Empirical Bayesian Method for the Estimation of Literacy Rate at Sub-district Level Case Study: Sumenep District of East Java Province

A.Tuti Rumiati[1], Khairil Anwar Notodiputro[2], Kusman Sadik[2], and I Wayan Mangku[3]

Abstract—**This paper discusses Bayesian Method of Small Area Estimation (SAE) based on Binomial response variable. SAE method being developed to estimate parameter in small area due to insufficiency of sample. The case study is literacy rate estimation at sub-district level in Sumenep district, East Java Province. Literacy rate is measured by proportion of people who are able to read and write, from the population of 10 year-old or more. In the case study we used Social Economic Survey (Susenas)data collected by BPS. The SAE approach was applied since the Susenas data is not representative enough to estimate the parameters at sub-district level because it's designed to estimate parameters in regional area (in scope of a district/city at minimum). In this research, the response variable being used was logit function trasformation of p_i (the parameter of Binomial distribution). We applied direct and indirect approach for parameter estimation, both using Empirical Bayes approach. For direct estimation we used prior distribution of Beta distribution and Normal prior distribution for logit function (p_i) and to estimate parameter by using numerical method, i.e integration Monte Carlo. For indirect approach, we used auxiliary variables which are combinations of sex and age (which is divided into five categories). Penalized Quasi Likelihood (PQL) was used to get parameter estimation of SAE model and Restricted Maximum Likelihood method (REML) for MSE estimation. Instead of Bayesian approach, we are also conducting direct estimation using classical approach in order to evaluate the quality of the estimators. This research gives some findings, those are: Bayesian approach for SAE model gives the best estimation because having the lowest MSE value compares to the other methods. For the direct estimation, Bayesian approach using Beta and logit Normal prior distribution give a very similar result to the direct estimation with classical approach since the weight of $\bar{y}_i$ is too large, which is about 0.905. It is also found that direct estimation using Bayesian approach with the Beta prior distribution gives better MSE than using logit normal prior distribution.**

Keywords—**SAE model, Bayesian approach, Binomial response, Monte Carlo integration, literacy rate**

Abstrak—Paper ini membahas metode pendugaan area kecil (Small Area Estimation: SAE) berbasis sebaran respon Binomial Metode SAE digunakan untuk pendugaan area kecil dimana jumlah contoh tidak cukup representatif untuk pendugaan area kecil tertentu. Studi kasus yang diambil adalah pendugaan angka melek huruf di wilayah kecamatan di Kabupaten Sumenep, Jawa Timur berbasis data Survai Ekonomi Nasional (Susenas) oleh BPS (2010). Susenas dirancang untuk pendugaan parameter di wilayah regional (minimal Kabupaten/kota) dan tidak cukup representatif untuk pendugaan parameter level kecamatan oleh karena itu digunakan pendekatan Metode SAE. Model SAE yang dibahas dalam penelitian ini menggunakan peubah respon fungsi logit (p_i) yang merupakan transformasi logit dari parameter Binomial p_i. Pendugaan parameter model SAE menggunakan Maximum Likelihood Estimator (MLE) dan pendugaan MSE menggunakan metode Restricted Maximum Likelihood (REML). Pendugaan parameter area kecil menggunakan pendekatan Bayes Empirik yaitu dengan mengaplikasikan integrasi numerik menggunakan metode Monte Carlo. Untuk membandingkan kualitas penduga, selain model SAE juga dilakukan pendugaan langsung melalui pendekatan klasik dan melalui pendekatan Bayes yang menggunakan sebaran prior Beta dan sebaran normal untuk fungsi logit (p_i). Dalam studi kasus digunakan peubah respon angka melek huruf yang diukur dari proporsi penduduk berusia 10 tahun ke atas yang bisa baca tulis, sedangkan peubah pembantu merupakan kombinasi antara jenis kelamin dan usia. Pengembangan model SAE menggunakan data Susenas 2010 dan untuk pendugaan area kecil digunakan data sensus penduduk tahun 2010. Dari penelitian ini diperoleh hasil bahwa pendugaan parameter menggunakan model SAE melalui pendekatan Bayes memberikan nilai pendugaan yang paling baik karena memiliki nilai MSE terendah dibandingkan metode lainnya. Metode pendugaan langsung enggunakan pendekatan Bayes memberikan hasil yang hampir sama dengan pendugaan klasik karena bobot untuk $\bar{y}_i$ terlalu besar yaitu sekitar 0,905. Nilai MSE untuk pendugaan Bayes menggunakan sebaran prior Beta sedikit lebih baik dibandingkan dengan menggunakan sebaran prior logit normal.

Kata Kunci—***Model SAE, pendekatan Bayes, peubah respon Binomial, integrasi Monte Carlo, angka melek huruf***

I. INTRODUCTION

Small Area Estimation (SAE), whereas the estimation method is model based, has been developed and used to estimate the small area parameters if the numbers of data from a certain small area are not representative enough to represent the population [1].

The problem of insufficiency data is found in most cases of small area estimation because most of many surveys are designed for larger area (regional or national scale).

In Indonesia, problems of insufficiency data for parameters estimation in small areas (sub-district or village) are also found. For example, if the estimation of small area is using the survey which are designed for national or regional scale (province to district/city) like National Sosial Economy Survey (Susenas) [2].

A. Tuti Rumiati is with Department of Statistics, FMIPA, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: agnestuti@yahoo.com

Khairil Anwar Notodiputo and Kusman Sadik are with Department of Statistics, FMIPA, Institut Pertanian Bogor, Bogor, Indonesia.

I Wayan Mangku is with Department of Mathematics, FMIPA, Institut Pertanian Bogor, Bogor, Indonesia.

One of the purposes of Susenas is to estimate Human Development Index (HDI), which is measuring the impact of regional development. HDI is measured in 3 basic dimensions, i.e. education component, health component and decent living standard which is measured by the purchasing capability [3]. Particularly for education component, literacy rate and average length of school period in a certain area are used to measure Education Index. The literacy rate is measured by reading and writing skill at 10-year-old-and-older people's. The Variable of Reading and writing skill is a binary variable.

SAE model for binary response variable has been developed by some researchers, such as SAE model applied to survey data in health sector which is based on combination of unit and area using hierarchical Bayesian approach [4], Empirical Bayesian method for binary data for small area estimation [5], SAE model estimation for binary data with Bayesian empirical and hierarchical estimation method [6], SAE model for proportion in business survey [7] and SAE for proportion estimation about labor (working, non-working, and non-working labor force) in Australia [8].

In this research we discuss about SAE model using Bayesian approach based on Binomial response variable. For the case study, we estimate literacy rate at small area (sub district) level based on Susenas data. Literacy rate, which is assumed to have a Binomial distribution with p_i parameter, was measured by the number of people of 10-year-old-or-older who can read and write.

Evaluation of quality estimators is done by comparing MSE estimation. To compare the quality of estimators, we also conducting direct estimation with classical and Bayesian approach using prior Beta distribution and logit normal distribution.

II. Method

Supose y_{ij} is defined as reading and writing ability of each individual in the population, which is a binary response variable. y_{ij} variable is assumed having Bernoulli distribution with p_i parameter, where y_{ij}=1 or 0 with the function of probability distribution as follows:

$$y_{ij} \mid p_i \overset{ind}{\sim} Bernoulli(p_i),$$

where, j=1,2,…,n_i; i=1,2,…,m

Small Area parameter to be estimated is small area proportion: $p_i = \overline{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$, whereas N_i is number of population in the i^th area. The sample proportion for the i^th area is $\hat{p}_i = \sum_{j=1}^{n_i} y_{ij} / n_i = y_i / n_i$, while $y_i = \sum_{j=1}^{n_i} y_{ij}$ has Binomial distribution with n_i and p_i parameters. The probability distribution function of y_i is:

$$f(y_i \mid p_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i} \tag{1}$$

or written as $y_i \mid p_i \overset{ind}{\sim} Binomial\ (n_i, p_i)$.

For the classical approach, $\hat{p}_i = \sum_j y_{ij} / n_i = y_i / n_i$ will be the direct estimator for proportion in the i^th area if the sampling selection is using simple random sampling.

Through Bayesian approach, the p_i parameter is considered as a random variable having a certain distribution. Bayesian approach for direct estimation method is based on two kinds of prior distribution, those are: 1) by assuming that p_i has Beta distribution with parameters α and β; and 2) by assuming that logit (p_i) = log [p_i/(1-p_i)] or probit $\Phi^{-1}(p_i)$ function has Normal distribution [7]. In this research, the prior used is Beta and logit-Normal distribution.

A. Parameter Estimation Using Beta Prior Distribution

In this case, p_i parameter to be considered as a random variable having Beta distribution with parameters α and β as the prior:

$$p_i \overset{iid}{\sim} Beta(\alpha, \beta); \alpha > 0, \beta > 0$$

The probability function of Beta distribution of p_i is:

$$f(p_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1} (1-p_i)^{\beta-1} \;\; ; \alpha > 0, \beta > 0 \tag{2}$$

Where Γ(.) is Gamma function.

Posterior distribution was obtained by deriving conditional distribution of $f(p_i|y_i,\alpha,\beta)$ from the joint probability distribution of $f(p_i,y_i,\alpha,\beta)$, which is multiplication of the Equation 1 and 2.

With known y_i, α and β, the posterior distribution of p_i is also as the form of Beta distribution $p_i \mid y_i, \alpha, \beta \overset{ind}{\sim} beta(y_i + \alpha, n_i - y_i + \beta)$ with probability distribution function:

$$f(y_i \mid \alpha, \beta) = \binom{n_i}{y_i} \frac{\Gamma(\alpha + y_i)\Gamma(\beta + n_i - y_i)}{\Gamma(\alpha + \beta + n_i)} x \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{3}$$

Bayesian estimator of p_i and the posterior variance are given by:

$$\hat{p}_i^B(\alpha, \beta) = E(p_i \mid y_i, \alpha, \beta) = \frac{y_i + \alpha}{n_i + \alpha + \beta} \tag{4}$$

and,

$$V(p_i \mid y_i, \alpha, \beta) = \frac{(y_i + \alpha)(n_i - y_i + \beta)}{(n_i + \alpha + \beta + 1)(n_i + \alpha + \beta)^2} \tag{5}$$

Empirical Bayes estimator for p_i is obtained by replacing α and β with their estimators, $\hat{\alpha}$ and $\hat{\beta}$. The estimators of α and β parameter can be found by applying two methods, by maximizing likelihood function or Maximum Likelihood (ML) method and using moment method.

Using ML estimators, $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ are obtained by maximizing likelihood function of l(α,β) from $y_i \mid \alpha, \beta \overset{ind}{\sim} Beta-binomial$ distribution. Since the closed-form for $\hat{\alpha}_{ML}$ and $\hat{\beta}_{ML}$ is not available, the Newton-Raphson or other iterative methods can be used to obtain parameter estimation of α and β [7].

The estimator for α and β was found by using moment estimator method as follows:

$$\hat{p} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \tag{6}$$

$$\frac{1}{\hat{\alpha} + \hat{\beta} + 1} = \frac{n_T s_p^2 - \hat{p}(1-\hat{p})(m-1)}{\hat{p}(1-\hat{p})\left[n_T - \sum_i n_i^2 / n_T - (m-1)\right]} \tag{7}$$

where, $s_p^2 = \sum_i (n_i / n_T)(\hat{p}_i - \hat{p})^2$, $n_T = \sum_i n_i$

By replacing α and β with $\hat{\alpha}$ and $\hat{\beta}$ into Equation 4, EB estimator of p_i is obtained, i.e

$$\hat{p}_i^{EB} = \hat{p}_i^{B}(\hat{\alpha},\hat{\beta}) = \hat{\gamma}_i \hat{p}_i + (1-\hat{\gamma}_i)\hat{p} \qquad (8)$$

where, $\hat{\gamma}_i = n_i/(n_i + \hat{\alpha} + \hat{\beta})$. As shown in formula (8), the empirical Bayes estimators of p_i ($\hat{p}_i^{EB}$)is the weighted average of $\hat{p}$. The bigger value of n_i, the bigger weight to be given to the $\hat{p}_i$. The estimation equation as mentioned above is similar to Fay-Heriot estimator for area-based model. $\hat{p}_i^{EB}$ estimator is close to unbiased for p_i if the valuae of m is big enough, because E($\hat{p}_i^{EB}$-p_i) will be close to zero.

To get *MSE* ($\hat{p}_i^{EB}$) estimator we used Jackknife method. This method will give unbiased estimator of *MSE* ($\hat{p}_i^{EB}$), i.e *mse* ($\hat{p}_i^{EB}$) [6] as follows:

$$mse\ (\hat{p}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i} \qquad (9)$$

where,

$$\hat{M}_{1i} = g_{1i}(\hat{\alpha},\hat{\beta},y_i) - \frac{m-1}{m}\sum_{l=1}^{m}\left[g_{1i}(\hat{\alpha}_{-l},\hat{\beta}_{-l},y_i - g_{1i}(\hat{\alpha},\hat{\beta},y_i)\right] \qquad (10)$$

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}\left(p_{i,-l}^{EB} - \hat{p}_i^{EB}\right)^2 \qquad (11)$$

with,

$$g_{1i}(\mu,\sigma,y_i) = V(p_i \mid y_i,\mu,\sigma) = E(p_i^2 \mid y_i,\mu,\sigma) - \left[\hat{p}_i^B(\mu,\sigma\right]^2 \qquad (12)$$

B. Parameter Estimation Using Prior Logit Normal Distribution

The second method of direct estimation using Bayesian opproach is by applying logit transformation for p_i parameter where logit (p_i) is considered has Normal distribution:

$$\text{logit}(p_i) = \log[p_i/(1-p_i)] \overset{iid}{\sim} N(\mu,\sigma^2) \qquad (13)$$

Implementation of Empirical Bayes method for the above model is more complex because there are no analytical forms for p_i estimator and posterior variance. Bayesian estimator of p_i, $\hat{p}_i^B$ (μ,σ) = E(p_i / y_i,μ,σ), can be defined as ratio between two expectation value [7], as follows:

$$\hat{p}_i^B(\mu,\sigma) = \frac{E\left[h_1(\mu+\sigma z)\exp\{h_2(y_i,\mu+\sigma z)\}\right]}{E\left[\exp\{h_2(y_i,\mu+\sigma z)\}\right]} \qquad (14)$$

where, $h_1(\mu+\sigma z_i) = p_i = \dfrac{e^{\mu+\sigma z_i}}{1+e^{\mu+\sigma z_i}}$ and

$$h_2(y_i,(\mu+\sigma z)) = (\mu+\sigma z)y_i - n_i\log(1+e^{\mu+\sigma z})$$

$$z_i \sim N(0,1)$$

In this research, we calculate the expectation value of nominator and denominator in the Equation 14 by generating z from normal distribution N(0,1) with n=500. For each value of z, the expectation value of each nominator and denominator is calculated using the following formula:

$$E\left[h_1(\mu+\sigma z)\exp\{h_2(y_i,\mu+\sigma z)\}\right] = \frac{1}{500}x\sum_a A_a \qquad (15)$$

$$E\left[\exp\{h_2(y_i,\mu+\sigma z)\}\right] = \sum_a^{500}\frac{1}{500}B_a \qquad (16)$$

where,

$$A_a = \frac{e^{\mu+\sigma z_a}}{1+e^{\mu+\sigma z_a}}x\exp(\mu+\sigma z_a)y_i - n_i\log(1+e^{\mu+\sigma z_a})x\frac{1}{\sqrt{2\pi}}e^{-1/2z_a^2}$$

$$B_a = \exp\left\{(\mu+\sigma z_a)y_i - n_i\log(1+e^{\mu+\sigma z_a})\right\}\text{x}\ \frac{1}{\sqrt{2\pi}}e^{-1/2z_a^2}$$

The estimator of μ and σ, which is $\hat{\mu}$ and $\hat{\sigma}$, can be obtained by applying ML method (by maximizing likelihood function *l(μ,σ)* through Newton-Raphson method) or moment method. The EB estimator of p_i, $\hat{p}_i^{EB}$ = $\hat{p}_i^{B}$ ($\hat{\mu}$,$\hat{\sigma}$) is obtained by replacing $\hat{\mu}$ and $\hat{\sigma}$ in equation 14. Posterior variance, *V(pi | yi,μ,σ)*, is defined by the following formula:

$$V(p_i \mid y_i,\mu,\sigma) = E(p_i^2 \mid y_i,\mu,\sigma) - \left[\hat{p}_i^B(\mu,\sigma\right]^2 = g_{1i}(\mu,\sigma,y_i) \qquad (17)$$

The posterior variance can be calculated as the expectation value from a function of standard normal distribution $Z \sim N(0,1)$.

The calculation of *mse* $_J$($\hat{p}_i^{EB}$) using ML estimator is quite complicated. For this reason, in this reasearch we prefer to use moment method[7], using formula:

$$\textstyle\sum_i y_i = n_T\hat{p} = n_T E[h_1(\mu+\sigma z)]$$

$$\sum_i (y_i^2 - y_i) = \left[\sum_i n_i(n_i-1)\right]E\left[h_1^2\ (\mu+\sigma z)\right] \qquad (18)$$

$\hat{\mu}_{-l}$ and $\hat{\sigma}_{-l}$ are the estimators obtained by deleting the l^{th} area from data.

Variance calculation can be done through Monte Carlo integration to calculate $E(p_i) = E[h_1(\mu + \sigma z)]$ and $E(p_1^2) = E[h_1^2(\mu + \sigma z)]$.

Moreover the Jackknife estimator of *MSE* ($\hat{p}_i^{EB}$) can be obtained through the same method as explained in Equation 9, 10 and 11.

C. Indirect Estimation (Model-based)

Every unit (individual) in the population can be catagorized into different mutually exclusive and exhaustive groups based on demographic status [6]. In this research we use combination of age (consist of 5 groups) and sex to catagorized each unit, therefore those two variables are used as covariates variables and p_{ij} is proportion of literate of of all 10-year-old-or-older people in each combination.

The logit (p_{ij}) can be modeled as logistic regression which relates the logit (p_{ij}) to its covariates (x_{ij}) with random area effect:

$$\text{logit}(p_{ij}) = x_{ij}^T\beta + \upsilon_i\ ,\ \upsilon_i \overset{iid}{\sim} N(0,\sigma_\upsilon^2) \qquad (19)$$

The model above is called logistic linear mixed model, which is a part of generalized linear mixed model. The covariate vector is assumed to be independent to the area i.

The proportion of literate people in area-i (p_i) is proportion of all 10-year-old-or-older people in the population which can read and write in any langguages. Thus, p_i can be decomposed into 2 components, they are literate people which are taken as samples and who are not to be taken as samples. Mathematically, p_i can be written with the following formula:

$$p_i = f_i\bar{y}_i + (1 - f_i)\bar{y}_i^* \tag{20}$$

where,
$f_i = n_i/N_i$,
$\bar{y}_i$ is sample average (proportion) and

$\bar{y}_i^* = \sum_{l \in s'_i} y_{il} / (N_i - n_i)$ is the average of the unit which are not taken as samples in the i^th^ area.

In Equation 20, $\overline{y}_i^*$ is unknown and will be estimated by using Bayesian estimation, that is:

$$\hat{p}_i^{*B} = E\left(p_i \mid y_i, \beta, \sigma_\upsilon\right) = \textstyle\sum_{l \in s'_i} \; p_{il} / (N_i - n_i) \tag{21}$$

where y_i is sample from the i^th^ area and $p_{il} = E(y_{il} / p_{il}, y_{il}, \beta, \sigma_v)$ for $l \in s'_i$ (s'_i are units in population whose are not taken as samples).

Bayesian estimator of $\overline{y}_i^*$ is obtained by replacing μ in Equation 13 with the logit function which is formed as linear model of x_{ij},as shown in Equation 19. Thus, Bayesian estimator of $\overline{y}_i^*$ is :

$$\hat{p}_i^{*B} = E\left(\sum_l p_{il} \mid y_i, \beta, \sigma_\upsilon\right) = \frac{E\left[(\sum_l p_{il} \exp\left\{h_i\left(\sum_{j \in s_i} x_{ij}^T y_{ij}, y_i, \sigma z, \beta\right)\right\}\right]}{E\left[\exp\left\{h_i\left(\sum_{j \in s_i} x_{ij}^T y_{ij}, y_i, \sigma z, \beta\right)\right\}\right]} \tag{22}$$

where,

$$h_i\left(\sum_{j \in s_i} x_{ij}^T y_{ij}, \sigma z, \beta\right) = \left(\sum_{j \in s_i} x_{ij}^T y_{ij}\right)\beta + (\sigma z) y_i - \sum_{j \in s_i} \log\left[1 + \exp\left(x_{ij}^T \beta + \sigma z\right)\right] \tag{23}$$

Bayesian estimator of p_i can be defined as:

$$\hat{p}_i^B = \hat{p}_i^B(\beta, \sigma_\upsilon) = f_i\bar{y}_i + (1 - f_i)\hat{p}_i^{*B} \tag{24}$$

Parameter estimation for β and σ_υ model can be done in various ways, including EM algorithm, MCMC, Penalized Quasi-Likelihood (PQL) and moment method [7]. Therefore, EB for p_i is $\hat{p}_i^{EB} = \hat{p}_i^{B}(\hat{\beta}, \hat{\sigma}_\upsilon)$. Since there is no closed form to get the above expectation value, the expectation value calculation is conducted by using numerical method.

Estimation of $MSE(\hat{p}_i^{EB})$ is conducted by applying Jackknife method, that is replacing $\hat{p}_i^{EB} = k_i(y_i, \hat{\mu}, \hat{\sigma})$ and $\hat{p}_{i,-l}^{EB} = k_i(y_i, \hat{\mu}_{-l}, \hat{\sigma}_{-l})$ into Equation 12 and 13, so that the value of $\widehat{M}_{1i}$, $\widehat{M}_{2i}$, and MSE are obtained.

D. Data

For aplication, we used two kinds of data, those are National Social Economy Survey (Susenas) data and population census data year 2010 for Sumenep District. The Susenas data was used for parameter estimation of the model and census data was used for estimating proportion of literate people of each subdistrict.

Three variables used from Susenas data are ability to read and write, age and gender. Number of sample of Susenas data in Sumenep district year 2010 is 2,307 (see Table 1).

The census data is used for estimating parameter in small area (kecamatan) as auxilliary information. Two variables to be used for estimation are gender and age. The number of people in Sumenep district based on the census data year 2010 is 884,003. Distribution of population in each subdistrict is shown on Table 1.

III. Results and Discussion

Based on Susenas data, the numbers of respondents who are 10 years old or older is 2,307 and the average proportion who are literate is about 77.6%. Figure 1 shows the proportion of people who can read and write in every sub-district

It is seen that sub-district of Batuputih and Talango are the areas which have the lowest literacy rate, while sub-district of Arjasa has the highest literacy rate and even higher than Sumenep sub-district. The other sub-districts which have literacy rate below the avarage are Gapura, Batang-batang, Dungkek, Guluk-guluk, Pasongsongan and Ambuten. The complete information is presented in Figure 1.

A. Direct Estimation

Direct estimation is used two approaches, classical and Bayesian approach. Through classical approach, direct estimation is conducted by applying ML estimation, with $\hat{p}_i = \sum_j y_{ij} / n_i = y_i / n_i$ formula.

For the Bayesian approach, we used Empirical Bayes method by using Beta and logit-normal distribution as a prior distribution

For the prior Beta distribution and applying moment method as explained in Equation 4 and 5, parameter estimate of α and β are $\hat{\alpha}$ = 6.007941 and $\hat{\beta}$ =1.735254. Empirical Bayes of $\hat{p}_i^{EB}$ is obtained by applying Equation 8. Furthemore for the prior logit-normal distribution as in Equation 14 and generating z from N(0,1) distribution with n=500, the expectation value of nominator and denominator in Equation 14 are obtained by calculating formula 15 and 16.

Calculation of estimated parameter (p_i) using direct estimation of the three methods, are shown in Figure 2 and Table 1.

For the direct estimation, Maximum Likelihood (ML) method gives almost similar result with Bayesian method using logit normal distribution but the beta-binomial approach is giving slightly different result. In the Empirical Bayesian estimation using logit-normal prior distribution, the weight for population component is quite small, so it doesn’t give significant affect to the Bayesian estimation. On the contrary, the weight of $\overline{y}_i$ in for the Beta prior distribution, $\hat{\gamma}_i = n_i/(n_i + \hat{\alpha} + \hat{\beta})$, is also quite large of about 0.905.

The MSE estimation using Jackknife method for direct estimation using Bayesian approach is shown in Figure 3. It can be seen that the two methods, using prior distribution of beta as well as logit normal distribution, do not give good accuracy in term of MSE measurement.

B. Indirect Estimation

For indirect estimation, we used logistic model with two auxilliary variables that are age and gender. Figure 4 are plot of proportion and logit function of proportion of leteracy data and with age for male and female group. With p_k be a proportion of literate population in kth category, k=1, …,5.

Figure 4 (a) describes the relation of p_k to age, and Figure 4 (b) describes the relation of logit(p_k) = log[p_k/(1-p_k)] value to age which is grouped into 5 categories; they are 10-30 years old, 30-40 years old, 40-50 years old, 50-60 years old, and above 60 years old.This figure is to show that literacy rate is affected by its covariates which are combination of ages and gender.

Based on the correlation test with α=5%, it was proven that the proportion of literacy is affected by age and sex. Therefore, sex and age variables can be used as auxiliary variables into SAE model for indirect estimation (model-based). Figure 4 shows that the older the age, the lower the proportion of literate people. Furthermore, the proportion of literate males tends to be larger than the female ones.

Through indirect estimation (with SAE model), model parameter estimation uses PQL method which is then used to estimate $\overline{y}_i^*$ through Equation 22. The estimation parameter of model (19) is as follows:

Constant	: 4.935
b1 = gender	: 0.788
b2 = age	: -0.915
sig varian	: 0.251

All estimated parameter are siginificant in the model with siginificant level of about <0.0001. In order to get the expectation value of nominator and denominator of the equation we used numerical integration of Monte Carlo method. Furthermore, Bayesian estimation of p_i parameter is calculated based on Equation 24. Using indirect method, the parameter estimate for the proportion of literate people in every sub-district are shown in Table 1 of the Appendix.

C. Comparation of direct and indirect estimation using logit function

Figure 5 is shown the graphic of estimate parameter of p_i (the proportion of 10-year-old-or-older literate people) using logit function based on direct and indirect method. As described by Figure 5, the direct estimation method through Empirical Bayesian approach gives significant differences result to the indirect estimation which are includes the two covariates, age and gender variables.

The presence of auxiliary variables, which are age and sex, affects p_i estimators, because the weight for the model component is more dominant than the weight for the direct estimation component due to the small sampling fraction ($f_i=n_i/N_i$).

The estimation of $MSE(\hat{p}_i^{EB})$ for the direct and indirect estimation are shown on Figure 6. The MSE values of direct estimation tend to be high. The MSE values for p_i estimation in sub-district of Batuputih are far higher than other sub-districts since the proportion of 10- year-old-or-older people is lower than the proportion in the other sub-districts. The MSE values for indirect estimation is the lowest one, so the indirect estimation using SAE model gives the best estimation.

IV. Conclusion

In The aplication using Sumenep regency data was shown that direct estimation through Bayesian approach gives a similar result to the classical approach. The reason are: 1) the weight of population component for empirical bayes estimation using logit-normal prior distribution is quite small, it's about 0.261% and 2) the weight of $\overline{y}_i$ in for the Beta prior distribution, $\hat{\gamma}_i$ = $n_i/(n_i+\hat{\alpha}+\hat{\beta})$, is quite large of about 0.905. It is due to very small $\hat{\alpha}$ and $\hat{\beta}$ compare to n_i whereas $\hat{\alpha}$= 6.007941, $\hat{\beta}$=1.735254 and in average n_i is about 83. Therefore, it doesn't give significant affect to the Bayesian estimation.

However, using Jackknife method, the MSE estimation for direct estimation with Bayesian approach using beta as the prior distribution as well as for logit normal distribution is quite high. It means that these two methods do not give good accuracy.

Based on Bayesian approach, model-based estimation gives the best MSE values among the three methods. The model-based estimation can give much lower MSE than direct approach with prior Beta and logit-Normal distribution. So we conclude that based on Bayesian approach, the indirect estimation or SAE model gives the best estimation.

The three methods give consistence result, that literacy rate of Batuputih sub-district is the worst while Arjasa subdistrict has the highest literay rate in Sumenep regency.

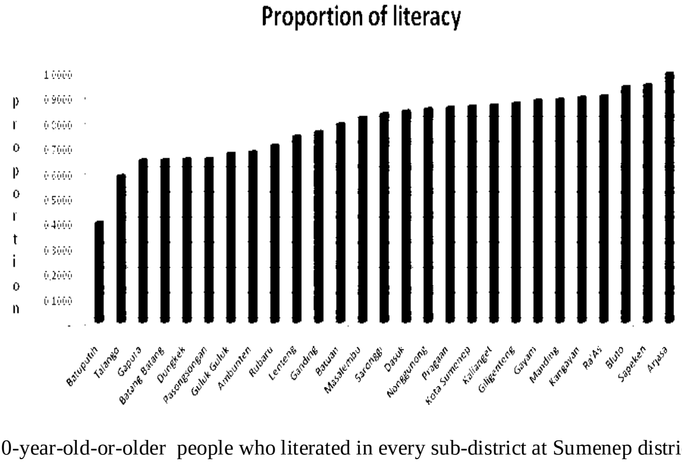


Figure 1. Proportion of 10-year-old-or-older people who literated in every sub-district at Sumenep district based on Susenas data, 2010

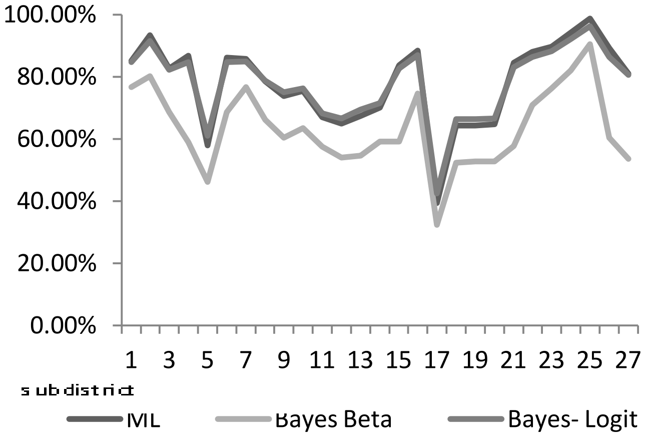


Figure 2. Plot of estimation result of parameter p_i (proportion of literate people aged 10 years old or older) with direct estimation method through classical (ML) and Bayesian approach

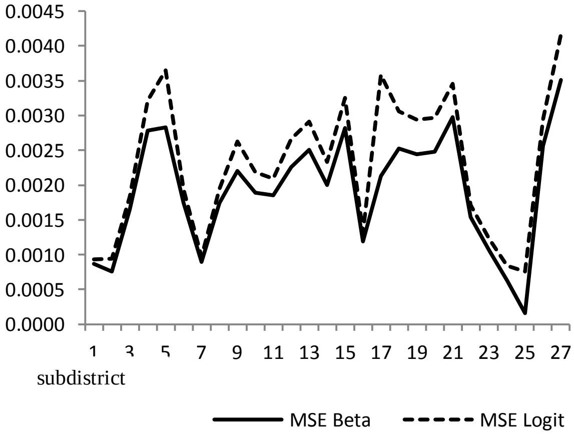
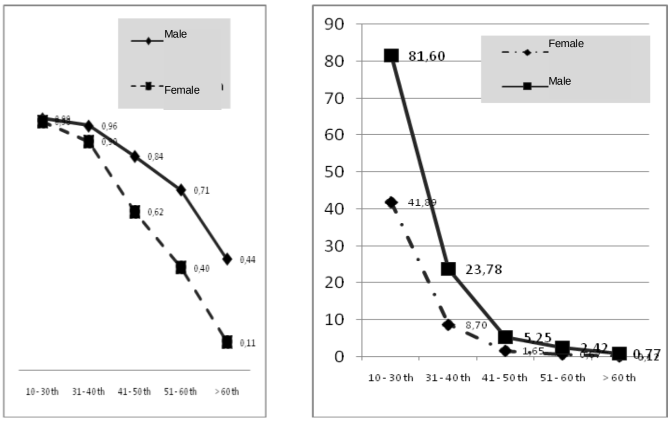


Figure 3. Plot for MSE estimation result to estimate parameter p_i (proportion of literate people aged 10 years old or older)



(a) (b)

Figure 4. The correlation between reading and writing ability with age based on the sex of population aged 10 years old or older in district of Sumenep

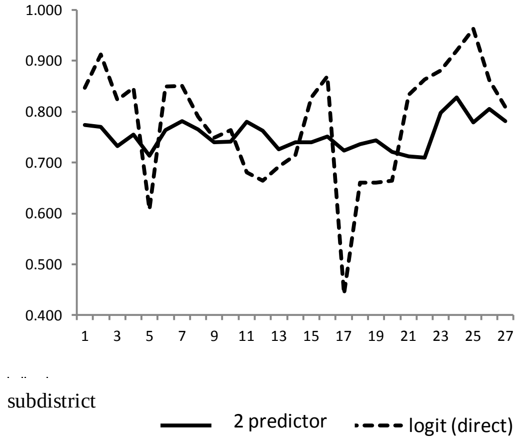


Figure 5. Plot for estimation result of parameter p_i (proportion of literate people aged 10 years old or older) with direct and indirect estimation method through Bayesian approach

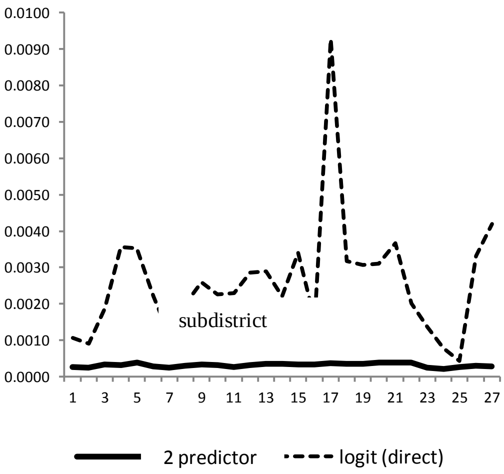


Figure 6. Plot for MSE estimation result to estimate parameter pi (proportion of literate people aged 10 years old or older)

TABLE 1.
PARAMETER ESTIMATION RESULT OF P_1 (PROPORTION OF LITERATE PEOPLE AGED 10 YEARS OLD OR OLDER) AND MSE IN SUBDISTRICTS

	Subdistrict	N	Direct	Logit		Beta		Model	
			pEB	pEB	MSE	pEB	MSE	pEB	MSE
1	Pragaan	51657	0,8521	0.8468	0.0011	0.8484	0.7840	0.7840	0.0002
2	Bluto	38456	0.9348	0.9123	0.0009	0.9228	0.7828	0.7828	0.0002
3	Saronggi	29270	0.8272	0.8234	0.0019	0.8231	0.7123	0.7123	0.0003
4	Giligenteng	22340	0.8684	0.8472	0.0036	0.8535	0.8275	0.8275	0.0003
5	Talango	32439	0.5802	0.6075	0.0035	0.5976	0.6972	0.6972	0.0004
6	Kalianget	32884	0.8615	0.8488	0.0022	0.8529	0.7212	0.7212	0.0004
7	Kota Sumenep	58880	0.8571	0.8507	0.0011	0.8529	0.7502	0.7502	0.0003
8	Batuan	10154	0.7865	0.7896	0.0020	0.7860	0.7768	0.7768	0.0001
9	Lenteng	48282	0.7375	0.7489	0.0026	0.7412	0.6858	0.6858	0.0004
10	Ganding	31254	0.7556	0.7634	0.0023	0.7575	0.7585	0.7585	0.0003
11	Guluk Guluk	44010	0.6696	0.6812	0.0023	0.6767	0.7488	0.7488	0.0003
12	Pasongsongan	36302	0.6489	0.6643	0.0029	0.6589	0.7573	0.7573	0.0003
13	Ambunten	31347	0.6750	0.6916	0.0029	0.6842	0.6688	0.6688	0.0005
14	Rubaru	31008	0.7010	0.7141	0.0022	0.7068	0.7476	0.7476	0.0003
15	Dasuk	25583	0.8372	0.8274	0.0034	0.8285	0.7800	0.7800	0.0002
16	Manding	24230	0.8837	0.8691	0.0016	0.8752	0.7502	0.7502	0.0003
17	Batuputih	37334	0.3950	0.4413	0.0093	0.4184	0.6749	0.6749	0.0004
18	Gapura	32170	0.6429	0.6609	0.0032	0.6544	0.7032	0.7032	0.0004
19	Batang Batang	44897	0.6437	0.6609	0.0031	0.6548	0.6706	0.6706	0.0004
20	Dungkek	32105	0.6471	0.6644	0.0031	0.6581	0.7240	0.7240	0.0003
21	Nonggunong	11686	0.8462	0.8327	0.0037	0.8352	0.7427	0.7427	0.0003
22	Gayam	28939	0.8806	0.8637	0.0020	0.8702	0.7057	0.7057	0.0004
23	Ra'As	30428	0.8977	0.8809	0.0014	0.8883	0.7817	0.7817	0.0003
24	Sapeken	33763	0.9412	0.9194	0.0008	0.9299	0.8383	0.8383	0.0002
25	Arjasa	49728	0.9874	0.9625	0.0004	0.9778	0.8663	0.8663	0.0001
26	Kangayan	17074	0.8919	0.8622	0.0033	0.8726	0.8207	0.8207	0.0003
27	Masalembu	17783	0.8108	0.8085	0.0042	0.8055	0.7172	0.7172	0.0003

REFERENCES

[1] R. E. Fay and R. A. Herriot, “Estimates of Income for Small Places: An Application of James Stain Procedures to Cencus Data”, *Journal of the American Statistical Association*, 1979.

[2] BPS, ”Kumpulan Metodologi Survei Sosial Tahun 2003-2005”, Badan Pusat Statistik (BPS), Jakarta, 2005.

[3] UNDP, “Concept and Measurement of Human Development”, *Human Development Report*, 1990.

[4] D. Malec, J. Sedransk, J. L. Moriarity, and F.B. Leclere, Small Area Inference for Binary Variables in National Health Interview Survey, *Journal of the American Statistical Association*, 1997.

[5] J. Jiang and P. Lahiri, *Empirical Best Prediction for Small Area Inference with Binary Data*. Annals of the Institute of Statistical Mathematics, 2001.

[6] J. N. K. Rao, *Small Area Estimation*, John Willey and Son, Inc. Publication, 2003.

[7] H. Chandra, R. Chambers, and N. Salvati, “Small Area Estimation of Proportions in Business Survey”. *Working Paper*. Centre for Statistical and Survey Methodology The University of Wollongong, 2009.

[8] H. J. Boonstra, B. Buelens, K. Leufkens, and M. Smeets, “Small Area Estimates of Labour Status in Dutch Municipalities”. Statistics Netherlands: The Hague/Heerlen, 2011.