Small Area Estimation of Poverty Using Remote Sensing Data (Case Study: Expenditure Per Capita Estimation of Very Poor Households in West Java, Indonesia)

Novia Permatasari[1], Bagaskoro Cahyo Laksono[1], Azka Ubaidillah[2]
[1] BPS-Statistics Indonesia
[2] Polytechnic of Statistics STIS

Abstract

Solving the problem of poverty, which is one of the main concerns of governments around the world, begins with providing accurate data to describe the population in poverty itself. The use of small-area estimation to estimate poverty in a small area is increasingly needed in order to get better poverty monitoring and policy making.

The success of small area model estimation depends on auxiliary variables used to make appropriate estimates and smaller variances. Numerous studies show that remote sensing with various advantages, such as objective measurement, low cost, frequent updates, and comprehensive area coverage, can be used as a covariate in small area models. However, studies using several remote sensing variables to estimate poverty are still limited, especially for very poor households whose expenditure per capita/month is less than 0.8 poverty line. This research aims to implement remote sensing data as auxiliary variables in a small area model to estimate expenditure per capita of very poor households in West Java, Indonesia.

The method used in this research is Small Area Estimation using Fay Herriot Model. Here we compare small area models using administrative data as a common auxiliary variable and remote sensing data (nighttime light, land surface temperature, air pollution, and spectral indices.

We found nighttime light data, which shows an area's social and economic activity, as a good auxiliary variable for expenditure per capita estimation of very poor households. Direct estimates, small area estimates of model using administrative data, and small area estimates of model using remote sensing data have similar patterns. Both small area models produced more accurate estimates for unsampled areas than direct estimates for sampled areas. Although the relative standard error is still slightly higher than the model using administrative data, remote sensing data is preferable because of the lower cost and more comprehensive coverage.

Overall, we show the potential of using remote sensing data as an auxiliary variable in the small area estimation model for poverty estimation. These small area estimates can be used to see poverty in detail, so the government can take policies to minimize very poor household and poverty in general. For further research, we can explore other remote sensing data showing poverty conditions in detail.

Keywords: big data, expenditure, poverty, small area, EBLUP

1. Background

Measuring poverty is crucial for creating programs and keeping track of progress towards eradicating poverty (Hersh et al., 2020). Poverty is calculated by comparing spending, as an approach of income data, to the poverty line that describes an individual's minimum basic needs. It requires accurate expenditure data to estimate poverty (Hill, 2021). By expenditure, the poor population can be divided into poor and very poor populations based on 0,8 of the poverty line. A very poor population that has expenditures less than 0,8 of the poverty line will be a priority to get government assistance.

The limitation of the currently existing official poverty and expenditure data is the level of estimations. Small area estimations are a popular method as the high demand for reliable statistics in small areas, such as low administrative levels (Kaban et al., 2022).

Small area estimation can generate more reliable estimates by borrowing strength from auxiliary variables without errors, such as administrative data and census data. (Gartina & Khitmah, 2020) (Syafira & Hajarisman, 2022) (Nurizza, 2021) (Hakim & Hajarisman, 2022) (Nirwana et al., 2022) (Maulana & Wulansari, 2021) shows that small area estimation can produce more reliable poverty indices and expenditure data in Indonesia. Small area estimation also provides estimation for non-sampling area (Maulana & Wulansari, 2021) (Wulansari et al., 2022) (Utami & Ubaidillah, 2022).

The success of small-area estimates strongly depends on the selected auxiliary variables (Kaban et al., 2022) (Molina & Rao, 2015). It needs accurate auxiliary variables. However, administrative data, Potential Village (Podes data), mostly consists of infrastructure data. Meanwhile, the population census is held once every 10 years. This data is irrelevant and does not represent the actual condition of poverty.

Besides low cost and frequent updates, big data can capture socio-economic conditions by objective measurement and comprehensive area coverage (Kaban et al., 2022) (Putri et al., 2022) (Pratesi et al., 2013). Big data, including satellite imagery data, reduces the cost of poverty measuring and provides information that is not obtained in traditional surveys (Hersh et al., 2020).

Big data and small area estimation is the best combination in some possible approaches: big data as a covariate in small area models and big data utility to validate small area estimates. A combination of NTL and administrative data is more effective than only administrative data as auxiliary variables of small area expenditure estimates (Kaban et al., 2022). Twitter data is also effective as an auxiliary variable to show small-area happiness index estimates (Aziz & Ubaidillah, 2021). This research aims to implement remote sensing data as auxiliary variables in a small area model to estimate expenditure per capita of very poor households in West Java, Indonesia.

2. Data used

2.1 Very Poor Expenditure

In Indonesia, the per capita consumption expenditure is collected by the national socio-economic survey (SUSENAS) and estimated using direct estimation. The direct estimation method estimates population parameters based solely on sample data obtained from that domain (Rao, 2003). This estimation method approach is design-based. The design-based estimation technique is used in data collection by the sampling design. The sampling design applied by Central Bureau Statistics to the Socio-Economic Survey is multistage sampling which is stated in the sampling scheme as follows:

Table 1. Socio-Economic Survey Sampling Scheme

Stage	units	Number of strata units h		Sampling Method	Sample Selection Opportunities	Sampling Fraction
		Population	Sample			
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	Census Block	N_h	n'_h	PPS-with replacement	$\frac{Z_{hi}}{Z_h}$	$n'_h \frac{Z_{hi}}{Z_h}$
		n'_h	n_h	Systematic	$\frac{1}{n'_h}$	$\frac{n_h}{n'_h}$
2	Household	M_{hi}^{up}	$\bar{m}$	Systematic	$\frac{1}{M_{hi}^{up}}$	$\frac{\bar{m}}{M_{hi}^{up}}$

Source: Socio-Economic Survey (Susenas) Sampling Methodology

District/city sampling fraction :

$$F = f_1 \times f_2 \times f_3 = n'_h \frac{Z_{hi}}{Z_h} \times \frac{n_h}{n'_h} \times \frac{\bar{m}}{M_{hi}^{up}}$$

Sampling fraction :

$$F = F_{kab} \times \frac{n_h^{prop}}{n_h^{kab}}$$

In addition, design-based techniques can be used by applying sampling formulas and weight values to data that the Central Bureau of Statistics has processed. Weight is used as a weight in making estimates. It is intended that the available samples can represent many populations. The steps are taken in preparing the weighing:

1. Build initial weight based on the sampling scheme
   Initial/base weight is the inverse of the sampling fraction, namely:

$$W^{design} = \frac{1}{F}$$

Design weight is built from the updated households and the initial target of the enumerator. Order design If the weight is good, it is necessary to control household updating activities.

2. Non-response adjustment weighted
   The non-response adjustment weight is used to revise the weight value based on actual enumeration at the census block and household level while maintaining the total probability value in the sampling frame.
3. Trimming weight
   Trimming aims to reduce the variation in weight between census blocks while still referring to the total weight control of the estimated total value.
4. Adjustments for household non-coverage
   It aims to control the estimation of the number of households based on projected data.
5. Secondary data control
   Secondary data control uses age and sex groups from population projection data. The age group is very dependent on the distribution of enumeration results.
6. Calibration of projection data
   The total number of projections is calibrated in the final weighting process.

For example, Y, is the variable under study, y_{ij} is the result value (characteristic) of the unit population to $j$which comes from a small area i $(i = 1, \dots, m; j = 1, \dots, N_i)$ and s_i is a sub-sample with n_i which is taken from a small area i, so that $s = s_1 \cup s_2 \cup \dots \cup s_m$and $n = \sum_{i=1}^{m} n_i$. Direct estimation can be calculated using the following formula (Rao and Molina, 2015):

$$\hat{Y}_i = \sum_{j=1}^{n_i} w_j y_{ij}$$

so that the formula for the average direct estimate can be written as follows:

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} w_j y_{ij}}{\sum_{j=1}^{n_i} w_j}$$

The unbiased variance estimator formula, namely:

$$Var\left(\hat{\bar{Y}}_i\right) = (1 - f_i)\frac{S_i^2}{n_i}; f_i = \frac{n_i}{N_i}$$

where $(1 - f_i)$is *the finite population correction factor* and

$$S_i^2 = \frac{\sum_{j=1}^{n_i} w_j}{\left(\sum_{j=1}^{n_i} w_j\right)^2 - \sum_{j=1}^{n_i} w_j{}^2} \sum_{j=1}^{N_i} w_j\left(y_{ij} - \bar{y}_i\right)^2; N_i \geq 2$$

if $N_i$not known, then $f_i$approached with$f = \frac{n}{N}$

Direct estimation requires a sample size that is large enough to meet the sample adequacy requirements to produce reasonably accurate estimates. When the sample size is too small, a direct estimation cannot be carried out because it results in poor accuracy. An estimate with poor accuracy cannot be considered in the policy-making process. In conditions like this, it is necessary to do indirect estimation to overcome this.

Table 1 shows the summary of direct estimates and relative standard error of expenditure of the poor population. Although all relative standard error is below 25%, there are four districts without samples, so there are no estimates: Bekasi, Depok, Bogor, and Banjar. Small area estimates are needed to get estimates in these non-sample districts.

Table 1. Summary of direct estimates of expenditure of the poor population in West Java, 2019

	Direct estimates	Relative Standard Error
Min	230,728	0.000
Median	284,261	1,624
Mean	281,781	2,086
Max	299,718	8,772
NA	4	4

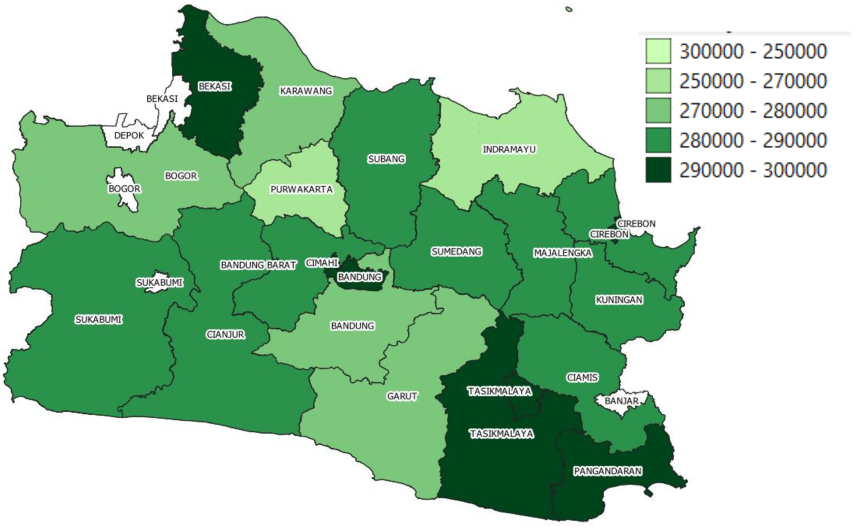

Figure 1. Direct estimates of expenditure of the poor population in West Java, 2019

2.2 Village Potential (*Podes Data*)

Podes collects a variety of information, both about the potential of the Village and data related to the vulnerabilities or challenges the Village faces. Podes provide information about employment, education, health, socio-culture, sports and entertainment, transportation, communication and information, economy, security, development and empowerment of village communities. Information related to vulnerabilities or challenges includes natural disasters, environmental pollution, social and health problems in the community, and security disturbances in villages. Podes are an essential instrument highly expected by the Central Government because the data generated from Podes can describe the current portrait of conditions in Villages, Districts, Regencies,/Cities throughout Indonesia. Apart from aiming to provide data on the Village's existence and development potential, including social, economic, and regional facilities and infrastructure, one of Podes objectives is to provide basic data for compiling statistics for the smallest area (Small Area Statistics). This study uses several variables from Village Potential, which are described in Table 2.

Table 2. Description and reference of auxiliary variables from village potential

Variables Description	Reference
Number of education facilities	(Adhitya et al., 2022)
Number of health facilities	(Poverty and Health, 2014)
Number of telecommunication facilities	(Silva & Zaenudeen, 2007)
Number of economic facilities	(Sugiharti & Primanthi, 2017)
Number of diseases	(Roberts, 2018)

2.3 Remote sensing data and zonal statistics

This study uses several satellite imagery variables, described in Table 3.

Table 3. Source of data and reference of auxiliary variables from satellite imagery data

Variables	Source of Data	Reference
Carbon Monoxide (CO)	Sentinel-5P NRTI CO: Near Real-Time Carbon Monoxide	(Putri et al., 2022)
Night Time Light (NTL)	V.2 VIIRS Nighttime Lights	(Putri et al., 2022)
Land Surface Temperature (LST)	MODIS Land Surface Temperature and Emissivity (MOD11)	(Putri et al., 2022)
Normalized Difference Vegetation Index (NDVI)	Sentinel-2 MSI: MultiSpectral Instrument, Level-2A	(Putri et al., 2022)
Normalized Difference Water Index (NDWI)	Sentinel-2 MSI: MultiSpectral Instrument, Level-2A	(Putri et al., 2022)
Normalized Difference Built-Up Index (NDBI)	Sentinel-2 MSI: MultiSpectral Instrument, Level-2A	(Putri et al., 2022)

Figure 2 shows districts with high expenditure estimates of poor people, Bekasi, Bandung, Tasikmalaya, and Pangandaran, have high CO, NTL, LST, and NDWI. Visually, CO, NTL, LST, and NDBI correlate positively to estimates of expenditure of poor people, while NDVI and NDWI have a negative correlation. The numerical value of satellite imagery of each district is calculated using zonal statistics using QGIS Application (*Zonal Statistics Plugin*, n.d.) (Njambi, 2022).

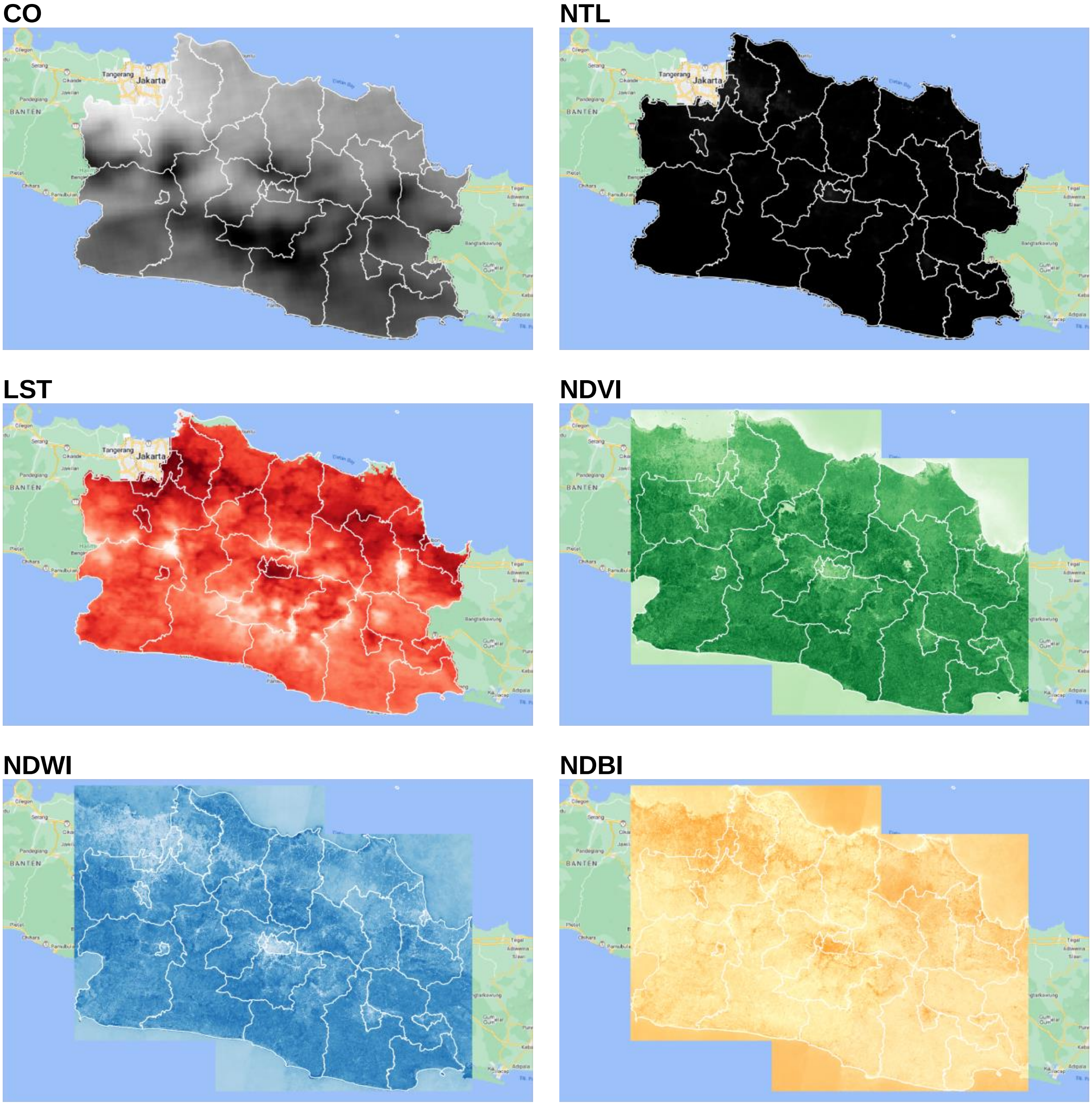

Figure 2. Visualization of the obtained data from NTL (nanowatts/cm2/sr), NDVI (index), BUI (index), NDWI (index), LST (Kelvin), CO (mol/m2), NO2 (mol/m2), and SO2 (mol/m2).

3. Method

3.1 Indirect estimation use Small Area Estimation (SAE)

Indirect estimation is an estimation process that utilizes information from other areas to estimate the parameter value of a particular area. This estimation is done because direct estimation often owns the sizeable standard error caused by the small sample size. The small area estimation (SAE) method is used to overcome this. According to Rao & Molina (2015), Small Area Estimation (SAE) is a statistical method for estimating subpopulation parameters with a small size. In SAE, there are two main problems. The problem is to use a small sample size in a small area to produce good parameter estimation. The second problem is estimating mean square error (MSE). The solution to the problem is by "borrowing information" from within the area, outside the area, and outside surveys (Pfeffermann, 2007).

There are two concepts mainly used for developing a small area parameter estimation model, namely :

1. Fixed Effect Models
   Variation of variable response within a small area can be fully described by relationships suitable variation from information addition.
2. Random Effect small area
   Diversity in a small specific area can't be explained by information addition.

A combination of both models forms a mixed model. The small area estimator differs from General Linear Mixed Model (GLMM).

Rao & Molina (2015) state that based on data availability, the small area model is divided into two basic models: the basic area and unit level. Basic area-level models are based on the availability information variable only on the area level. In comparison, basic unit-level models are based on the availability of information variable individually (variable information accompaniment available at the unit level). In this research, the basic model to focus on is the basic area-level model.

The basic model of the connecting area level estimator direct with supporting data from other domains in a small area with covariate from the area concerned that is $\boldsymbol{x}_i^T = (x_{1t}, \dots, x_{pi})$. Parameter small area to be estimated that is θ_i. A linear model that can explain the connection is (Rao & Molina, 2015):

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_i u_i, i = 1,2, \dots, m$$

Where:

$\boldsymbol{\beta}$:$(\beta_0, \beta_1, \dots, \beta_p)^{\boldsymbol{T}}$ is vector coefficient regression sized $(p+1) \times 1$

b_i : a known positive constant

u_i : *random effect* from a *small area,* assumed $u_i \sim N(0, \sigma_u^2)$

m : number of observations / small area

An exciting conclusion about population got assumed. For mark prediction, direct $\hat{\theta}_i$known and can be written as follows:

$$\hat{\theta}_i = \theta_i + e_i, i = 1,2, \dots, m$$

Where $e_i$is the assumed sampling error$e_i \sim N(0, \psi_i)$

If both models are put together, so will form *Fay-Heriot Model* with equality as follows:

$$\hat{\theta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + b_i u_i + e_i, i = 1,2, \dots, m$$

3.2 Fay Herriot Model

The Fay-Heriot model is used in application small area estimation. Suppose the Fay-Heriot model is as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ is the vector from observation sized $n \times 1$,$\boldsymbol{X}$ And $\boldsymbol{Z}$ is matrix sized $m \times p, u$is a vector of *the random effects area,* and $e$is a sampling *error which* is :

$$\boldsymbol{u}\ iid\ (0, \boldsymbol{G}), \qquad \boldsymbol{e}\ iid\ (0, \boldsymbol{R})$$

Where $\boldsymbol{G} = \boldsymbol{I}_m\sigma_u^2$and $\boldsymbol{R} = \boldsymbol{I}_m\sigma_e^2$, where $\boldsymbol{I}_m$ It is a matrix identity. Matrix covariance from $\boldsymbol{y}$ stated with $\boldsymbol{\Omega} = \mathbf{ZGZ}^{\mathrm{T}} + \boldsymbol{R}$

The most common approach to getting those parameters is the BLUP proposed by Henderson (1953) or EBLUP.

3.3 Best Linear Unbiased Predictor (EBLUP)

The Best Linear Unbiased Predictor (BLUP) proposed by Henderson (1953) was used For get estimate *random effects* on *linear mixed models,* where

$$\boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

example $\hat{\boldsymbol{\mu}}$ is the estimate of $\boldsymbol{\mu}$ *unbiased* Where $E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu}$ and *the Mean Square Error* (MSE) of $\hat{\boldsymbol{\mu}}$ is

$$MSE(\hat{\boldsymbol{\mu}}) = E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^2$$

With minimizing $MSE(\hat{\boldsymbol{\mu}})$ then the BLUP estimator is obtained as follows:

$$\tilde{\boldsymbol{\mu}}^H = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$$

Where

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\boldsymbol{\delta}) = (\boldsymbol{X^T\Omega^{-1}X})^{-1}\boldsymbol{X^T\Omega^{-1}y}$$

Which is the *Best Linear Unbiased Estimator* (BLUE) from $\boldsymbol{\beta}$**,** and

$$\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\boldsymbol{\delta}) = \boldsymbol{GZ^T\Omega^{-1}}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$$

is *the random effect area* estimator for each area with component variance $\delta = (\sigma_u^2, \sigma_e^2)$. Then the BLUP model is expressed by:

$$\tilde{\boldsymbol{\mu}}^H = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{ZGZ^T\Omega^{-1}}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$$

With matrix covariance, $\boldsymbol{G}(\boldsymbol{\delta}) = \boldsymbol{I}_m\sigma_u^2$*the sampling error* $\boldsymbol{R} = \boldsymbol{I}_m\sigma_e^2$matrix and matrix covariance variable $\boldsymbol{y}$ that is $\boldsymbol{\Omega} = \mathbf{ZGZ}^{\mathrm{T}}\boldsymbol{R}$

Equality also got _ written down as follows:

$$\begin{aligned}\tilde{\mu}_i^H(\sigma_u^2) &= \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} + \frac{\sigma_u^2 b_i^2}{\sigma_u^2 b_i^2 + \sigma_e^2}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) \\ &= \gamma_i y_i + (1 - \gamma_i)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}, \qquad i = 1, \dots, m\end{aligned}$$

Where

$$\gamma_i = \frac{\sigma_u^2 b_i^2}{\sigma_u^2 b_i^2 + \sigma_e^2}$$

Where $b_i$is a positive constant and $\widetilde{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$**,** where

$$\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\sigma_u^2) = \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{\boldsymbol{x_i x_i^T}}{\sigma_u^2 b_i^2 + \sigma_e^2}\right]^{\boldsymbol{-1}} \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{\boldsymbol{x_i} y_i}{\sigma_u^2 b_i^2 + \sigma_e^2}\right]$$

3.4 MSE from BLUP

The MSE of the BLUP estimator is as follows:

$$MSE(\tilde{\mu}_i^H) = E(\tilde{\mu}_i^H - \mu_i)^2$$

Example $\hat{\mu}$ is the estimate of $\mu$the *unbiased* Where $E(\hat{\mu}) = \mu$and MSE of $\hat{\mu}$is

$$MSE(\tilde{\mu}_i^H) = E(\tilde{\mu}_i^H - \mu)^2 = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2), \qquad i = 1, .., m$$

Where

$$g_{1i}(\sigma_u^2) = \frac{\sigma_u^2 b_i^2 \sigma_e^2}{\sigma_u^2 b_i^2 + \sigma_e^2} = \gamma_i \sigma_e^2$$

And

$$g_{2i}(\sigma_u^2) = (1 - \gamma_i)^2 \boldsymbol{x_i^T} \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{\boldsymbol{x_i x_i^T}}{\sigma_u^2 b_i^2 + \sigma_e^2}\right]^{\boldsymbol{-1}} \boldsymbol{x_i}$$

Where

$$\gamma_i = \frac{\sigma_u^2 b_i^2}{\sigma_u^2 b_i^2 + \sigma_e^2}$$

The BLUP estimator is highly component-dependent variance σ_u^2 assumed to be known. In practice, σ_u^2 this is not known and must be estimated. The estimation method σ_u^2that can be used is *the Restricted Maximum Maximum Likelihood* (REML). Method Fixed *Restricted Maximum Likelihood* (REML). *Unbiased,* though the sample is small compared to this method's *Maximum Likelihood* (ML) (Rao & Molina, 2015).

3.5 Empirical Best Linear Unbiased Predictor (EBLUP)

BLUP is a very component-dependent variance σ_u^2which must be known. If it is not known, then an estimation of the variance component is carried out, σ_u^2one of which is the *Restricted Maximum Likelihood* (REML) method. Such models are called *Empirical Best Linear Unbiased Predictor* (EBLUP), as follows:

$$\hat{\mu}_i^H(\hat{\sigma}_u^2) = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} + \frac{\hat{\sigma}_u^2 b_i^2}{\hat{\sigma}_u^2 b_i^2 + \sigma_e^2}(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})$$
$$= \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)\boldsymbol{x_i^T} \widehat{\boldsymbol{\beta}}, \qquad i = 1, \dots, m$$

Where

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2 b_i^2}{\hat{\sigma}_u^2 b_i^2 + \sigma_e^2}$$

Where $b_i$is a positive constant and $\widetilde{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$**,** where

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\hat{\sigma}_u^2) = \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{\boldsymbol{x_i x_i^T}}{\hat{\sigma}_u^2 b_i^2 + \sigma_e^2}\right]^{\boldsymbol{-1}} \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{\boldsymbol{x_i} y_i}{\hat{\sigma}_u^2 b_i^2 + \sigma_e^2}\right]$$

3.6 MSE from EBLUP

The MSE of the EBLUP estimator is as follows:

$$MSE(\hat{\mu}_i^H) = E(\hat{\mu}_i^H - \mu_i)^2 = E(\hat{\mu}_i^H - \tilde{\mu}_i^H + \tilde{\mu}_i^H - \mu_i)^2$$
$$= E(\hat{\mu}_i^H - \tilde{\mu}_i^H)^2 + E(\tilde{\mu}_i^H - \mu_i)^2 + 2E(\hat{\mu}_i^H - \tilde{\mu}_i^H)(\tilde{\mu}_i^H - \mu_i)$$

Because

$$E(\hat{\mu}_i^H - \tilde{\mu}_i^H)(\tilde{\mu}_i^H - \mu_i) = 0$$

And

$$E(\tilde{\mu}_i^H - \mu_i)^2 = MSE(\tilde{\mu}_i^H)$$

so

$$MSE(\hat{\mu}_i^H) = MSE(\tilde{\mu}_i^H) + E(\hat{\mu}_i^H - \tilde{\mu}_i^H)^2$$

component $E(\hat{\mu}_i^H - \tilde{\mu}_i^H)^2$referred to as $g_{3i}(\delta)$(Rao and Molina, 2015), where

$$g_{3i}(\delta) = tr\left[\left(\frac{\partial b_i^T}{\partial \delta}\right)\Omega_i\left(\frac{\partial b_i^T}{\partial \delta}\right)\overline{\Omega}(\hat{\delta})\right]$$

Then MSE from EBLUP can be written as follows (Prasad and Rao, 1990):

$$MSE(\hat{\mu}_i^H) \approx g_{1i}(\delta) + g_{2i}(\delta) + g_{3i}(\delta)$$

Because it's estimated to use REML, then can write down become

$$MSE(\hat{\mu}_i^H) = g_{1i}(\hat{\delta}) + g_{2i}(\hat{\delta}) + g_{3i}(\hat{\delta})$$

3.5 SAE with information cluster

Research conducted by Annisa (2014) modified the EBLUP equation for non-sampled areas by including cluster information in the equation $\hat{\theta}_i^{EBLUP} = \boldsymbol{x_i^T}\widetilde{\boldsymbol{\beta}}$ which aims to produce better estimates of non-sampled areas. Haris (2019) developed the SAE clustering function to estimate non-sample areas. In the process, clustering is based on the accompanying variable X which is carried out in all research areas, both sampled and non-sampled. Furthermore, all research areas will be obtained into specific clusters. This can be used to improve the estimate of the unsampled area. Cluster information is obtained by including the random effect of the area in the EBLUP equation. The random effect of the area will be searched for in each cluster which functions to obtain the value of the random effect of the area in a particular area. The random effect of the area obtained is then averaged, the results of the average are entered into the model as an estimator of the random effect of the area for non-sampled areas. The following is the formula for adding the average random effect of the area:

$$\bar{\hat{v}}k = \frac{1}{n_k}\sum_{i=1}^{n_k} \hat{v}_i$$

where $\bar{\hat{v}}k$ is the average random effect of the area in the k-cluster, nk is the number of sample areas in the k-cluster, and $\hat{v}i$ is the random effect of the i-th sample area. The above equation is then substituted into the EBLUP equation $\hat{\theta}_i^{EBLUP} = \boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}}$ then an estimation equation will be obtained using the EBLUP method with cluster information. The following formula is formed:

$$\hat{\theta}_i^{EBLUP} = \boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}} + \bar{\hat{v}}k$$

where $\boldsymbol{x}_i^T$ is the matrix of the accompanying variables, $\boldsymbol{\beta}$ is the regression coefficient vector of size $(p + 1) \times 1$, $\bar{\hat{v}}k$ the average random effect of the area in the k-cluster.

4. Result

Auxiliary variables used in the model were highly correlated with target variables, selected using stepwise regression, and had significant p-value. EBLUP Model using administrative data and satellite imagery data are shown in Table 4 and Table 5. All of the coefficient's p-values are below the significance level, alpha=5%. It shows that the number of health workers, the number of wired telephones, the number of markets, the number of restaurants, and the NTL index significantly affect the per capita expenditure of poor people.

Table 4. Parameter estimates of EBLUP model using administrative data

Variables	Description	coefficients	Standard error	t-value	p-value
	(Intercept)	1.260367e+01	1.143228e-02	1102.463725	0.000000e+00
X1	Number of health workers	-3.278722e-05	9.963765e-06	-3.290645	9.995795e-04
X2	Number of wired telephone	1.352227e-06	2.942691e-07	4.595205	4.323235e-06
X3	Number of market	3.967958e-05	1.847121e-05	2.148185	3.169908e-02
X4	Number of restaurant	-1.014014e-04	2.971599e-05	-3.412351	6.440509e-04

Table 5. Parameter estimates of EBLUP model using remote sensing data

Variables	Description	coefficients	Standard error	t-value	p-value
	(Intercept)	12.550505779	0.008142072	1541.438771	0.00000000

X1	NTL_mean	0.003378719	0.001293983	2.611101	0.00902513

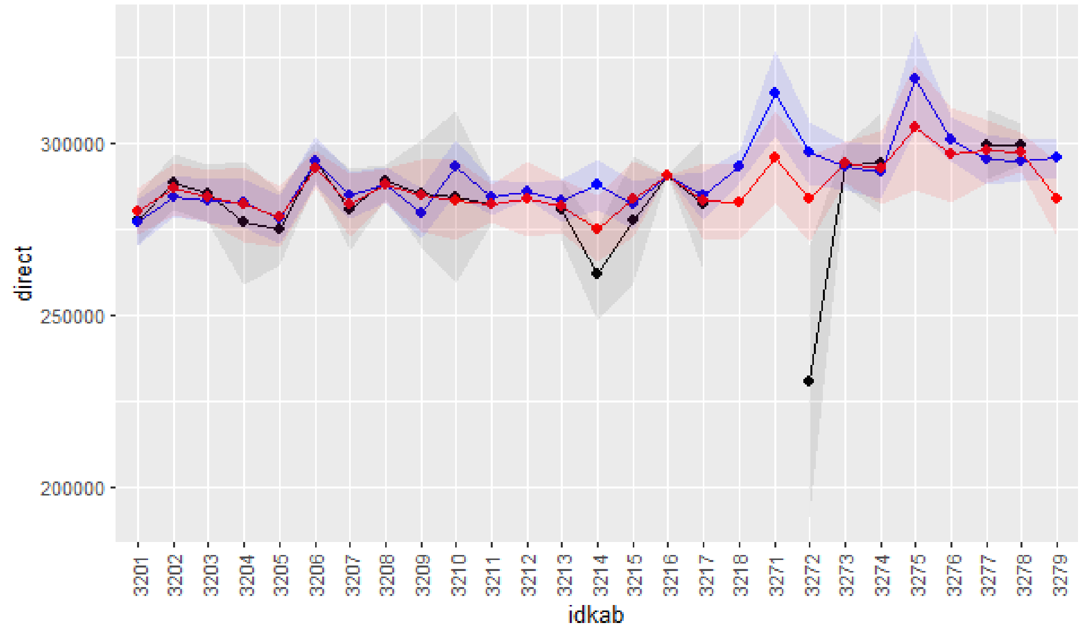

Figure 3. Estimation of direct estimates, EBLUP using administrative data, and EBLUP using satellite imagery data

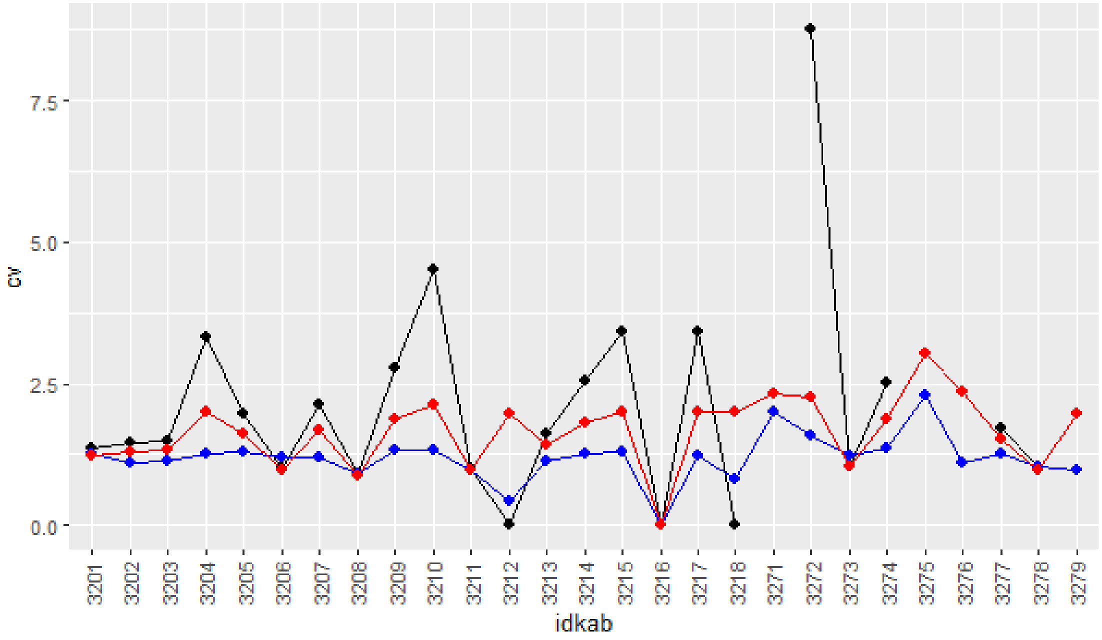

Figure 4. RSE of direct estimates, EBLUP using administrative data, and EBLUP using satellite imagery data

The direct estimates, EBLUP using administrative data and EBLUP using satellite imagery data, along with their confidence interval, are plotted in Figure 3. It shows that direct estimates, small area estimates of the model using administrative data, and small area estimates of models using remote sensing data have similar patterns. The estimates using EBLUP are more stable than direct estimates; there is no outlier such as Sukabumy city. Observing the confidence interval shows that the confidence interval of EBLUP is shorter than direct estimates, but still overlaps. It shows that both SAE models use administrative and satellite imagery data to be more reliable than direct estimates. Both small area models also produced estimations for the unsampled area, which are Bekasi, Depok, Bogor, and Banjar.

Figure 4 plots the RSE of direct estimates, EBLUP using administrative data, and EBLUP using remote sensing data. It shows that the RMSEs of the EBLUP estimators in both models are smaller than the RMSEs of the direct estimates for almost all areas. Small area estimates using EBLUP can decrease the RSE below 5% in all areas. Compared to EBLUP using satellite imagery, EBLUP using administrative data is slightly lower.

The average expenditure of a model using satellite imagery is 283.965 rupiahs, closer to direct estimates than the average of a model using administrative data, 288.015 rupiahs. Minimum and maximum value of EBLUP is slightly higher than direct estimates, and the EBLUP using satellite imagery is closer than using administrative data.

The average and maximum of relative standard error of both EBLUP models are below the RSE of direct estimates. The EBLUP using administrative data decreased the maximum of RSE to 2.289 percent, slightly under the RSE of EBLUP using satellite imagery data in 3.043. Although the relative standard error is still slightly higher than the model using administrative data, remote sensing data is preferable because of the lower cost and more comprehensive coverage.

Table 6. Summary of expenditure estimates and their relative standard error

	Estimates			Relative Standard Error		
	Direct	EBLUP1	EBLUP2	Direct	EBLUP1	EBLUP2
Min	230728	276942	275226	0.000	0.000	0.000
Median	284261	288015	283965	1.624	1.231	1.817
Mean	281781	290424	287196	2.086	1.185	1.648
Max	299718	318662	304528	8.772	2.289	3.043
NA	4			4		

5. Conclusion

Bost small area estimation models, using administrative data such as Podes data and remote sensing data, can produce more reliable estimates than direct estimates. They also produce estimates for non-sample area using small area estimates using information clusters. Using podes data, the number of health workers, the number of wired telephones, the number of markets, and the number of restaurants are significant in EBLUP model, meanwhile for remote sensing data only NTL is significant. The relative standard error of the model using Podes is slightly lower than using remote sensing data. Although the relative standard error is still slightly higher than the model using administrative data, remote sensing data is preferable because of the lower cost and more comprehensive coverage.

References

Adhitya, B., Prabawa, A., & Kencana, H. (2022). Analisis Pengaruh Pendidikan, Kesehatan, Sanitasi dan Rata-Rata Jumlah Anggota Keluarga Per Rumah Tangga terhadap Kemiskinan di Indonesia. *Ekonomis: Journal of Economics and Business*, *6*(1). 10.33087/ekonomis.v6i1.501

Annisa R., Kurnia A., Indahwati. (2014). Cluster Information of Non-Sampled Areain Small Area Estimation. IOSR Journal of Mathematics 10(1): 15-19.

Aziz, S. D., & Ubaidillah, A. (2021). Big Data for Small Area Estimation: Happiness Index with Twitter Data. *Proceedings of 2021 International Conference on Data Science and Official Statistics (ICDSOS)*, *2021*(1). 10.34123/icdsos.v2021i1.248

Gartina, D., & Khitmah, L. (2020). Pendugaan Kemiskinan Menggunakan Small area Estimation dengan Pendekatan Emperical Best Linear Unbiased Prediction (EBLUP). *Jurnal Statistika Universitas Muhammadiyah Semarang*, *8*(2). 10.26714/jsunimus.8.2.2020.159-165

Hakim, A. H., & Hajarisman, N. (2022). Pendugaan Rata-rata Pengeluaran Per Kapita Menurut Kabupaten/Kota di Provinsi Jawa Barat Melalui Empirical Best Linear Unbiased Prediction dalam Pendugaan Area Kecil. *Bandung Conference Series: Statistics*, *2*(2). 10.29313/bcss.v2i2.4747

Haris, Faisal. (2019). Kajian MSE Area Tidak Tersampel pada Small Area Estimation. [Skripsi]. Jakarta: Politeknik Statistika STIS.

Hersh, J., Engstorm, R., Mann, M., Martin, L., & Mejia, A. (2020). *Mapping Income Poverty in Belize Using Satellite Features and Machine Learning*. Inter-American Development Bank. 10.18235/0002345

Hill, H. (2021). What's Happened to Poverty and Inequality in Indonesia over Half a Century? *Asian Development Review*, *38*(1), 68-97. https://doi.org/10.1162/adev_a_00158

Kaban, P. A., Nasution, B. I., Caraka, R. E., & Kurniawan, R. (2022). Implementing night light data as auxiliary variable of small area estimation. *Communications in Statistics - Theory and Methods*. 10.1080/03610926.2022.2077963

Maulana, M. w., & Wulansari, I. Y. (2021). Implementasi Empirical Best Linear Unbiased Prediction Fay-Herriot dalam Menduga Rata-Rata Pengeluaran per Kapita Level Kecamatan di Provinsi Jawa Timur dengan Tambahan Informasi Cluster. *Seminar Nasional Official Statistics 2021*, *2021*(1). 10.34123/semnasoffstat.v2021i1.1051

Molina, I., & Rao, J. N. K. (2015). *Small Area Estimation*. Wiley.

Nirwana, M., Sunengsih, N., & Hendrawati, T. (2022). Small Area Estimation Untuk Pengeluran Per Kapita Kabupaten Pesisir Barat Dengan Metode Empirical Best Linear Unbiased Predictor. *E-Journal BIAStatistics*, *16*(2), 70-83. 10.1234/bias.v16i2.164

Njambi, R. (2022). An introduction to zonal statistics. Retrieved 2023, from https://up42.com/blog/an-introduction-to-zonal-statistics

Nurizza, W. A. (2021). Penerapan Model Fay-Herriot pada Small Area Estimation Studi Simulasi Pengeluaran Per Kapita Level Kabupaten/Kota Provinsi Kalimantan Timur Tahun 2020. *Buletin Statistika dan Aplikasi Terkini*, *1*(1).

Pfeffermann, D. (2007). Small Area Estimation-New Developments and Directions. *International Statistical Review/ Volume 70*, *Issue 1*, 125-143.

Poverty and Health. (2014, August 25). World Bank Group. Retrieved February 27, 2023, from https://www.worldbank.org/en/topic/health/brief/poverty-health

Pratesi, M., Pedreschi, D., Giannotti, F., Marchetti, S., Salvati, N., & Maggino, F. (2013). Small area model-based estimators using big data sources. European Commission.

Putri, S. R., Wijayanto, A. W., & Sakti, A. D. (2022). Developing Relative Spatial Poverty Index Using Integrated Remote Sensing and Geospatial Big Data Approach: A Case Study of East Java, Indonesia. International Journal of Geo-Information, 11(5), 275. 10.3390/ijgi11050275

Roberts, S. (2018, January 10). Key Facts: Poverty and Poor Health. Health Poverty Action. Retrieved February 27, 2023, from https://www.healthpovertyaction.org/news-events/key-facts-poverty-and-poor-health/

Silva, H. d., & Zaenudeen, A. (2007). Poverty reduction through telecom access at the 'Bottom of the Pyramid. Centre for Poverty Analysis Annual Symposium on Poverty Research in Sri Lanka.

Sugiharti, L., & Primanthi, M. R. (2017). The Determinants of Poverty: Case of Indonesia. *Global Journal of Business5 and Social Science Review (GJBSSR)*, *5*(3), 58-68.

Syafira, A., & Hajarisman, N. (2022). Penerapan Penduga Area Kecil Melalui Metode Empirical Best Linear Unbiased Prediction (EBLUP) untuk Estimasi Persentase Penduduk Miskin Level Kabupaten/Kota di Provinsi Jawa Barat. *Bandung Conference Series: Statistics*, *2*(2). 10.29313/bcss.v2i2.4507

Utami, M. S., & Ubaidillah, A. (2022). Pendugaan Persentase Rumah Tangga yang Memiliki Akses Terhadap Air Minum Layak, Sanitasi Layak, serta Rumah Layak Huni dan Terjangkau pada Level Kecamatan Di Provinsi Papua Tahun 2019 Menggunakan Model Fay Herriot Multivariat. *Seminar Nasional Official Statistics 202*, *2022*(1). 10.34123/semnasoffstat.v2022i1.1498

Wulansari, J., Permatasari, N., & Ubaidillah, A. (2022). Pendugaan Area Kecil Persentase Anak-anak Usia Kurang dari 18 Tahun yang Hidup di Bawah Garis Kemiskinan Tingkat Kabupaten/Kota di Indonesia Tahun 2020. *Seminal Nasional Official Statistics 2022*, *2022*(1). 10.34123/semnasoffstat.v2022i1.1467

Zonal Statistics Plugin. (n.d.). QGIS Documentation. Retrieved 2023, from https://docs.qgis.org/2.18/en/docs/user_manual/plugins/plugins_zonal_statistics.html