Hierarchical Bayes Small Area Estimation for Gender Parity Index in Education: Case Study East Java Province

Rifdatun Ni'mah[1], Nur Iriawan[2]

[1]Department of Statistics, Sepuluh Nopember Institute of Technology
Surabaya 60111 Indonesia
rifdatun11@mhs.statistika.its.ac.id

[2]Department of Statistics, Sepuluh Nopember Institute of Technology
Surabaya 60111 Indonesia
nur_i@statistka.its.ac.id

Abstract: ***Gender gaps can be identified through indicators i.e. Gender Parity Index (GPI). It is calculated based on the ratio of the School Participation Rate (SPR) of women to men at every level of education. The direct estimate of GPI is obtained from the result of the survey which has a small sample size for each level of education. An estimation variable with a few available samples can be done by Small Area Estimation (SAE). Our applied Hierarchical Bayesian (HB) approach for SAE using model linking log linier to estimate the IPG on the three levels of education in East Java province. Markov Chain Monte Carlo (MCMC) with Gibbs sampling methods are consider to solve the computational aspect. HB SAE under the Fay - Herriot model is obtained using some auxiliary variable such as a number average of household members, expenditure average per capita per month, the number of building schools and region's education budgets. There are two significant results. First, HB has a better strength in explaining the variability in areas with small sample sizes and could reduce CV's direct survey estimates about ± 29.21%. Second, cross validation's result shows that the HB SAE model for all levels of education is fit to the actual observations and reliable.***

Keywords: Small Area, Hierarchical Bayes, MCMC, Gender Parity.

1. Introduction

Gender Parity Index (GPI) is one of the indicators that used to measure the gaps in gender issue and the achievement of gender-based development in the education, health and economy. It is a ratio that capture the performance of women to men. The goal of development effort is to make GPI close to one. It means that the development provides equal opportunities between men and women. GPI measured by participation rates is for national and provincial scale. Information for region/city can't be obtained. Participation rate itself is calculated using data from Survei Sosial Ekonomi Nasional (Susenas — National Social Economic Survey). The unit sampling for Susenas is household and its sample size is available for regions level from now on. But its sample size will be small when we break it down to group age level of education.

Problems sample size requires a more complex approach than using a simple direct estimation. A special estimator i.e. estimators that borrow information from other related areas in space or time or through an expected supplementary information related to the observed variables, can be formed. This estimator is named as a small area estimator. Small Area Estimation (SAE) is an important topic due to the increasing demand for reliable small area statistics even when only a few samples are available for these areas [1]. Many recent studies that have applied these techniques to obtain reliable estimates of variable interest for small areas [2]-[7]. Some research showed that Bayes approach is reliable for a variety of cases contexts [8]-[10]. We consider applying a mismatched linking model to improve the estimation [8]. References [11] suggesting various mismatched linking models using HB approach e.g. log linear linking model. It is useful to avoid negative estimates on the observed variables.

The purpose of this paper is to obtain reliable model-based-estimate for IPG at regions level within province for each level of education. We explain about the area level model under the Fay-Herriot model to assume that the sampling variance is unknown and propose HB approach regarding to log linear linking model in section 2. We also explain about checking model using posterior predictive. We identify some auxiliary variables that used for model in section 3. The comparison between direct estimates and model-based estimates for IPG at each level of education is conducted in section 4. Checking the goodness of fit of the model was also conducted to determine the accuracy and reliability of the model-based estimates. In the last section, we offer some conclusions.

2. Small Area Estimation

2.1 Area level model

Let θ_i denoted the area parameter of interest for the i-th area, where i=1, ...,m and m is the total number of areas. Area level model assumes that assumes θ_i is related to area level; auxiliary variable through $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ through a linear model.

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the $p \times 1$ vector of unknown regression parameters and v_i are area random effects. There are several assumptions for v_i Such as being independent and identical with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$. Normality of v_i

may also be assumed. Model (1) is referred to as a linking model for θ_i.

Basic area level model also assumes that the direct estimates y_i of variables interest θ_i can be calculated when the sample size of the area $n_i > 1$ [12]. The following model for y_i is

$$y_i = \theta_i + \varepsilon_i \quad (2)$$

where ε_i is the sampling error associated with the direct estimates. The required assumptions for random variables ε_i are independent and normally distributed with mean $E(\varepsilon_i|\theta_i) = 0$ and variance sampling $V(\varepsilon_i|\theta_i) = \sigma_i^2$. Model (2) is called a sampling model for direct estimates y_i.

The combination between models (1) and (2) will lead to area level linear mixed model that well-known as Fay-Herriot model [13].

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i \quad (3)$$

The sampling variance is usually assumed to be known in the model. In general, it can be estimated directly from the survey data. Reference [14] proposed a Hierarchical Bayes approach to address the estimation of σ_i^2.

2.2 Linking model

One of the alternative linking models that can be used is log linear models [15].

$$\log(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta} + v_i \quad (4)$$

The sampling model (2) and linking model linking (4) can be presented in a HB framework as follows

$$y_i|\theta_i \sim N(\theta_i, \sigma_i^2) \quad (5)$$

$$\log(\theta_i)|\boldsymbol{\beta}, \sigma_v^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2) \quad (6)$$

The linking models (4) implies that the mean of small area conditionally has a log normal distribution. The posterior mean can be obtained as the HB estimator and the posterior variance as the measure of uncertainty for the estimator.

2.3 Hierarchical Bayes Approach

HB framework can be established in two models under the assumption that the sampling variance σ_i^2 is known and unknown [16]. The following HB framework under Fay-Herriot model with assuming that the sampling variance is unknown and estimated using direct unbiased estimator s_i^2 is

$$y_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2) \quad (7)$$

$$d_i s_i^2|\sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2 \quad (8)$$

$$\theta_i|\boldsymbol{\beta}, \sigma_v^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2) \quad (9)$$

where $d_i = n_i - 1$.

Priors for the parameters $\boldsymbol{\beta}, \sigma_v^2, \sigma_i^2$ are

$$\pi(\boldsymbol{\beta}) \propto 1 \quad (10)$$

$$\pi(\sigma_v^2) \sim IG(a_0, b_0) \quad (11)$$

$$\pi(\sigma_i^2) \sim IG(a_i, b_i) \quad (12)$$

where a_i, b_i $(0 \le i \le m)$ are chosen for a very small constant to describe the vagueness of σ_v^2 and σ_i^2. Let $N(.)$ denoted for normal distribution and $IG(.)$ for the inverse gamma distribution [14].

The linking model defined in (4) to (6) could change the probability function (9) becomes

$$\log(\theta_i)|\boldsymbol{\beta}, \sigma_v^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2) \quad (13)$$

We can use Gibbs sampling method [17] with Metropolis Hastings algorithm [18] to find the posterior mean and variance. More detailed explanation can be found in [12].

We compute the Hierarchical Bayesian estimate using Markov Chain Monte Carlo (MCMC) with apply Gibbs sampling method [17]. The Gibbs sampling only samples from the full conditional distributions. Therefore we obtained the joint distribution of formed joint distribution of $\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\beta}$ and σ_v^2 under the conditions (7) through (13). The joint distribution of $\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2, \boldsymbol{\beta}$ and σ_v^2 takes the form

$$[\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\sigma}^2] \cdot [\log(\boldsymbol{\theta})|\boldsymbol{\beta}, \sigma_v^2] \cdot [\boldsymbol{\sigma}^2] \cdot [\boldsymbol{\beta}] \cdot [\sigma_v^2] \quad (14)$$

where

$$[\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\sigma}^2] \cdot [\log(\boldsymbol{\theta})|\boldsymbol{\beta}, \sigma_v^2] \cdot [\boldsymbol{\sigma}^2] \cdot [\boldsymbol{\beta}] \cdot [\sigma_v^2]$$
$$= \prod_{i=1}^{m} [y_i|\theta_i, \sigma_i^2] \cdot [d_i s_i^2|\sigma_i^2] \cdot [\log(\theta_i)|\boldsymbol{\beta}, \sigma_v^2] \cdot [\sigma_i^2] \quad (15)$$

We focus to obtain the full conditional distribution for posterior $[\theta_i|\boldsymbol{y}]$. From the Gibbs sampling perspective, there are $(2m + 2)$-variable problem together with $\boldsymbol{\beta}$ and σ_v^2. We identify the forms of the full conditionals based on (13) as follows

$$[\log(\boldsymbol{\theta})]\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_v^2| = N(\boldsymbol{\theta}^*, \boldsymbol{\tau}^*) \quad (16)$$

where

$$\theta_i^* = \frac{y_i\sigma_v^2 + \mathbf{x}_i'\boldsymbol{\beta}\sigma_i^2}{\sigma_v^2 + \sigma_i^2} \quad (17)$$

$$\tau_i^* = \frac{\sigma_v^2\sigma_i^2}{\sigma_v^2 + \sigma_i^2} \quad (18)$$

We can change the parameter $\boldsymbol{\tau}^*$ into a simpler form, namely coefficient of variance partition γ_i

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2} \quad (19)$$

So that, the equation (17) and (18) can be transformed into a simple form.

$$\theta_i^* = \gamma_i y_i + (1 - \gamma_i)\mathbf{x}_i'\boldsymbol{\beta} \quad (20)$$

$$\tau_i^* = \sigma_i^2 \gamma_i \quad (21)$$

The full conditional distribution of $[\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\sigma^2},\sigma_v^2]$ can be expressed as the following

$$\left[\boldsymbol{\beta}\middle|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\sigma}^2,\sigma_v^2\right] = \left[\boldsymbol{\beta}\middle|\boldsymbol{\theta},\sigma_v^2\right]$$
$$= N\left(\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_j^{'}\right)^{-1}\left(\sum_{i=1}^{m}x_i\log(\theta_i)\right),\sigma_v^2\left(\sum_{i=1}^{m}x_i x_j^{'}\right)^{-1}\right) \quad (22)$$

For both component of variance, the variance of random effect σ_i^2 and the variance of sampling error $\boldsymbol{\sigma_v^2}$, full distribution conditional for both of them can be expressed as follows

$$\left[\boldsymbol{\sigma}^2\middle|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma_v^2\right] = \left[\sigma_i^2\middle|\mathbf{y},\boldsymbol{\theta}\right] = \prod_{i=1}^{m}\left[\sigma_i^2\middle|y_i,s_i^2,\theta_i\right]$$
$$= IG\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i - \log(\theta_i))^2 + d_i s_i^2}{2}\right) \quad (23)$$
$$\left[\sigma_v^2\middle|\mathbf{y},\boldsymbol{\theta},\boldsymbol{\sigma}^2,\boldsymbol{\beta}\right] = \left[\sigma_v^2\middle|\boldsymbol{\theta},\boldsymbol{\beta}\right]$$
$$= IG\left(a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^{m}\left(\log(\theta_i) - \mathbf{x}_i^{'}\boldsymbol{\beta}\right)^2\right) \quad (24)$$

Since all conditional distributions are formed, we can apply the implementation of Gibbs sampling.

2.4 Cross Validation

It is a technique for assessing how the results of a statistic-al analysis will be generalized to an independent data set. The purpose why these statistics are often used is to predict and estimate how accurately a predictive model will emerge in practice [19]-[21]. It's used to compare the performance of the different prediction models' methods. Leave-one-out cross validation involving the observations from the original data as training data and the remaining observations as the testing data. This method is repeated as many as the number of observations in the sample so that each observation is used once as the validation data.

Checking the HB estimates are applied by replicating the posterior predictive. We apply leave-one-out cross validation. The results of the replication is a probability value called Bayesian p-value. Bayesian p-value present the probability that the results of simulation can be more extreme than observational data. They are measured by

$$P_b = \Pr\left[\mathbf{y}^{rep}\middle|\boldsymbol{\theta} \geq \mathbf{y}\middle|\boldsymbol{\theta}\right] \quad (25)$$

Probability values that are too high or too low indicates lack of fit. A model is suspected if the tail probability close to 0 or 1 which indicated that the pattern of replication is not true if the model is correct [22]. If $P_B \leq 0.10$ or $P_B \geq 0.90$ is necessary to note about the lack of fit of the model. For the MCMC, the posterior distribution formed by all observations should not change much if one observation is removed (leave-one-out).

3. Methodology

3.1 Data and Variables

We collected data indicators that needed to measure the participation rate. Data is obtained from Susenas, Census and other sources with basic year 2010. The variable interest is GPI for three level of education. The population for the first levels are people in the age group 7 to 12 years. The second and third levels of education are the age group 13 to 15 years and 16 to 18 years old.

The auxiliary variables are determined by two factors, the internal factors and the external factors. The internal factors are the number average of family members and the expenditure average per capita per month. The external factors that used are the number of school buildings and the amount of region the education budget. There are 38 regions in East Java province that identified as area level for the model.

3.2 Procedure

GPI is estimated using HB approach. It requires a fairly complex computational processes. We apply Gibbs sampling method for the Markov Chain Monte Carlo. We use WinBUGS to help us solve the computational of HB estimate. To obtain a convergent value estimate, we try using the number of updates performed N = 50,000, T = 20 and the number of burn-in sample B = 200. The greater the iteration is done so then it will be getting closer to obtain a convergent value. Model-based estimation for GPI using HB approach is compared with the direct estimates. The evaluation process is seen from the residue, coefficient of variation and cross validation. Evaluation using a cross validation method applied leave-one-out on the posterior predictive.

4. Result

The convergence of parameters can be achieved after auxiliary variables used in the model undergoing a process of transformation. The process of computing using the original observational data has not been able to produce convergent estimates. So that data have been transformed by standardize. Standardize is done by reducing the data with the mean and dividing by the standard deviation of each variable.

Table 1: Estimate of parameters HB

Parameter	*Level 1*	*Level 2*	*Level 3*
β_1	* 0.0014 (±0.1191)	*-0.0003 (±0.1206)	*-0.0103 (±0.1231)
β_2	*-0.0028 (±0.1440)	*-0.0208 (±0.1218)	* 0.0313 (±0.1355)
β_3	* 0.0053 (±0.2200)	*-0.0036 (±0.1904)	* 0.0330 (±0.2219)
β_4	*-0.0055 (±0.1928)	* 0.0086 (±0.1886)	*-0.0682 (±0.2219)
β_0	1.0060 (±0.1070)	1.0060 (±0.1174)	0.9262 (±0.1185)

* Confidence interval 95% consists zero number

Table 1 shows that all variables used in this case have not been able to have a significant influence. It applies to all levels of education were observed. Only the intercepts that have the most significant role in the model indirectly. A rough idea of the GPI in the province of East Java for all levels of education can be seen from the value of the

intercept model. GPI for the first and second lower secondary levels close to 1. It shows that the proportion of women participating in the education levels one and two almost equal and/or greater than male participation. Intercept for third level of education is under number 1. The proportion of female participation in the third level of education is lower than men.

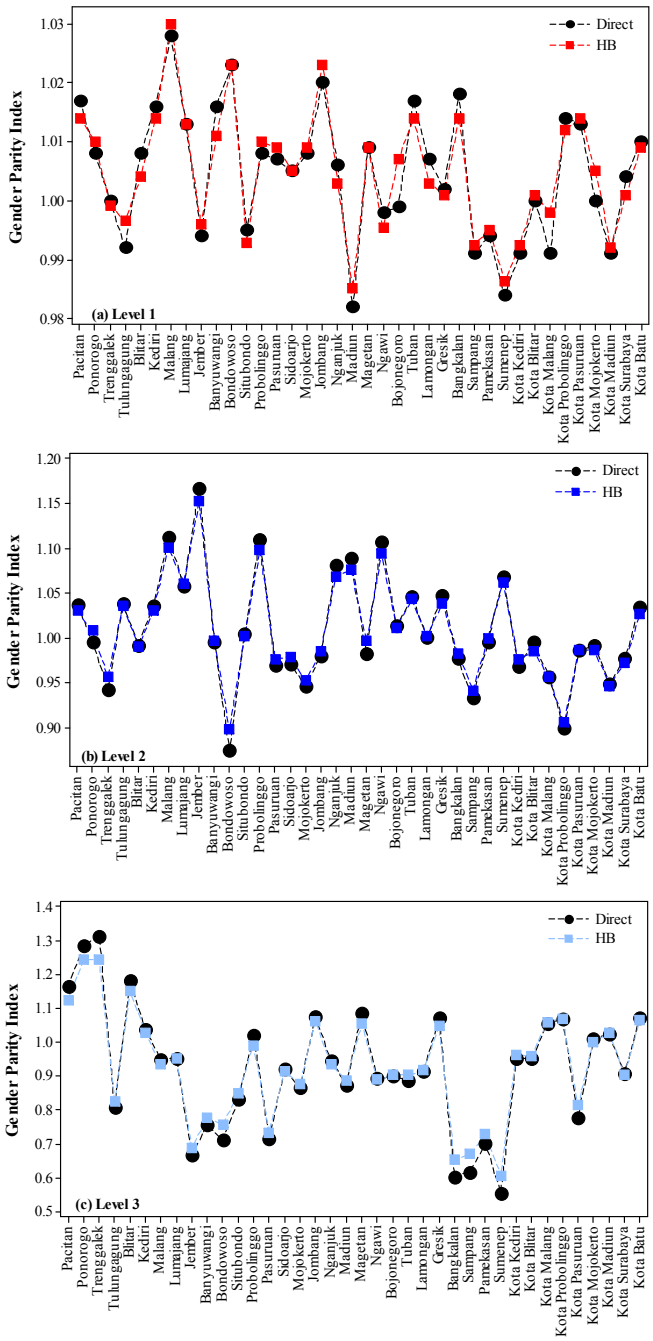

Figure 1. Comparison GPI between direct estimates and HB

Figure 1 shows a comparison estimates using either direct estimation or HB approach. The movement pattern of HB estimates for all level of education are able to follow the pattern of direct estimation. The random effect variance for each level sequentially from the first to the third level of education is 2.530 (± 0.4917), 2.397 (± 0.4723) and 2.324 (± 0.4600). In the figure 1, the sample size for the first levels to the third levels are decreases.

The residuals of HB estimates are presented in Figure 2. The range of residual on the first level is spread at ± 0.0061. The region that has the greatest residual on the estimation of GPI for the first level is Bojonegoro district. Residual between direct estimates and HB on the second level of education are in the range of ± 0.0192. The number of residual on the second level is larger than the first level of education. HB estimate for Bondowoso district on the second level has the highest residual rate. HB's residual for the third levels spread around the points ± 0.0613. The residual for the third levels higher than the first and second levels. GPI estimation results for Trenggalek districts have the greatest residual value. It is about 0.07095. Direct estimate for Trenggalek district is 1.313 whereas HB estimate is 1.2420.

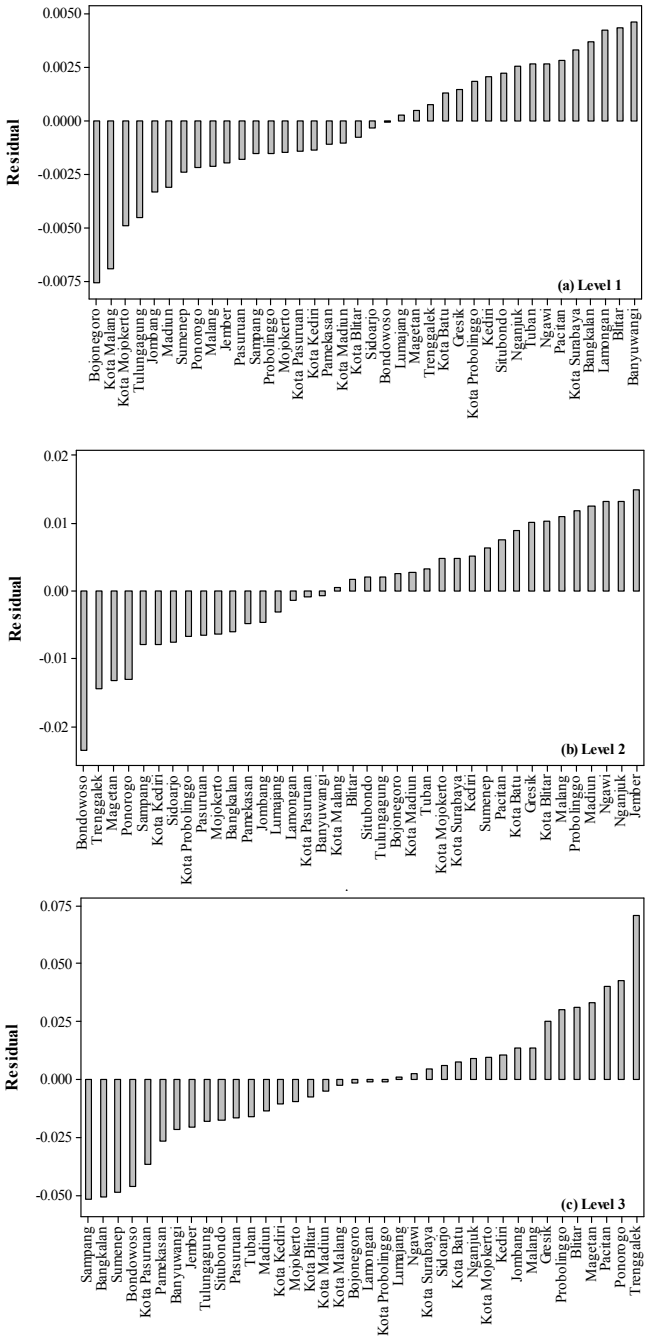

Figure 2. Residual's HB toward direct estimates

The coefficient of variation is presented by district/city with the largest sample size to the smallest sample size. The coefficient of variation was calculated to see the relative variability of the estimation. The coefficient of variation of the HB is calculated by dividing the standard deviation by the posterior mean. HB is able to provide smoother

coefficient of variation when compared with direct estimates. HB is able to reduce the coefficient of variation of 29.17% on average in the first levels. The coefficient of variation produced by HB consistently delivering smaller value relative variation than direct estimation. HB is able to reduce the coefficient of variation direct estimates on the second level by an average of 29.02% with a range of 28.01% to 30.79%. For the third level of education, HB is able to reduce the coefficient of variation about 29.44% of direct estimates.

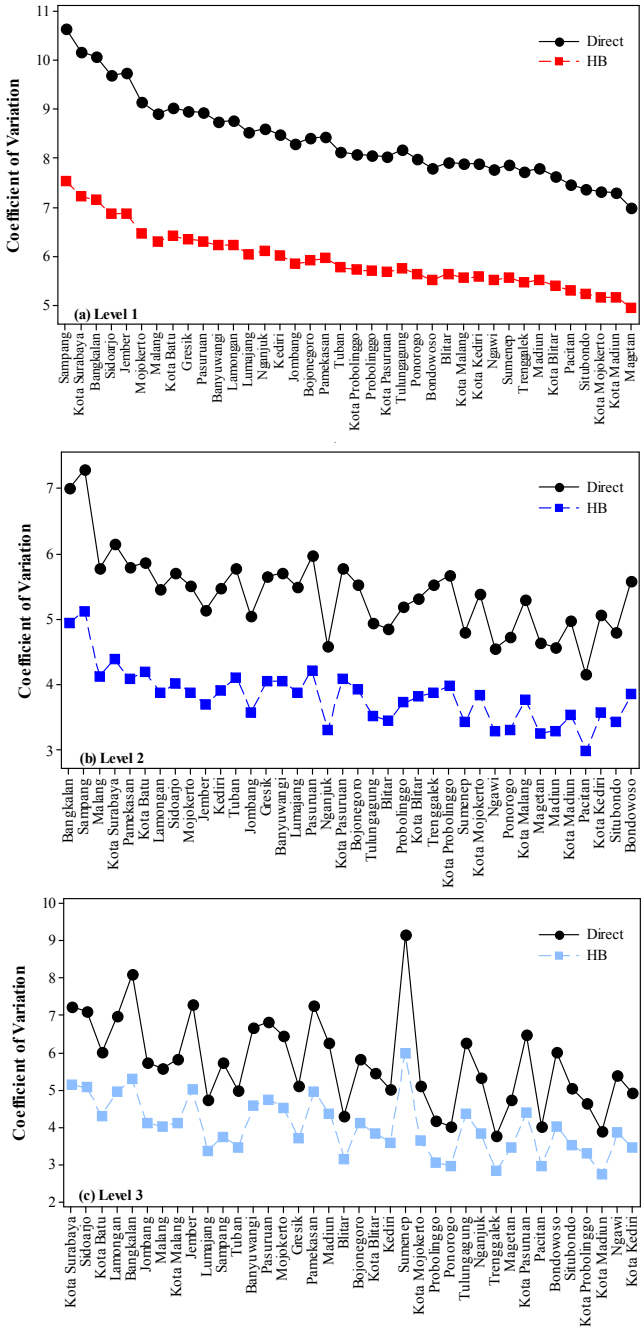

Figure 3. CV's comparison between direct estimates and HB estimates

The highest reduction of the coefficient of variation is 29.74% for first level. The greatest reduction occurred in the Bojonegoro district caused the considerable deviation between the direct estimate and HB. The same thing happens in second level of education. Bondowoso district has the greatest residual. It is about 30.79%. Unlike the other levels, the magnitude of the deviation that occurred in the Trenggalek district for the third level are indicated because of HB is only able to reduce the coefficient of variation of the direct estimate about 24.93%.

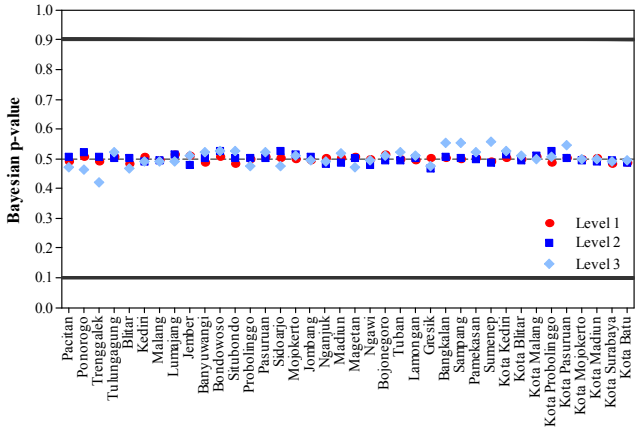

Figure 4. Bayesian *p-value* untuk Prediksi Posterior HB

Checking model is performed using posterior predictive. Posterior probabilities for the predictions for all districts/cities in East Java are in the safe range of numbers. The resulting probability value did not approach the lower and upper extreme values. This indicates that HB estimate and the indirect modelling that obtained for GPI in the first to third levels are reliable and in accordance with the observations.

5. Conclusion

The results shows that only the intercepts are able to exercise significant influence, while auxiliary variables have not been able to have a significant influence. The measurement of the relative variability expressed in the coefficient of variation. It suggests that HB estimate is able to capture the variability of the small sample size well. The coefficient of variation of the HB gives a smaller value of diversity when compared to the direct estimates for all levels of education. The HB is able to reduce the coefficient of variability of direct estimates by an average of ± 29.21% for all levels of education. Testing models using posterior predictive shows that all levels of education have the appropriate estimation model.

For further research, it is needed to adding the auxiliary variables for anticipate the absence of correlations between the observation variables and no information entered during the formation of indirect model. Implementation of the other SAE approach and the linking model such as proposed by [11] can be done to get the best estimate and do a comparison for the same case study.

References

[1] D. Pfefferman, "Small area estimation-New developments and direc-tions," *International Statistical Review*, vol. 70, pp. 125-143, 2002.

[2] C. Elbers, J. Lanjouw, and P. Lanjouw, "Micro-level estimation of poverty and inequality," *Econometrica*, vol. 71, pp. 355-364, 2003.

[3] E. V. Slud and T. Maiti, "Small area estimation based on survey data from a left-censored fay-herriot model," *Journal of Statistical Planning and Inference*, vol. 141, no. 11, pp. 3520-3535, 2011.

[4] R. Chambers and N. Tzavidis, "M-quantile models for small area estimation," *Biometrika*, vol. 93, pp. 255-268, 2006.
[5] C. Quintano, R. Castellano, and G. Punzo, "Estimating poverty in the italian provinces using small area estimation models," *Metodoloski zvezki*, vol. 4, no. 1, pp. 37-70, 2007.
[6] E. Fabrizi, M. R. Ferrante, and S. Pacei, "Small area estimation of average household income based on unit level models for panel data," *Survey Methodology*, vol. 33, no.2, pp. 187-198, 2007.
[7] S. Song, "Small area estimation of unemployment: from feasibility to implementation," in New Zealand Association of Economists Confe-rence, 2011.
[8] Y. You, J. Rao, and P. Dick, "Benchmarking hierarchical bayes small area estimators with application in census undercoverage estimation," in *Proc. SSC of the Survey Methods Section*, 2002, pp. 81-86.
[9] J. Rao, "Small area estimation: methods and applications. Applications of small area estimation techniques in the social sciences," Mexico City: Iberoamerican University, 2012.
[10] L. Mohadjer, J. N. Rao, B. Liu, T. Krenzke,and W. Van de Kerckhove, "Hierarchical bayes small area estimates of adult literacy using un-matched sampling and linking models," *Journal of the Indian Society of Agricultural Statistics*, vol. 66, pp. 55-63, 2012.
[11] M. Trevisani and N. Torelli, N. "Hierarchical bayesian models for small area estimation with count data," Trieste: Universita Degli Studi di Trieste, 2007.
[12] Y. You and J. Rao, "Small area estimation using unmatched sampling and linking model," *The Canadian Journal of Statistics*, vol. 30, pp. 3-15, 2002.
[13] R. Fay and R. A. Herriot, "Estimation of income for small places: an application of james-stein procedures to census data," *Journal of the American Statistical Association*, vol. 74, pp. 268-277, 1979.
[14] Y. You and B. Chapman, "Small area estimation using area level models and estimated sampling variances," *Survey Methodology*, vol. 32, pp. 97-103, 2006.
[15] G. S. Datta, M. Ghosh, R. Steorts, and J. Maples, "Bayesian benchmarking with applications to small area estimation," Washington D.C., United State: U.S. Census Bureau, 2009.
[16] Q. M. Zhou and Y. You, "Hierachical bayes small area estimation for the canadian community health survey," in *Proc. SSC of the Survey Methods Section*, 2008.
[17] A. Gelfand and A. Smith, "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, pp. 398-409, 1990.
[18] S. Chip and E. Greenberg, "Understanding the metropolis-hasting algorithm," *The American Statistician*, vol. 94, pp. 327-335, 1995.
[19] P. A. Devijver, "Pattern recognition: a statistical approach," London: Prentice-Ppl, 1982.
[20] S. Geisser, "Predictive inference," New York: Chapman and Ppl, 1993.
[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence 2*, pp. 1137-1143, 1995.
[22] A. Gelman, J. B. Carlin, H. Stern, and D. B. Rubin, "Bayesian data analysis," London: Chapman and Ppl, 1995.

Author Profile

Rifdatun Ni'mah was born in Gresik, Indonesia, in 1990. She received the B.S. in Statistics from Sepuluh Nopember Institute of Technology in 2011; she is taking her Master degree in Statistics from Sepuluh Nopember Institute of Technology.

Nur Iriawan received the B.S. degree in statistics from Sepuluh Nopember Institute of Technology, Indonesia in 1986 and the M.S. degree in Computer Science from Indonesia University and Maryland University, U.S in 1990. He joined Sandwich Program for his Master degree. He received the Ph.D. degree from School of Maths. and Stats., Curtin University of Technology, Perth-Western Australia in 1999. He is currently a professor and lecturer in statistics department of Sepuluh Nopember Institute of Technology.