ISSTEC 2009
The First International Seminar on Science and Technology
JANUARY 24, 2009

ISBN : 978 – 979 – 19201 – 0 – 0

Indonesia Case Small Area Estimation

Anang Kurnia[1], Khairil A. Notodiputro[2], Asep Saefuddin[2], and I Wayan Mangku[3]

[1,2]Department of Statistics, [3]Department of Mathematics, Institut Pertanian Bogor
Jl. Meranti Wing 22 Level 4, Kampus IPB Darmaga, Bogor 16680, Indonesia
[1] e-mail contact: anangk@ipb.ac.id

Abstract

The demand for reliable statistic in smaller regions such as sub-district area is inevitable as a basis for a good planning and effective decision-making processes. The Statistics Indonesia has put many efforts to meet this demand using direct estimation. However, there are some instances in which direct estimation fails to produce estimates with the required precision due to the limited number of effective sample size. The increasing demand for small area estimates has motivated the need to develop more reliable methods for producing small area estimates with higher precision. In this paper we describe some SAE techniques for Indonesia case. In particular, we extend the EBLUP model for the logarithmic scale interested variable and we modified the Prasad-Rao (1990) MSE for this problem. An application to BPS data demonstrates the satisfactory performance of the methods.

Keywords: Empirical best linear unbiased prediction, generalized additive mixed model, nonparametric model, non-linear transformation.

Introduction

Formerly small area statistic in Indonesia is produced using census data which were broken down into smaller domain such as province or district. These data has been published in Province Statistic (*Propinsi dalam Angka*) or District Statistic (*Kabupaten dalam Angka*).

It is uneasy to determine how the small area statistic was first applied in Indonesia. However, the literature showed that small area estimation was first introduced by Smeru Research Institute in collaboration with BPS (Statistics Indonesia) to produce poverty maps at three provinces in Indonesia (DKI Jakarta, East Java and East Kalimantan) as a pilot project that has been done on 2001 – 2003. The study used various data sources, such as: (1) SUSENAS 1999 - Consumption and Core Module, (2) Population CENSUS 2000, and (3) PODES (Village Census) 2000.

There are two main problem faced in applied of small area estimation concept in Indonesia case, especially for SUSENAS and PODES data. The first is model pattern of auxiliary variables and interested variable which it do not linear or parametric model. Secondly, the magnitude of ratio between small area variances with sampling error variances.

To overcome the first problem, Kurnia and Notodiputro (2007) have proposed the generalized additive mixed models. Kurnia, Notodiputro and Ibrahim (2007) also used a nonparametric approach for this problem. Further, the second problem, Kurnia, Sartono and Wulandari (2007) try to use GREG approach.

SUSENAS, the National Socio-Economic Survey, is a nationally representative household survey, covering all areas of the country. One part of the SUSENAS is conducted every year, collecting information on the characteristics of over 200,000 households and 800,000 individuals. This part of the SUSENAS is known as the Core SUSENAS. Another part of the SUSENAS is conducted every three years, specifically collecting information on very detailed consumption expenditure from around 65,000 households. This is known as the Consumption Module of SUSENAS. The sample households are a randomly selected as a subset of the 200,000 households in the Core SUSENAS sample of the same year.

PODES, meanwhile, is a complete enumeration of village data throughout Indonesia. The information collected through this village census includes village characteristics such as size of area, population, infrastructure and local industries. The information is obtained from official village documents as well as interviews with village officials. The PODES survey is usually conducted three times in every ten years, usually prior to and as a preparation for an agricultural census, an economic census, and a population census.

Some SAE Techniques for Indonesia Case Data

Design-Based Methods

Direct estimator which is derived from only on domain-specific sample data and it is usually unbiased, but they may have large variances. In the context of design-based methods (direct estimator), a sampling design is used to select a sample s from population U with probability $p(s)$. We also consider that U_i denotes a subpopulation (or domain). In the absence of auxiliary population information, the domain total Y_i is estimated by :

ISBN : 978 – 979 – 19201 – 0 – 0

$$\hat{Y}_i = \hat{Y}(y_i) = \sum_{j \in s} w_j y_{ij} = \sum_{j \in s_i} w_j y_{ij}$$

where s_i is the sample of elements belonging to domain U_i ; w_j = design weights which depend on s and element-j ($j \in s$). The weight w_j may be interpreted as the number of elements in the population represented by the sample element-j. An important choice is $w_j = 1/\pi_j$; $\pi_j = \sum_{\{s: j \in s\}} p(s)$, j = 1,2...$N_i$. On the other hand, in the absence of auxiliary population information, the domain mean $\bar{Y}_i$ is estimated by : $\hat{\bar{Y}}_i = \frac{\hat{Y}_i}{\hat{N}_i}$

If domain-spesific auxiliary information is available; i.e. a domain total of the auxiliary information $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T = Y(\mathbf{x}_i)$ is known then the domain total Y_i can be estimated by generalized regression estimator (GREG) which is written by :

$$\hat{Y}^*_{iGR} = \hat{Y}_i + (X_i - \hat{X}_i)^T \hat{B}_i$$

where $\hat{X}_i = \hat{Y}(x_i) = \sum_{j \in s_i} w_j x_{ij}$ with $x_{ij} = x_j$ if $j \in U_i$ and $x_{ij}=0$ otherwise, $\hat{B}_i$ is the solution from:

$$\left(\sum_{j \in s} w_j x_{ij} x_{ij}^T / c_j \right) \hat{B}_i = \sum_{j \in s} w_j x_{ij} y_{ij} / c_j$$

The GREG estimator $\hat{Y}^*_{iGR}$ is not approximately unbiased although the domain sample size is large. As a result we can use modified GREG estimator to produce an unbiased or approximately unbiased estimator for Yi which borrow information values of variable interest: y ; from outside the spesific domain. The modified GREG estimator is:

$$\tilde{Y}_{iGR} = \hat{Y}_i + (X_i - \hat{X}_i)^T \hat{B} = \sum_{j \in s} \tilde{w}_{ij} y_j$$

where

$$\tilde{w}_{ij} = w_j a_{ij} + (X_i - \hat{X}_i) \left(\sum_s w_j x_j x_j^T / c_j \right)^{-1} \left(w_j x_j / c_j \right) \quad ;$$

a_{ij} is the domain indicator variable

The modified GREG estimator is approximately unbiased as the overall sample size increasses even if the domain sample size is small. This estimator is also called *survey regression estimator*.

The GREG estimator can also be written as

$$\tilde{Y}_{iGR} = X_i^T \hat{B} + \sum_{j \in s_i} w_i e_i \, .$$

The first term $X_i^T \hat{B}$ is the synthetic-regression estimator and the second term $\sum_{j \in s_i} w_j e_j$ approximately corrects the bias in synthetic estimator. An evaluation of GREG estimator and application for Indonesia case data could be see in Kurnia, Sartono and Wulandari (2007).

Linear Mixed Model Based Methods

We consider the following Fay-Herriot model (see Fay and Herriot, 1979) for the basic area level model

$y_i = x_i'\beta + \upsilon_i + e_i$

where υ_i and e_i are independent with $\upsilon_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$ for i = 1, 2, ..., k. We assume that β and A unknown but D_i (i = 1, 2, ..., k) are known.

The best predictor (BP) of $\theta_i = x_i'\beta + \upsilon_i$ if β and A known is given by

$\hat{\theta}_i^{BP} = \hat{\theta}_i(y_i|\beta, A) = x_i'\beta + (1 - B_i)(y_i - x_i'\beta)$

where $B_i = D_i/(A + D_i)$ for i = 1, 2, ..., k. Let $\hat{\theta}_i^{BP} = \hat{\theta}_i(y_i|\beta, A)$ is also Bayes estimator of θ_i under the following Bayesian models:

(i) $y_i|\theta_i \sim N(\theta_i, D_i)$

(ii) $\theta_i \sim N(x_i'\beta, A)$ is prior distribution for θ_i , i = 1, 2, ..., k.

The Bayes estimator is given from the posterior distribution

$$(\theta_i|y_i, \beta, A) \sim N\left(\left(\frac{y_i}{D_i} + \frac{x_i'\beta}{A} \right) \left(\frac{1}{D_i} + \frac{1}{A} \right)^{-1} \right)$$

$$= N\left(x_i'\beta + \frac{A}{A + D_i}(y_i - x_i'\beta), \frac{AD_i}{A + D_i} \right)$$

Based on the formulation, we could view that

$\hat{\theta}_i^{EB} = E(\theta_i|y_i, \beta, A) = x_i'\beta + (1 - B_i)(y_i - x_i'\beta)$

where $MSE(\hat{\theta}_i^{EB}) = Var(\theta_i|y_i, \beta, A) = \frac{AD_i}{A + D_i} = (1 - B_i)D_i = g_{1i}(A)$. The estimator $\hat{\theta}_i^{BP}$ are equivalent with $\hat{\theta}_i^{EB}$ for cases that are normally distributed.

When A is known, β could be estimated using the weighted maximum likelihood method

$$\log L(\beta, V) = -\frac{1}{2} \log|V| - \frac{1}{2}(Y - X\beta)' V^{-1}(Y - X\beta)$$

where $V = Diag(A + D_1, A + D_2, ..., A + D_k)$.

Let $\beta^* = \hat{\beta}_i(A) = (X'V^{-1}X)^{-1} X'V^{-1}Y$ and by replacing β with β^* in the $\hat{\theta}_i^{BP}$, we get the best linear unbiased predictor (BLUP) of θ_i given by

ISBN : 978 – 979 – 19201 – 0 – 0

$$\hat{\theta}_i^{BLUP} = \hat{\theta}_i(y_i|A) = x_i'\beta^* + (1 - B_i)(y_i - x_i'\beta^*)$$

Ghosh and Rao (1994) describe the $MSE(\hat{\theta}_i^{BLUP}) = g_{1i}(A) + g_{2i}(A)$, where

$$g_{1i}(A) = \frac{AD_i}{A + D_i} = (1 - B_i)D_i, \text{ and}$$

$$g_{2i}(A) = D_i^2/(A + D_i)\,[x_i'(X'V^{-1}X)^{-1}\,x_i]$$
$$= D_i\,(1 - B_i)\,[x_i'(X'V^{-1}X)^{-1}\,x_i] \quad \text{for } i = 1, 2, \ldots, k.$$

However, in practice both β and A are unknown. If we replace β by $\hat{\beta}$ and A by $\hat{A}$ in the BLUP ($\hat{\theta}_i^{BLUP}$) estimator, we get the empirical best linear unbiased predictor (EBLUP)

$$\hat{\theta}_i^{EBLUP} = \hat{\theta}_i(y_i|\hat{A}) = x_i'\hat{\beta} + (1 - \hat{B}_i)(y_i - x_i'\hat{\beta})$$

If defined MSE of $\hat{\theta}_i^{EBLUP}$ is $MSE(\hat{\theta}_i^{EBLUP}) = E(\hat{\theta}_i^{EBLUP} - \theta_i)^2 = Var(\hat{\theta}_i^{EBLUP}) + (Bias\ \hat{\theta}_i^{EBLUP})^2$, Kacker and Harville (1984) reformulated it as

$$MSE(\hat{\theta}_i^{EBLUP}) = MSE(\hat{\theta}_i^{BLUP}) + E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2$$
$$= H_{1i}(A) + H_{2i}(A)$$

where $H_{1i}(A) = MSE(\hat{\theta}_i^{BLUP}) = g_{1i}(A) + g_{2i}(A)$ and $H_{2i}(A) = E(\hat{\theta}_i^{EBLUP} - \hat{\theta}_i^{BLUP})^2$. Leading term $g_{1i}(A)$ lead to large reduction in MSE relative to the MSE of the direct estimator, $g_{2i}(A)$ is due to estimating of β and $H_{2i}(A)$ is due to estimating A. Prasad and Rao (1990) used the Taylor series method to estimate $g_{1i}(A)$, $g_{2i}(A)$ and $H_{2i}(A)$. The MSE estimator of $\hat{\theta}_i$

$$MSE(\hat{\theta}_i^{EBLUP})^{PR} = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2\,g_{3i}(\hat{A})$$

where $g_{3i}(\hat{A}) = \frac{2D_i^2}{k^2(A + D_i)^3}\sum_{j=1}^{k}(A + D_j)^2$. The $MSE(\hat{\theta}_i)^{PR}$ is identical to the Bayes risk as defined by Butar and Lahiri (2003).

Some evaluation and application of linear mixed models in SAE for Indonesia cases data present in Kurnia and Notodiputro (2006a, 2006b), Kurnia and Notodiputro (2005a, 2005b) and Handayani anda Kurnia (2006).

The GAMM Approach

Rao (2003) gives extensive review of the most commonly used estimators, including synthetic and composite estimator, empirical best unbiased linear predictors, empirical Bayes and hierarchical Bayes approach. All of them in use for small area estimation based on parametric approach. In this chapter we propose generalized additive mixed model (GAMM) for SAE. The GAMM approach has significant advantages over its parametric approach to model auxiliary variable, and then we adopt this approach to application in small area estimation.

We consider an extension of the Fay-Herriot model for the basic area level model

$y_i = x_i'\beta + \upsilon_i + e_i$, $i = 1, 2, \ldots, k$

where β is coefficient regression parameters, υ_i are random effect area, and e_i are sampling errors. We also assume $e_i \sim (0, D_i)$, $\upsilon_i \sim (0, A)$ and that they are independent. D_i is usually assumed to be known, see Rao (2003).

We assume that y_i and x_i are related by a smooth function m(.). Let X be the random vector of predictors, thus

$y_i = m(x_i) + \upsilon_i + e_i$, $i = 1, 2, \ldots, k$

where $\upsilon_i|X \sim (0, \upsilon(x_i))$, $e_i \sim (0, D_i)$, and e_i and υ_i are independent. The small area mean functions

$\theta_i(x_i) = m(x_i) + \upsilon_i$

is linear combination of mean $m(x_i)$ and the random effects υ_i. We can use an estimator of the mean function using a linear smoother such as smoothing splines, regression splines, and local polynomial regression. For detail discussion of these methods, see Hastie and Tibshirani (1990).

If we use Kernel smoothing function to estimate $m(x_i)$, the best predictor for small area means θ_i can be written as

$E(\theta_i|y_i) = \gamma_i\, y_i + (1 - \gamma_i)\, \hat{m}_h(x_i)$

where $\gamma_i = \upsilon(x_i) / (\upsilon(x_i) + D_i)$. To approximate MSE, we substitute $x_i'\beta$ in linear mixed model with $\hat{m}_h(x_i)$.

$$mse(\hat{\theta}_i) = \frac{D_i\hat{\sigma}_u^2}{D_i + \hat{\sigma}_u^2} + (1 - \hat{\gamma})^2\, mse\left(\hat{m}_h(x_i)\right)$$
$$+ 2D_i^2\left(\hat{\sigma}_u^2 + D_i\right)^{-3} mse\left(\hat{\sigma}_u^2\right)$$

Please see Kurnia and Notodiputro (2007) for detail discussion an evaluation and application of this method in Indonesia case data.

A Nonparametric Approach

Some authors have started to adopt the nonparametric approach in SAE. Zheng and Little (2004) propose a model-based estimator for cluster sampling, in which the regression model combines a spline model with a random effect for the cluster. Opsomer et.al. (2008) show how the inclusion of a spatial spline can improve the fit relative to a model which only uses a random effect for the small areas, as would be done in traditional small area estimation.

For brief description of methodology, we will closely follow the description in Opsomer et.al.

ISBN : 978 – 979 – 19201 – 0 – 0

(2008) and Ruppert, Wand and Carroll (2003). Consider first the simple model

$y_i = m_0(x_i) + \varepsilon_i$

where $\varepsilon_i \sim (0, \sigma^2_\varepsilon)$.

If the $m_0(.)$ is to be estimated using P-splines, the approximated it is

$$m_0(x; \beta, \gamma) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_k \gamma_k (x - \kappa_k)^p_+$$

where p is the degree of the spline, $(x)^p_+$ is the function that $x^p \mathbf{I}_{(x>0)}$, $\kappa < \ldots < \kappa_K$ is a set of fixed knots and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_K)'$ the coefficient vectors for the "parametric" and the "spline" portions of the model. If K is sufficiently large, the class of functions $m(x; \beta, \gamma)$ is very large and can approximate most smooth function $m_0(.)$ with a very high degree of accuracy, even for p is small.

The spline function uses the truncated polynomial spline basis $\{1, x, \ldots, x^p, (x - \kappa_1)^p_+, \ldots, (x - \kappa_K)^p_+\}$ to approximate the function m_0. Opsomer et.al (2008) were explained that other bases are also possible and might be preferable to the truncated polynomials.

In P-spline regression, K is typically taken to be large relative to the size of the dataset. Hence, the model $m_0(x; \beta, \gamma)$ is potentially over-parameterized. This issue is avoided by putting a penalty on the magnitude of the spline parameter γ. For a given dataset $\{(x_i, y_i): i = 1, 2, \ldots n\}$, this is done by defining the regression estimators as the minimizers over β and γ of

$$\sum_i (y_i - m(x_i; \beta, \gamma)^2 + \lambda_\gamma \gamma' \gamma$$

where λ_γ is a fixed penalty parameter. Ruppert, Wand and Carrol (2003) treat the λ as random effect in a linear mixed model specification.

Since the P-spline and the small area estimation models can be viewed as random effects models, it is natural to try to combine both into a nonparametric small area estimation framework based on linear mixed model regression.

Specifically, suppose there are M small areas, $U_1, \ldots, U_M$, for which estimates are to be constructed. Define $d_{it} = I_{(i \in U_m)}$, and for each observations I, let $\mathbf{d}_i = (d_{i1}, \ldots, d_{iM})'$. Let $\mathbf{Y} = (y_1, \ldots, y_n)'$,

$$X = \begin{pmatrix} 1 & x_1 & \ldots & x_1^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^p \end{pmatrix},$$

$$Z = \begin{pmatrix} (x_1 - \kappa_1)^p_+ & \ldots & (x_1 - \kappa_k)^p_+ \\ \vdots & \vdots & \vdots \\ (x_n - \kappa_1)^p_+ & \ldots & (x_1 - \kappa_k)^p_+ \end{pmatrix}$$

and $D = (\mathbf{d}_1', \ldots, \mathbf{d}_n')'$. We assume that the data follow the model

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\upsilon + \varepsilon$

where

$\gamma \sim (\mathbf{0}, \Sigma_\gamma)$ with $\Sigma_\gamma \cong \sigma^2_\gamma \mathbf{I}_K$

$\upsilon \sim (\mathbf{0}, \Sigma_\upsilon)$ with $\Sigma_\upsilon \cong \sigma^2_\upsilon \mathbf{I}_M$

$\varepsilon \sim (\mathbf{0}, \Sigma_\varepsilon)$ with $\Sigma_\varepsilon \cong \sigma^2_\varepsilon \mathbf{I}_n$

and each of the random components is assumed independent of the others.

The model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\upsilon + \varepsilon$ includes both the spline function, which can be thought of as a nonparametric mean function specification $\mathbf{X}\beta + \mathbf{Z}\gamma$, and includes the small area random effects $\mathbf{D}\upsilon$. For the purpose of fitting this model and using the appropriate amount of smoothing for the spline, it is convenient to continue to treat $\mathbf{Z}\gamma$ as random effect term, so that

$\text{Var}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\,\Sigma_\gamma\,\mathbf{Z}' + \mathbf{D}\,\Sigma_\upsilon\,\mathbf{D}' + \Sigma_\varepsilon$

If the variances of the random components are known, the standard results from BLUP theory guarantee that, given the model specifications $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\upsilon + \varepsilon$, the GLS estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

and the predictors

$$\hat{\gamma} = \Sigma_\gamma \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$\hat{\upsilon} = \Sigma_\upsilon \mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

are optimal among all linear estimator/predictors. For a given small area U_m, we will assume that we are interested in predicting

$$\bar{y}_m = \bar{x}_m\hat{\beta} + \bar{z}_m\hat{\gamma} + e_m\hat{\upsilon},$$

which is a linear combination of the GLS estimator and the BLUP. The mean square error (MSE) estimator of interested parameter was calculated pursuant to Prasad and Rao (1990) technique. Please see Kurnia, Notodiputro, and Ibrahim (2007) for evaluation and application of this method in Indonesia case data.

EBLUP for Logarithmic Transformation in SAE

The parameter of interest in this study is finite population mean, $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$ where N_i is population number in each small area and y_{ij} is interested variable. We define a linear mixed model for log-transformed of interested variable as

$$log(Y_{ij}) = \beta_0 + \sum_k \beta_k X_{ki} + \upsilon_i + \varepsilon_{ij}$$

ISBN : 978 – 979 – 19201 – 0 – 0

where ε_{ij} are *iid* from $N(0,\sigma^2)$, v_i are iid form $N(0,\sigma_v^2)$, while β_k, σ^2 and σ_v^2 are unknown parameter.

The EBLUP of mean of $log(Y_{ij})$ can we write as $\hat{\theta}_i^{EBLUP} = X_i\hat{\beta} + \gamma_i(\hat{\theta}_i^{D} - X_i\hat{\beta})$, where $\gamma_i = \hat{\tau}^2/(\hat{\tau}^2 + \frac{\hat{\sigma}^2}{n_i})$ and $\hat{\theta}_i^{D} = \frac{1}{n_i}\sum_j \log(Y_{ij})$ is direct estimator in log-transformed. The $MSE(\hat{\theta}_i^{EBLUP})$ = $g_1 + g_2 + 2g_3$ (Prasad-Rao formula). Furthermore, the EBLUP approach for total finite population ($\hat{Y}_i$) as back-transformed of the $\hat{\theta}_i^{EBLUP}$ could be write $\hat{Y}_i = \exp(\hat{\theta}_i^{EBLUP + \frac{msep}{2}})$ and the $MSE(\hat{Y}_i) = e^{msep}(e^{msep} - 1)e^{2\hat{\theta}_i^{EBLUP}}$ where *msep* is $MSE(\hat{\theta}_i^{EBLUP})$

Consider to the naïve back-transformed model, by simple back-transformation, the population total Y_i is given by $\hat{T}_{Ai} = \sum_{s(i)} Y_{ij} + \sum_{r(i)} e^{x_i\hat{\beta}}$, where $\hat{\beta} = (X_s'V^{-1}X_s)^{-1}X_s'V^{-1}(logY_s)$ and $V = \sigma_i^2 I + \tau^2 I$. However the estimate is bias.

Let the other estimator for finite population total Y_i

$\hat{T}_{Bi} = \sum_{s(i)} Y_{ij} + \sum_{r(i)} e^{x_i\hat{\beta} + \frac{1}{2}V_i}$

Under the normality of error, we have $\hat{\beta} \sim N(\beta, (X_s'V^{-1}X_s)^{-1})$, $\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k-1)}$, $\frac{(m-1)\hat{\tau}^2}{\tau^2} \sim \chi^2_{(m-1)}$, $Cov(\sigma^2, \tau^2) = 0$, $Cov(\sigma^2, \beta) = 0$, and $Cov(\tau^2, \beta) = 0$. The MGF for $\log(Y_{ij})$ is $E(e^{t\log(Y_{ij})}) = exp(x_i\beta t + \frac{1}{2}t^2V_i)$ and then for t = 1 we have $E(Y_{ij}) = exp(x_i\beta + \frac{1}{2}V_i)$ and for t = 2 we have $E(Y_{ij}^2) = exp(2x_i\beta + 2V_i)$.

Therefore

$$Var(Y_{ij}) = E(Y_{ij}^2) - (E(Y_{ij}))^2$$
$$= e^{(2x_i\beta + 2V_i)} - e^{2x_i\beta + V_i}$$
$$= e^{V_i}(e^{V_i} - 1)e^{2x_i\beta}$$

$$Var(\hat{T}_i - T_i) = Var(\sum_{s(i)} Y_{ij} + \sum_{r(i)} e^{x_i\hat{\beta} + \hat{V}_i/2} - \sum_{s(i)} Y_{ij})$$
$$= Var(\sum_{r(i)} e^{x_i\hat{\beta} + \hat{V}_i/2} - \sum_{r(i)} Y_{ij})$$
$$= Var(\sum_{r(i)} e^{x_i\hat{\beta} + \hat{V}_i/2}) + Var(\sum_{r(i)} Y_{ij})$$

$Var(\sum_{r(i)} \hat{Y}_{ij}) = \sum_{j \in r(i)} \sum_{k \in r(i)} Cov(\hat{Y}_{ij}, \hat{Y}_{ik})$

and by Taylor Series, we write

$Cov(\hat{Y}_{ij}, \hat{Y}_{ik}) \approx \left(\frac{\partial}{\partial\theta}\hat{Y}_{ij}\right) Var(\hat{\theta}) \left(\frac{\partial}{\partial\theta}\hat{Y}_{ik}\right)$.

$Var(\hat{T}_i - T)$ is similar with $g_1 + g_2$ in Prasad and Rao (1990).

Furthermore, we can measure the uncertainty of the EBLUP by its mean squared prediction error, defined as $MSPE(\hat{T}_i(\hat{\tau}^2)) = E(\hat{T}_i(\hat{\tau}^2) - T_i)^2$.

We write

$$MSPE(\hat{T}_i(\hat{\tau}^2)) = MSPE(\hat{T}_i(\tau^2)) + E(\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2))^2$$
$$+ 2E\left((\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2))(\hat{T}_i(\tau^2) - T_i)\right)$$

We assumed the cross-product in right side of is zero, and then we can write

$MSPE(\hat{T}_i(\hat{\tau}^2)) = MSPE(\hat{T}_i(\tau^2)) + E(\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2))^2$

and $MSPE(\hat{T}_i(\tau^2)) = Var(\hat{T}_i - T_i)$.

By the Taylor series expansion of $\hat{\tau}^2$ around τ^2, we obtain

$\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2) = (\hat{\tau}^2 - \tau^2)\frac{\partial \hat{T}_i(\tau^2)}{\partial \tau^2} + r$

where

$\frac{\partial \hat{T}_i(\tau^2)}{\partial \tau^2} = \frac{\partial \sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2 + \tau^2/2}}{\partial \tau^2} = \frac{1}{2}e^{\tau^2/2}\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2}$.

Therefore,

$\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2) = (\hat{\tau}^2 - \tau^2)\frac{1}{2}e^{\tau^2/2}\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2} + r$.

Thus, we obtain the approximation of the second term in $MSPE(\hat{T}_i(\hat{\tau}^2))$ as :

$$E(\hat{T}_i(\hat{\tau}^2) - \hat{T}_i(\tau^2))^2 =$$
$$E\left[(\hat{\tau}^2 - \tau^2)\frac{1}{2}e^{\tau^2/2}\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2}\right]^2$$
$$= \frac{1}{4}e^{\tau^2}\left(\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2}\right)^2 E[(\hat{\tau}^2 - \tau^2)]^2$$
$$= \frac{1}{4}e^{\tau^2}\left(\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2}\right)^2 \nabla(\hat{\tau}^2, \hat{\sigma}^2)$$

where $\nabla(\hat{\tau}^2, \hat{\sigma}^2)$ (asymptotic variance of $\hat{\tau}^2$) is

$\tau^4 var(\hat{\sigma}^2) + \sigma^4 var(\hat{\tau}^2) - 2\sigma^2\tau^2 cov(\hat{\sigma}^2, \hat{\tau}^2)$,

thus the second-order MSPE approximation of $\hat{T}_i(\hat{\beta}, \hat{\tau}^2)$ is given by

$mspe(\hat{T}_i(\hat{\beta}, \hat{\tau}^2)) = var(Y_{ij}) + var(\hat{T}_i - T_i) - \frac{1}{4}e^{\tau^2}\left(\sum_{r(i)} e^{x_i\hat{\beta} + \sigma_i^2/2}\right)^2 \nabla(\hat{\tau}^2, \hat{\sigma}^2)$

The based-model simulation study was based on characteristic from a sample of 105 sub-districts in

ISBN : 978 – 979 – 19201 – 0 – 0

Bogor. The finite population data was generated with following model:

$$\log(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} + \beta_8 x_{8ij} + v_i + \varepsilon_{ij}$$

where ε_{ij} are *iid* from $N(0,\sigma^2)$. v_i are iid form $N(0,\sigma_v^2)$, while β_k, σ^2 and σ_v^2 is specified pursuant to parameter of case Bogor Data.

Table 1 present the relative bias and relative variance based on generated data where:

1. $\hat{Y}_{1i} = \exp\left(\hat{\theta}_i^{EBLUP+\frac{MSE}{2}}\right)$
2. $\hat{Y}_{2i} = \exp\left(\hat{\theta}_i^{D+\frac{n_i VAR}{2}}\right)$ with correction
3. factor n_i and $\hat{\theta}_i^D = \frac{1}{n_i}\sum_j \log(y_{ij})$
4. Direct = direct estimator
5. $\hat{Y}_{3i} = \sum_{s(i)} y_{ij} + \sum_{r(i)} e^{x_{ij}b+s^2/2}$
6. $\hat{Y}_{4i} = \sum_{s(i)} y_{ij} + \sum_{r(i)} \hat{k}_i^{-1} e^{x_{ij}b+s^2/2}$;
7. $\hat{k}_i = 1 + \frac{1}{2}\left(s^2 x_{ij}(X_s'X_s)^{-1}x_{ij} + \frac{s^4}{2n_i}\right)$

Table 1. Relative bias and relative MSE

Relative Bias

AREA	Y1	Y2	DIRECT	Y3	Y4
1	-0.01221	0.00525	0.00448	-0.01070	-0.01082
2	-0.01228	0.00001	-0.00104	-0.01027	-0.01039
.					
.					
.					
104	-0.01280	0.01093	0.01008	-0.01052	-0.01065
105	-0.00947	0.01841	0.01759	-0.00844	-0.00846
Average	-0.01273	0.00133	0.00048	-0.01103	-0.01110

Relative MSE

AREA	Y1	Y2	DIRECT	Y3	Y4
1	130.3774	576.7893	571.4011	115.1015	115.9054
2	169.5672	613.1814	613.8662	141.1582	142.0855
.					
.					
.					
104	199.4080	1116.6233	1103.0773	148.2057	149.4829
105	86.5147	947.4726	932.2026	66.8503	67.0247
Average	137.1999	761.9239	758.7267	113.5510	114.1676

Note:

1. Relative bias = $E\left(\frac{\hat{Y}-Y}{Y}\right) \approx \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{Y}_r - Y}{Y}\right)$
2. $MSE = E(\hat{Y}-Y)^2 \approx \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2$ and Relative MSE = $\frac{MSE}{Y}$

The empirical case evaluation used PODES 2005 and SUSENAS 2005 data especially for Bogor District and Municipality. Variable of interest is mean of per capita expenditure which is calculated from food expenditure and non-food expenditure for each sub-district. Status of sub-district, percentage of farmer, percentage of farmer worker, percentage of household using electrics, percentage of rice field area, percentage of household poor guarantee health, distance to city central, and time to city central are used as auxiliary variables. Table 2 exhibits the results of estimation.

Table 2. Estimation for per capita expenditure (IDR) for Bogor Data

Sub-District	Y1	Y2	Direct	Y3	Y4	RRMSEY1	RRMSEY2	RRMSE	RRMSEY3	RRMSEY4
3201010002	442,736.06	537,340.83	527,540.99	350,779.17	349,174.04	9.23799	11.62282	12.26599	6.76057	7.09316
3201010010	347,917.72	366,680.76	364,311.54	412,781.13	411,065.04	9.22425	9.33135	9.96789	6.77341	7.10832
.										
.										
3271060003	447,224.85	450,592.34	449,200.82	552,539.24	552,143.47	9.11540	7.10029	7.37147	9.72940	9.96547
3271060005	588,332.61	681,712.05	697,474.72	503,423.31	502,650.03	6.72459	7.92574	12.83002	10.72529	10.94225
3271060008	504,504.74	571,792.61	566,478.32	545,364.29	544,948.74	9.11688	11.46578	14.48496	8.60141	8.86715
3271060009	469,433.67	508,955.33	505,627.58	488,970.81	486,866.36	9.22883	9.41511	10.09923	6.63896	6.97958

ISBN : 978 – 979 – 19201 – 0 – 0

Sub-District	Y1	Y2	Direct	Y3	Y4	RRMSEY1	RRMSEY2	RRMSE	RRMSEY3	RRMSEY4
3271060011	380,877.06	385,861.88	386,722.42	505,751.84	505,305.94	9.12073	8.44087	10.36220	6.47718	6.82744

References

Butar, F. B. and Lahiri, P. 2003. On measures of uncertainty of empirical Bayes small area estimators. Model selection, model diagnostics, empirical Bayes and hierarchical Bayes, Journal of Statistical Planning and Inference, 112, pp: 63–76.

Chambers, R and Chandra, H. 2006. Improved Direct Estimator for Small Area. S3RI Methodology Working Paper M06/07.

Fay, R.E. and Herriot, R.A. 1979. "Estimates of income for small places: an application of James-Stein procedures to Census data". Journal of the American Statistical Association, Vol. 74, p:269-277.

Handayani, D. dan Kurnia, A. 2006. Pendekatan Empirical Bayes pada Pendugaan Nilai Tengah Populasi Terhingga dalam Kasus Small Area Estimation. JMAP (ISSN 1412-8632) Vol. 5 No. 2.

Kackar, R.N. and Harville, D.A. 1984. Approximations for standard errors of estimations of fixed and random effects in mixed linear models. Journal of the American Statistical Association, 79, pp: 853-862.

Kurnia, A., Notodiputro, K.A. and Ibrahim, N.A. 2007. A Nonparametric Approach in Small Area Estimation. Proceeding at the ICCS-IX 2007, 12 - 14 December 2007. Universiti of Malaya, Syah Alam – Kuala Lumpur.

Kurnia, A., Sartono, B. dan Wulandari, R. 2007. Pengaruh Misspesifikasi Desain Survey pada Pendugaan Area Kecil dengan Pendekatan *Generalized Regression.* Prosiding pada Seminar Nasional Matematika dan Pendidikan Matematika, Universitas Negeri Yogyakarta, 24 November 2007.

Kurnia, A and Notodiputro, K.A. 2007. Generalized Additive Mixed Models for Small Area Estimation. Proceeding at the 2nd International Conference on Mathematical Sciences 2007 (ICoMS-2007), 28 - 29 May 2007. Universiti Teknologi Malaysia.

Kurnia, A. and Notodiputro, K.A. 2006a. EB-EBLUP MSE Estimator on Small Area Estimation with Application to BPS'Data. *Proceeding at The First International Conference on Mathematics and Statistics (*ICoMS-1*)*, Bandung, June 19-21, 2006

Kurnia, A. and Notodiputro, K.A. 2006b. Penerapan Metode Jackknife dalam Pendugaan Area Kecil = *The Jackknife Approach in Small Area Estimation. Forum Statistika dan Komputasi ISSN 0853-8115 Vol. 11 No. 1.*

Kurnia, A. and Notodiputro, K.A. 2005a. Aplikasi Metode Bayes pada Small Area Estimation. *Proceeding at The National Seminar on Statistics VII.* ITS Surabaya: November, 26 2005.

Kurnia, A. and Notodiputro, K.A. 2005b. Generalized Linear Mixed Model on Small Area Estimations. *Forum Statistika dan Komputasi* (ISSN 0853-8115) *Vol. 10 No. 2.*

Opsomer, J. D., et. al. 2008. Nonparametric Small Area Estimation using Penalized Spline Regression. Journal Royal Statistics Society Series B, 70, p:265–286.

Prasad, N.G.N. and Rao, J.N.K. 1990. "The Estimation of Mean Squared Errors of Small Area Estimators". Journal of American Statistical Association, 85, p:163-171.

Rao, J.N.K. 2003. Small Area Estimation, New York : John Wiley and Sons.

Ruppert, R., M. Wand, and R. Carroll. 2003. Semiparametric Regression. Cambridge University Press.

Saei, A. and Chambers, R. 2003. Small Area Estimation: A Review of Methods Based on the Application of Mixed Models. S3RI Methodology Working Paper M03/16.

Zheng, H. and R. J. A. Little. 2004. Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. Survey Methodology 30, p:209–218.