# Clustering Information of Non-Sampled Area in Small Area Estimation of Poverty Indicators



# Clustering Information of Non-Sampled Area in Small Area Estimation of Poverty Indicators

**V Y Sundara[1], A Kurnia[1]* and K Sadik[1]**

[1]Department of Statistics, Bogor Agricultural University, Indonesia

Email: anangk@apps.ipb.ac.id

**Abstract.** Empirical Bayes (EB) is one of indirect estimates methods which used to estimate parameters in small area. Molina and Rao has been used this method for estimates nonlinear small area parameter based on a nested error model. Problems occur when this method is used to estimate parameter of non-sampled area which is solely based on synthetic model which ignore the area effects. This paper proposed an approach to clustering area effects of auxiliary variable by assuming that there are similarities among particular area. A simulation study was presented to demonstrate the proposed approach. All estimations were evaluated based on the relative bias and relative root mean squares error. The result of simulation showed that proposed approach can improve the ability of model to estimate non-sampled area. The proposed model was applied to estimate poverty indicators at sub-districts level in regency and city of Bogor, West Java, Indonesia. The result of case study, relative root mean squares error prediction of empirical Bayes with information cluster is smaller than synthetic model.

## 1. Introduction

Survey is a method to collect data which has several advantages such as less time, less budget and other resources. Most of national surveys provide limited information and low precision in predicting the small area level [1]. Direct estimate of parameters in area with small sample size would have large variance and even cannott produce estimates when there is no sample unit for those area [2]. The science of small area estimation is developed to solve this problem. The indirect estimate or model-based in small area estimation was first introduced by Fay and Heriot [3] using anempirical Bayes (EB) method. The problem of the ordinary EB for non-sampled area using synthetic model is ignoring the area effects. Anisa et al (2014) studied the effect of adding the cluster information empirical unbiased best linear prediction on non-sampled area to produce better predictions [4]. The simulation study showed the use of factor analysis in clustering has increased the average percentage of accuracy particularly when the Ward method is implemented [5].

Poverty is one issue that became the focus of governments in many countries. Statistics Indonesia uses poverty indicator which is proposed by Foster, Greer and Thorbecke. The indicators are head count index - $P_0$ (the percentage of population below the poverty line), poverty gap index - $P_1$(the average size of each expenditure gap of the poor to the line poverty) and poverty severity index - $P_2$(provide a picture of the spread of expenditure among the poor).

In this paper, simulation study is carried out to evaluate the performance of the proposed models or cluster information with ordinary empirical Bayes or Molina and Rao method. This paper also presents application of the proposed model using data from Statistics Indonesia to estimate poverty indicators at sub-districts level in regency and city of Bogor.

## 2. Direct Estimation and Empirical Bayes Estimation of Poverty Indicators

Direct estimation is a method parameter estimation which is only based on sample [6]. Direct estimation has a large variance for small sample size. Direct estimation of FGT poverty measures can be defined as follows

$$P_{ai} = \frac{1}{N_i}\sum_{j=1}^{N_i} P_{\alpha ij}\,, \quad i = 1, \dots, m \tag{2.1}$$

with $P_{\alpha ij}$defined as

$$P_{aij} = \left(\frac{z - E_{ij}}{z}\right)^{\alpha} I\left(E_{ij} < z\right), \qquad j = 1, \dots, N_i, \qquad \alpha = 0,1,2 \tag{2.2}$$

where $I\left(E_{ij} < z\right) = 1$ if $E_{ij} < z$ (person under poverty) and $I\left(E_{ij} < z\right) = 0$ if $E_{ij} \geq z$ (person not under poverty). For α = 0 is the proportion of the population below the poverty line, called the head count index ($P_0$). For α = 1 represents the poverty gap index ($P_1$), and for α = 2 is the poverty severity index ($P_2$). Estimates of FGT poverty measure for each small area $i$ is as follows

$$\hat{P}_{ai} = \frac{1}{n_i}\sum_{j=1}^{n_i} P_{\alpha ij}\,, \quad i = 1, \dots, m, \quad \alpha = 0,1,2 \tag{2.3}$$

Molina and Rao proposed EB method to estimate poverty measures in a small area [1]. This method assumes that welfare variable can be transformed to follows a normal distribution. Let $Y_{ij} = T(E_{ij})$ is a transformation on welfare, and $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$. Thus FGT poverty measure for each small area i in equation (2.2) can be written as follows

$$P_{aij} = \left(\frac{z - T^{-1}\left(Y_{ij}\right)}{z}\right)^{\alpha} I\left(T^{-1}\left(Y_{ij}\right) < z\right), \qquad j = 1, \dots, N_i, \qquad \alpha = 0,1,2 \tag{2.4}$$

Empirical Bayes estimation for the poverty measure $P_{ai}$ formulated as follows [1]:

$$\hat{P}_{\alpha i}^{EB} = \frac{1}{N_i}\left[\sum_{j \in s_i} P_{\alpha ij} + \sum_{j \in r_i} \hat{P}_{\alpha ij}^{EB}\right] \tag{2.5}$$

where $s_i$ is the sampled unit and $r_i$ the non-sampled unit. $\hat{P}_{\alpha ij}^{EB}$ is an empirical Bayes estimators of $P_{\alpha ij} = h_\alpha(Y_{ij})$ are defined as follows:

$$\hat{P}_{aij}^{EB} = E_{\boldsymbol{y_r}}\left[h_\alpha\left(Y_{ij}\right)|\boldsymbol{y_s}\right] = \int_{IR} h_\alpha(y) f_{Y_{ij}}(y|\boldsymbol{y_s})\, d_y, \quad j \in r_i \tag{2.6}$$

where $f_{Y_{ij}}(y|\boldsymbol{y_s})$ is a conditional probability density function of $Y_{ij}$. According to Molina and Rao [1], the complexity of the functions $h_\alpha(.)$, there is not explicit expression for the expectation in (2.6), but it can be approximated by Monte Carlo. For this, generate $L$ replicates of $\boldsymbol{y}_r$ from the distribution of $\boldsymbol{y}_{ir}|\boldsymbol{y}_{is} \sim N(\boldsymbol{\mu}_{r|s}, \boldsymbol{V}_{r|s})$ or generate univariate value $Y_{ij}^{(l)}$ from the nested error linear regression model as follows:

$$Y_{ij}^{(l)} = \boldsymbol{x}_{ij}\hat{\beta} + \hat{u}_i + \varepsilon_{ij}\,, \quad u_i \sim N\left(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i)\right), \quad \varepsilon_{ij} \sim N(0, \hat{\sigma}_e^2) \tag{2.7}$$

where $\hat{\gamma}_i = \sigma_u^2(\sigma_u^2 + \sigma_e^2/n_i)^{-1}$ and $n_i$ is the sample size. Then, an approximate to the best predictor of $Y_{ij}$ is

$$\hat{P}_{\alpha ij}^{EB} \approx \frac{1}{L}\sum_{l=1}^{L} h_\alpha\left(\boldsymbol{y}_{ij}^{(l)}\right), j \in r_i \tag{2.8}$$

For $n_i = 0$ or domain $i$ is not sampled, then $Y_{ij}^{(l)}$ for $j = 1, \dots, N_i$ are generated by bootstrap from $Y_{ij} = x_{ij}'\hat{\beta} + u_i^* + e_{ij}^*$ where $u_i^* \sim iidN(0, \hat{\sigma}_u^2)$ and $e_{ij}^* \sim iidN(0, \hat{\sigma}_e^2)$ with assume $u_i^*$ is independent of $e_{ij}^*$. Formula (2.6) is used to get the estimator $\hat{P}_{\alpha ij}^{EB}$ of $P_{aij}$ and the EB estimator $P_{ai}$ is defined as

$$\hat{P}_{\alpha i}^{EB} = \frac{1}{N_i}\sum_{j=1}^{N_i}\hat{P}_{\alpha ij}^{EB} \tag{2.9}$$

The estimator (2.9) is essentially a synthetic estimator since no sample observations are available from domain $i$ if $n_i = 0$

## 3. Cluster Analysis

Cluster analysis is a multivariate technique to classify object based on similarities or dissimilarities. Characteristics between object can be measured by Euclidean distance, Mahalanobis distance, and others. There are two approaches in cluster method, hierarchical and non-hierarchical methods. Hierarchical method is used when the number of cluster is unknown. While non-hierarchical methods is used when the number of cluster is known [7]. Cluster analysis can be used when there is non-sampled area. The addition of cluster information to non-sampled areas shows that in general has a better prediction [4].

## 4. Proposed Model

The basic model which is used in this study is a nested error regression model. This model referred to Molina and Rao model. This study use unit level with $i$ and $j$ respectively denotes area and unit of sampled area, while $i^*$ and $j^*$ respectively denotes area and unit of non-sampled area. Response variable denoted by $y_{ij}$, auxiliary variable denoted by $x_{ij}$, random effect of area denoted by $u_i$, and sampling error for sampled area denoted by $e_{ij}$. Prediction of non-sampled area $(\hat{y}_{i^*j^*})$ is obtained from auxiliary variables of non-sampled area$(x_{i^*j^*})$. The Molina and Rao model for one auxiliary variable can be written as follows:

a) Model for population :

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij} \tag{4.1}$$

b) Prediction model for sampled area:

$$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_i \tag{4.2}$$

c) Prediction model for non-sampled area :

$$\hat{y}_{i^*j^*} = \hat{\beta}_0 + \hat{\beta}_1 x_{i^*j^*} \tag{4.3}$$

The cluster information model is modifying basic model by adding average the random effects of area on each clusters into prediction model for non-sampled area. The average of random effects of area is defined by $\bar{\hat{u}}_{(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k}\hat{u}_i$ with $m_k$ is the number of sample area on the $k^{\text{th}}$ cluster. The cluster information model can be written as follows:

a) Model for population :

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_i + e_{ijk} \tag{4.4}$$

b) Prediction model for sampled area :

$$\hat{y}_{ijk} = \hat{\beta}_0 + \hat{\beta}_1 x_{ijk} + \hat{u}_i \tag{4.5}$$

c) Prediction model for non-sampled area :

$$\hat{y}_{i^*j^*k} = \hat{\beta}_0 + \hat{\beta}_1 x_{i^*j^*k} + \bar{\hat{u}}_{(k)} \tag{4.6}$$

## 5. Simulation Study

A simulation study has been carried out to study the performance of the proposed model of non-sampled area. We simulate a population of size $N = 502$, composed of $M = 46$ areas with each area is 6-16 units. Response variable for a population generated by nested error regression. Auxiliary variables generated by the uniform distribution $U(0,10)$ with intercept $\beta_0 = 1.0$ and regression coefficient $\beta_1 = 0.5$. The population is divides into three clusters, with each cluster consisting of 18, 19 and 9 areas respectively. The intercept and regression coefficients for each cluster are shown in Table 1. Random effects of area $(u_i)$ and sampling error $(e_{ij})$ respectively were generated from

normal distribution $u_i \sim iid\ N(0\ ,0.15)$ and $e_{ijk} \sim iid\ N(0\ ,0.50)$. The response variable $y_{ijk}$ were generate from nested error linear regression model.

**Table 1**. Intercept and Regression Coefficients on Each Cluster

| Cluster | $\beta_{0k}$ | $\beta_{1k}$ |
|---|---|---|
| 1 | 3.0 | 6.5 |
| 2 | 2.0 | 3.5 |
| 3 | 1.0 | 2.5 |

There are three cluster areas on population that has been generated with three non-sampled area. Population is shown in Figure 1.

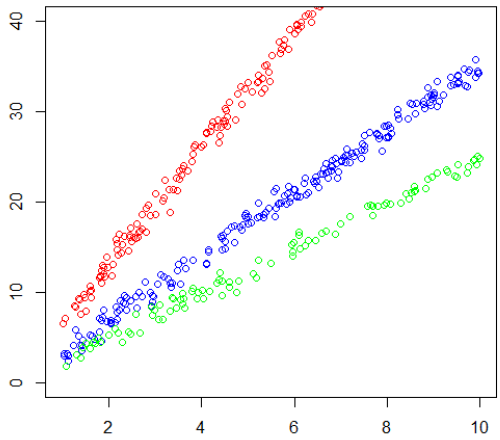


**Figure 1.** Population's Simulation Study.

Simulation study evaluated based on relative bias (RB) and relative root mean squares error (RRMSE). The formulas can be defined as follows:

$$RB_i = \frac{1}{ul}\sum_{i=1}^{ul}\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)$$

$$RRMSE_i = \frac{1}{\theta_i}\sqrt{\frac{1}{ul}\sum_{i=1}^{ul}\left(\hat{\theta}_i - \theta_i\right)^2}$$

This process was repeated $ul = 1000$ times. Relative bias and relative root mean squares error of empirical Bayes of $P_0$, $P_1$, and $P_2$ for non-sampled area are shown in Table 2 and Table 3 respectively.

**Table 2**. Relative Bias of EB estimates of Poverty Indicators

| Area | EB (Molina and Rao) | | | EB (Cluster Information) | | |
|---|---|---|---|---|---|---|
| | $P_0$ | $P_1$ | $P_2$ | $P_0$ | $P_1$ | $P_2$ |
| 13 | 0.895 | 1.614 | 2.452 | -0.203 | -0.411 | -0.574 |
| 23 | -0.227 | -0.150 | -0.251 | -0.143 | -0.066 | -0.183 |
| 37 | -0.321 | -0.172 | -0.154 | -0.163 | -0.030 | -0.090 |
| Mean | 0.116 | 0.431 | 0.682 | -0.170 | -0.169 | -0.282 |

**Table 3**. Relative Root Mean Squares Error of EB estimates of Poverty Indicators

| Area | EB (Molina and Rao) | | | EB (Cluster Information) | | |
|---|---|---|---|---|---|---|
| | $P_0$ | $P_1$ | $P_2$ | $P_0$ | $P_1$ | $P_2$ |
| 13 | 0.899 | 1.623 | 2.468 | 0.214 | 0.419 | 0.579 |
| 23 | 0.230 | 0.158 | 0.257 | 0.160 | 0.090 | 0.195 |
| 37 | 0.322 | 0.181 | 0.171 | 0.168 | 0.085 | 0.136 |
| Mean | 0.484 | 0.654 | 0.965 | 0.181 | 0.198 | 0.303 |

In this simulation study, it can be seen empirical Bayes method with cluster information has a smaller relative bias and relative root mean squares error than empirical Bayes Molina and Rao (synthetic model) for non-sampled area.

## 6. Case Study

In this study we used data National Socio-Economic Survey 2013 and Village Potential 2014 in regency and city of Bogor. There are 40 sub-districts in regency of Bogor and 6 sub-district in city of Bogor. In regency of Bogor, there are 3 non-sampled areas($n_i = 0$): Megamendung, Tanjungsari and Parung Panjang sub-districts. The response variable is average expenditure per capita per month from National Socio-Economic Survey data.

Auxiliary variables obtained from Village Potential data. There are six selected auxiliary variables from Village Potential 2014. The auxiliary variables are proportion of the number of villages with the main source of income in agriculture, proportion of the number of villages with the main income source processing industry, proportion of villages with the main income source of the field of large trading / retail and restaurants, proportion of villages with the main source of income in services, proportion of store or grocery shop, and proportion of restaurant.

The problem is arises when estimates non-sampled sub districts. Molina and Rao method used a synthetic model that ignores the effect of area. In this study is added synthetic model with mean effect of area in each cluster sub-district. Auxiliary data is used to cluster sub-district and used Euclidean distance and Ward method. Sub-districts in Regency and city of Bogor are divided into three clusters and each cluster is shown in Table 4.

Characteristics sub-district in first cluster is a group of sub-districts with the main source of income in the large or retail trade and restaurants. Second cluster is group of sub districts with the main source of income in agriculture. The last cluster is group of sub-districts with education, economics, and health facilities are adequate

Poverty line 2013 that used in regency of Bogor is 271970 and in the city of Bogor is 360518. Direct estimate of non-sampled area ($n_i = 0$) can't be made. Summary statistics for empirical Bayes estimate both estimates and relative root mean squares error of $P_0$, $P_1$ and $P_2$ for non-sampled area are shown in Table 5 and Table 6 respectively.

**Table 4.** Cluster of Sub-districts in Regency and City of Bogor.

| Cluster | Number of Sub-districts | Sub-districts |
|---|---|---|
| 1 | 18 | Nanggung, Leuwiliang, Leuwisadeng, Pamijahan, Ciampea, Dramaga, Ciomas, Tamansari, Caringin, Ciawi, Cisarua, Sukaraja, Tanjungsari, Jonggol, Kelapa Nunggal, Tajur Halang, Gunung Sindur, Tanah Sereal |
| 2 | 19 | Cibungbulang, Tenjolaya, Cijeruk, Cigombong, Megamendung, Babakan Madang, Sukamakmur, Cariu, Citeureup, Kemang, Ranca Bungur, Parung, Ciseeng, Rumpin, Cigudeg, Sukajaya, Jasinga, Tenjo, Parung Panjang |
| 3 | 9 | Cileungsi, Gunung Putri, Cibinong, Bojong Gede, Bogor Selatan, Bogor Timur, Bogor Utara, Bogor Tengah, Bogor Barat |

**Table 5.** Empirical Bayes Estimates of Poverty Indicators in Regency and City of Bogor

| Sub-districts | Sample Size | Direct | **EB (Molina and Rao)** | | | **EB (Cluster Information)** | | |
|---|---|---|---|---|---|---|---|---|
| | | | $P_0$ | $P_1$ | $P_2$ | $P_0$ | $P_1$ | $P_2$ |
| Megamendung | 0 | - | 0.113 | 0.026 | 0.009 | 0.140 | 0.034 | 0.012 |
| Tanjungsari | 0 | - | 0.150 | 0.034 | 0.012 | 0.151 | 0.032 | 0.011 |
| Parung Panjang | 0 | - | 0.140 | 0.033 | 0.011 | 0.180 | 0.042 | 0.015 |

**Table 6.** RRMSE predictions of Poverty Indicators in Regency and City of Bogor

| Sub-districts | EB (Molina and Rao) | | | EB (Cluster Information) | | |
|---|---|---|---|---|---|---|
| | $P_0$ | $P_1$ | $P_2$ | $P_0$ | $P_1$ | $P_2$ |
| Megamendung | 0.333 | 0.945 | 1.059 | 0.288 | 0.719 | 0.786 |
| Tanjungsari | 0.393 | 0.982 | 1.337 | 0.424 | 1.111 | 1.534 |
| Parung Panjang | 0.460 | 0.356 | 0.484 | 0.354 | 0.287 | 0.359 |
| Mean | 0.396 | 0.761 | 0.960 | 0.356 | 0.705 | 0.893 |

It can be seen empirical Bayes method with cluster information has a smaller relative root mean squares error predictions than empirical Bayes Molina and Rao for non-sampled area.

## 7. Conclusion

Direct estimation cannott be implemented when the area is not selected as the sample units, while Empirical Bayes method is used to solve that problem. Both simulation study and case study show that empirical Bayes method with cluster information is better than empirical Bayes method proposed by Molina and Rao on non-sampled area in terms of relative root mean squares error. Cluster information can improve predictive ability of non-sampled area by modifying the synthetic model.

## Acknowledgments

The authors gratefully acknowledge to Statistics Indonesia for their support and provide data in this research, to Penelitian Unggulan Sesuai Mandat Divisi (PUD) which has funded this research, and to Indonesia Endowment Fund for Education (LPDP) which has funded study of the first author.

**References**

[1] Molina I and Rao J N K 2010 *The Canadian J. of Statistics* **38** 369-385
[2] Sadik K 2009 *Forum Statistika dan Komputasi* **15** 8-13
[3] Fay R E and Herriot R A 1979 *J. of The American Statistical Association* **74** 269-277
[4] Anisa R, Kurnia A and Indahwati 2014 *IOSR J. of Mathematics* **10** 15-19
[5] Wahyudi, Notodiputro K A, Kurnia A and Anisa R 2016 *AIP Conf. Proc.* **1707** 080017-10 doi: 10.1063/1.4940874
[6] Rao J N K 2003 *Small Area Estimation* (New York: John Wiley and Sons)
[7] Johnson R A and Wichern D W 2007 *Applied Multivariate Statistical Analysis 6th Edition* (London: Prentice-Hall)