HL model has good flexibility.

Figure 1 shows the bias values of the direct and model estimators. Both estimators have the same bias value tendency. The comparison of the two values is entirely accurate. The HL estimator has a lower bias value than the DE estimator in all areas. The bias value of the HL estimator is relatively minimal, ranging between 0.0088 and 0.056. From this value, it can be said that the HL proportion estimator is almost unbiased. However, it

appears that the HL estimator tends to be slightly overestimated.

Figure 2 shows the MSEP values of the direct estimator (MSEP $\hat{p}_i^{DE}$) and the HL estimator (MSEP $\hat{p}_i^{HL}$). In the majority, the MSEP value of the HL estimator was lower than that of the direct estimator of MSEP. From this value and the bias value of the HL estimator, which is smaller than the direct estimator, it can be said that the SAE-BB-HL model can improve the precision of proportion estimation.

Table 2. Fixed effects coefficients

	Estimate	Parameter
(Intercept)	-7.20	0
z1	-0.01	1
z2	1.10	1
z3	-0.01	1

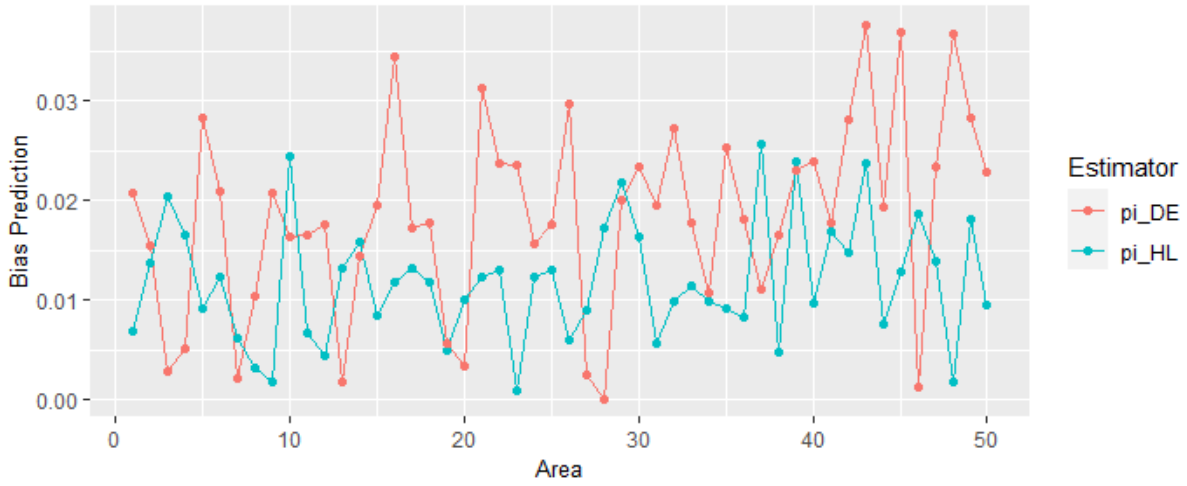

Figure 1. Bias proportion of direct estimation ($\hat{p}_i^{DE}$) and SAE-BB-HL estimation ($\hat{p}_i^{HL}$)

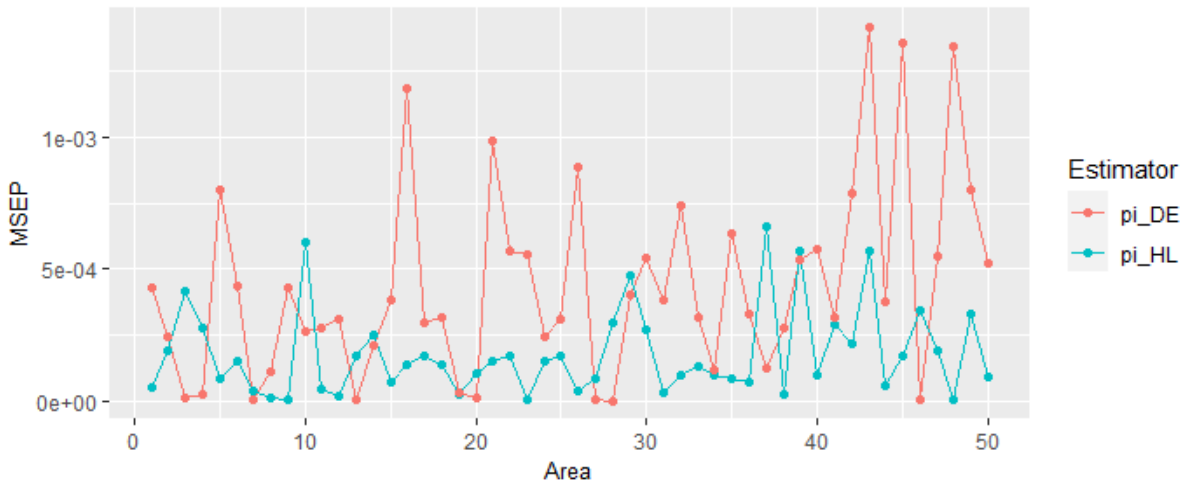

Figure 2. MSEP proportion of direct estimation ($\hat{p}_i^{DE}$) and SAE-BB-HL estimation ($\hat{p}_i^{HL}$)

Figure 3. Direct estimates $(\hat{p}_i^{DE})$ and SAE-BB-HL estimates $(\hat{p}_i^{HL})$of illiteracy prediction

4.3. Prediction of Illiteracy Proportion per Sub-district in Bengkulu Province

The SAE-BB-HL model predicts illiterate data at the sub-district level in Bengkulu Province. The prediction results are presented in Figure 3, with predictions from direct estimators as a comparison. Figure 3 shows that the SAE-BB-HL proportion estimator has the same predictable pattern as the direct estimator. The SAE-BB-HL model has enough flexibility. In addition, the predictive value of the SAE-BB-HL model tends to be more significant when compared to the direct estimator.

In Bengkulu Province, there are 128 sub-districts divided into ten cities/regencies. Predictions of the proportion of illiteracy vary across districts. Bengkulu City has the majority of sub-districts with lower illiteracy prediction values compared to other Regencies. This is following the situation that the city of Bengkulu is a city center in Bengkulu Province which has better education and health facilities. The economy in Bengkulu City is also quite good. This proves that the city of Bengkulu is an office and shopping center with adequate facilities and infrastructure.

On the other hand, several sub-districts were predicted to have illiterate people by direct estimators. However, through the SAE-BB-HL model, it was predicted that there were illiterate people in these sub-districts. On average, according to the predictions of the SAE-BB-HL model, the direct estimator also predicts that Bengkulu City has a minor proportion of illiterates compared to other regencies.

5. Conclusions

This paper revealed that the prediction bias and MSEP resulting from the SAE-BB-HL model were lower than that of the direct estimates. This implied that the SAE-BB-HL model tended to increase the precision and accuracy of proportion estimates. However, the simulation study showed that the estimates of fixed parameters, as well as the overdispersion parameter, were biased. Further research is required to develop an estimation method to increase the precision and accuracy of the SAE-BB-HL model. Based on the empirical data it was also revealed that the predictive value of the illiteracy rate at the sub-district level in Bengkulu Province was greater than the direct estimate. This empirical evidence amplified the importance of future research toward increasing the accuracy and precision of the SAE-BB-HL model.

Acknowledgments

We would like to thank all those who have helped provide valuable suggestions for improving the quality of this paper and our gratitude to the Ministry of Education and Culture of the Republic of Indonesia for National Postgraduate Education Scholarships (BPP-DN) 2019 and funding the 2022 Doctoral Dissertation Research. Thanks also to Statistics Indonesia (BPS) for The SUSENAS and PODES data.

REFERENCES

[1] J. N. K. Rao and I. Molina, Small Area Estimation: Second Edition. 2015.

[2] X. A. Harrison, “A comparison of observation-level randomeffect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution,” PeerJ, vol. 2015, no. 7, 2015, doi: 10.7717/peerj.1114.

[3] I. L. Do Ha, Y. Lee, and J. K. Song, “Hierarchical likelihood

approach for frailty models," Biometrika, vol. 88, no. 1, 2001, doi: 10.1093/biomet/88.1.233.

[4] J. N. K. Rao, Small Area Estimation. Canada: John Wiley & Sons, Inc., 2003.

[5] E. Tanur and A. Kurnia, "Small area estimation for autoregressive model with measurement error in the auxiliary variable," Commun. Math. Biol. Neuros, vol. 83, no. 1, 2022, doi: https://doi.org/10.28919/cmbn/7577.

[6] N. H. Pusponegoro, A. Kurnia, K. A. Notodiputro, A. M. Soleh, and E. T. Astuti, "Area specific effects selection of small area estimation for construction of regional consumer price indices in Indonesia," Commun. Math. Biol. Neuros, vol. 90, no. 1, 2021, doi: https://doi.org/10.28919/cmbn/6747.

[7] Y. Lee, J. A. Nelder, and Y. Pawitan, Generalized linear models with random effects: Unified analysis via H-likelihood. 2006.

[8] J. Jiang, Linear and Generalized Linear Mixed Models and Their Applications. New York: Springer, 2007.

[9] W. R. Bell, H. C. Chung, G. S. Datta, and C. Franco, "Measurement error in small area estimation : Functional versus structural versus naïve models," Surv. Methodol., vol. 45, no. 1, pp. 61–80, 2019.

[10] J. P. Burgard, M. D. Esteban, D. Morales, and A. Pérez, "A Fay–Herriot model when auxiliary variables are measured with error," Test, vol. 29, no. 1, pp. 166–195, 2020, doi: 10.1007/s11749-019-00649-3.

[11] J. P. Burgard, M. D. Esteban, D. Morales, and A. Pérez, "Small area estimation under a measurement error bivariate Fay–Herriot model," Stat. Methods Appl., vol. 30, no. 1, pp. 79–108, 2021, doi: 10.1007/s10260-020-00515-9.

[12] J. Najera-Zuloaga, D. J. Lee, and I. Arostegui, "A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes," Biometrical J., vol. 61, no. 3, pp. 600–615, 2019, doi: 10.1002/bimj.201700251.

[13] L. M. R. Ybarra and S. L. Lohr, "Small area estimation when auxiliary information is measured with error," Biometrika, vol. 95, no. 4, pp. 919–931, 2008, doi: 10.1093/biomet/asn048.

[14] J. G. Skellam, "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable between the Sets of Trials," J. R. Stat. Soc. Ser. B, vol. 10, no. 2, 1948, doi: 10.1111/j.2517-6161.1948.tb00014.x.

[15] R. L. Prentice, "Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors," J. Am. Stat. Assoc., vol. 81, no. 394, 1986, doi: 10.1080/01621459.1986.10478275.

[16] M. I. Lora and J. M. Singer, "Beta-binomial/gamma-poisson regression models for repeated counts with random parameters," Brazilian J. Probab. Stat., vol. 25, no. 2, 2011, doi: 10.1214/10-BJPS118.