RESEARCH ARTICLE | OCTOBER 17 2018

Modeling malaria incidence in Bengkulu province using small area estimation FREE

Etis Sunandi

AIP Conf. Proc. 2021, 060014 (2018)
https://doi.org/10.1063/1.5062778

Modeling Malaria Incidence in Bengkulu Province Using Small Area Estimation

Etis Sunandi[1,a)]

[1]*Department of Mathematics, University of Bengkulu, Bengkulu, 38125, Indonesia*

[a)]Corresponding author: esunandi@unib.ac.id

Abstract. The purpose of this study is to model malaria incidence in Bengkulu province using small area estimation. The model used is the Log-Normal model. In this research, the method is applied to estimate parameters in Small Area Estimation using Hierarchical Bayesian and direct estimation methods. The data used was collected by the Bureau of Statistics (BPS). The results show that the parameters for the discrete data in the Small Area Estimation, which is the average estimation of the Log-Normal prior function, are more accurate than the direct estimation. Other results are obtained from these estimates of the Hierarchical Bayesian estimator have a *trend* (tendency) which is equal to the direct estimator. It means that both methods generate consistent estimators.

Keywords: Estimation, log-normal model, tendency.

INTRODUCTION

Malaria is one of the deadliest diseases in the world. Malaria is still a global problem, especially in developing countries. This issue is one of the points discussed and stipulated in the Millennium Development Goals (MDGs) agreement 2015 on the 6th point of resistance against HIV/AIDS, Malaria, and other diseases. Similarly, in the WHO's 60th meeting on 18 May 2007, there has been a global commitment to the elimination of malaria for every country.[1]

In Indonesia, about 35 percent of the population live in areas at risk of being infected with malaria and as many as 38 thousand people are reported to die per year due to severe malaria Plasmodium falciparum. Malaria incidences occur almost every year in various endemic areas of Indonesia. Some areas have been categorized as red zone areas of malaria, such as East Nusa Tenggara, West Nusa Tenggara, Maluku, North Maluku, Central Kalimantan, Bangka Belitung, Riau Islands and Bengkulu.[2]

Bengkulu Province is one of the transmigration areas outside Java-Bali which is endemic to malaria. According to INFODATIN 2016, Bengkulu province is in 10th rank with Annual Parasite Incidence (API), which is equal to 2.77. This means that malaria incidence rate in Province of Bengkulu in 2016 amounted to 2.77 per 1000 population. One of the endemic areas is South Bengkulu District. Malaria in South Bengkulu District was in 2nd rank of 10 prioritized health issues. The malaria morbidity rate over the past five years showed an increasing trend in cases each month.[3]

Small Area Estimation (SAE) is a statistical technique for estimating subpopulation parameters of small sample size. The notion of the small area not only refers to an area/region (in a geographical sense), but also refers to a portion of the population defined according to certain criteria, for example by sex, race, age and others.[4] Small Area Estimation (SAE) is an indirect estimation of a parameter in a relatively small area in the survey sampling. This estimation is applied when the direct estimation is not able to provide sufficient accuracy for the sample size in the small area, so the generated statistics will have a large variance or the predictions cannot be made because they are not represented in the survey.[5] The estimation of parameters of SAE is based on the model of a small area that

The 8th Annual Basic Science International Conference
AIP Conf. Proc. 2021, 060014-1–060014-7; https://doi.org/10.1063/1.5062778
Published by AIP Publishing. 978-0-7354-1739-7/$30.00

requires additional information that has a relationship with the variables being observed, also known as auxiliary variables.

The SAE method has been widely introduced by Rao and Molina, including Best Linear Unbiased Prediction (BLUP), Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB), and Hierarchical Bayes (HB). EBLUP and BLUP are methods for continuous data.[6] Meanwhile, EB and HB are methods for binary data or count data. The Bayes concept is used because the estimation of domains with small samples desperately needs supporting information, whether derived from previous research, even from a subjective or specific assessment of each domain/area. Modeling SAE is being widely studied by applying Bayesian Estimation, because it is more profitable to have a Mean Square Error (MSE) value smaller than BLUP or EBLUP.

There are many studies on Small Area Estimation for count data such as Trevisani and Torelli,[7] They discuss a number of model specifications for estimating small area counts and their relative merits are illustrated. They conducted a laboratory study for target areas. Simulated data were generated by assuming population characteristics of interest as well as survey sampling design as known. In one set of experiments, the numbers of employed/unemployed from census data were utilized; in others, population characteristics were varied. Clement provides a critical review of the main advances in small area estimation (SAE) methods in recent years with application to disease mapping.[8] The review discusses in detail earlier developments of small area estimating methods in the field of disease mapping of the terminology which we refer to as "Extensions." Illustrative examples of the application of Small Area Estimation (SAE) to disease mapping are presented. Hajarisman et al. discussed two-level Poisson Bayes and developed a model using two different prior distributions.[9] Based on some of the research mentioned above, therefore the writer has been interested to research ***MODELING MALARIA INCIDENCE IN BENGKULU PROVINCE USING SMALL AREA ESTIMATION.***

METHODOLOGY

Basic Model of Small Area Estimation

A small area is defined as a subset of the population such that its size is small where a variable is concerned. The classic approach to estimate small area parameters is known as a direct estimator. This method does not have sufficient precision due to the smallness of the sample size used to get these estimates. Therefore, the indirect estimation method has been developed so as to use the strength of the area nearby and data sources outside the area for which the statistics were obtained. There are two connection models in the indirect estimation: implicit and explicit connector model.[4] This liaison model is used to connect small areas with other small areas. Model implicit connections are used on design-based estimates sampling. This model produces predictors of varying degrees. The design is relatively small compared with the direct estimator.

There are two main ideas used in SAE, namely, the assumption that the diversity in the small area of the response variable can be explained entirely by the corresponding diversity relationship in the supplementary information, called the fixed effect model. Second, assuming the specific diversity of the small area cannot be explained by the additional information is called the small area random effects (random effect) model. The combination of these two models to form a mixed model (mixed model) is called the Fay–Herriot model.[4] Other mixed linear models are EBLUP, EB, and HB methods, which include many uses for small area estimates.

Hierarchical Bayesian (HB) Method

Hierarchical Bayes is referred to as fully Bayes because it provides subjective prior and empirical prior distributions (based on data), whereas EB only includes the empirical prior distribution. In addition, the advantages of HB are that it is relatively clearer and more appropriate for inferencing problems, its computations are also relatively easy using Markov Chain Monte Carlo (MCMC) techniques, and it is also able to cope with models with random effects following distributions other than the normal distribution.[9]

In a *direct estimation,* estimation of the mean (θ_i) can be determined by the following formula:

$$\theta_i = \frac{y_i}{e_i}$$

Here, the y_i form binary data or data counts in the area to-^{i} which want to be noticed, $e_i = n_i \left(\sum_i y_i / \sum_i n_i \right)$ is the expected value of the number of cases from the data area to-i. the constant n_i is equal to the number of samples taken from the area to^{-i}, $i = 1, 2, \ldots, m$ with m the number of selected areas that are small areas.

Poisson-Lognormal Model

Parameter estimation using Hierarchical Bayesian (*HB*) method with the Poisson-Log Normal model is as follows:[4]

i. $y_i|\theta_i \sim iid\ Poisson(e_i\theta_i)$

ii. $\xi_i = log(\theta_i)|\mu, \sigma^2 \sim iid\ N(\mu, \sigma^2)$

iii. $f(\mu, \sigma^2) \propto f(\mu)f(\sigma^2)$ (μ and σ^2 are independent) with

iv. $f(\mu) \propto 1\ ;\ \frac{1}{\sigma^2} \sim G(a, b)\ ;\ a \geq 0\ , b > 0$

(1)

The model (i) in equation (1) is a variable sample distribution of y_i. Parameter θ_i is mean to be estimated in this study by using the log function in the model (ii). To perform an estimation of the mean (θ_i), there must first be estimated $\hat{\mu}$ and $\hat{\sigma}^2$. The proper prior distribution for $\hat{\sigma}^2$ is an Inverse Gamma with specified parameters a and b.

The mean for the Hierarchical Bayes model is estimated as the mean $E(\theta_i|y)$ with variance $V(\theta_i|y)$ of the joint posterior distribution:

$$f(\theta_i|\boldsymbol{y}) = \int_{\hat{\mu}} \int_{\hat{\sigma}^2} f(\theta_i, \hat{\mu}, \hat{\sigma}^2|\boldsymbol{y})\ d\hat{\sigma}^2 d\hat{\mu} \tag{2}$$

The posterior distribution of the equation (2) is an *open form*, then an alternative solution that can be used is to calculate the posterior quantity through numerical integration. One of the methods that can be used is Markov Chain Monte Carlo (MCMC). The purpose of MCMC is to establish a chance of the Markov chain until eventually towards a certain posterior distribution. Posterior spread calculation results in samples of posterior magnitudes. Finally, the parameters of posterior distribution can be predicted.

The famous MCMC procedure is conditional Gibbs (Gibbs Conditionals). The conditional Gibbs form for the Log-Normal is:[4]

$$\begin{aligned} [\hat{\mu}|\theta_i, \sigma^2, y_i] &\sim N\left(\frac{1}{m}\sum_{i=1}^{m} \xi_i, \frac{\sigma^2}{m}\right) \\ [\hat{\sigma}^2|\theta_i, \hat{\mu}, y_i] &\sim G\left(\frac{m}{2} + a, \frac{1}{2}\sum_i (\xi_i - \hat{\mu})^2 + b\right) \\ f(\theta_i|\hat{\mu}, \hat{\sigma}^2, y_i) &\propto \theta_i^{y_i - 1} exp\left[-e_i\theta_i - \frac{1}{2\hat{\sigma}^2}(\xi_i - \hat{\mu})^2\right] \end{aligned} \tag{3}$$

Estimation of parameters $\hat{\mu}$ and σ^2 comes directly from (i) and (ii) on (3). Meanwhile, part (iii) equation (3) is expressed as $f(\theta_i|\hat{\mu}, \hat{\sigma}^2, y_i) \propto k(\theta_i)h(\theta_i|\hat{\mu}, \hat{\sigma}^2)$, where:

$h(\theta_i|\hat{\mu}, \hat{\sigma}^2) \propto g'(\theta_i) exp\left\{\frac{-(\xi_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right\}$ with $g'(\theta_i) = \frac{\partial g(\theta_i)}{\partial \theta_i}$ and $g(\theta_i) = \log(\theta_i)$ $k(\theta_i) = exp\left(-e_i\theta_{i\,i}\right)\theta_i^{y_i}$

In accordance with (3), $\hat{\mu}$ and $\hat{\sigma}^2$ follow the standard multivariate normal and Gamma distributions. So, the values of both parameters can be estimated by generating a random sample. Rao and Molina state that the average value of Hierarchical Bayes will be assumed through the MH (Gibbs sampling Metropolis-Hasting) algorithm, as follows:[6]

1. Generate $\xi_i \sim iid\ N(\hat{\mu}, \hat{\sigma}^2)$ then search for the value $\theta_i^{(0)} = g^{-1}(\xi_i)$
2. Calculate $\theta_i^* = \frac{1}{\theta_i^{(0)}} exp\left\{\frac{-(\xi_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right\}$
3. Calculate $a(\theta_i^{(d)}, \theta_i^*) = \min\left\{k(\theta_i^*)/k(\theta_i^{(d)}), 1\right\}; d = 0, 1, \dots, D$
4. Generate u from the uniform distribution (0,1).
5. Set $\theta_i^{(d+1)} = \theta_i^*$ if $u \leq a(\theta_i^{(d)}, \theta_i^*)$
6. Repeat steps 3 through 4, until *D* is obtained.

After the MH simulation, the following proportional estimates are obtained: $\left\{\theta_1^{(d)}, \dots, \theta_m^{(d)}; d = 1, \dots, D\right\}$. Then, the posterior being observed can be calculated. The estimator of the mean of the HB is

$$\hat{\theta}_i^{HB} \approx \frac{1}{D}\sum_{k=1}^{D} \theta_i^{(d)}$$

while the proportion of variance of estimators of the hierarchical Bayes ($V(\theta_i^{HB}|\hat{\theta})$) is

$$V(\hat{\theta}_i^{HB}|\hat{\theta}) = \frac{1}{D-1}\sum_{d=1}^{D}\left(\theta_i^{(d)} - \hat{\theta}_i^{HB}\right)^2.$$

On the other hand, characteristics of $(\hat{\theta}_i^{HB})$ are MSE and bias. MSE is a scale to measure the variance of small area estimators. Meanwhile, the bias is the difference between the expectation of the estimator and the parameter. The smaller the MSE and bias, the more valid and accurate the parameter estimator. It is mathematically written as follows:

$$\text{Bias}\left(\hat{\theta}_i^{HB}\right) = \frac{1}{D}\sum_{q=1}^{D}\left[\theta_i^{(q)} - \hat{\theta}_i^{HB}\right]$$

$$MSE\left(\hat{\theta}_i^{HB}\right) = \frac{1}{D}\sum_{q=1}^{D}\left[\theta_i^{(q)} - \hat{\theta}_i^{HB}\right]^2 = V(\hat{\theta}_i^{HB}|\hat{\theta}) + \left\{\text{Bias}\left(\hat{\theta}_i^{HB}\right)\right\}^2$$

Root Mean Square Error (RMSE) has been employed as a measurement of the accuracy and validity of the Hierarchical Bayes mean estimator. RMSE is defined as the square root of the average of the squared differences between the actual proportions and the estimator. Thus, the most accurate estimator will have the smallest RMSE, which is zero. The RMSE is given by $RMSE = \sqrt{MSE}$.

RESULTS AND DISCUSSION

The *Hierarchical Bay ice* method (HB) for small area estimation is used for indirect estimates. This study uses Village Potential data in 2014 for the number of malaria patients in Bengkulu Province. Of all the villages in Bengkulu Province, there are 30 villages detected with malaria incidence. These villages are then used as object of observation in this research. In the end there are compared the parameter estimators resulting from the direct

estimators and those resulting from the Hierarchical Bayes method. The best estimator is selected based on the value of the RMSE.

From the data of Clement, it is known that malaria is common in Suka Merindu Village in the Putri Hijau subdistrict of North Bengkulu district (60 incidences).[8] Meanwhile, the lowest incidence of malaria is in Papahan Village in the Kinal subdistrict of Kaur district (1 incidence). There are 10 cases in each village in Bengkulu Province in 2014. In addition, 75% of the villages in Bengkulu Province in 2014 had 19 malaria incidences. More details are provided in Table 1.

TABLE 1. Summary of statistics of Malaria incidence in Bengkulu Province [8]

Number of malaria Incidence (2014)	
Mean	16.53
Median	10:00
Minimum	1.00
Maximum	60.00
Q1	5.50
Q3	19.50

The estimation results (Table 2) show that 75% of the villages have an average of 2.135. Meanwhile, 25% of the villages have an average of 0.571. In addition, the HB estimator is seen to have a *trend* (tendency) which is equal to that of the direct estimation (Fig. 1). It means that both methods generate consistent estimators.

In addition, the HB method has also estimated the variance and MSE of the mean estimators. In terms of summary RMSE of HB estimators (Table 2), all of the villages have a variance of the direct estimators (DE) of the value of the variance of HB estimators. However, in general, the estimator of the HB variance is smaller than the direct estimator. So, it can be concluded that the estimation of the mean using a Poisson-Log Normal model HB is better than direct estimation.

TABLE 2. Comparison of HB and DE Statistics

	$\tilde{\theta}_i^{HB}$				$\tilde{\theta}_i^{DE}$			
	tetai	**var (tetai)**	**bias**	**rmse**	**tetai**	**var (tetai)**	**bias**	**rmse**
Mean	1.591	0.002	-0.017	0.029	1.618	0.315	0.000	0.427
Minimum	0.000	0.000	-0.309	0.000	0.051	0.001	0.000	0.036
Maximum	6.520	0.032	0.003	0.333	6.521	2.772	0.000	1.665
Q1	0.532	-	-	-	0.571	-	-	-
Q3	2.135	-	-	-	2.132	-	-	-

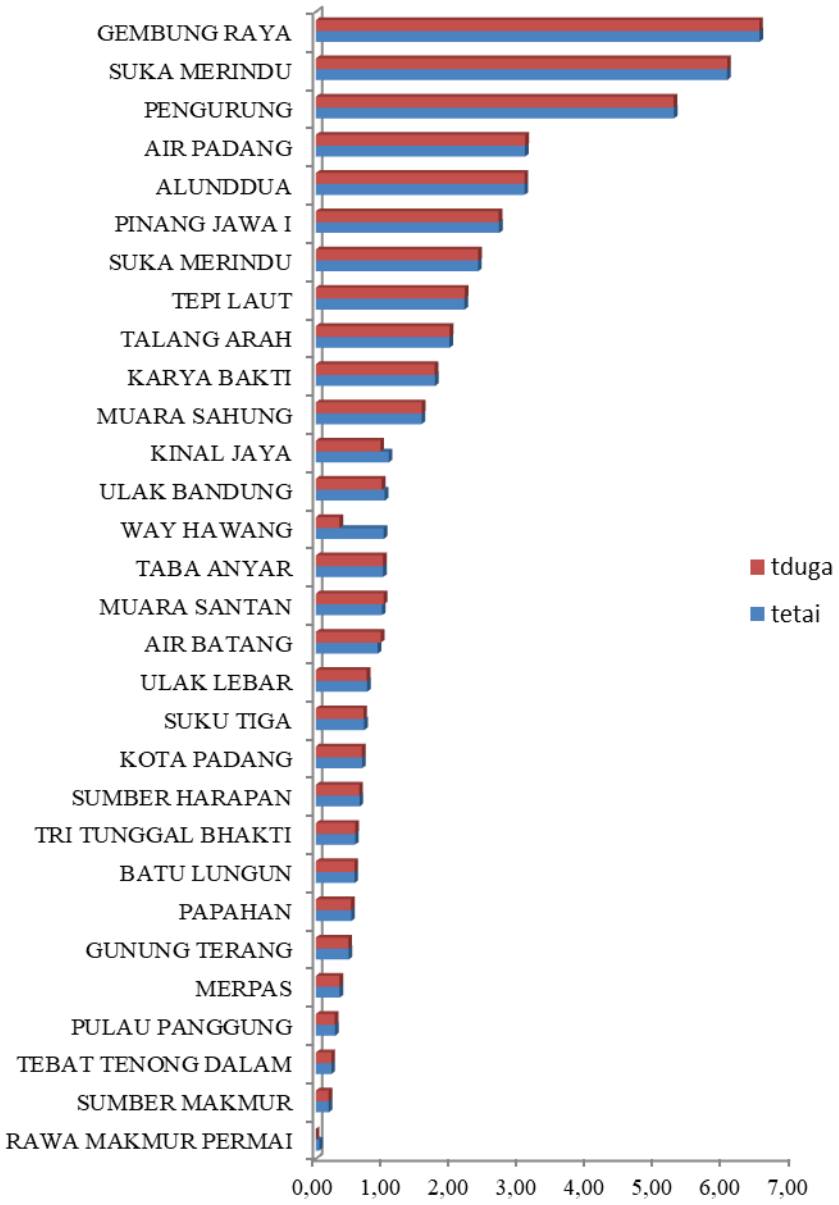

FIGURE 1. Comparison of Direct Estimators and HB Estimators.

CONCLUSIONS

According to BPS data, some villages had ten incidences of malaria in 2014. The biggest malaria incidence occurred at Suka Merindu Village in the Putri Hijau subdistrict of North Bengkulu district, that is, 60 incidences. Meanwhile, the lowest incidence of malaria is in Papahan Village in the Kinal subdistrict of Kaur district (1 incidence). In addition, 75% of the villages in Bengkulu Province in 2014 had 19 incidences of malaria. The results obtained from these estimates of the Hierarchical Bayesian (HB) estimator have a *trend* (tendency) which is equal to that of the direct estimator. It means that both methods generate consistent estimators. On the other hand, the

variance of the HB estimator is smaller than that of the direct estimator. Thus, it can be concluded that the estimation of the mean using the Poisson-Log Normal model HB is better than direct estimation.

ACKNOWLEDGEMENTS

We like to show our gratitude to the committee of the 8th Annual Basic Science International Conference 2018 for accepting our manuscript to be published in this conference proceeding.

REFERENCES

1. S. Manumpa. J. Period. Epidemiol. **4**, 338-348 (2016).
2. Darmiah, Baserani, K. Abdul, Isnawati and S. Yuniarti, J. Health Epidemiol. Commun. Dis. **3**, 36-41 (2017).
3. R. Mayasari, A. Diana, and S. Hotnida. Media of Health Res. Dev. **23**, 158-164 (2013).
4. M. Ghosh and J. N. K, Rao. Statistic. Sci. **9**, 55-93 (1994).
5. N. G. N Prasad and J. N. K Rao, J. Am. Statistic. Assoc. **85**, 163-171 (1990).
6. J. N. K Rao and I. Molina, *Small Area Estimation* (John Wiley & Sons, New York, 2015), pp. 1-442.
7. M. Trevisani and N. Torelli, Open J. Statistics **7**, 2161-7198 (2017).
8. E. P. Clement, Int. J. Probabil. Statistics **3**, 15-22 (2014).
9. N. Hajarisman, A. K. Mutaqin and A. I. Ahmad. Statistika **12**, 81-91 (2012).