International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

International Journal of
Sciences:
Basic and Applied
Research

ISSN 2307-4531
(Print & Online)

Published by:
GSSRR

Visit: www.gssrr.org

Gini Ratio Prediction by Estimating the Components Based on the Ybarra-Lohr Model Small Area Estimation with Estimated Sampling Variance

Nadya Avicena[a], Anang Kurnia[b*], I Made Sumertajaya[c]

[a,b,c]*IPB University, Jl. Raya Dramaga, Babakan, Dramaga District, Bogor City, West Java, Indonesia*

[b]*Email: anangk@apps.ipb.ac.id*

Abstract

Gini ratio is one of the tools used to measure income inequality, so it is necessary to know the value of Gini ratio to a smaller regional level such as a subdistrict. According to Badan Pusat Statistik (BPS), the components of the Gini ratio are the average per capita expenditure and the relative frequency of households for each expenditure class in the subdistrict. Per capita expenditure data available through SUSENAS is designed to obtain national statistics down to the district level so that estimates are made for the level of subdistrict expenditure classes. Direct estimation for a small sample can cause significant standard errors therefore Small Area Estimation (SAE) with Logarithm Transformation is used to estimate the average per capita expenditure for each subdistrict expenditure class in Depok City 2020. The Ybarra-Lohr area-level model was used because of the availability of auxiliary data with measurement error. Previously, the sampling variance required for estimating the average per capita expenditure was estimated by comparing several estimation methods. As sampling variance estimation method, probability distribution produces an estimate of the average per capita expenditure with the smallest RRMSE, with a random effect variance and goodness of Ybarra-Lohr model are $\hat{\sigma}_v^2$ = 0.686 and R^2= 0.929. The best result of the average per capita expenditure estimation for each expenditure class is used to obtain Gini ratio for each subdistrict in Depok City 2020.

Keywords: Sampling Variance; Gini Ratio; saeme; EBLUP Log Transform.

----

* Corresponding author.

1. Introduction

Economic inequality or inequality in terms of income is a fundamental problem in Indonesia. Income inequality is a striking difference in income in society [1]. One of the measuring tools used to measure income inequality is the Gini ratio, which values 0 to 1. The smaller Gini ratio, the more evenly distributed the income distribution is. Various efforts have been made to reduce the Gini ratio by the government. However, these efforts will not be right on target if the location of inequality is not known to the smallest regional level. Gini ratio up to sub-district level is not available. The Gini ratio at the subdistrict level can be calculated using the following formula [2]:

$$R_i = 1 - \sum_{j=1}^{m_i} f_{pj}\left(F_{cj-1} + F_{cj}\right),\ \ i = 1, \dots, l$$

Where R_i is an estimator of Gini ratio of each i-th subdistrict, f_{pj} is the relative frequency of households in the j-tj expenditure class, and F_{cj} is the cumulative frequency of the proportion of expenditure class, so that f_{pj}, F_{cj-1}, and F_{cj} are components of the Gini ratio.

Per capita expenditure average of each expenditure class as a numerator for the proportion of expenditure can be calculated using data from the National Socio-Economic Survey (SUSENAS) conducted by BPS. The relative frequency of households in each expenditure class for each subdistrict can also be calculated using SUSENAS data by dividing the number of households on the j-th and i-th expenditure class with the number of households in the i-th subdistrict. However, per capita expenditure data available through SUSENAS is designed to obtain national statistics down to the district level, so the results of this survey are not suitable for direct estimation because large standard errors will arise, which is caused by inadequate sample sizes.

The method developed to deal with this problem is the Small Area Estimation (SAE), which uses additional information obtained from similar areas, previous surveys conducted in the same area, and other variables related to the variable that wants to predict [3]. This kind of estimation is called indirect estimation. One of the methods often used in small area estimation is Empirical Best Linear Unbiased Prediction (EBLUP). Reference [3] states that there are two types of basic SAE models used as the basis for the EBLUP model, namely the area-level model and the unit-level model. Both models are based on the availability of supporting data used. The unit-level model is used if the available auxiliary data correspond individually to the response data. In contrast, the area-level model is used if the auxiliary data is only available at a certain area-level. The auxiliary variables used in SAE ideally can explain the diversity between small areas and do not contain errors. However, in reality the expected auxiliary variables are often unavailable or do not match conceptually and period [4], so Reference [5] examines the SAE model that uses auxiliary variables with measurement error. In addition, standard SAE models such as EBLUP are sometimes not able to explain the data well because of the tight linearity assumptions of the model. Therefore Reference [6] made improvements by modifying EBLUP by first transforming the response data before using SAE to estimate the parameter’s variable of concern. In addition to using the p-vector of the area-level auxiliary variables, the known sampling variance (SV) assumption is also used. Usually, the SV estimated directly from the sample can be unreliable [7]. Reference [8] compared SV estimation methods to see their effect on small area estimation precision.

This study will estimate per capita expenditure average for each class of expenditure in each subdistrict as a component of the subdistrict Gini ratio. The skewed to the right distribution of per capita expenditure tend not to meet the normality assumption in the EBLUP. The corresponding auxiliary variables contain errors, and the sampling variance is estimated using several methods (direct estimation, probability distribution, and bootstrap). The area-level model using auxiliary variables with measurement errors with logarithmic transformations on the per capita expenditure response needs to be done to predict the Gini ratio of each subdistrict in the city of Depok.

2. Methodology

2.1. Gini Ratio

According to Reference [2], the Gini ratio measures the level of inequality in expenditure as a proxy for population income. The Gini coefficient is based on the Lorenz curve, a cumulative expenditure curve that compares the distribution of a certain variable (e.g. income) with a uniform distribution representing the cumulative percentage of the population. The Gini coefficient ranges from 0 to 1. If the Gini coefficient is 0 it means perfect equality, while if it is 1, it means perfect inequality. Changes in the Gini ratio indicate a change in the distribution of population expenditure. The Gini ratio has decreased, which means that the distribution of population expenditures has improved. Gini ratio can be calculated through the following formula:

$$R_i = 1 - \sum_{j=1}^{m_i} f_{pij}\left(F_{cij-1} + F_{cij}\right),\ \ i = 1, \dots, l \qquad (1)$$

where R_i is Gini ratio of the i-th subdistrict, f_{pij} is the relative frequency of households in the j-th expenditure class, F_{cij-1} is the cumulative frequency of the proportion of expenditures in the (j-1)-th expenditure class, and F_{cij} cumulative frequency of the proportion of expenditures in the j-th expenditure class. Based on equation (1), the Gini ratio value of each sub-district is calculated based on the relative frequency of households in the j-th expenditure class and the proportion of expenditures for the j-th expenditure class in the i-th subdistrict. The class of expenditure, in this case, is the expenditure interval used by BPS in Depok City in Figures (2021) as follows:

Table 1: BPS Expenditure Class.

Class Code	Expenditure Interval (Rupiah)
1	300.000 – 499.999
2	500.000 – 749.999
3	750.000 – 999.999
4	1.000.000 – 1.499.999
5	≥ 1.500.000

Source: Reference [9]

The definition of per capita expenditure according to Reference [9] is the cost incurred for the consumption of all household members for a month divided by the number of household members. The pattern of expenditure can be used as a tool to measure the level of welfare of the population, where changes in its composition are used as an indicator of changes in the level of welfare. In the 2020 National Socio-Economic Survey (SUSENAS) data, per capita expenditure is available for each household unit, so the average per capita expenditure for each expenditure class is calculated based on the following equation:

$$\bar{y}_{ij} = \frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{N_{ij}}, \qquad i = 1, \dots, l\,;\ j = 1, \dots, m_i \tag{2}$$

Where $\bar{y}_{ij}$ is per capita expenditure average in the j-th expenditure class and i-th subdistrict, y_{ijk} is k-th household expenditure per capita, j-th expenditure class and i-th sub-district, and N_{ij} is the number of households in the j-th expenditure class and i-th subdistrict. The proportion of expenditure for each class of expenditure in each subdistrict is:

$$c_{ij} = \frac{\bar{y}_{ij} N_{ij}}{\sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} y_{ijk}}, \qquad i = 1, \dots, l \tag{3}$$

Where c_{ij} is the proportion of expenditure per capita for the j-th and i-th subdistrict. Meanwhile, the relative frequency of households according to expenditure class can be defined based on the following equation:

$$f_{pij} = \frac{p_{ij}}{p_i}, \ \ i = 1, \dots, l\,;\ j = 1, \dots, m_i \tag{4}$$

Where f_{pij} is the relative frequency of households in the j-th and i-th sub-districts of expenditure class, p_{ij} is the number of households in the j-th expenditure class and i-th subdistrict, and p_i is the number of households in the i-th subdistrict.

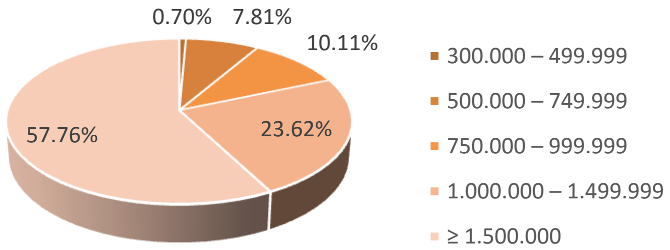

Figure 1: Percentage of Depok City Population.

by Per Capita Expenditure Class a Month in 2020 [9]

The percentage of population by per capita expenditure class in Figure 1 is obtained by direct estimation based on SUSENAS 2020 data. Direct estimation can be done because the survey is designed to obtain national statistics up to the district/city level. Larger sample size is needed to get the population percentage according to the expenditure class of each subdistrict in Depok. To increase the sample size, the average expenditure per capita in the j-th expenditure class for each sub-district ($\bar{y}_{ij}$) will be estimated using the Small Area Estimation (SAE).

2.2. Small Area Estimation (SAE)

A survey is conducted to directly estimate a parameter for a population with a certain area or domain level with a large number of samples. If a direct estimator is used for a smaller domain, but the available samples for that domain are not large enough, then the standard error is large [10]. This problem can be handled by small area estimation by adding information about the same parameter in another small area with similar characteristics, past values, or variables that correlate with the observed variable. This estimation technique is also known as an indirect estimation. The statistical approach for model-based indirect estimators is divided into two, implicit and explicit models. Explicit models include unit-level models and area-level models [3]. An Area-level based model is based on the availability of auxiliary variable data that exists only for a certain area-level. It is assumed that the variable of interest is a function of the average of the response variables, $\theta_i = g(\bar{Y}_i)$ for a certain $g(\,.\,)$ which relates to the auxiliary data of a certain small area $x_i = (x_{1i}, \dots, x_{pi})^T$ and follow the linear model as follows:

$$\theta_i = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} + z_i v_i, i = 1, \dots, m \qquad (5)$$

Where z_i is a known positive constant and $\beta = (\beta_1, \dots, \beta_p)^T$ is a regression coefficient of size $p \times 1$, while v_i is a random effect area which is assumed to have an identical and independent distribution with $E_m(v_i) = 0$ and $V_m(v_i) = \sigma^2{}_v, \sigma^2{}_v \geq 0$. The estimator θ_i can be found by assuming that a direct estimator of $\widehat{\theta_\iota}$ exists:

$$\widehat{\theta_\iota} = \theta_i + e_i, i = 1, \dots, m$$

Where $e_i \sim N(0, \sigma^2_{e_i})$ and $\sigma^2_{e_i}$ are known. So, from the above two equations we get:

$$\widehat{\theta_\iota} = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} + z_i v_i, + e_i, i = 1, \dots, m \qquad (6)$$

which is a special form of mixed linear model or Fay-Herriot model in small area estimation.

2.3. Empirical Best Linear Unbiased Prediction (EBLUP)

The models used in the area-level are:

$$\widehat{\theta_\iota} = \theta_i + e_i = \widehat{\theta_\iota} = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} + z_i v_i, + e_i, i = 1, \dots, m$$

Where $\boldsymbol{x_i}$ is the area-level variable and z_i is a design matrix. The technique for solving the model to obtain a

BLUP for $\theta_i = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} + z_i v_i$ has been developed by Reference [11], assuming σ_v^2 is known. The BLUP estimator for θ_i is as follows:

$$\hat{\theta}_i^{BLUP} = \boldsymbol{x_i}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}} + \gamma_i(\widehat{\theta_\iota} - \boldsymbol{x_i}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}})$$

$$\hat{\theta}_i^{BLUP} = \gamma_i\widehat{\theta_\iota} + (1 - \gamma_i)\boldsymbol{x_i}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}}$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2)$ and $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient with Generalized Least Square (GLS) where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T V^{-1} X})^{-1}\boldsymbol{X^T V^{-1}}\hat{\boldsymbol{\theta}}$ **.** The BLUP method developed by Reference [11] assumes that the components of the variance of random effects are known in the linear mixed model, while in reality the components of this variance are unknown. As a result, the variance of random effects must be estimated. Reference [12] used the maximum likelihood (ML) method and the restricted maximum likelihood (REML) method. Estimation of σ_v^2 using either the ML or REML methods was performed using the Fisher Scoring algorithm. The EBLUP estimator by replacing the value of σ_v^2 with its estimator $\hat{\sigma}_v^2$ is as follows:

$$\hat{\theta}_i^{EBLUP} = \gamma_i\widehat{\theta_\iota} + (1 - \gamma_i)\boldsymbol{x_i}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}}$$

2.4. Area-level Model SAE using Auxiliary Variable with Measurement Error

SAE introduced by Reference [13] assumes that there is no error in the auxiliary variables used in the model. Reference [14] then examines the Fay-Herriot model using auxiliary variables with measurement errors. The Ybarra-lohr model is a modification of the Fay-Herriot model when the presence of errors is not ignored. Ybarra-Lohr modified the Fay-Herriot model to the following model:

$$\hat{\theta}_i = \hat{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta} + r_i(x_i\hat{\boldsymbol{x}}_{\boldsymbol{i}}) + e_i \qquad (7)$$

where,

i = 1, 2, …, m is the index of small area

$\hat{\theta}_i$ = direct estimator of the i-th small area parameter θ_i

x_i = vector of auxiliary variables in the i-th small area

β = regression coefficient vector

$r_i(x_i\hat{x}_i)$ = $(x_i - \hat{x}_i)\beta + v_i$

v_i = random effect of the i-th small area, assumed to be independently and identically distributed, $v_i \sim N(0, \sigma_v^2)$ and $\sigma_v^2 \geq 0$

e_i = sampling error, $e_i \sim N(0, \sigma_{e_i}^2)$

2.5. EBLUP with Logarithm Transformation

A logarithmic transformation is defined in the linear mixed model as follows:

$$\widehat{\theta_i^L} = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta} + z_i v_i, + e_i$$

where $\widehat{\theta_i^L} = \frac{1}{n_i}\sum_{j\epsilon s(i)} \log(y_{ij})$, sampling error $e_i \sim N(0, \sigma_e^2)$, area random effect $v_i \sim N(0, \sigma_v^2)$. Reference [6] explains that if we follow the standard EBLUP theory, EBLUP with a mean value of $\log(y_{ij})$, then the estimator for θ_i can be written as follows:

$$\hat{\theta}_i^{EBLUP*} = \hat{\gamma}_i \widehat{\theta_i^L} + (1 - \hat{\gamma}_i) x_i^T \hat{\beta}$$

With $\hat{\beta}$ obtained based on the weighted least squares method for the regression parameter β of the mixed linear model, where $\hat{\gamma}_i = \sigma_v^2/(\sigma_v^2 + \sigma_e^2)$. Because what is expected is the actual estimator for the mean value in each i-th area, the lognormal distribution is used to perform the back transformation of the model. It is assumed that $\hat{\theta}_i^{EBLUP*}$ is normally distributed. Reference [6] formulates the actual estimator for the mean or estimator of the logarithm transformation EBLUP $\hat{\theta}_i^{AK-EBLUP}$ for the i-th area as follows:

$$\hat{\theta}_i^{AK-EBLUP} = \exp(\hat{\theta}_i^{EBLUP*} + \frac{1}{2}\hat{v}_i^{EBLUP*}) \qquad (8)$$

Then the MSE estimator $\hat{\theta}_i^{AK-EBLUP}$ is as follows:

$$MSE\left(\hat{\theta}_i^{AK-EBLUP}\right) = e^{\hat{v}_i^{EBLUP*}}(e^{\hat{v}_i^{EBLUP*}} - 1)e^{2\hat{\theta}_i^{EBLUP*}} \qquad (9)$$

where $\hat{v}_i^{EBLUP*}$ is MSE of $\hat{\theta}_i^{EBLUP*}$.

2.6. Sampling Variance Estimation Method

Area-level SAE model is a model based on the availability of auxiliary variable data that exists only for a certain area-level. The model uses a p-vector of the area-level supporting variables with the assumption that the sampling variance (SV) is known. Usually, the SV estimated directly from the sample can be unreliable [7]. Reference [8] compared SV estimation methods to see their effect on small area estimation precision. Some methods used include direct estimation, probability distribution, and bootstrap. In the direct estimation method, the following equation is used:

$$s_i^2 = \frac{\sum(y_{ij} - \bar{y}_i)^2}{n_i - 1}, i = 1, \dots, m; \quad j = 1, \dots, n_i \qquad (10)$$

where $\bar{y}_i$ and n_i are the average and number of samples in i-th area. The probability distribution method, assuming normality of the calculated parameters, SV is estimated based on:

$$(n_i - 1)s_i^2 \sim \sigma_i^2 \chi^2_{(n_i-1)}, \quad i = 1, \dots, m \tag{11}$$

where σ_i^2 is SV in i-th area and s_i^2 is calculated based on equation (10). In the bootstrap method, 1000, 10000, and 50000 resamplings are performed and then estimate the SV using equation (10) which has been described.

2.7. Analysis Procedure

This study uses data from the National Socio-Economic Survey (SUSENAS 2020) with 987 sample households spread across 11 sub-districts in Depok City, West Java. The stages carried out in this research are as follows:

1. Data Preparation.

   a. Define the output class.

   b. The data used as response variables and direct estimators in this case is the average per capita expenditure at the area-level (expenditure class).

   c. The data used as a supporting variable is the average building area for each expenditure class.

   d. Checking normality assumptions for response variables.

2. Estimating per capita expenditure average of each expenditure class.

   a. Estimating the average expenditure per capita directly for each expenditure class based on equation (2).

   b. Estimating sampling variance with direct estimation method based on equation (10), probability distribution based on equation (11), and bootstrap (resampling 1000, 10000, and 50000).

   c. Estimating per capita expenditure average based on the logarithmic EBLUP transformation using an auxiliary variable with measurement error and sampling variance from the estimated results of step 2b.

   d. Comparing the best Relative Root Mean Square Error (RRMSE) of each estimated per capita expenditure result.

$$RRMSE = \frac{\sqrt{MSE_{ij}}}{\hat{\hat{y}}_{ij}} \times 100\%$$

3. Predict directly the relative frequency of households for each class of expenditure based on equation (4).
4. Calculating the Gini ratio of each sub-district based on equation (1) using the estimated average expenditure per capita with the best RRMSE and the relative frequency of households.

3. Result and Discussion

3.1. Selection of Expenditure Class Intervals, Auxiliary Variable, and Normality Assumption of Response Variables

The calculation of the sub-district Gini ratio based on equation (1) is carried out at the expenditure class level, so the estimation of the average per capita expenditure will also be carried out at the expenditure class level. The selection of expenditure class intervals is based on BPS expenditure class intervals in table 1 and the distribution of per capita expenditure data for Depok City 2020. So, the expenditure class intervals used in this study are as follows:

for each Expenditure Class in Depok City 2020 (in hundreds of thousands of rupiah)

Table 3: Summary of Per Capita Expenditure Average.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.774	19.550	27.714	37.618	122.142

Table 4: Expenditure Class Interval Used.

Class Code	Expenditure Class (Rupiah)
1	<500.000
2	500.000 – 749.999
3	750.000 - 999999
4	1.000.000 – 1.499.999
5	1.500.000 – 2.499.999
6	2.500.000 – 3.499.999
7	3.500.000 – 4.499.999
8	>=5.000.000

The auxiliary variables were selected based on their correlation with the response variables. From the variables contained in the SUSENAS data, the variable average building area (RLB) is considered the most correlated variable with the response variable with a correlation of 0.345 and p-value 0.001843 in the Pearson correlation test. After selecting the appropriate auxiliary variables, checking the normality assumption is carried out on the per capita expenditure response variables. Per capita expenditure as a response variable in this study has not met the normality assumption, so a log transformation is needed to meet the assumptions and presented in the QQ plot below.

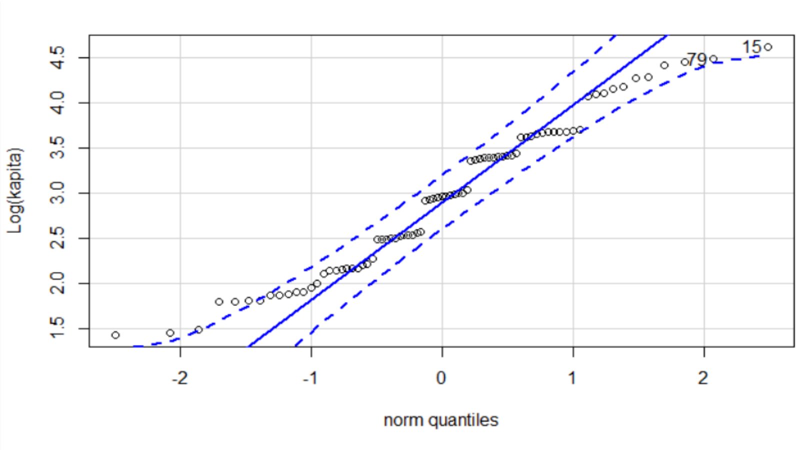

Figure 2: Log Transform QQ Plot Per Capita Expenditure Average, Depok City Expenditure Class.

3.2. Result of Direct Estimation of Per Capita Expenditure Average

Expenditure per capita in the SUSENAS data is available at the unit (household) level. To estimate the per capita expenditure average in each expenditure class in all subdistricts directly, it is calculated based on equation (2), and the results are presented in the table below.

Table 5: Direct Estimation of Per Capita Expenditure Average.

Subdistrict	Class Code	n	f_{pij}	$\hat{y}_{\text{direct}}$	Var_{direct}
10 (Sawangan)	101	1	0.0167	0.416	0
	102	3	0.05	0.698	0.002
	103	4	0.0667	0.846	0.015
	104	9	0.15	1.262	0.018
	105	23	0.3833	2.004	0.024
	106	9	0.15	2.855	0.013
	107	1	0.0167	3.743	0
	108	10	0.1667	8.576	0.356
11 (Bojongsari)	112	4	0.08	0.598	0.012
	113	4	0.08	0.823	0.011
	114	13	0.26	1.284	0.014
	115	15	0.3	1.933	0.024
	116	6	0.12	3.053	0.011
	117	4	0.08	3.728	0.006
	118	4	0.08	10.147	0.481
20 (Pancoran mas)	202	5	0.0391	0.602	0.018
	203	11	0.0859	0.875	0.010
	204	15	0.1172	1.25	0.014
	205	39	0.3047	1.957	0.016
	206	23	0.1797	2.956	0.010
	207	15	0.1172	3.954	0.008
	208	20	0.1563	7.266	0.205
21 (Cipayung)	211	2	0.0333	0.442	0.015
	212	2	0.0333	0.669	0.015
	213	5	0.0833	0.902	0.009
	214	10	0.1667	1.257	0.014
	215	13	0.2167	2.077	0.018
40 (Cimanggis)	402	4	0.0301	0.665	0.002
	403	4	0.0301	0.968	0.002
	404	19	0.1429	1.203	0.018
	405	28	0.2105	2.014	0.024
	406	14	0.1053	2.976	0.010
	407	16	0.1203	3.94	0.007
	408	48	0.3609	8.299	0.213
41 (Tapos)	412	4	0.0339	0.649	0.012
	413	7	0.0593	0.845	0.006
	414	20	0.1695	1.219	0.013
	415	30	0.2542	1.943	0.025
	416	26	0.2203	2.923	0.005
	417	12	0.1017	3.951	0.004
	418	19	0.161	6.071	0.024
50 (Beji)	501	1	0.0083	0.425	0
	502	7	0.0583	0.656	0.007
	503	4	0.0333	0.862	0.006
	504	14	0.1167	1.208	0.013
	505	29	0.2417	1.989	0.027
	506	28	0.2333	2.973	0.010
	507	9	0.075	3.895	0.005
	508	28	0.2333	7.162	0.153
60 (Limo)	602	1	0.0345	0.74	0
	604	5	0.1724	1.253	0.016
	605	10	0.3448	1.906	0.029
	606	8	0.2759	3.022	0.011
	607	3	0.1034	3.934	0.000

Subdistrict	Class Code	n	f_{pij}	$\hat{y}_{\text{direct}}$	Var_{direct}
	216	11	0.1833	3.056	0.003
	217	6	0.1	3.878	0.010
	218	11	0.1833	6.515	0.075
30 (Sukmajaya)	302	3	0.02	0.609	0.010
	303	12	0.08	0.873	0.008
	304	18	0.12	1.204	0.017
	305	39	0.26	1.885	0.016
	306	16	0.1067	3	0.007
	307	21	0.14	4.052	0.005
	308	41	0.2733	6.381	0.069
31 (Cilodong)	312	1	0.0156	0.612	0
	313	1	0.0156	0.912	0
	314	4	0.0625	1.301	0.014
	315	20	0.3125	1.865	0.021
	316	14	0.2188	3.119	0.008
	317	3	0.0469	3.778	0.002
	318	21	0.3281	6.035	0.088
	608	2	0.069	5.865	0.027
61 (Cinere)	612	4	0.0533	0.645	0.014
	613	4	0.0533	0.87	0.009
	614	6	0.08	1.207	0.026
	615	13	0.1733	1.845	0.028
	616	12	0.16	2.886	0.010
	617	10	0.1333	3.98	0.008
	618	26	0.3467	8.86	0.282

The table above shows that the sample in the six expenditure classes consists of only one household, which causes the sampling variance to be zero. The sampling variance in other expenditure classes is also close to zero, which is caused by the value of per capita expenditure is almost the same across households in each expenditure class. This makes the sampling variance calculated based on equation (2) unreliable if used to estimate expenditure per capita using the log transformation EBLUP method. In practice, smoothed sampling variance estimates are used in the Fay-Herriot model and then considered as known variances [15].

3.3. Estimating Per Capita Expenditures with EBLUP Log Transform and Predicted Sampling Variance

Reference [8] used several methods to estimate the sampling variance and compared their effects on the precision of small area estimation. The method used is direct estimation, probability distribution, bootstrap, and bayes. In this study, the sampling variance will be estimated using a direct estimator, probability distribution, and bootstrapping. To estimate per capita expenditure using the results of the estimated sampling variance and with the help of auxiliary variables with measurement errors, first the random and fixed components are estimated in the Ybarra-Lohr model with the log transformation EBLUP. The analysis was carried out using R software the help of the saeme package. The best prediction of per capita expenditure will be selected based on the comparison of RRMSE values.

a) Ybarra-Lohr Model Results and Prediction of Per Capita Expenditure with Direct Estimation Variance

The direct estimation method produces a sampling variance which is considered as a known variance to estimate the random and fixed components of the Ybarra-Lohr model and to estimate per capita expenditure for each class of expenditure in all subdistricts in Depok City.

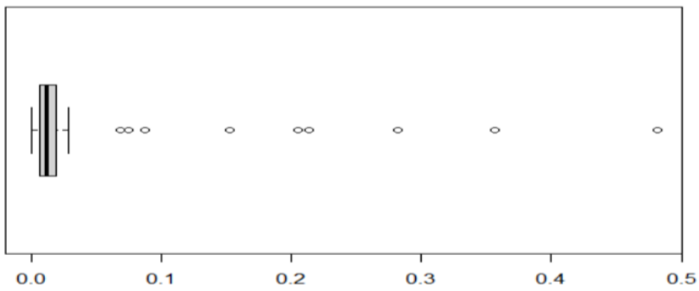

Figure 3: Boxplot of Direct Estimation's SV.

and Random Effects with Variance of Direct Estimation

Table 6: Estimate of Regression Coefficients.

	$\hat{\beta}$	SE	*p-value*
Intercept	2.862	0.095	1.23×10^{-196}
RLB	0.00022	0.000143	1.17×10^{-1}

$\hat{\sigma}_v^2$ = 0.657 R^2= 0.772

(in hundreds of thousands of rupiah)

Table 7: Summary of Predicted Per Capita with Variance of Direct Estimation.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.822	19.280	24.116	37.566	86.614

b) Ybarra-Lohr Model Results and Prediction of Per Capita Expenditure with Probability Distribution Variance

The results of sampling variance with a probability distribution based on equation 11 are considered as known variances for use in estimating the random and fixed components of the Ybarra-Lohr model, and estimating per capita expenditure for each class of expenditure in all subdistricts in Depok City. In contrast to other methods, RLB used as a supporting variable significantly affects per capita expenditure with =10%.

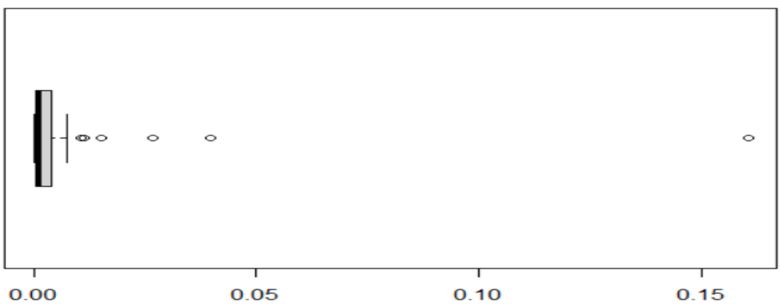

Figure 4: Boxplot of Probability Distribution's SV.

and Random Effects with Variance of Probability Distribution

Table 8: Estimate of Regression Coefficients.

	$\hat{\beta}$	SE	*p-value*
Intercept	2.889	0.096	6.795 x 10^{-199}
RLB	0.000223	0.00012	7.17 x 10^{-2}

$\hat{\sigma}_v^2$ = 0.686 R^2= 0.929

(in hundreds of thousands of rupiah)

Table 9: Summary of Predicted Per Capita with Variance of Probability Distribution.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.746	19.331	25.934	37.587	88.517

c) Ybarra-Lohr Model Results and Prediction of Per Capita Expenditure with Bootstrap Variance

Resampling to get sampling variance with bootstrap was done with three types of replications (1000, 10000, and 50000). Each of these results is used as a known variance to estimate the random and fixed components of the Ybarra-Lohr model, and to estimate per capita expenditure for each expenditure class in all subdistricts in Depok City.

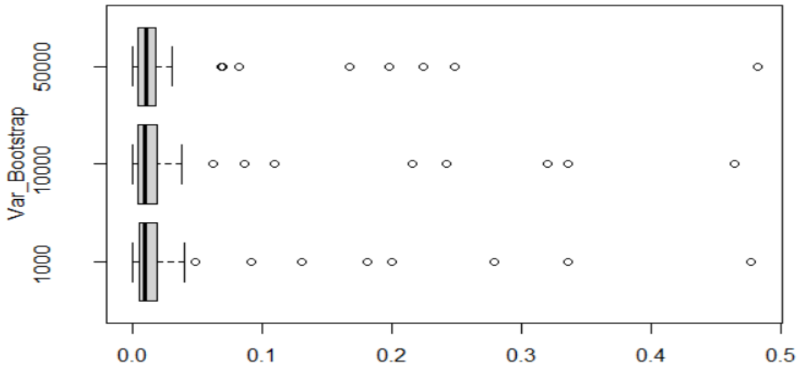

Figure 5: Boxplot of Bootstrap's SV with 1000, 10000. and 50000 replicates.

and Random Effects with Variance of Bootstrap 1000 replicates

Table 10: Estimate of Regression Coefficients

	$\hat{\beta}$	SE	*p-value*
Intercept	2.865	0.0956	5.187 x 10^{-197}
RLB	0.000224	0.00014	1.1867 x 10^{-1}

$\hat{\sigma}_v^2$ = 0.66 R^2= 0.794

and Random Effects with Variance of Bootstrap 10000 replicates

Table 11: Estimate of Regression Coefficients.

	$\hat{\beta}$	SE	*p-value*
Intercept	2.861	0.0956	1.123 x 10^{-196}
RLB	0.000225	0.00014	1.093 x 10^{-1}

$\hat{\sigma}_v^2$ = 0.66 R^2= 0.77

and Random Effects with Variance of Bootstrap 50000 replicates

Table 12: Estimate of Regression Coefficients.

	$\hat{\beta}$	SE	*p-value*
Intercept	2.87	0.0958	5.492 x 10^{-197}
RLB	0.000221	0.00015	1.64 x 10^{-1}

$\hat{\sigma}_v^2$ = 0.662 R^2= 0.847

(in hundreds of thousands of rupiah)

Table 13: Summary of Predicted Per Capita with Variance of Bootstrap 1000 Replicates.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.820	19.289	24.414	37.261	86.609

(in hundreds of thousands of rupiah)

Table 14: Summary of Predicted Per Capita with Variance of Bootstrap 10000 Replicates.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.803	19.302	24.414	37.566	86.891

(in hundreds of thousands of rupiah)

Table 15: Summary of Predicted Per Capita with Variance of Bootstrap 50000 Replicates.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.161	8.825	19.279	24.645	37.567	85.068

The selection of the estimated per capita expenditure results used as a component of the Gini ratio is based on the Relative Root Mean Square Error (RRMSE).

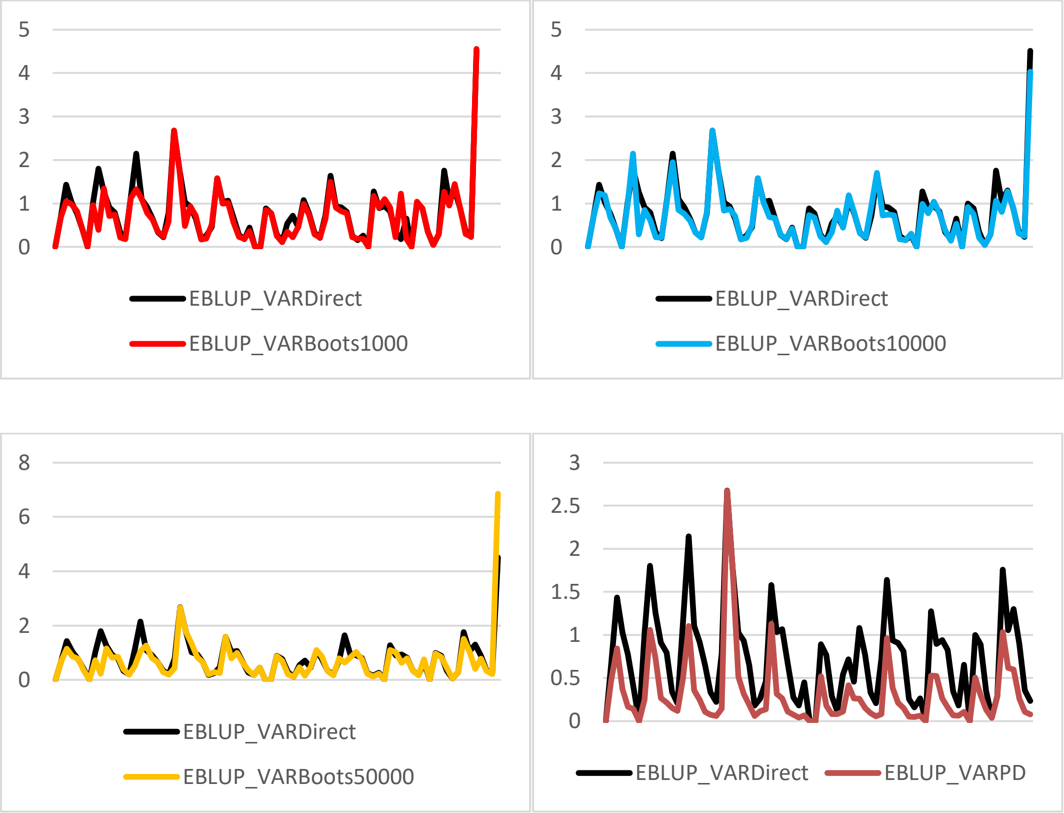

Figure 6: RRMSE Comparison of Ybarra-Lohr direct estimating variance, Ybarra-Lohr probability distribution variance, and Ybarra-Lohr bootstrap variance.

The figure above compares the RRMSE for 79 expenditure classes in all sub-districts in Depok City. From the comparison of the four methods, the estimated 50000 replicate bootstrap variance results have the smallest RRMSE in 30 classes of expenditure. The estimated 10000 replicate bootstrap variance results have the smallest RRMSE in 26 expenditure classes, and the 1000 replicate bootstrap variance estimates have the smallest variance in 15 expenditure classes.

The estimated variance with a probability distribution has the smallest RRMSE in the 66 classes of expenditure, while the RRMSE from the direct estimate of variance tends to be the largest among other methods. Then the result of the estimated per capita expenditure used to predict the Gini ratio is the result of the prediction of the Ybarra-Lohr variance probability distribution model.

3.4. Gini Ratio Prediction

After obtaining the estimated results for the average per capita expenditure and the relative frequency of households for each expenditure class, the Gini ratio for each sub-district can be calculated based on the estimated results. An example of the calculation for one subdistrict (Sawangan) can be seen in table 16.

Table 16: Gini ratio calculation for Sawangan subdistrict.

Kecamatan	Kelas Pengeluaran	f_{pij}	c_{ij}	F_{cij}	$F_{cij} + F_{cij-1}$	$f_{pij}(F_{cij-1} + F_{cij})$	R
10 (Sawangan)	1 (<500000)	0.0167	0.0024	0.002	0.0024	0.00004	0.39
	2 (500000 – 749999)	0.0500	0.0122	0.015	0.0171	0.00085	
	3 (750000 – 999999)	0.0667	0.0198	0.034	0.0491	0.00328	
	4 (1000000 – 1499999)	0.1500	0.0663	0.101	0.1353	0.02029	
	5 (1500000 – 2499999)	0.3833	0.2685	0.369	0.4700	0.18019	
	6 (2500000 – 3499999)	0.1500	0.1496	0.519	0.8882	0.13322	
	7 (3500000 – 4499999)	0.0167	0.0218	0.541	1.0595	0.01766	
	8 (>=5000000)	0.1667	0.4593	1	1.5407	0.25678	

The results of the Gini ratio calculation for each sub-district in the Depok city with the results of estimating the average indirect per capita expenditure will be compared with the results of the calculation with the direct estimation of the average per capita expenditure, and the results can be seen in Figure 4.

Based on the picture below, the Gini ratio in each subdistrict with per capita expenditure due to the estimated SAE log transformation tends to be smaller than the Gini ratio with the estimated per capita expenditure directly. In the Gini ratio with per capita expenditure as a direct estimate, Sawangan and Bojongsari subdistricts are the largest (0.41), while Limo subdistricts are the lowest (0.26). For the Gini ratio with per capita expenditure as a result of the SAE log transformation prediction, Sawangan and Cinere subdistricts became the largest (0.39), while the subdistrict with the lowest Gini ratio was Limo subdistrict (0.25).

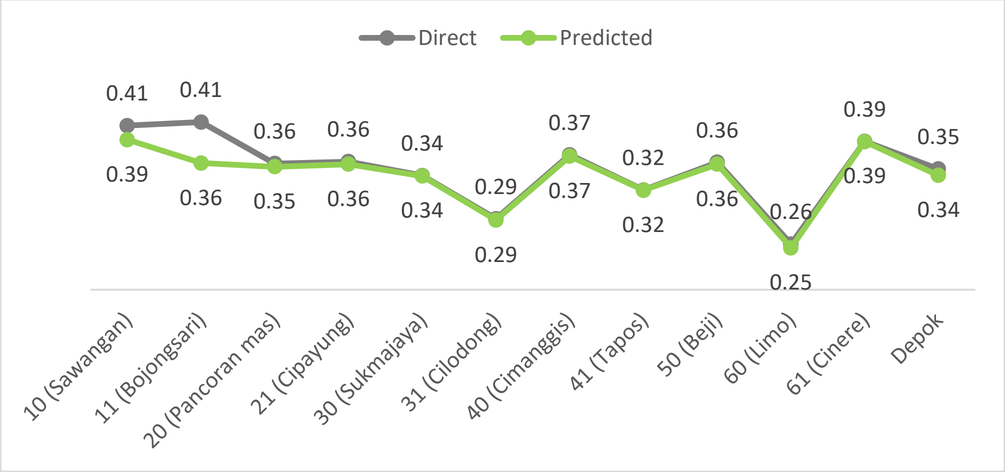

Figure 4: Comparison Plot of the Gini Ratio for each subdistrict in the Depok City (2020).

4. Conclusion

Estimates of per capita expenditure for each class of expenditure in all subdistricts in Depok City 2020 are directly and indirectly. Indirect estimation uses the EBLUP log transformation method, an auxiliary variable with measurement error and sampling variance estimated first. Sampling variance estimated using a probability distribution produces an estimate of per capita expenditure with the smallest RRMSE compared to other variance sampling estimation methods with the variance and goodness of the Ybarra-Lohr model being $\hat{\sigma}_v^2 =$ 0.686 and R^2= 0.929.

The estimated per capita expenditure as a component of the Gini ratio with the smallest RRMSE are used to predict the Gini ratio of each subdistrict in Depok City 2020. The predicted value of the subdistrict Gini ratio calculated by direct and indirect per capita expenditure is not much different, but the small sample size in each expenditure class makes indirect estimates better.

References

[1] M. P. Todaro. *Pembangunan Ekonomi di Dunia Ketiga*. Jakarta : Ghalia Indonesia. 2003.

[2] Badan Pusat Statistik. “Gini Rasio” Internet: https://sirusa.bps.go.id, [Mar. 2021]

[3] J. N. K. Rao. *Small Area Estimation*. New Jersey : John wiley and Sons, Inc. 2003.

[4] H. Chandra, U. C. Sud, and Y. Gharde. “Small Area Estimation Using Estimated Population Level Auxiliary Data”. *J Communications in Statistics-Simulation and Computation*, vol. 44, pp. 1197-1209, Oct. 2004.

[5] M. Komalasari. “Kajian Pendugaan Area Kecil Menggunakan Peubah Penyerta yang Mengandung Galat (Studi Kasus: Rata-rata Lama Sekolah di Kabupaten Kampar).” thesis, Institut Pertanian Bogor, Indonesia, 2019.

[6] A. Kurnia. “Prediksi Terbaik Empirik untuk Model Transformasi Logaritma di dalam Pendugaan Area Kecil dengan Penerapan pada Data SUSENAS.” dissertation, Institut Pertanian Bogor, Indonesia, 2009.

[7] Y. You and B. Chapman. “Small Area Estimation using Area-level Models and Estimated Sampling Variances”, *Survey Methodology,* vol: 32, pp. 97-103. 2006.

[8] M. S. Kermanshahi, Y. Mehrabi, A. Kavousi, A. R. Baghestan, and F. M. Nasrabadi. “Effects of Sampling Variance Estimation Methods on Precision of Small Area Estimation.”, *JP Journal of Biostatistics,* vol: 14, pp. 93-106. 2017.

[9] Badan Pusat Statistik. *Kota Depok dalam Angka 2021*. Depok: Badan Pusat Statistik. 2021.

[10] M. Ghosh and J. N. K. Rao. “Small Area Estimation: An Appraisal”, *Statistical Science,* vol: 9, pp. 55-98. 1994.

[11] C. R. Henderson. “Best Linear Unbiased Estimation and Prediction under a Selection Model”, *Biometrics,* vol: 31, pp. 423-447. 1975.

[12] D. A. Harville. “Maximum Likelihood Approach to Variance Component Estimation and Related Problems”. *Journal of The American Statistical Association*. vol: 73, pp. 724-731. 1977.

[13] R. E. Fay and R. A. Herriot. “Estimates of Income for Small Places: An Aplication of James-Stein Procedures to Census Data”. *Journal of The American Statistical Association,* vol: 74, pp. 269-277, 1979

[14] L. M. R. Ybarra and S. L. Lohr. “Small Area Estimation when Auxiliary Information Measured with Error.” *Biometrika*. vol: 95, pp. 919-931. 2008.

[15] Y. You. “Small Area Estimation using Area-level Models with Model Checking and Aplications”, in *Proc*. Statistical Society of Canada 2008 Proceedings of The Survey Methods Section. 2008.