Jurnal Matematika Vol. 15, No.2, Desember 2025, pp. 138-145 ISSN: 1693-1394
Article DOI: 10.24843/JMAT.2025.v15.i02.p193

Small Area Estimation to Estimate the Percentage of Simple Housing Needs in Buleleng Regency

Luh Devi Maharani M[1], I Komang Gde Sukarsa[2*], I Wayan Sumarjaya[3]

[1,2,3]*Program Studi Matematika, Fakultas MIPA-Universitas Udayana, Indonesia*

Email: gedesukarsa@unud.ac.id[2]

*Corresponding Author

ABSTRACT

Data is a source of information to support the decision-making process of the object under study so that the availability of data is important to fulfill. Survey as one of the techniques used to provide data has weaknesses such as area parameters outside the design cannot be explained by the statistics generated. Small area estimation (SAE) is an indirect estimator method that can overcome the problems of survey data. SAE is used to estimate population or subpopulation parameters with limited sample size to produce statistics that are as precise as direct estimators. The general approach used in SAE is EBLUP. The topic in this research is the need for simple houses in Buleleng Regency. By estimating the percentage of simple housing needs for each sub-district in Buleleng through direct estimation and EBLUP, the results show that the estimation value of the direct estimator is not more precise than EBLUP indicated by the MSE output of the direct estimator which is greater than EBLUP. Based on the output of the estimation results with the EBLUP method, the results show that the highest and lowest percentage of housing needs are in Sukasada Sub-district at 67.05% and Banjar Sub-district at 37.19%, respectively.

Keywords: Small Area Estimation; Direct Estimation; Indirect Estimation; Empirical Best Linear Unbiased Prediction (EBLUP); Mean Squared Error (MSE).

I. Introduction

Data is a crucial element in any study, as it provides information about specific objects or subjects that are being investigated. The availability of data is a primary consideration in the analytical process, as it shapes the decisions that are ultimately made. The choice of analytical method is also a crucial aspect of this process, as it influences the reliability of the conclusions that are drawn. In the field of statistics, a plethora of data analysis methods exists to address a multitude of issues. Two prominent examples are the census and survey, which are the most frequently utilized methods by Central Bureau of Statistics (BPS) to generate official statistics and data. The resulting statistics serve as a valuable informational resource for a diverse range of stakeholders.

Surveys are a more prevalent methodology than censuses. In essence, a survey is a technique employed on a predetermined object, which implies that the statistics yielded are limited to objects that have been predetermined (1), or that the estimated parameters are only capable of explaining the specified objects. The object in question encompasses demographic or socio-demographic areas or regions (2,3). This raises the question of whether the official statistics produced are able to accommodate areas not involved in the survey, given the continued acceleration of data needs in an effort to optimize policies in the public and private sectors. This situation is referred to as data limitation, or in the context of survey research, as

small area. Small areas are also referred to as areas that are unable to produce precise estimates through the direct approach due to insufficient data size (3).

The direct approach, or direct estimation, is a design-based technique that employs survey weights and associated inferences (3). Direct estimation techniques are founded upon data sourced from the pertinent area; thus, in small areas with constrained data availability, it may result in elevated standard error values (1). Statistical methods offer a viable approach to addressing these limitations through a model-based approach, also known as indirect estimation. The indirect method represents an alternative solution for cases of data limitations, utilizing auxiliary variables that connect information between areas through a model (2). Auxiliary variables can be obtained from other areas of the same survey, previous surveys, or related variables (2).

Two approaches commonly employed in small area estimation (SAE) are direct and indirect estimation. SAE is a method used to estimate small area parameters (3). This method has been widely applied in several large countries to address issues related to health, agriculture, income, and poverty. In practice, SAE employs data from a population to estimate parameters from smaller subpopulations (4). The indirect approach is based on two basic models: the basic area-level model (Type A), also referred to as the Fay-Herriot model (2), and the basic unit-level model (Type B).

There are multiple approaches to indirect estimation, including empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). EBLUP represents a non-Bayesian approach, whereas EB and HB are Bayesian approaches. Parameter estimation using the Bayes method necessitates prior information that is seldom available, rendering non-Bayes methods a potential solution (5). The EBLUP method, which is applicable to continuous data types, represents a refinement of the BLUP approach (3). Furthermore, the EBLUP model represents a development of the Fay-Herriot model, as outlined by (3).

The case study in this research is related to the necessity for the construction of simple houses in Buleleng Regency. A house is one of three primary necessities for a decent quality of life. As indicated on the official website of REI, the increase in productive generations implies an increase in housing needs. Buleleng remains the regency with the highest population growth rate in Bali Province, at 2.33%, according to the 2020 census results (6). Consequently, the need for affordable, decent housing in Buleleng is still high when compared to other regencies in Bali Province

II. Methods

This research project is situated within Buleleng Regency, with the data used as the response variable obtained through a direct survey conducted by the researchers themselves. This involved several respondents in each of the nine sub-districts within Buleleng Regency.

Moreover, the variable employed in this study is the percentage of simple housing needs, which serves as the variable of interest or response variable (Y). The percentage of simple housing needs is calculated by dividing the number of households that do not own a house by the total number of sampled households in each subdistrict. The auxiliary variables (X_i) include the number of poor people (X_1), data on affordable housing prices (X_2), average family size (X_3), the area of each subdistrict (X_4), and population density (X_5).

The indirect approach in Small Area Estimation (SAE) is based on two basic models, which are contingent on the availability of auxiliary variables

Basic Area Level Model (Type A)

This model is employed when the auxiliary variables are only available at the area level. It is also referred to as the Fay-Herriot model (2):

$$\hat{\theta}_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + b_i v_i + e_i, i = 1, \dots, m \quad (1)$$

where,

$\hat{\theta}_i$: Estimated small area parameter

$\boldsymbol{z}_i^T$: Auxiliary variables

$\boldsymbol{\beta}$: Regression coefficient

b_i : Known positive constants

v_i : Small area random effect

e_i : Sampling error

m : Number of observation

Basic Unit Level Model (Type B)

This model is employed when the auxiliary variable is at the unit level and corresponds with the variable of interest.

$$\theta_{ij} = \boldsymbol{z}_{ij}^T \boldsymbol{\beta} + \nu_i + e_{ij} \quad (2)$$
$$j = 1, \dots, N_i; i = 1, \dots, m$$

Empirical Best Linear Unbiased Prediction (EBLUP)

The EBLUP method is applicable to continuous data types and represents a refinement of the BLUP approach. Since estimating the variance component of random effects is often unknown in practice, it must be estimated through sample data. The EBLUP model represents a further development of the Fay-Herriot model:

$$\hat{\theta}_i^H = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \boldsymbol{z}_i^T \widehat{\boldsymbol{\beta}} \quad (3)$$

where,

$\hat{\gamma}_i = \frac{\sigma_v^2 b_i^2}{\psi_i + \sigma_v^2 b_i^2}$,

σ_v^2 : Small area random effect variance

$\boldsymbol{z}_i^T$: Auxiliary variables

$\widehat{\boldsymbol{\beta}}$: Regression coefficient

The following is a description of the stages of data analysis employed in this study.

1. Perform direct estimation of the response variable with the formula:

$$\hat{\bar{Y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (4)$$
$$i = 1, \dots, m; j = 1,2, \dots, n_i$$

where:

$\hat{\bar{Y}}_i$: The estimated mean percentage of basic housing requirements in the i-th subdistrict,

n_i : denotes the number of households in the i-th sub-district,

y_{ij}: the percentage of need for the j-th simple house in the i-th sub-district,

m : number of sub district.

2. Test the normality assumption on the response variable using the Kolmogorov-Smirnov Test.
3. Selecting auxiliary variables that are significantly correlated with the response variable

using the Pearson Product Moment Correlation Test.
4. Perform indirect estimation with the EBLUP method.
5. Calculating the mean squared error of the direct and indirect estimates with the Bootstrap method.
6. Comparing the MSE calculation results.

III. Results and Discussion

Prior to commencing the analytical process, comprising direct and indirect estimation, a data exploration phase is initiated. The ensuing section presents the findings of the data exploration undertaken in the present study.

Table 1. Data Description

Desc	Y	X_1	X_2	X_3	X_4	X_5
Min	29.73	23359	100.8	3.25	46.94	274
1st Q	34.69	29098	126.0	3.36	97.68	508
Med	50.00	37041	134.4	3.38	118.24	593
Mean	52.12	38200	136.7	3.39	151.76	880
3rd Q	61.54	47606	144.0	3.41	172.93	856
Max	85.71	57860	168.0	3.53	356.57	3233

a. The Percentage of Simple Housing Needs (Y)
   The data on the percentage of simple housing needs indicates that the highest level of housing needs is 85.71% in Kubutambahan Sub-district, while the lowest is 29.73% in Tejakula Sub-district.
b. The Number of Poor People (X_1)
   It is established that the district with the highest concentration of impoverished individuals is Gerokgak, with a total of 57,860 individuals classified as poor. Conversely, Busungbiu is identified as the district with the lowest prevalence of poverty, with a total of 23,359 individuals classified as poor.
c. Affordable Housing Price (X_2)
   The lowest price of inexpensive housing units is IDR108,000,000 in the Busungbiu subdistrict, while the highest price is IDR168,000,000 in the Buleleng and Sukasada subdistricts.
d. Average Family Size (X_3)
   The average number of family members is highest in Sukasada and lowest in Banjar.
e. Luas Wilayah (X_4)
   The largest sub-district is Kecamatan Gerokgak and the smallest is Kecamatan Buleleng.
f. Population Density (X_5)
   The subdistrict with the highest population is Buleleng, while the subdistrict with the lowest population is Busungbiu.

3.1 Direct Estimation of the Percentage of Simple House Needs in Buleleng Regency

Table 2. Simple House Needs in Buleleng Regency

Subdistrict	Number of KK	House Needs (%)
Gerokgak	30,389	34.69
Seririt	30,236	50.00
Busungbiu	15,996	61.54
Banjar	28,076	42,22

Sukasada	26,625	74.42
Buleleng	46,703	34.67
Sawan	25,668	56.10
Kubutambahan	21,877	85.71
Tejakula	23,350	29.73

The results indicate that Kubutambahan is the subdistrict with the highest level of housing need, with a percentage of 85.71%, and Tejakula is the subdistrict with the lowest level of housing need, with a percentage of 29.73%. These results indicate that the 85.71% figure in Kubutambahan sub-district represents the number of families lacking a simple housing structure, or approximately 18,750 families. Similarly, in Tejakula sub-district, as many as 6,942 households lack a simple house.

Moreover, the assumption of normality for the response variable data was tested using the Kolmogorov-Smirnov approach. The resulting p-value (0.9665) was found to be greater than the alpha value (0.05), indicating that the response variable data follows a normal distribution. Consequently, the EBLUP method can be employed in this context.

3.2 EBLUP Estimation of Percentage of Simple House Needs in Buleleng Regency

Prior to estimating with the EBLUP method, it is first necessary to determine the auxiliary variables that will be used. An appropriate auxiliary variable can be identified through the significance coefficient value between the response variable and the auxiliary variable. In this study, the Pearson Product Moment (PPM) Correlation Test was employed, and the following results were obtained:

Table 3. Correlation Coefficient

Variabel	Koefisien Korelasi	Intepretasi
$\boldsymbol{Y}$ **dan** $\boldsymbol{X_1}$	0.4	Korelasi rendah
$\boldsymbol{Y}$ **dan** $\boldsymbol{X_2}$	0.2	Korelasi diabaikan
$\boldsymbol{Y}$ **dan** $\boldsymbol{X_3}$	0.5	Korelasi rendah
$\boldsymbol{Y}$ **dan** $\boldsymbol{X_4}$	0.1	Korelasi diabaikan
$\boldsymbol{Y}$ **dan** $\boldsymbol{X_5}$	0.3	Korelasi rendah

The results above are interpreted in accordance with the methodology proposed by (7), as outlined in (8). In light of the aforementioned results, it was determined that variable X_3 is the most suitable auxiliary variable. The percentage of simple housing needs in Buleleng Regency (Y) exhibits a moderate correlation with the variable average number of family members in Buleleng Regency (X_3). This finding is also supported by the research of (9), which indicates that the number of households is a significant factor influencing housing demand.

Once the accompanying variables have been determined, the subsequent step is to estimate the variance component of the random effect (ν_i) using the restricted maximum likelihood (REML) technique. Previously, it was established that the fundamental area-level model, also known as the Fay-Herriot model in small areas, is:

$$\hat{\theta}_i = \boldsymbol{z}_i^T \boldsymbol{\beta} + b_i \nu_i + e_i, i = 1, \dots, m \quad (5)$$

with $\boldsymbol{z}_i^T$ is $p \times 1$ vector of area-level auxiliary variables, ν_i is small area random effect that assumed to be independent and identically distributed $\nu_i \sim (0, \sigma_v^2)$, e_i is sampling error with $e_i \sim (0, \psi_i)$ and ψ_i is known, ν_i and e_i are independent.

Moreover, the optimal linear unbiased prediction (BLUP) of θ_i under the assumption of known σ_v^2 is given by (2):

$$\tilde{\theta}_i^H = \boldsymbol{z}_i^T\tilde{\boldsymbol{\beta}} + \nu_i$$
$$\tilde{\theta}_i^H = \boldsymbol{z}_i^T\tilde{\boldsymbol{\beta}} + \gamma_i\left(\hat{\theta}_i - \boldsymbol{z}_i^T\tilde{\boldsymbol{\beta}}\right)$$

where $\gamma_i = \frac{\sigma_\nu^2 b_i^2}{\psi_i + \sigma_\nu^2 b_i^2}$;

$\psi_i = MSE\left(\widehat{\theta_i}\right) = {}^{S_i^2}/_{n_i}, i = 1, \ldots, m;$

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_\nu^2) = \left[\sum_{i=1}^m \boldsymbol{z}_i\,\boldsymbol{z}_i^T \Big/ \left(\psi_i + \sigma_\nu^2 b_i^2\right)\right]^{-1}\left[\sum_{i=1}^m \boldsymbol{z}_i\hat{\theta}_i \Big/ \left(\psi_i + \sigma_\nu^2 b_i^2\right)\right] \quad (6)$$

In practice, it is challenging to ascertain σ_v^2, therefore, it must be estimated initially using sample data (10) with the REML method.

The subsequent step is to employ the Fay-Herriot model to simulate the limited area in question. Initially, one must estimate the Fay-Herriot coefficient ($\tilde{\boldsymbol{\beta}}$), the random effect ($\nu_i$), and the variance of the random effect (σ_v^2). Utilising the R software, the value of $\sigma_v^2 = 164.4$, with the following regression coefficient outcomes:

Table 4. Regression Coefficient

Variable	Coefficient	*Standard Error*	*p-value*
Intercept	-286.02	228.73	0.251
$\boldsymbol{X_3}$	99.75	67.65	0.183

The results indicate that the p-value of variable X_3 is greater than the significance threshold ($\alpha = 0.05$), thereby demonstrating that variable X_3 does not exert a statistically significant influence on the response variable. Moreover, a normality test was conducted on the residual variance of the random effect (ν_i), and the results demonstrated that the p-value was less than the significance threshold (α), specifically $4.441 \times 10^{-16} < 0.05$. This indicates that the distribution of the residual variance of the random effect does not adhere to a normal distribution. This indicates that the resulting model is invalid, and thus the EBLUP model cannot be used to estimate the percentage value of simple housing needs in Buleleng Regency. However, the analysis will continue until the final stage of this study. The resulting Fay-Herriot model is:

$$\hat{Y} = -286.02 + 99.75X_3 \quad (7)$$

where:

$\hat{Y}$: The percentage of simple housing needs in Buleleng Regency (%)

X_3 : Average of family size

The model above is interpreted as follows:

1. The constant value produces a negative value of -286.02, which indicates that when the variable average number of family members in Buleleng Regency (X_3) is 0 or constant and does not change, the percentage of simple housing needs in Buleleng Regency is -286.02%. This can be interpreted as a decrease in housing needs by 286.02%.
2. The coefficient value generated by the variable "average number of family members in Buleleng Regency" (X_3) is 99.75. This indicates that an increase of one unit in the variable "average number of family members in Buleleng Regency" will result in a 99.75% increase in the percentage of housing needs in Buleleng Regency.

Moreover, the same program, specifically the *R* program, provides an additional EBLUP estimation method, namely the prediction of the EBLUP value through the small area model obtained. The results of the alternative method are presented in the following table.

Table 5. EBLUP Estimation

Subdistrict	Number of KK	House Needs (%)
Gerokgak	30,389	49.99
Seririt	30,236	40.39
Busungbiu	15,996	65.98
Banjar	28,076	37.19
Sukasada	26,625	67.05
Buleleng	46,703	48.92
Sawan	25,668	51.03
Kubutambahan	21,877	54.25
Tejakula	23,350	54.25

The results indicate that Sukasada is the subdistrict with the highest level of housing need, with a percentage of 67.05%, while Banjar is the subdistrict with the lowest level of housing need, with a percentage of 37.19%. These results indicate that 67.05% of households in Sukasada lack access to simple housing, equating to approximately 17,852 households. Similarly, 10,441 households in Banjar are also without adequate housing.

3.3 Mean Squared Error (MSE)

The results of the mean square error (MSE) calculation for both the direct and indirect estimation methods are presented in the line diagram below.

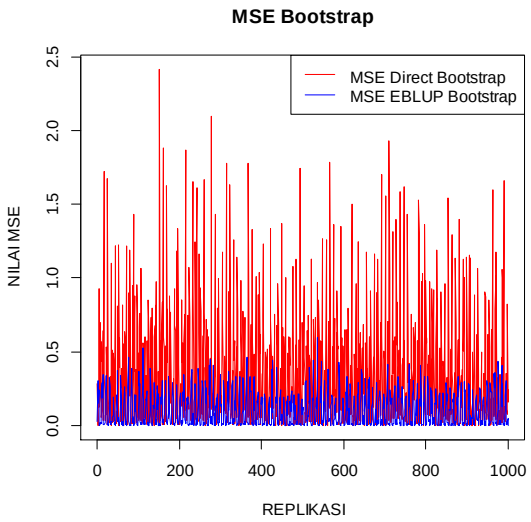

Figure 1. *Mean Square Error* Direct Estimation and EBLUP

The figure above demonstrates that, in accordance with the Bootstrap approach, the MSE of the direct estimator is greater than that of the EBLUP estimator. The minimum and maximum values produced by the direct estimator and EBLUP estimator are 3.211111×10^{-6} and 2.415613, and 1.234568×10^{-9} and 0.5959566, respectively. Therefore, it can be stated that the MSE of the EBLUP Bootstrap is less than that of the direct Bootstrap.

IV. Conclusion

The results of this study indicate that while the EBLUP method produces a lower Mean Squared Error (MSE) than the direct estimator , empirical evidence suggests its performance is not inherently superior in this specific case because the normality assumption for random effect variance residuals was not met, resulting in an invalid model. Nevertheless, consistent

with numerous SAE-related studies showing that indirect methods effectively reduce MSE, the EBLUP estimates are determined to be more precise than direct estimates for this study. Direct estimation identifies Kubutambahan as having the highest housing needs and Tejakula the lowest, whereas the EBLUP approach identifies Sukasada with the highest needs and Banjar with the lowest. Future research should consider utilizing unit-level models (Type B) if unit-level auxiliary variables are available, and address current limitations by testing for outliers or employing Spatial EBLUP (SEBLUP) methods to account for spatial influences.

References

[1] Aminah AS, Pawitan G, Tantular B. Empirical best linear unbiased prediction method for small areas with restricted maximum likelihood and bootstrap procedure to estimate the average of household expenditure per capita in Banjar Regency. In: AIP Conference Proceedings. 2017. p. 1–7.
[2] Rao JNK. Small Area Estimation. New Jersey: John Wiley & Sons, Inc.; 2003.
[3] Rao JNK, Molina I. Small area estimation. New Jersey: John Wiley & Sons, Inc.; 2015.
[4] Sunandi E, Agustina D, Fransiska H. Estimating the Poverty level in the Coastal Areas of Mukomuko District Using Small Area Estimation: Empirical Best Linear Unbiased Prediction Method. In: International Conference on Statistics and Analytics 2019. 2020.
[5] Amaliana L, Lusiana ED. Penerapan Metode Empirical Best Linear Unbiased Prediction (EBLUP) pada Model Fay-Herriot Small Area Estimation (SAE). In: Prosiding Seminar Nasional Integrasi Matematika dan Nilai Islami. 2017. p. 312–9.
[6] Statistik Indonesia 2023 - Badan Pusat Statistik Indonesia [Internet]. [cited 2025 Dec 30]. Available from: https://www.bps.go.id/id/publication/2023/02/28/18018f9896f09f03580a614b/statistik-indonesia-2023.html
[7] Hinkle D, Wiersma W, Jurs S. Applied statistics for the behavioral sciences [Internet]. 2003 [cited 2025 Dec 31]. Available from: https://library.wur.nl/WebQuery/titel/1944963
[8] Kencana EN. Metode Statistika pada Kepariwisataan dan Hospitaliti. Denpasar: Pustaka Larasan; 2024.
[9] Ayuningtyas FJ, Astuti IP. Faktor Penentu Permintaan Rumah Tinggal di Provinsi Daerah Istimewa Yogyakarta. Jurnal Ekonomi & Studi Pembangunan. 2018;19(1):85–90.
[10] Widiarti, Periwi RR, Sutrisno A. Perbandingan Mean Squared Error (MSE) Metode Prasad-Rao dan Jiang-Lahiri-Wan Pada Pendugaan Area Kecil. In: Seminar Nasional Teknoka. 2017. p. 56–60.

.