Empirical Best Linear Unbiased Prediction Method with K-Medoids Cluster for Estimate Per Capita Expenditure of Sub-District Level

Irlandia Ginanjar, *Member IAENG,* Septie Wulandary, Toni Toharudin

***Abstract*—One of the data needed is per capita expenditure of sub-district level. The National Socio-Economic Survey (*Susenas*) obtain per capita expenditure data. However, due to limited samples, *Susenas* cannot get statistical information down to the sub-district or village level from this survey. Therefore, a model that includes additional data (census data) to estimate sub-district levels uses Small Area Estimation (SAE) modelling. In this study, the Empirical Best Linear Unbiased Prediction (EBLUP) method estimates the per capita expenditure of sub-district levels in Jambi Province in 2018. The EBLUP modelling also applies cluster information to estimate per capita expenditure in the sub-districts that are not surveyed (non-sampled area). The K-Medoids Cluster method is used to get sub-district clusters based on their characteristics. The mean of area random effects per cluster adds to the prediction model for the non-sampled area. This study aims to compare and evaluate the direct estimation method and EBLUP method to estimate sub-district level per capita expenditure based on Relative Root Mean Square Error (RRMSE). The result shows that EBLUP estimation produces more accurate estimates than direct estimation. This research also estimates the per capita expenditure of non-sampled areas using the EBLUP method by applying K-Medoids Cluster information.**

***Index Terms*—Small Area Estimation, EBLUP, K-Medoids Cluster, Per Capita Expenditure**

I. Introduction

BPS-Statistics Indonesia is a Non-Ministerial Government Institution whose role is to provide data for the government and society in Indonesia. There are two types of data collected by BPS: data obtained from censuses or self-conducted surveys (primary data) and data obtained from other institutions (secondary data). The National Socio-Economic Survey (*Susenas*) is a survey with a household approach carried out by BPS. This survey collects household social and economic information, including monthly per capita expenditure data. The survey conducts twice a year; the March *Susenas* produces regency/municipality estimation and the September *Susenas* for the province estimation. *Susenas* can obtain the minor estimate, which is the regency/municipality estimation number. However, due to limited samples, *Susenas* cannot get statistical information down to the sub-district or village level from this survey.

The studies that examine the process of estimating or predicting values for each sub-district in Indonesia include: Jaya et al. [1] who predicted malaria risk for each sub-district in Bandung City using a spatial model. In addition, Jaya and Chadidjah [2] compared the simultaneous (SAR) and conditional autoregressive (CAR) models for the number of diarrhea cases in each sub-district in Bandung City. However, apart from having a spatial relationship between sub-districts, all sub-districts must be surveyed for the two methods used in these studies. Thus, this study estimates each sub-district value in Indonesia using the Small Area Estimation (SAE) method.

According to Giusti et al. [3], the solution to provide data to the smallest level is by increasing the sample size so that the direct estimation method becomes more reliable or using the Small Area Estimation (SAE) method. However, increasing the sample size will increase the survey cost, so the SAE method is more likely to be an alternative solution for the problem. There are several methods in SAE, including Best Linear Unbiased Prediction (BLUP), Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB), and Hierarchical Bayes (HB). The BLUP and EBLUP methods are limited to continuous response variables, while EB and HB can be used more broadly for continuous, binary, or count response variables [4].

There are several studies on estimating per capita expenditure. Salma et al. [5] estimated per capita expenditure of sub-districts using the Robust EBLUP method in Bogor Regency. Ubaidillah et al. [6] conducted a Multivariate Fay-Herriot model for small area estimation applied to Indonesia's household consumption per capita expenditure. Dedianto and Wulansari [7] carried out the study using the Pseudo EBLUP method in estimating household expenditure in East Java Province. In this study, the researchers use the EBLUP method to add cluster information for per capita expenditure data based on the March 2018 *Susenas* results.

In Indonesia, 575 sub-districts were not surveyed in the March 2018 *Susenas*. In Jambi, two sub-districts were not

Manuscript received June 16, 2021; revised April 19, 2022. This work was supported by the Acceleration of Associate Professor Research Universitas Padjadjaran (RPLK) 2021 Number 1959/UN6.3.1/PT.00/2021, and BPS-Statistics of Jambi Province.

I. Ginanjar is an assistant professor in Department of Statistics, Universitas Padjadjaran, West Java, 45363, Indonesia, e-mail: (irlandia@unpad.ac.id).

S. Wulandary is a statistician in BPS-Statistics of Jambi Province, Jambi, 36122, Indonesia, e-mail: (septie@bps.go.id).

T. Toharudin is an associate professor in Department of Statistics, Universitas Padjadjaran, West Java, 45363, Indonesia, e-mail: (toni.toharudin@unpad.ac.id).

selected as samples. Then problems arise when estimating parameters for sub-districts that are not surveyed (non-sampled area). The standard EBLUP estimation method for the non-sampled area uses the global synthetic model. The synthetic model ignores the area's random effects because there are no random effects in non-sampled areas [8]. The absence of area random effects in the synthetic model causes a bias in estimating non-sampled areas.

Previous studies used cluster information to estimate the parameters in non-sampled areas. Anisa [9] stated that the addition of cluster information in the model provides small Relative Bias (RB) and Relative Root Mean Square Error (RRMSE). Wahyudi et al. [10] compared some cluster methods and concluded that the Ward method was better. The simulation study conducted by Susanti et al. [11] concluded that the cluster method was better than the Nearest Neighbor (NN) method since it provided a smaller Absolute Relative Bias (ARB) and RRMSE. There has been no research on the EBLUP method with K-Medoids Cluster information for actual data compared with the direct estimation method.

Based on the explanation above, this study compares the direct estimation method and the EBLUP method in estimating per capita expenditure of sub-districts in Jambi based on RRMSE. In addition, this research also estimates the per capita expenditure of non-sampled areas using the EBLUP method by applying K-Medoids Cluster information.

II. Methods

A. Data

This study uses secondary data from BPS-Statistics Indonesia, which are:

1. The response variable, which is the per capita expenditure of sub-districts in Jambi Province (139 sub-districts) from the March *Susenas* 2018.
2. Facility and infrastructure data available in sub-districts in Jambi Province (141 sub-districts) from the 2018 Village Potential Census (*Podes*) as predictor variables.

This research is a case study for all sub-districts in regencies/municipalities in Jambi in 2018. The open-source R software conducts the data processing. The research variables are presented in Table I.

B. Direct Estimation Method

According to Rao and Molina [4], in the survey context, direct estimation estimates population parameters in an area based only on samples in that area. Per capita expenditure from a sub-district is calculated using the formula:

$$\bar{y}_i = \frac{\sum_{j=i}^{n_i} w_{ij} y_{ij}}{\sum_{j=i}^{n_i} w_{ij}}, i = 1,..,m,\ j = 1,\dots,n_i, \quad (1)$$

where:

$\bar{y}_i$: the per capita expenditure in *i*-th sub-district,
y_{ij} : *j*-th household per capita expenditure in *i*-th sub-district,
w_{ij}: *j*-th household weighting factor in *i*-th sub-district obtained from *Susenas* sampling design,
n_i : the number of households in *i*-th sub-district,
m : the number of sub-districts.

The calculation of w_{ij} written in BPS book [12].

TABLE I
Research Variable Candidates

Response Variables	
Y	Per capita expenditure in each sub-district
Predictor Variables	
X_1	Number of populations
X_2	Number of buildings along the river banks
X_3	Number of colleges
X_4	School facility ratio
X_5	Number of Community Health Center (*Puskesmas*)
X_6	Number of Polyclinics
X_7	Health facility coverage
X_8	Doctor coverage
X_9	Health worker coverage
X_{10}	Disability coverage
X_{11}	Midwife ratio
X_{12}	Number of grocery stores
X_{13}	Number of stalls
X_{14}	Percentage of villages using LPG fuel
X_{15}	Number of farming villages

The direct estimation variance formula is defined as follows:

$$var(\hat{\bar{y}}_i) = \frac{\sum_{j=i}^{n_i} w_{ij}(w_{ij}-1)(y_{ij}-\bar{y}_i)^2}{\left(\sum_{j=i}^{n_i} w_{ij}\right)^2}. \quad (2)$$

According to Hindmarsh [13], the direct estimation $\bar{y}_i$ is an unbiased estimator so that $MSE(\bar{y}_i) = var(\bar{y}_i)$.

An estimator for the parameter of a sub-population can be obtained directly based on samples in that sub-population (direct estimator) [14]. However, that estimator is unbiased but has a large variance because the sample size is small. Therefore, the small area estimation method with the indirect estimation can be an alternative solution.

C. Small Area Estimation (SAE)

Small Area Estimation (SAE) can improve the sample size of surveys by borrowing the strength of neighboring areas and the relation between auxiliary/predictor variables and the variables of interest [4]. Therefore, the availability of the data from predictor variables will significantly determine success in making the SAE model. Based on the availability of predictor variables, the SAE model is divided into two types, which are:

i) Area level model, used when predictor variables are available only at the area level. This model uses the response variable obtained from direct estimation in small areas and the basic area level of predictor variables. This model is called the Fay-Herriot model.
ii) Unit level model, used when information is available at the unit level for both response and predictor variables. This model is known as the Battese, Harter, and Fuller models.

D. Empirical Best Linear Unbiased Prediction (EBLUP)

This research uses the basic area level of the Empirical Best Linear Unbiased Prediction (EBLUP) method. The

estimation method is Restricted Maximum Likelihood (REML) which assumes normality. According to Rao and Molina [4], the REML method produces an unbiased parameter estimation, so the parameter estimation in this study uses the REML methods. Therefore, area-based EBLUP can be written as follows:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + z_i v_i + e_i. \quad (3)$$

The equation (3) is known as the Fay-Herriot model with $z_i = 1$. Rao and Molina [4] stated the EBLUP estimator as follows:

$$\begin{aligned}\hat{\mu}_i &= \mathbf{x}_i^T\hat{\boldsymbol{\beta}} + \hat{v}_i \\ &= \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}},\end{aligned} \quad (4)$$

where

$$\hat{\boldsymbol{\beta}} = \left[\frac{\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^T}{\hat{\sigma}_v^2 z_i^2 + \psi_i}\right]^{-1}\left[\frac{\sum_{i=1}^m \mathbf{x}_i y_i}{\hat{\sigma}_v^2 z_i^2 + \psi_i}\right], \quad (5)$$

$$\hat{v}_i = \hat{\gamma}_i\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right), \quad (6)$$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2 z_i^2}{\hat{\sigma}_v^2 z_i^2 + \psi_i}.$$

Mean Square Error (MSE) is used to see the estimation accuracy of the resulting model. The Mean Square Error (MSE) of the EBLUP model is [4]:

$$\mathrm{MSE}(\hat{\mu}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \quad (7)$$

where:

$$g_{1i}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 z_i^2 \psi_i}{\hat{\sigma}_v^2 z_i^2 + \psi_i} = \hat{\gamma}_i\psi_i,$$

$$g_{2i}(\hat{\sigma}_v^2) = (1-\hat{\gamma}_i)^2\mathbf{x}_i^T\left[\frac{\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^T}{\hat{\sigma}_v^2 z_i^2 + \psi_i}\right]^{-1}\mathbf{x}_i,$$

$$g_{3i}(\hat{\sigma}_v^2) = \frac{{\psi_i}^2 z_i^2}{(\psi_i + \hat{\sigma}_v^2 z_i^2)^3}\bar{V}(\hat{\sigma}_v^2),$$

and $\bar{V}(\hat{\sigma}_v^2) = m^{-2}\sum_{i=1}^m(\psi_i + \hat{\sigma}_v^2 z_i^2)^2/z_i^4$ is an asymptotic variance of $\hat{\sigma}_v^2$.

The steps of the research using the EBLUP method were as follows:

1. Conducting a direct estimation for the response variable (y_i), for 139 of 141 sub-districts in Jambi, as in equations (1) and equations (2) for the variance y_i.
2. Estimating the variance component σ_v^2 using the REML method through the fisher scoring iteration procedure. The estimation process uses R software with the "sae" package.
3. Estimating $\hat{\boldsymbol{\beta}}$ according to equations (5) and $\hat{v}_i$ according to equation (6).
4. Selecting predictor variables using the backward elimination method with the following procedure:
   a. Modelling all predictor variables in the EBLUP model based on the values of $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ obtained from step 3.
   b. Excluding predictor variables with p-values exceeding 0.05 in stages, then modelling again.
   c. Selecting and setting the predictor variables used in the model based on the smallest AIC value.
5. Testing the normality assumption of residual; the response variable transforms if it is not normal. Then repeating steps 2 to 4 with the transformed data. In this study, *Y* is transformed into lognormal.
6. Estimating the log of per capita expenditure in each sub-district $(\hat{\mu}_i)$ according to equations (4) using selected predictor variables.
7. Calculating the EBLUP MSE ($\mathrm{MSE}(\hat{\mu}_i)$) according to equation (7) using the "sae" package.

A reverse transformation is performed after obtaining an estimated log of per capita expenditure for each sub-district. Because an actual estimator for the mean is expected in the *i*-th area, the lognormal distribution is used to reverse the equation (4) [15]. Furthermore, it assumes that $\hat{\mu}_i$ has a normal distribution. So, the actual value estimator for the mean or the logarithmic transformation estimator EBLUP $(\hat{\mu}_i^{TL\ EBLUP})$ for the *i*-th area is [15]:

$$\hat{\mu}_i^{TL\ EBLUP} = exp\left(\hat{\mu}_i + \frac{1}{2}\mathrm{MSE}(\hat{\mu}_i)\right), \quad (8)$$

where $\mathrm{MSE}(\hat{\mu}_i)$ is an estimator for Mean Square Error (MSE) of $\hat{\mu}_i$.

The MSE estimator for the mean estimator in equation (8) can be obtained as follows:

$$\mathrm{MSE}(\hat{\mu}_i^{TL\ EBLUP}) = e^{\mathrm{MSE}(\hat{\mu}_i)}\left(e^{\mathrm{MSE}(\hat{\mu}_i)} - 1\right)e^{2\hat{\mu}_i}. \quad (9)$$

E. Relative Root Mean Square Error (RRMSE)

Relative Root Mean Square Error (RRMSE) values can compare direct and indirect estimation results. However, according to Kish [16], the absolute value of the variance from several measurements cannot be used directly for comprising each other.

A measurement comparison result would be more meaningful if presented in the relative form as the coefficient of variation (RRMSE). RRMSE values are obtained according to the following formula:

$$\mathrm{RRMSE}(\theta_i) = \frac{\sqrt{\mathrm{MSE}(\theta_i)}}{\theta_i} \times 100\%, \quad (10)$$

where θ_i is the parameter to be estimated.

F. EBLUP with Cluster Information

Model is developed by adding the *k*-th cluster information to the EBLUP model to estimate the non-sampled area. The cluster technique used in this study is the K-Medoids Cluster. According to Patel and Singh [17], the K-Medoids Cluster is suitable for extensive data containing outliers.

In this study, the sampled and non-sampled areas are grouped based on the selected *X* variables (eight variables).

Then in the sampled area, the random effects of the area obtained from step (3) of the EBLUP method calculation are averaged per cluster. Finally, the mean of area random effects per cluster is entered into the model to estimate area random effects for the non-sampled area.

The mean of area random effects per cluster is as follows:

$$\bar{\hat{v}}_{(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k} \hat{v}_i, \quad (11)$$

where m_k is the number of sample areas on the k-th cluster. Prediction model for non-sampled area is as follows:

$$\hat{\mu}_{i*k} = \mathbf{x}_{i*k}^T \hat{\boldsymbol{\beta}} + \bar{\hat{v}}_{(k)}, \quad (12)$$

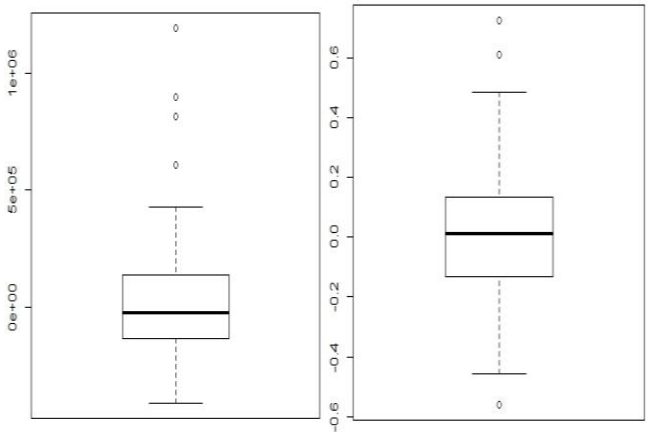
with $i * k$ is the non-sampled area on the k-cluster and $\bar{\hat{v}}_{(k)}$ is the average area random effects on the k-th cluster.

The steps of estimating the non sampled sub-districts were as follows:

1. Applying the K-Medoids Cluster method to the sub-district's population data. All sub-districts are grouped into four clusters based on selected predictor variables, using the "ClusterR" package.
2. In the sampled area, the known components $\hat{v}_i$ are averaged in each cluster according to equation (11).
3. Estimating per capita expenditure logs in the non-sampled area using the EBLUP model by adding cluster information $(\hat{\mu}_{i*k})$ according to equation (12).

III. Data Analysis and Results

The EBLUP estimator uses a linear mixed model in its estimation [15]. The linear mixed model assumes a normal distribution, so the EBLUP model's residual must have a normal distribution. Based on Figure 1.a., the boxplot of the residual is asymmetrical. The Anderson-Darling normality test results in a significance value of 0.00008, meaning that the residuals are not normal distribution with a level significance of 5%.



a) Before Transformation b) After transformation

Fig. 1. Residual Boxplot Before and After Transformation

The Y variable transforms into a lognormal (log) to have a normal residual distribution. In Figure 1.b. it appears that the distribution of residual after transformation is symmetrical and does not extend to the right. The Anderson-Darling normality test results in a significance value of 0.6608, meaning that the residual follows the normal distribution at the 5% significance level.

A. Selection of Predictor Variables

After identifying the distribution of the response variable, the next step is selecting predictor variables using the EBLUP model with a backward elimination procedure. This procedure, modelling, begins by using all the predictor variables into the model. Then the modelling is done again by eliminating insignificant parameters with a 5% significance level.

The best model based on the parameter significance value and the smallest AIC is model 4. Therefore, model 4 is the best model for constructing per capita expenditure estimates of sub-district levels in Jambi Province. Eight predictor variables in Model 4, namely the number of population (X_1), number of colleges (X_3), school facility ratio (X_4), number of polyclinics (X_6), doctor coverage (X_8), health worker coverage (X_9), disability coverage (X_{10}), and midwife ratio (X_{11}).

EBLUP modelling involves all predictor variables (Table II) and eight significant predictor variables generated in the model. Furthermore, modelling is carried out by gradually releasing insignificant variables evaluated based on the AIC value, as shown in Table III.

TABLE II
PARAMETER ESTIMATION $\boldsymbol{\beta}$ WITH EBLUP METHOD FOR 15 PREDICTOR VARIABLES

Estimator	Coefficient Value	Standard Error	t-value	p-value
$\hat{\beta}_0$	13.634	0.209	65.178	0.000*
$\hat{\beta}_1$	-6.439e-06	1.955e-06	-3.294	0.001*
$\hat{\beta}_2$	-9.381e-06	6.956e-05	-0.135	0.893
$\hat{\beta}_3$	0.065	0.023	2.856	0.004*
$\hat{\beta}_4$	0.700	0.312	2.245	0.025*
$\hat{\beta}_5$	0.011	0.007	1.628	0.104
$\hat{\beta}_6$	0.012	0.003	3.767	0.000*
$\hat{\beta}_7$	4.410e-04	2.709e-04	1.628	0.104
$\hat{\beta}_8$	3.908e-05	1.400e-05	2.792	0.005*
$\hat{\beta}_9$	-0.002	5.933e-04	-2.798	0.005*
$\hat{\beta}_{10}$	5.631e-04	3.437e-04	1.638	0.101
$\hat{\beta}_{11}$	1.121	0.416	2.696	0.007*
$\hat{\beta}_{12}$	0.002	0.001	1.603	0.109
$\hat{\beta}_{13}$	-1.945e-04	9.980e-05	-1.949	0.051
$\hat{\beta}_{14}$	-0.004	0.002	-1.711	0.087
$\hat{\beta}_{15}$	-0.004	0.004	-0.927	0.354

*Note: * significant at the 0.05 significance level*

TABLE III
AIC VALUE IN THE SELECTION PROCESS OF PREDICTOR VARIABLES

Model	Predictor Variables Candidates	AIC
1	$X_1, X_2, X_3, X_4, \ldots, X_{15}$	-16.21
2	$X_1, X_3, X_4, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	-19.30
3	$X_1, X_3, X_4, X_6, X_8, X_9, X_{10}, X_{11}, X_{13}$	-19.10
4	$X_1, X_3, X_4, X_6, X_8, X_9, X_{10}, X_{11}$	-19.63

B. EBLUP Model

Based on Table IV, all predictor variables are significant in the model. There are two predictor variables, which are the number of population (X_1) and the health worker coverage (X_9), which have a negative relationship (not unidirectional) with sub-district per capita expenditure. In

comparison, six other predictor variables have a positive (unidirectional) relationship. So, the EBLUP model obtained is:

$$\log(\hat{y}_i) = 13{,}3289 - 5{,}8557 \times 10^{-06} X_{1i} + 0{,}0609 X_{3i} + 0{,}9279 X_{4i} + 0{,}0088 X_{6i} + 3{,}4252 \times 10^{-05} X_{8i} - 0{,}0016 X_{9i} + 6{,}4263 \times 10^{-04} X_{10i} + 1{,}1506 X_{11i} + \hat{v}_i. \quad (14)$$

Equation (14) explains that the log per capita expenditure in each sub-districts in Jambi Province will be higher if universities, school facilities, polyclinics, doctors, people with disabilities, and midwives are included. Conversely, the log per capita expenditure for each sub-district in Jambi will be lower with more population and health workers.

TABLE IV
THE ESTIMATION OF EBLUP PARAMETERS WITH SELECTED PREDICTOR VARIABLES

Estimator	Coefficient Value	Standard Error	t-value	p-value
$\hat{\beta}_0$	13.329	0.072	186.377	0.000*
$\hat{\beta}_1$	-5.856e-06	1.607e-06	-3.645	0.000*
$\hat{\beta}_3$	0.061	0.023	2.698	0.007*
$\hat{\beta}_4$	0.928	0.277	3.357	0.001*
$\hat{\beta}_6$	0.009	0.003	3.353	0.001*
$\hat{\beta}_8$	3.425e-05	1.367e-05	2.505	0.012*
$\hat{\beta}_9$	-0.002	5.521e-04	-2.844	0.005*
$\hat{\beta}_{10}$	6.426e-04	3.239e-04	1.984	0.047*
$\hat{\beta}_{11}$	1.151	0.391	2.947	0.003*

*Note: * significant at the 0.05 significance level*

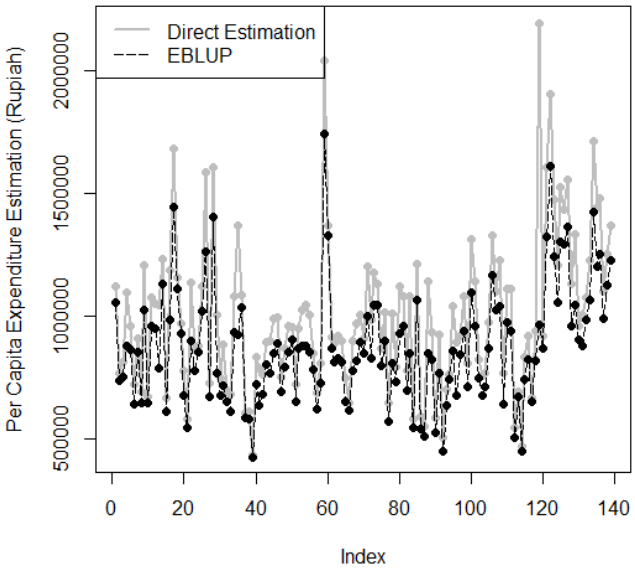


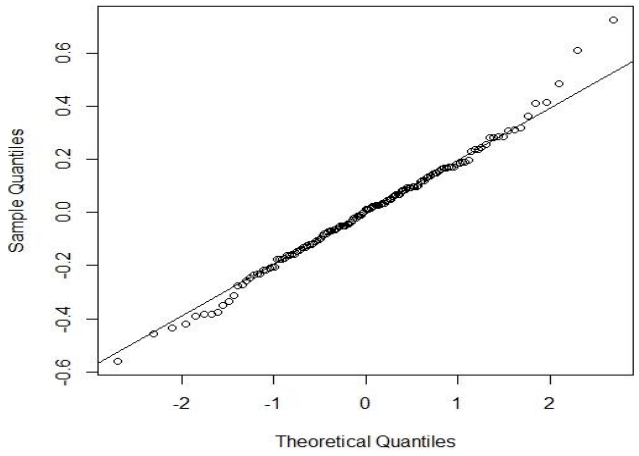
Fig. 2. Per Capita Expenditure Estimation (Rupiah) of Sub-district Level Using Direct Estimation Method and EBLUP Method

Then the reverse transformation on per capita expenditure per sub-district is calculated using equation (9). The results of per capita expenditure for each sub-district which are transformed back are then compared with the results of direct estimation, as shown in Figure 2. In Figure 2, the per capita expenditure estimation using the EBLUP method is lower than the direct estimation method.

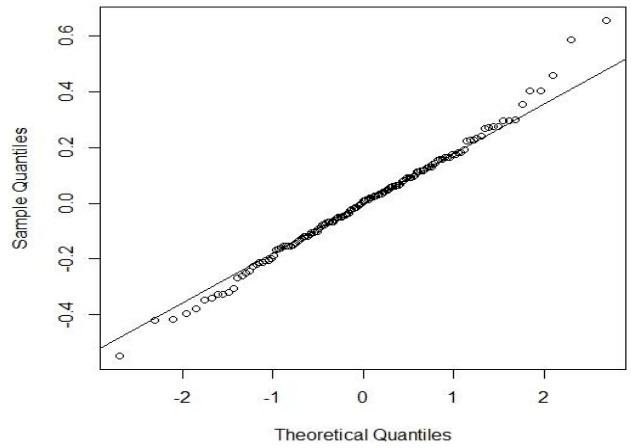
C. *EBLUP Model Diagnostics*

After obtaining the best EBLUP model and estimating per capita expenditure for each sub-district, the next step is to conduct a diagnostic model by examining the assumption of normality in the residuals and area random effects. Normality test through visualization of normal Q-Q plot is shown in Figure 3. The figure shows that the residual points are along the main diagonal line, so the residual and area random effects are normally distributed.

The normality test uses the Anderson-Darling test. The p-value for the residuals is 0.661, while the random effect of the area is 0.699. It concludes that the two variables are normal distribution. It means that the EBLUP model used in this study has met the assumptions of normality of residuals and area random effects.



a) Normal Q-Q Plot of Residuals



b) Normal Q-Q Plot of Area Random Effect

Fig. 3. Normal Q-Q Plot of Residuals and Area Random Effects

D. *RRMSE Direct Estimation and EBLUP Model Comparison*

Figure 4 shows that the EBLUP model gives a lower RRMSE than direct estimation. The RRMSE of the EBLUP model has a value of less than 25% in all sub-districts. On the other hand, the RRMSE of direct estimation has a very high value for Bathin II Pelayang and Tabir Ilir.

The RRMSE comparison between direct estimation and the EBLUP method is present in Figure 5. Based on the boxplot, direct estimation has a more comprehensive RRMSE range than the EBLUP method. Although there are

still outliers in both methods, outliers in the EBLUP model are pretty close to the boxplot tail. Thus, the EBLUP method produced a smaller diversity than the direct estimation method.

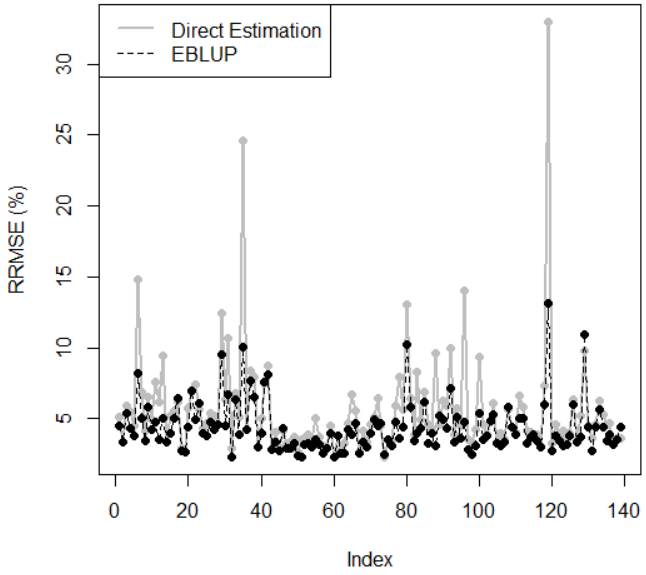


Fig. 4. RRMSE (%) Direct Estimation and EBLUP Model

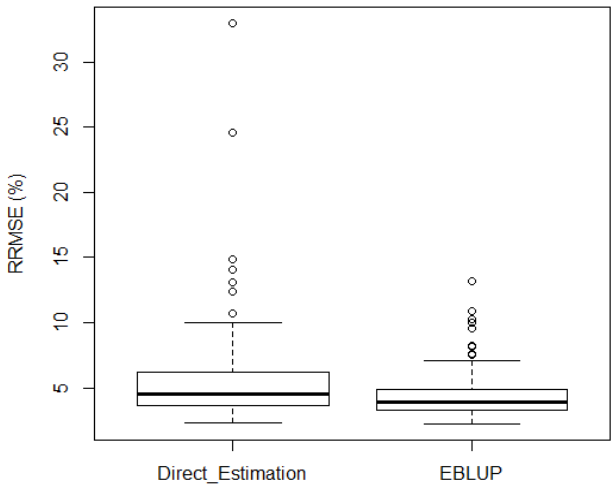


Fig. 5. Boxplot RRMSE (%) Direct Estimation and EBLUP Model

E. Estimation of Non-Sampled Sub-Districts

Per capita expenditure estimation in sub-districts uses cluster information by K-Medoids Cluster technique. In this clustering process, the EBLUP model uses eight significant predictor variables. The clustering process is used for all sub-districts (141 sub-districts) by determining the number of clusters in four clusters. Table V shows the results of grouping sub-districts in Jambi.

Analysis of the characteristics of each cluster uses the statistical average of the predictor variables. Cluster one includes sub-districts, with most of the villages being agricultural villages. In this cluster, school and health facilities are adequate, but health personnel such as midwives, doctors, etc. are inadequate.

The second cluster includes sub-districts, mainly the capital of regencies and sub-districts in Jambi City. This second cluster has the most population, tertiary institutions, an adequate number of school facilities, good and easily accessible health and health workforce facilities, and are industrial, commercial and economic center areas.

The third cluster has the smallest average population, limited school and health facilities, low health workforce coverage, and most sub-districts in this cluster are areas with challenges to access. Finally, the fourth cluster includes sub-districts with the most buildings along the river, where most of the population are engaged in agriculture and plantations.

After forming the sub-district clusters, the next step is to use the known components $\hat{v}_i$, then average per cluster. As shown in Table VI, information of $\bar{\hat{v}}_{(k)}$ is then entered into the EBLUP model as an estimator of the area random effects for non-sampled sub-districts.

The two non-surveyed sub-districts in this study are Pangkalan Jambu and Jujuhan Ilir. These sub-districts are a member of the third cluster. The estimations of per capita expenditure from EBLUP modelling with additional cluster information are in Table VII. Based on Table VII, the per capita expenditure estimation for Pangkalan Jambu is Rp. 656,725 and Jujuhan Ilir is Rp. 775,570.

TABLE V
SUB-DISTRICTS GROUPING IN JAMBI WITH K-MEDOIDS CLUSTER

Cluster	Cluster Members
1	Danau Kerinci, Air Hangat Timur, Depati VII, Kayu Aro, Gunung Tujuh, Bahar Selatan, Sungai Gelam, Kumpeh, Maro Sebo, Taman Rajo, Geragai, Bathin III, Danau Teluk.
2	Bangko, Sarolangun, Muara Bulian, Jambi Luar Kota, Tungkal Ilir, Rimbo Bujang, Rimbo Tengah, Kota Baru, Alam Barajo, Jambi Selatan, Paal Merah, Jelutung, Talanaipura, Danau Sipin, Jambi Timur.
3	Gunung Raya, Sitinjau Laut, Jangkat, Jangkat Timur, Muara Siau, Lembah Masurai, Tiang Pumpung, Batang Masumai, Sungai Manau, Renah Pembarap, Pangkalan Jambu, Tabir Ilir, Tabir Lintas, Tabir Barat, Batang Asai, Limun, Cermin Nan Gedang, Bathin VIII, Muara Sabak Barat, Kuala Jambi, Berbak, Sadu, Tungkal Ulu, Renah Mendaluh, Muara Papalik, Pengabuan, Bram Itam, Seberang Kota, Kuala Betara, Serai Serumpun, VII Koto Ilir, Bathin II Babeko, Rantau Pandan, Muko-Muko Bathin, Bathin III Ulu, Limbur Lubuk, Jujuhan Ilir, Kumun Debai, Sungai Penuh, Sungai Bungkal, Pesisir Bukit.
4	Bukit Kerman, Batang Merangin, Keliling Danau, Air Hangat, Air Hangat Barat, Gunung Kerinci, Siulak, Siulak Mukai, Kayu Aro Barat, Pamenang, Pamenang Barat, Renah Pamenang, Pamenang Selatan, Bangko Barat, Nalo Tantan, Tabir, Tabir Ulu, Tabir Selatan, Tabir Timur, Margo Tabir, Pelawan, Singkut, Pauh, Air Hitam, Mandiangin, Mersam, Maro Sebo Ulu, Bathin XXIV, Muara Tembesi, Bajubang, Maro Sebo Ilir, Pemayung, Mestong, Sungai Bahar, Bahar Utara, Kumpeh Ulu, Sekernan, Mendahara, Mendahara Ulu, Dendang, Muara Sabak Timur, Rantau Rasau, Nipah Panjang, Merlung, Batang Asam, Tebing Tinggi, Senyerang, Betara, Tebo Ilir, Muara Tabir, Tebo Tengah, Sumay, Tengah Ilir, Rimbo Ulu, Rimbo Ilir, Tebo Ulu, VII Koto, Pelepat, Pelepat Ilir, Tanah Sepenggal, Tanah Sepenggal Lintas, Tanah Tumbuh, Bathin II Pelayang, Jujuhan, Pasar Jambi, Pelayangan, Tanah Kampung, Pondok Tinggi, Hamparan Rawang, Koto Baru.

TABLE VI
THE MEAN OF AREA RANDOM EFFECTS FOR EACH CLUSTER

Cluster	$\bar{\hat{v}}_{(k)}$
1	-0.0569
2	-0.0345
3	-0.0153
4	0.0258

TABLE VII
PER CAPITA EXPENDITURE ESTIMATION IN NON-SAMPLED AREA

Sub-district	Per Capita Expenditure (Thousand Rupiah)	Cluster
Pangkalan Jambu	656.725	3
Jujuhan Ilir	775.570	3

IV. CONCLUSION

A. Conclusion

Estimating sub-districts per capita expenditure using the EBLUP method generates smaller RRMSE than the direct estimation method. It concludes that this method can improve the estimation parameters obtained using the direct estimation method. Besides, per capita expenditure in non-surveyed sub-districts can be obtained using the EBLUP method with cluster information.

B. Suggestion

The model used in this study is the area level model (Fay-Herriot). The variable of interest is continuous data transformed into a normal distribution. Further research can be done with a unit-level model with no normal distribution or using SAE robust techniques.

In estimating non-sample sub-districts with cluster information, further development of other SAE models can use cluster information. In addition, other cluster techniques can also be developed for further research, for example, using fuzzy clustering techniques.

ACKNOWLEDGEMENT

The authors would like to thank Amiek Chamami, SST, M.Stat and Faisal Haris, SST for data collection and support.

REFERENCES

[1] IGNM. Jaya, Y. Andriyana, and B. Tantular, "Spatial Prediction of Malaria Risk with Application to Bandung City, Indonesia," *IAENG International Journal of Applied Mathematics*, vol.51, no.1, pp 199-206, 2021.

[2] IGNM. Jaya, and A. Chadidjah, "Spatial Autoregressive in Ecological Studies: A Comparison of the SAR and CAR Models," *Engineering Letters*, vol. 29, no.1, pp 207-212, 2021.

[3] C. Giusti, S. Marchetti, M. Preseti, and N. Salvati, "Robust Small Area Estimation and Oversampling in Estimation of Poverty Indicators," *Survey Research Methods*. vol. 6, pp: 155-163, 2012.

[4] JNK. Rao, and I. Molina, *Small Area Estimation Second Edition*, New York: John Wiley & Sons, 2015.

[5] A. Salma, K. Sadik, and KA. Notodiputro, "Small Area Estimation of Per Capita Expenditures Using Robust Empirical Best Linear Unbiased Prediction (REBLUP)," *AIP Conference Proceedings*, vol. 1827, pp 020027, 2017.

[6] A. Ubaidillah, KA. Notodiputro, A. Kurnia, and IW. Mangku, "Multivariate Fay-Herriot Models for Small Area Estimation with Application to Household Consumption per Capita Expenditure in Indonesia," *Journal of Applied Statistics*, vol. 46, issue. 15, pp 2845-2861, 2019.

[7] D. Dedianto, and IY. Wulansari, 2018. "Aplikasi Small Area Estimation (SAE) Metode Pseudo-EBLUP pada Official Statistics di Indonesia," *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol.10, no. 2, pp 33-38, 2018.

[8] A. Saei, and R. Chambers, "Empirical Best Linear Unbiased Prediction for Out of Sample Area". *Southampton Statistical Sciences Research Institute Working paper*, M05/03, 2005.

[9] R. Anisa, A. Kurnia, and Indahwati, "Cluster Information of Non-Sampled Area in Small Area Estimation," *IOSR Journal of Mathematics*, vol. 10, issue. 1, pp: 15-19, 2014.

[10] Wahyudi, KA. Notodiputro, A. Kurnia, and R. Anisa, "A Study of Area Clustering using Factor Analysis in Small Area Estimation," *AIP Conference Proceedings*, vol. 1707, pp. 080017, 2016.

[11] AN. Susanti, K. Sadik, and A. Kurnia, "A Comparison of Cluster Method and Nearest Neighbor Method for Non-sample Area in the Small Area Estimation," *International Journal of Scientific Research in Science, Engineering and Technology*, vol.4, no. 9, pp 463-368, 2018.

[12] BPS (Badan Pusat Statistik), *Buku Pedoman Kepala BPS Provinsi, Kepala Bidang Statistik Sosial, dan Kepala BPS Kabupaten/Kota, SUSENAS Maret 2018*, Jakarta: BPS. pp: 14-18, 2018.

[13] DM. Hindmarsh, "Small area estimation for health surveys", *Doctor of Philosophy thesis*, School of Mathematics and Applied Statistics, University of Wollongong, 2013.

[14] KA. Notodiputro, and A. Kurnia, "Pendekatan General Linear Mixed Model pada Small Area Estimation," *Forum Statistika dan Komputasi*, vol. 10, no.2, pp 12-16, 2005.

[15] A. Kurnia, *Prediksi Terbaik Empirik untuk Model Transformasi Logaritma di Dalam Pendugaan Area Kecil dengan Penerapan pada Data SUSENAS*, Bogor: Institut Pertanian Bogor, 2009.

[16] L. Kish, *Survey Sampling*, New York: John Wiley & Sons, 1965.

[17] A. Patel, and P. Singh, "New Approach for K-mean and K-medoids Algorithm," *International Journal of Computer Applications Technology and Research*, vol. 2, issue 1, pp 1-5, 2013.

Irlandia Ginanjar (M'20) received his PhD in Mathematics at the Institut Teknologi Bandung, in 2017. He is a Lecturer at the Department of Statistics, Universitas Padjadjaran. His research interests are related to multivariate statistical analysis, sampling, computer science, and marketing. He has published research papers on national and international journals and conference proceedings; some indexed in Scopus. Moreover, he acts as head of statistics department and head of the research group in big data analysis.

Septie Wulandary received her Master's degree in Applied Statistics at the Department of Statistics, Universitas Padjadjaran. She is a Statisticians at the BPS-Statistics of Jambi Province, Jambi. Her research interests are related to regression, multivariate statistical analysis, data mining, and computer science. She has published research papers on national journals.

Toni Toharudin currently Associate Professor at the Department of Statistics, Universitas Padjadjaran, Indonesia. Toni does research in Statistics. He earned his Master of Science from KU Leuven, Belgium (2004-2005) and Doctoral degree from The University of Groningen (2007-2010). Moreover, he acts as head of the research group in sociometric and time series.