Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2022, 2022:83
https://doi.org/10.28919/cmbn/7577
ISSN: 2052-2541

SMALL AREA ESTIMATION FOR AUTOREGRESSIVE MODEL WITH MEASUREMENT ERROR IN THE AUXILIARY VARIABLE

ERWIN TANUR[1,2], ANANG KURNIA[1,*]

[1]Department of Statistics, IPB University, Bogor 16680, Indonesia

[2]Education and Training Centre, Statistics Indonesia, Jakarta 12620, Indonesia

Abstract: Small Area Estimation is a good method for estimating parameters with a limited number of samples or none at all. The method's development is continuously carried out in line with the development of types of data encountered in research. One of developments is in estimating parameters for the case of panel data with auxiliary variables containing measurement errors. This condition is often encountered in the use of survey data. One of most useful surveys in Indonesia about this issue is *Susenas* or the National Socio-Economic Survey. Since 2015, the *Susenas* has been implemented in two periods a year, that is in March and September. In March, data is collected with a representative sample size for an estimate at up to the district/city level. As for the *Susenas* in September, the data collected is less representative for an estimate at up to the provincial level. The September data collection object is part of the March data collection object, thus some repeated sample units are found in the September and March data. A variable of concern in this study is the average consumption per capita that has an asymmetrical distribution. One approach for this case is the lognormal distribution-based modeling. The use of information that has measurement error as an auxiliary variable in the form of a random variable is deemed capable of producing a better estimate. For

[*]Corresponding author
E-mail address: anangk@apps.ipb.ac.id
Received July 1, 2022

the repeatedly-obtained data, a first-order autoregressive model approach is applied. In this study, a Small Area Estimation method was developed to handle a small sample size under the repeated data condition, as well as the use of information in the March period as an auxiliary variable with measurement errors.

Keywords: first-order autoregressive; small area estimation; Susenas; measurement error.

2010 AMS Subject Classification: 97K80, 91G70.

1. INTRODUCTION

The general survey design challenge is to draw conclusions at a specified level of precision at minimal cost, or to achieve the best precision at a given cost. In addition, a lower response rate to the survey may increase the risk of non-response bias. The influential factor is a proper sample selection, which provides an accurate response in a survey or research. When the response to a survey differs from the actual expected response, a measurement error occurs. Besides, a measurement bias occurs when the response tends to differ from the true value in one direction. The measurement error and bias should be considered and minimized at the survey design stage [14].

The assumption in the Small Area Estimation model is that auxiliary variables are measured without error [9]. Therefore, the commonly-used auxiliary information is data from the census and administrative or registry data. The development of a Small Area Estimation model with measurement errors occurs when the auxiliary variable in the Small Area Estimation is assumed to contain no errors as it comes from the census or administrative data. However, in reality, the census and administrative data are often times completely unavailable and up-to-date to serve as auxiliary information. Accordingly, the solution is to use survey data as auxiliary information in the Small Area Estimation model despite a consequence that the use of survey data may contain sample errors. Modeling the Small Area Estimation with a measurement error is necessary to minimize a bias in the model.

In order to meet the government's needs of data, especially in shorter time intervals (from once to twice a year or more), since 2011, the Statistics Indonesia has made changes in the implementation of *Susenas* or the National Socio-Economic Survey. Important changes in the 2011 *Susenas* implementation, which were continued until 2014, include: (1) data collection is carried out four times a year, from the previous twice a year; (2) consumption data is collected in all the periods of data collection (to produce representative poverty rate figures at up to district/city level), from the previous once a year, except in the year the consumption/expenditure module changes, to twice that year to produce provincial and national poverty rates [18].

In general, direct estimation is a classic approach to estimate small-area parameters based on the application of a design-based model [7]. In relation to the 2015 *Susenas* data collection, the small sample size in September became a consideration of the Small Area approach in making estimates. The direct estimation method on a small area does not have sufficient accuracy due to the small size of samples taken to obtain the estimates. Direct estimation of a small area is an unbiased estimator but it has a large variance as it is obtained from small sample size [7].

Fuller (1987) stated that measurement error affects the slope of the regression curve and the presence of measurement error weakens the regression coefficient [19]. The measurement error causes bias in parameter estimation in statistical models, the detection of relationships between variables is lost, and analysis of graphical models becomes difficult due to the fact that the data depiction is unreal [15].

Sample data from a survey can be used to obtain direct estimates for a large area. For example, estimates of the average monthly household income in a sub-district are based solely on survey data available or obtained from that sub-district. In a research related to income and personal wealth, an asymmetrical observation response, which is skewing to the right, is often found. Karlberg (2000) estimated the total population of survey variables with very long skewing from small sample size [4]. Doing so by applying the direct method would be problematic for two reasons: (i) when there are no extreme values in the sample, the prediction is too small, and (ii) if there are extreme values in the example, the prediction becomes very large. The traditional method of dealing with outliers usually compensates for outliers in the example, so the reason point (ii) is avoided, while a small negative bias in point (i) remains. Thorburn (1991) suggested to use the lognormal superpopulation model to estimate in such cases [1].

Muchlisoh et al. (2017) [17], developed a Small Area Estimation model with a first-order autoregressive random effect which was the development of a model by Rao-Yu (1994) [8]. In estimation using the model by Rao-Yu (1994) [8], the variance of the sampling error is assumed to be known, but in practice, it is frequently unknown. Datta et al. (2002) made an estimate by applying the Jackknife method and several smoothing techniques [6]. Esteban et al. (2012) implemented the Generalized Variance Function (GVF) approach [10].

Battese et al. (1988) [5] and Prasad and Rao (1990) [13] used a unit-level nested error linear regression model in which the covariate was not subject to measurement error. When the auxiliary information is available at the individual level, the underlying model is called a nested error regression model. Ghosh et al. (2006) proposed a nested error linear regression population model with an area-level covariate, x, which is a subject with a measurement error [11].

Tanur et al. (2018) [2] applied the 2015 *Susenas* data to the model in Torabi et al. (2009) [12]. This study applied Small Area Estimation to improve the accuracy of the estimated average consumption per capita from the September data collection by using information from the March period. In this study, the model approach by Ghosh et al. (2006) [11] was compared with the model by Torabi et al. (2009) [12] to estimate the average consumption per capita in West Java Province in 2015.

Tanur et al. (2021) conducted a study that developed a method of estimating small areas in populations that were asymmetrical with the auxiliary variables from the survey results [3]. Developing the Small Area Estimation method is important to improve the effectiveness of the small sample size for the September period, so that the estimator value for per-capita expenditure or consumption is obtained by using information available in the March survey.

The remaining part of the paper is organized as follows: Section 2 discusses the research objectives; Section 3 describes the basic model in detail; Section 4 describes the development model in detail; Section 5 describes the results of the model evaluation; Section 6 discusses the results of the application of the developed model; and Section 7 concludes the paper.

The 2015 *Susenas* data collection case shows that the September data collection is part of samples that become the object of the March data collection (partially repeated). In this case, results of the March and September data collection of *Susenas* are interrelated panel surveys. Information from the March data collection can also be used as an auxiliary variable that has measurement errors in making estimates. It is necessary to develop an alternative for small area estimation for partially repeating data (autoregressive model) with the auxiliary variables for the March data collection as variables containing measurement errors. More details can be seen in Figure 1.

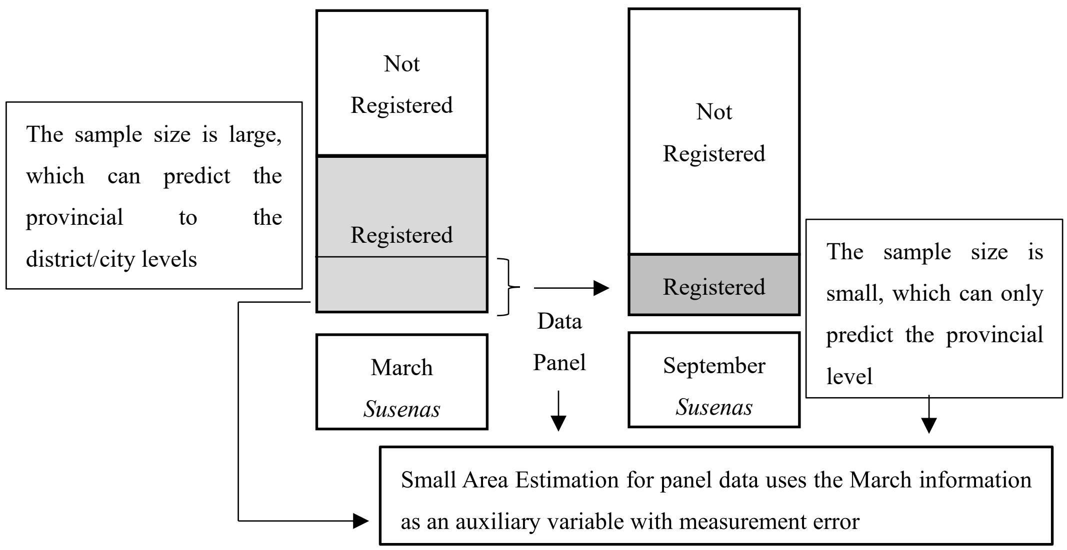

Figure 1. Research Problem Formulation

From the background and problems presented, the research question is on how to obtain a small-area estimation that uses the September data collection by considering the effect of time element, with March as the first period (t1) and September as the second period (t2).

2. RESEARCH OBJECTIVES

This study aimed to obtain a Small Area Estimation model by considering the condition of partial panel data at the unit level and the presence of non-repeating sample elements as auxiliary variables, as well as the use of information with measurement errors because it was the result of a survey. The model development was carried out under a condition in which some sub-districts (as the unit level in the study) were recorded twice. This was identified as a panel case or AR(1) by considering information on other sub-districts recorded in March only. Information from the March data collection would be used as an auxiliary variable for the area level that had measurement errors in the two periods.

The September data of *Susenas* could be identified as some examples that become the object of the March data collection (partially repeated). There were units (sub-districts) recorded twice and some were only recorded once in the two periods of the 2015 *Susenas* data collection. In addition, there was information obtained from the March *Susenas* data collection for the district/city level that could be used as additional information. However, because the information from the March *Susenas* is the result of survey data, it is assumed that there is a measurement error in its use. It is necessary to develop an alternative for Small Area Estimation for partially repeating data (panel) by using information on auxiliary variables containing measurement errors. This study applied a lognormal approach in estimating the average consumption per capita for the first-order autoregressive model by considering the occurrence of measurement errors in the auxiliary variables. Due to the limited information from the September survey results, making Small Area Estimation is the right approach to this problem.

3. BASIC MODEL

The basic model in this estimation refers to the model proposed by Muchlisoh (2017), hereinafter referred to as the SAE-AR1 model [17].

$$y_{itj} = \mathrm{x}'_{itj}\beta + v_i + u_{it} + \mathrm{e}_{itj}, \tag{1}$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1,$$

The β component is the coefficient of the auxiliary variable. The notation i is a small area index which is defined to move from 1 to m (the number of small areas). The notation t is a time index that moves from 1 to T. The notation j is the sample unit in the i-th small area at the t-th time, which moves from 1 to n_{it} (the number of sample units in each small area). The component v_i

is the random effect of area i which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_v^2 $\left(v_i \sim N(0, \sigma_v^2)\right)$. The u_{it} component is a time-area random effect which is assumed to follow a first-order autoregressive process in every small area i. The component ε_{it} is the error of u_{it} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_ε^2 $\left(\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)\right)$. The e_{itj} component is the error of y_{itj} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_e^2 $\left(\mathrm{e}_{itj} \sim N(0, \sigma_\mathrm{e}^2)\right)$. The ρ component is an autoregressive coefficient with an absolute value of less than 1. The random effects of v_i, u_{it} and e_{itj} are assumed to be independent.

The next basic model is a Small Area Estimation model with measurement errors on the auxiliary variables proposed by Tanur et al. (2021) [3].

$$\mathrm{y}_{ij}^* = \mathrm{X}_{ij}' \beta + \mathrm{w}_i \alpha + v_i + \mathrm{e}_{ij}, \tag{2}$$

with: $\mathrm{W}_i = \mathrm{w}_i + \eta_i$.

where $\mathrm{y}_{ij}^* = log\left(\mathrm{y}_{ij}\right)$, y_{ij} is the value of the research response variable for the j-th unit in the i-th area. The variable X_{ij} is a auxiliary variable at the unit level in the i-th area (fixed effect). The variable w_i is an unknown true area-specific covariate with respect to y_{ij}, where w_i follows a normal distribution with mean μ_w and variance σ_w^2. The variable W_i is the auxiliary variable with measurement error. The component η_i is the measurement error in the auxiliary variable, with η_i assumed to follow a normal distribution with a mean of 0 and a variance of σ_η^2. The component v_i is an area random effect, where v_i is assumed to follow a normal distribution with a mean of 0 and a variance of σ_v^2. The e_{ij} component is the model error assumed to follow a normal distribution with a mean of 0 and a variance of σ_e^2. The components v_i, e_ij and w_i are assumed to be independent. The coefficient of fixed effect is denoted by β, and the coefficient of the auxiliary variable which has measurement error is denoted by α.

To answer the research problems in this discussion, a combination of the two equations is carried out. This results in a small area estimation model for the first-order autoregressive random effect with measurement error on the covariates (SAE-AR1-ME).

4. DEVELOPMENT MODEL

The variable of concern in this study is the average consumption per capita, which is first transformed into logarithmic form. The development model proposed in this study is:

$$\mathrm{y}_{itj}^{*} = \mathrm{X}_{itj}'\beta + \mathrm{w}_{it}\alpha + v_i + u_{it} + \mathrm{e}_{itj}, \tag{3}$$

with: $\mathrm{W}_{it} = \mathrm{w}_{it} + \eta_{it}$, $\mathrm{y}_{itj}^{*} = log(y_{itj})$.

The response variable y_{itj} is the j-th sample unit in a small area i at time t which is assumed to have a relationship with a vector of the auxiliary variables X_{itj}, which is assumed to be available for each population unit in a small area i. The β component is the coefficient of the auxiliary variable which is constant. The notation i is a small area index which is defined to move from 1 to m (the number of small areas). The notation t is a time index that moves from 1 to T. The notation j is the sample unit in the i-th small area at the t-th time, which moves from 1 to n_{it} (the number of sample units in each small area). The variable w_{it} is the unknown true area-specific covariate for time t-th, corresponding to $\boldsymbol{y}_{itj}$, with the mean μ_w and variance σ_w^2. While W_{it} is the auxiliary variable with measurement error for the t-th time. The component η_{it} is the measurement error in the auxiliary variable for time t, η_{it} is assumed to follow a normal distribution with an average of 0 and a variance of σ_η^2. The v_i component is the random effect of the i-th area which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_v^2. The u_{it} component is a time-area random effect which is assumed to follow a first-order autoregressive process in every small area i. The component ε_{it} is the error of u_{it} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_ε^2. The e_{itj} component is the error of y_{itj} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ_e^2. The ρ component is an autoregressive coefficient with an absolute value of less than 1. The random effects of w_{it}, v_i, u_{it} and e_{itj} are assumed to be independent.

Equation (3) can be written in matrix form as:

$$\mathbf{y}^* = \mathbf{X\beta} + \mathbf{Z}1\mathbf{w}\alpha + \mathbf{Z}2\boldsymbol{v} + \mathbf{Z}3\boldsymbol{u} + \mathbf{e} \tag{4}$$

where, $\mathbf{y}^* = (\mathbf{y}^*_{i1}, \dots, \mathbf{y}^*_{mT})'$ is a response variable vector of size $\mathrm{nt} \times 1$, with $\mathbf{y}^*_{it} = \left(y^*_{i11}, \dots, y^*_{iTn_{it}}\right)'$ size $\mathrm{n}_i\mathrm{t} \times 1$. $\mathbf{X} = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{mT})'$ matrix of constant variables of size $\mathrm{nt} \times \mathrm{p}$, with $\mathbf{X}_{it} = \left(\mathrm{x}_{i11}, \dots, \mathrm{x}_{iTn_{iT}}\right)'$, size $\mathrm{n_i t} \times \mathrm{p}$. $\boldsymbol{\beta} = \left(\beta_{\mathbf{1}}, \dots, \beta_{\boldsymbol{p}}\right)'$ is a constant vector of coefficient variables of size $\mathrm{p} \times 1$, α is the value of the coefficient of the auxiliary variable with a measurement error of size 1×1. $\mathbf{w} = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{mt})'$ is a vector of auxiliary variables with a measurement error of size $\mathrm{mt} \times 1$ with $\mathbf{w}_{it} = \left(\mathrm{w}_{i11}, \dots, \mathrm{w}_{itn_{it}}\right)'$. $\boldsymbol{v} = (\boldsymbol{v}_{i1}, \dots, \boldsymbol{v}_{mt})'$ is a random effect vector of area size $\mathrm{mt} \times 1$. $\mathbf{u} = (\mathbf{u}_{i1}, \dots, \mathbf{u}_{mt})'$ is a random effect vector of area size $\mathrm{mt} \times 1$, $\mathbf{e} = (\boldsymbol{e}_{i1}, \dots, \boldsymbol{e}_{mt})'$ is a model error vector of size $\mathrm{nt} \times 1$, with $\mathbf{e}_{it} = \left(\mathrm{e}_{i11}, \dots, \mathrm{e}_{itn_{it}}\right)'$. $\mathbf{Z}1 = \mathbf{I}_m \otimes \mathbf{Z}1_i$, $\mathbf{Z}2 = \mathbf{I}_m \otimes \mathbf{Z}2_i$, $\mathbf{Z}3 = \mathbf{I}_m \otimes \mathbf{Z}3_i$. Z1, Z2 and Z3 components are size $n \times \mathrm{mt}$, where $\mathbf{Z1}_i = \mathbf{Z2}_i = \mathbf{1}_{n_i}$ and $\mathbf{Z3}_i = \left(\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)$ with $\otimes$ is kronecker multiplication, $\mathbf{I}_m$ is an identity matrix of size $\mathrm{m} \times \mathrm{m}$, $\mathbf{1}_{n_i}$ is a vector of size n_i where all the elements are valuable 1, $\mathbf{1}_{n_{it}}$ is -a vector of size n_{it} where all the elements are valuable 1 and $\mathbf{I}_T$ is an identity matrix of size $\mathrm{T} \times \mathrm{T}$.

Assume that the element u_{it} in (3) is stationary, then the expected value and the structure of the variance matrix for the vector $\mathbf{u}$ are:

$$E(\mathbf{u}) = 0 \text{ and } Cov(\mathbf{u}) = \boldsymbol{G}_{3i} = \sigma_\varepsilon^2 \boldsymbol{\Gamma}\ , \tag{5}$$

where $\boldsymbol{\Gamma}$ is a symmetric matrix of size $T \times T$ with element to $(t.t')$ worth $\frac{\rho^{|t-t'|}}{1-\rho^2}$. Matrix $\boldsymbol{\Gamma}$ is:

$$\boldsymbol{\Gamma} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \rho \\ \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix}. \tag{6}$$

In equation (4), $\mathbf{e}_{it}$ is assumed to be independent so that the expected value and the variance matrix structure of $\mathbf{e}_i = (\mathbf{e}'_{i1}, \dots, \mathbf{e}'_{iT})'$ are:

$$E(\mathbf{e}_i) = 0 \text{ and } Cov(\mathbf{e}_i) = \boldsymbol{R}_i = \sigma_\mathrm{e}^2 \mathbf{I}_{n_i}. \tag{7}$$

If $\mathrm{G}_{1i} = Cov(\boldsymbol{v}_i) = \sigma_v^2$ and $\mathrm{G}_{2i} = cov(\boldsymbol{w}_i) = \sigma_w^2$, and it is assumed that $\boldsymbol{v}_i$, $\mathbf{u}_i, \mathbf{e}_i$ and $\mathbf{w}_i$ are independent, the variance matrix of $\mathbf{y}_i^*$ in equation (3):

$$\mathbf{V}_i = Cov(\mathbf{Z}_{1i}\boldsymbol{v}_i) + cov(\mathbf{Z}_{2i}\boldsymbol{w}_i\alpha) + Cov(\mathbf{Z}_{3i}\mathbf{u}_i) + Cov(\mathbf{e}_i),$$

$$\mathbf{V}_i = \mathbf{Z}_{1i}Cov(\boldsymbol{v}_i)\mathbf{Z}'_{1i} + \alpha^2\mathbf{Z}_{2i}cov(\boldsymbol{w}_i)\mathbf{Z}'_{2i} + \mathbf{Z}_{3i}Cov(\mathbf{u}_i)\mathbf{Z}'_{3i} + Cov(\mathbf{e}_i),$$

$$\mathbf{V}_i = \mathbf{Z}_{1i}\mathrm{G}_{1i}\mathbf{Z}'_{1i} + \alpha^2\mathbf{Z}_{2i}\mathrm{G}_{2i}\mathbf{Z}'_{2i} + \mathbf{Z}_{3i}\mathbf{G}_{3i}\mathbf{Z}'_{3i} + \mathbf{R}_i,$$

$$\mathbf{V}_i = \mathbf{Z}_{1i}\sigma_v^2\mathbf{Z}'_{1i} + \alpha^2\mathbf{Z}_{2i}\sigma_w^2\mathbf{Z}'_{2i} + \mathbf{Z}_{3i}\sigma_\varepsilon^2\boldsymbol{\Gamma}\mathbf{Z}'_{3i} + \sigma_{\mathrm{e}}^2\mathbf{I}_{n_i},$$

$$\mathbf{V}_i = \sigma_v^2\mathbf{Z}_{1i}\mathbf{Z}'_{1i} + \alpha^2\sigma_w^2\mathbf{Z}_{2i}\mathbf{Z}'_{2i} + \sigma_\varepsilon^2\mathbf{Z}_{3i}\boldsymbol{\Gamma}\mathbf{Z}'_{3i} + \sigma_{\mathrm{e}}^2\mathbf{I}_{n_i},$$

$$\mathbf{V}_i = \sigma_v^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \alpha^2\sigma_w^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \sigma_\varepsilon^2\left(\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)\boldsymbol{\Gamma}\left(\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)' + \sigma_{\mathrm{e}}^2\mathbf{I}_{n_i},$$

$$\mathbf{V}_i = \sigma_v^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \alpha^2\sigma_w^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \sigma_\varepsilon^2\left(\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)\left(\boldsymbol{\Gamma}\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)' + \sigma_e^2\mathbf{I}_{n_i},$$

$$\mathbf{V}_i = \sigma_v^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \alpha^2\sigma_w^2\mathbf{1}_{n_i}\,\mathbf{1}'_{n_i} + \sigma_\varepsilon^2\left(\mathbf{I}_T\boldsymbol{\Gamma}\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\mathbf{1}'_{n_{it}}\right) + \sigma_e^2\mathbf{I}_{n_i},$$

$$\mathbf{V}_i = \sigma_v^2\mathbf{J}_{n_i} + \alpha^2\sigma_w^2\mathbf{J}_{n_i} + \sigma_\varepsilon^2\left(\boldsymbol{\Gamma} \otimes \mathbf{J}_{n_{it}}\right) + \sigma_e^2\mathbf{I}_{n_i}. \tag{8}$$

The components $\mathbf{1}_{n_i}$ and $\mathbf{1}_{n_{it}}$ are vectors of size n_i and n_{it}, respectively, whose elements are all 1. The components $\mathbf{J}_{n_i}$ and $\mathbf{J}_{n_{it}}$ are square matrices of size $n_i \times n_i$ and $n_{it} \times n_{it}$, where all elements are 1.

In equation (3) it is assumed that $\boldsymbol{v}_i \sim iidN(0.\sigma_v^2)$ and $\boldsymbol{w}_i \sim iidN(\mu_w, \sigma_w^2)$, so that the expected values and the variance matrix structure of the vectors $\boldsymbol{v} = (v_1, \dots, v_m)'$ and $\mathbf{w} = (\mathrm{w}_1, \dots, \mathrm{w}_m)'$:

$$E(\mathbf{v}) = 0 \text{ and } Cov(\mathbf{v}) = \mathbf{G}_1 = \sigma_v^2\mathbf{I}_m, \tag{9}$$

$$E(\mathbf{w}) = \mu_w \text{ and } cov(\mathbf{w}) = \mathbf{G}_2 = \sigma_w^2\mathbf{I}_m. \tag{10}$$

Furthermore, in equation (4), $\mathbf{u}$ and $\mathbf{e}$ are assumed to be independent, so that the expected value and the structure of the variance matrix of the vector $\mathbf{u} = (u'_1, \dots, u'_m)'$ and $\mathbf{e} = (e'_1, \dots, e'_m)'$ respectively are:

$$E(\mathbf{u}) = 0 \text{ and } Cov(\mathbf{u}) = \mathbf{G}_3 = \mathbf{I}_m \otimes \sigma_\varepsilon^2\boldsymbol{\Gamma}, \tag{11}$$

$$E(\mathbf{e}) = 0 \text{ and } cov(\mathbf{e}) = \mathbf{R} = \mathbf{I}_m \otimes \sigma_e^2\mathbf{I}_{n_i} = \sigma_e^2\mathbf{I}_n, \tag{12}$$

where $n = \sum_{t=1}^{m} n_i$ and $\mathbf{I}_n$ is an identity matrix of size $n \times n$.

Assuming $\mathbf{v}, \mathbf{w}$, $\mathbf{u}$, and $\mathbf{e}$ are independent of each other, then the variance matrix of $\mathbf{y}^*$ in equation (4) is:

$\mathbf{V} = Cov(\mathbf{Z}1\boldsymbol{v}) + Cov(\mathbf{Z}2\mathbf{w}\alpha) + Cov(\mathbf{Z}3\boldsymbol{u}) + Cov(\mathbf{e})$,

$\mathbf{V} = \mathbf{Z}1 Cov(\boldsymbol{v})\mathbf{Z}1' + \alpha^2\mathbf{Z}2 Cov(\mathbf{w})\mathbf{Z}2' + \mathbf{Z}3 Cov(\boldsymbol{u})\mathbf{Z}3' + Cov(\mathbf{e})$,

$\mathbf{V} = \mathbf{Z}1\mathbf{G}_1\mathbf{Z}1' + \alpha^2\mathbf{Z}2\mathbf{G}_2\mathbf{Z}2' + \mathbf{Z}3\mathbf{G}_3\mathbf{Z}3' + \mathbf{R}$,

$\mathbf{V} = (\mathbf{I}_m \otimes \mathbf{Z}1_i)\sigma_v^2\mathbf{I}_m(\mathbf{I}_m \otimes \mathbf{Z}1_i)' + \alpha^2(\mathbf{I}_m \otimes \mathbf{Z}2_i)\sigma_w^2\mathbf{I}_m(\mathbf{I}_m \otimes \mathbf{Z}2_i)' + (\mathbf{I}_m \otimes$
$\mathbf{Z}3_i)(\mathbf{I}_m \otimes \sigma_\varepsilon^2\boldsymbol{\Gamma})(\mathbf{I}_m \otimes \mathbf{Z}3_i)' + \mathbf{R}$,

$\mathbf{V} = \sigma_v^2(\mathbf{I}_m \otimes \mathbf{Z}1_i\mathbf{Z}1_i{}') + \alpha^2\sigma_w^2(\mathbf{I}_m \otimes \mathbf{Z}2_i\mathbf{Z}2_i{}') + \sigma_\varepsilon^2(\mathbf{I}_m \otimes \mathbf{Z}3_i\boldsymbol{\Gamma}\mathbf{Z}3_i{}') + \mathbf{I}_m \otimes \sigma_e^2\mathbf{I}_{n_i}$,

$\mathbf{V} = \mathbf{I}_m \otimes \left(\sigma_v^2\mathbf{Z}1_i\mathbf{Z}1_i{}' + \alpha^2\sigma_w^2\mathbf{Z}2_i\mathbf{Z}2_i{}' + \sigma_\varepsilon^2\mathbf{Z}3_i\boldsymbol{\Gamma}\mathbf{Z}3_i{}' + \sigma_e^2\mathbf{I}_{n_i}\right)$,

$\mathbf{V} = \mathbf{I}_m \otimes \left(\sigma_v^2\mathbf{J}_{n_i} + \alpha^2\sigma_w^2\mathbf{J}_{n_i} + \sigma_\varepsilon^2\left(\boldsymbol{\Gamma} \otimes \mathbf{J}_{n_{it}}\right) + \sigma_e^2\mathbf{I}_{n_i}\right)$,

$$\mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_i. \tag{13}$$

Equation (4) produces the equations for $E(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u})$ and $Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u})$ which are written in matrix form, respectively:

$$E(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) = \mathbf{X}\boldsymbol{\beta} + \mu_w\alpha\mathbf{1}_n + \mathbf{Z}2\boldsymbol{v} + \mathbf{Z}3\boldsymbol{u}, \tag{14}$$

$$Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) = \alpha^2\sigma_w^2\mathbf{1}_n + \sigma_e^2\mathbf{1}_n, \tag{15}$$

$\mathbf{Z}2 = \mathbf{I}_m \otimes \mathbf{Z}2_i$, $\mathbf{Z}3 = \mathbf{I}_m \otimes \mathbf{Z}3_i$. The $\mathbf{Z}2$ and $\mathbf{Z}3$ components of a matrix of size $n \times \mathrm{mt}$, where $\mathbf{Z2}_i = \mathbf{1}_{n_i}$ and $\mathbf{Z3}_i = \left(\mathbf{I}_T \otimes \mathbf{1}_{n_{it}}\right)$ with $\otimes$ is kronecker multiplication, $\mathbf{I}_m$ is an identity matrix of size $\mathrm{m} \times \mathrm{m}$, $\mathbf{1}_{n_i}$ is a vector of size n_i where all elements are 1, $\mathbf{1}_{n_{it}}$ is a vector of size n_{it} where all elements are 1,$\mathbf{1}_n$ is a vector of size n where all elements are 1, and $n = \sum_{t=1}^{m} n_i$, $\mathbf{I}_T$ is an identity matrix of size T×T.

From equations (14) and (15), then $E(\mathbf{y}|\boldsymbol{v},\boldsymbol{u})$ or $\hat{\mathbf{y}}$ is found by doing the back transformation, it is obtained as follows:

$$\hat{\mathbf{y}} = E(\mathbf{y}|\boldsymbol{v},\boldsymbol{u}) = exp[E(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) + 0.5Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u})], \tag{16}$$

$$\hat{\mathbf{y}} = exp[(\mathbf{X}\boldsymbol{\beta} + \mu_w\alpha\mathbf{1}_n + \mathbf{Z}2\boldsymbol{v} + \mathbf{Z}3\boldsymbol{u}) + 0.5(\alpha^2\sigma_w^2\mathbf{1}_n + \sigma_e^2\mathbf{1}_n)]. \tag{17}$$

According to the estimation of the model parameters from the available sample data, equation (17) can be written:

$$\hat{\mathbf{y}} = exp\left[\left(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mu}_w\hat{\alpha}\mathbf{1}_n + \mathbf{Z}2\hat{\boldsymbol{v}} + \mathbf{Z}3\hat{\boldsymbol{u}}\right) + 0.5(\hat{\alpha}^2\hat{\sigma}_w^2\mathbf{1}_n + \mathbf{1}_n\hat{\sigma}_e^2\mathbf{1}_n)\right]. \tag{18}$$

Parameters that are of concern in this study are obtained from the sum of the values of observations that are members of the sample, with index (s), and unit values that are not members of the sample, with index (r). Estimated average total observations based on area is:

$$\bar{\mathbf{Y}}_{\mathrm{i}} = \frac{1}{N_i}\left(\sum_{(s)} \boldsymbol{y} + \sum_{(r)} \boldsymbol{y}\right). \tag{19}$$

Equation (19) is then estimated in the form of equation:

$$\hat{\bar{\mathbf{Y}}}_{\mathrm{i}} = \frac{1}{N_i}\left(\sum_{(s)} \boldsymbol{y} + \sum_{(r)} \hat{\mathbf{y}}\right). \tag{20}$$

To estimate the parameter $\bar{\mathbf{Y}}_{\mathrm{i}}$ $\left(\hat{\bar{\mathbf{Y}}}_{\mathrm{i}}\right)$, it is obtained by the following steps:

a. Predict the variance component ($\hat{\sigma}_w^2$, $\hat{\sigma}_v^2$, $\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_e^2$) using the REML (restricted maximum likelihood) method approach,

b. Predict ρ $(\hat{\rho})$ in the equation $\boldsymbol{u}_{it} = \rho\boldsymbol{u}_{i.t-1} + \boldsymbol{\varepsilon}_{it}$ by regressing the residual value obtained.

c. Predict the variance of u_{it}, with the form of equation:

$$\hat{\sigma}_u^2 = \hat{\sigma}_\varepsilon^2 \big/ (1 - \hat{\rho}^2), \tag{21}$$

d. Predicting coefficients of random variables containing measurement errors, $\hat{\alpha}$, with $\hat{\alpha} = k_w^{-1}\, \hat{\alpha}_{OLS_t}$, where $\hat{\alpha}_{OLS} = \frac{S_{yw}}{S_{ww}}$, $S_{yw} = \frac{1}{2}\sum_{i\epsilon s}(y_i - \bar{y}_s)(w_i - \bar{w}_s)$, $S_{ww} = \frac{1}{2}\sum_{i\epsilon s}(w_i - \bar{w}_s)^2$, $k_w = \frac{\hat{\sigma}_w^2}{\hat{\sigma}_w^2 + \hat{\sigma}_\eta^2}$. $\hat{\sigma}_\eta^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(W_{it} - \bar{W}_{it})^2}{n_T - m}$, $n_T = \sum_{i=1}^{m} n_i$

e. Calculate the average value of a random variable that has a measurement error (μ_w), which is estimated by the equation $\hat{\mu}_w = \frac{1}{m}\sum \bar{W}_{it}$.

f. Calculate variance value, $\mathbf{V}\left(\mathrm{y}_{itj}^*\right)$, where:

$$\mathbf{V}\left(\mathrm{y}_{itj}^*\right) = \hat{\alpha}^2\hat{\sigma}_w^2 + \hat{\sigma}_v^2 + \hat{\sigma}_u^2 + \hat{\sigma}_e^2, \tag{22}$$

g. Estimate the value $\boldsymbol{\beta}$, with the equation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y}^* - \mathbf{X}'\mathbf{V}^{-1}\hat{\alpha}\mu_w), \tag{23}$$

h. Predict the random effect of $\boldsymbol{v}_i$, with the equation:

$$\widehat{\boldsymbol{v}}_i = \hat{\gamma}_{i(v)}\left(\overline{\boldsymbol{y}}_i^* - \overline{\mathbf{X}}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{w}_i\widehat{\alpha}\right), \tag{24}$$

where $\hat{\gamma}_{i(v)} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\alpha}_t^2\hat{\sigma}_w^2 + \hat{\sigma}_e^2/n_i + \hat{\sigma}_\varepsilon^2}$,

i. Predict the random effect of time-area $\boldsymbol{u}_{it}$, with the equation:

$$\widehat{\boldsymbol{u}}_{it} = \hat{\gamma}_{i(u)}\left(\overline{\boldsymbol{y}}_i^* - \overline{\mathbf{X}}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{w}_i\widehat{\alpha}\right), \tag{25}$$

where $\hat{\gamma}_{i(u)} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_v^2 + \hat{\alpha}_t^2\hat{\sigma}_w^2 + \hat{\sigma}_e^2/n_i + \hat{\sigma}_\varepsilon^2}$

j. Estimating the value of variance on the condition that v and u are known $Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u})$, is obtained in the form of equation:

$$Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) = \alpha^2\sigma_w^2\mathbf{1}_n + \sigma_e^2\mathbf{1}_n, \tag{26}$$

k. Calculate the expected value of $\mathbf{y}^*$ provided $\boldsymbol{v}_i$ and $\boldsymbol{u}_{it}$ are known,

$$E(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) = \mathbf{X}\boldsymbol{\beta} + \mu_w\alpha\mathbf{1}_n + \mathbf{Z}2\boldsymbol{v} + \mathbf{Z}3\boldsymbol{u}, \tag{27}$$

l. Predict the value of $\widehat{\boldsymbol{y}}$ obtained from:

$$\widehat{\boldsymbol{y}} = exp[E(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u}) + 0.5Var(\mathbf{y}^*|\boldsymbol{v},\boldsymbol{u})], \tag{28}$$

m. Calculate the average value of the district/city:

$$\widehat{\widehat{\mathbf{Y}}}_{\mathrm{i}} = \frac{1}{N_i}\left(\sum_{(s)}\boldsymbol{y} + \sum_{(r)}\widehat{\boldsymbol{y}}\right), \tag{29}$$

n. Calculate the provincial average:

$$\widehat{\widehat{\mathrm{Y}}} = \frac{1}{N}\sum_{i=1}^{m}\widehat{\widehat{\mathbf{Y}}}_{\mathrm{i}} * N_i, \tag{30}$$

where $\widehat{\widehat{\mathrm{Y}}}$ is the estimator for the average consumption per capita at the provincial level, $\widehat{\widehat{\mathbf{Y}}}_i$ is the estimator for the average consumption per capita at the district/city level, N is the total population in the province and N_i is the population in each district/city.

o. Next calculate the value of mse (mean square of error):

$$mse\left(\widehat{\widehat{\mathbf{Y}}}_{\mathrm{i}}\right) = B^{-1}\sum_{b=1}^{B}\left(\widehat{\widehat{\boldsymbol{Y}}}_{i*}^{H}(b) - \widehat{\widehat{\boldsymbol{Y}}}_{i*}(b)\right)^2 \tag{31}$$

5. MODEL EVALUATION

As previously discussed, the form of the proposed development model is in accordance with equation (3), is:

$$y^*_{itj} = X'_{itj}\beta + w_{it}\alpha + v_i + u_{it} + e_{itj},$$

with: $W_{it} = w_{it} + \eta_{it}$, $y^*_{itj} = log(y_{itj})$.

The response variable y_{itj} is the j-th sample unit in a small area i at time t which is assumed to have a relationship with a vector of auxiliary variables X_{itj}, which is assumed to be also available for each population unit in a small area i. The β component is the coefficient of the auxiliary variable which is constant. The notation i is a small area index which is defined to move from 1 to m (the number of small areas). The notation t is a time index that moves from 1 to T. The notation j is the sample unit in the i-th small area at the t-th time, which moves from 1 to n_{it} (the number of sample units in each small area). The variable w_{it} is the unknown true area-specific covariate for time t, corresponding to $\boldsymbol{y}_{itj}$, with the mean μ_w and variance σ^2_w. While W_{it} is the auxiliary variable with measurement error for the t-th time. The component η_{it} is the measurement error on the auxiliary variable for t-th time, η_{it} is assumed to follow a normal distribution with an average of 0 and a variance of σ^2_η. The v_i component is the random effect of the i-th area which is assumed to follow a normal distribution with a mean of 0 and a variance of σ^2_v. The u_{it} component is a time-area random effect which is assumed to follow a first-order autoregressive process in every i-th small area. The component ε_{it} is the error of u_{it} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ^2_ε. The e_{itj} component is the error of y_{itj} which is assumed to follow a normal distribution with a mean of 0 and a variance of σ^2_e. The ρ component is an autoregressive coefficient with an absolute value of less than 1. The random effects of w_{it}, v_i, u_{it} and e_{itj} are assumed to be independent. Simulations are conducted to further evaluate the goodness of the proposed model.

Simulation data is generated according to the characteristics of the data used in the application, namely the population generated with variables that have measurement errors. Simulations are carried out to observe the effect of autoregressive coefficient (ρ), area random effect (v), auxiliary variable (w) on the estimation results. When generating the data population, a combination of values is given;

- Three types of autoregressive coefficient values (ρ), that are small, medium and large.
- Three types of area random effect variance values (σ_v^2), that are small, medium and large.
- Three types of value for the variance auxiliary variable (σ_w^2), that are small, medium and large.

The evaluation is carried out by comparing the results of scenarios determined against the two Small Area Estimation models applied, namely:

- Small Area Estimation of the autoregressive model with measurement error in the auxiliary variable (SAE-AR1-ME).
- Small Area Estimation of the autoregressive model without measurement error in the auxiliary variables (SAE-AR1).

The evaluation of the model is then performed by comparing the value of the goodness of the model in each estimation methods carried out (SAE-AR1-ME and SAE-AR1), including Relative Bias (RB), Root Mean Square of Error (RMSE) and Coefficient of Variation (CV) for each area and also calculated the average value of each size for the entire area.

5.1. Simulation

At this stage, the population data generation process would be carried out, as well as sample data collection for 1000 replications for further analysis using the SAE-AR1-ME and SAE-AR1 models. The stages of the simulation process for each scenario are as follows:

a. Designing population data, X, for the first and second periods

   i. Determining the number of areas (i), with $i = 1, \ldots, m,$ $(\mathrm{m} = 27)$

   - Specifying the unit size in the area, each of which is 40, 44, 31, 31, 36, 39, 26, 31, 40, 26, 26, 31, 29, 16, 29, 23, 16, 10, 6, 7, 29, 5, 12, 11, 3, 10 and 4. The unit size in this area is made for two periods with the same value for each period $(n_{it}, t = 1{,}2)$.
   - To describe the condition of a small area, the number of population in each area (N_i) is determined for two periods. The population size is set at 100 times the number of unit sample sizes in each area for two periods. So that the total sample population for

each period is 4000, 4400, 3100, 3100, 3600, 3900, 2600, 3100, 4000, 2600, 2600, 3100, 2900, 1600, 2900, 2300, 1600, 1000, 600, 700, 2900, 500, 1200, 1100, 300, 1000 and 400.

- Generating population data X_{1t} $(t = 1,2)$ which is assumed to follow a uniform distribution with a range of values from 0 to 0.03.
- Generating population data X_{2t} $(t = 1,2)$ which is assumed to follow a uniform distribution with a range of values from 0 to 0.0003.

ii. Determining the value of the coefficients of the fixed auxiliary variable, among others; $\beta_0 = 0.96$, $\beta_1 = -9.34$, $\beta_2 = 4186.77$ and the coefficient of the random auxiliary variable, $\propto = 0.92$.

iii. Determining the value of the autoregressive coefficient (ρ) that is small (0.1), medium (0.5) and large (0.9).

iv. Setting the value of the area random effect variance (σ_v^2), that is small (0.006), medium (0.06) and large (0.6).

v. Setting the value of the variance of the auxiliary variable with measurement errors (σ_w^2), that is small (0.002), medium (0.02) and large (0.2).

vi. Setting the variance value for the error from u_{it} i.e. $\sigma_\varepsilon^2 = 0.01$,

vii. Setting the variance value for the error of y_{itj}, which is $\sigma_e^2 = 0.12$.

viii. Calculating the value of the time-area random effect variance σ_u^2, where: $\sigma_u^2 = \sigma_\varepsilon^2 \big/ (1-\rho^2)$.

ix. Setting the mean value of w_i, $\mu_w = 13$.

x. Generating the value of $\mathbf{w}_i$ for the first t as much area as the distribution for $w_{i,t1}$ is $w_{i,t1} \sim N(\mu_w, \sigma_w^2)$.

xi. Determining the value of w_i for the second t is the same as w_i for the first t.

xii. Setting the value of the auxiliary variable with $X_{itj} = X_{1t}, X_{2t}$.

xiii. Generating the area random effect value $\boldsymbol{v}_i$ for the first t following a normal distribution, $\boldsymbol{v}_{i,t1} \sim (0, \sigma_v^2)$.

xiv. Determining the value of $\boldsymbol{v}_i$ for the second t is the same as $\boldsymbol{v}_i$ for the first t.

xv. Generating an error value from u_{it}, i.e. ε_{it} assumed $\varepsilon_{it} \sim iidN(0, \sigma_{\varepsilon}^2)$.

xvi. Generating a time area random effect value for the first t, i.e. $u_{it_1} \sim N(0, \sigma_u^2)$.

xvii. Determining the value of the random effect of the time area for the second t, where $u_{it_2} = \rho u_{i.t_1} + \varepsilon_{it}$.

xviii. Generating model error value, e_{itj} with $e_{itj} \sim N(0, \sigma_e^2)$.

xix. Calculating the value of y_{itj} for two periods;

$$y_{itj} = \exp\left(\mathrm{X}'_{itj}\beta + w_{it}\alpha + v_i + u_{it} + e_{itj}\right).$$

xx. Calculating the $\bar{y}_i$ value (mean by area) of the population over two periods:

$$\bar{y}_{i\_p} = \frac{1}{N_i}\left(\sum y_{ij}\right).$$

b. Estimating small area parameters.

i. Taking n_{it} random samples of y_{itj_s} for each area.

ii. Evaluation of sample data using the indirect method.

Repeating step b.i as many as B large numbers, i.e. $B = 1000$.

a) $\mathrm{w_i}$ random (Empirical Best Linear Unbiased Prediction/EBLUP Measurement Error) in area.

a.1. Obtaining the value of the variable of concern based on the sample data obtained in each iteration ($\bar{y}_{i\_s\_indirect\_SAE\_AR1\_ME}$).

a.2. Calculating the value of RB, RMSE, and CV.

$$RB_{i\_SAE\_AR1\_ME} = \frac{1}{B}\sum_{l=1}^{B}\frac{\bar{y}_{i\_s\_indirect\_SAE\_AR1\_ME(l)} - \bar{y}_{i\_p}}{\bar{y}_{i\_p}},$$

$$RMSE_{i\_SAE\_AR1\_ME} = \sqrt{\frac{1}{B}\sum_{l=1}^{B}\left(\bar{y}_{i\_s\_indirect\_SAE\_AR1\_ME(l)} - \bar{y}_{i\_p}\right)^2},$$

$$CV_{i\_SAE\_AR1\_ME} = \frac{RMSE_{i\_SAE\_AR1\_ME}}{\bar{y}_{i\_p}} \times 100\%.$$

b) $\mathrm{w_i}$ fixed (EBLUP) by using the auxiliary variable without measurement error.

b.1. Obtaining the value of the variable of concern based on the sample data found in each repetition ($\bar{y}_{i\_s\_indirect\_SAE\_AR1}$).

b.2. Calculating the value of RB, RMSE, dan CV.

$$RB_{i\_SAE\_AR1} = \frac{1}{B}\sum_{l=1}^{B}\frac{\bar{y}_{i\_s\_indirect\_SAE\_AR1(l)}-\bar{y}_{i\_p}}{\bar{y}_{i\_p}},$$

$$RMSE_{i\_SAE\_AR1} = \sqrt{\frac{1}{B}\sum_{l=1}^{B}\left(\bar{y}_{i\_s\_indirect\_SAE\_AR1(l)} - \bar{y}_{i\_p}\right)^2},$$

$$CV_{i\_SAE\_AR1} = \frac{RMSE_{i\_SAE\_AR1}}{\bar{y}_{i\_p}} \times 100\%.$$

Analysis were performed in the SAE-AR1-ME and SAE-AR1 models. The value of the goodness of the model being compared includes relative bias (RB), the root mean square error (RMSE) and coefficient of variance (CV) and average is also calculated.

5.2. SIMULATION RESULTS

The results of the simulation stages were obtained for each measure of the goodness of the model (Table 1 – Table 3). Table 1 describes the simulation results of the average relative bias ($\overline{\text{RB}}$), Table 2 describes the simulation results of the average root mean square error ($\overline{RMSE}$), while Table 3 explains of the simulation results of the average coefficient of variance ($\overline{CV}$) of the SAE-AR1-ME and SAE-AR1 methods. Simulations were run on the population generated with measurement errors on the auxiliary variables.

a. Relative Bias

Unbiased means that the expected value of the estimator is the same as the predicted parameter, for example the parameter $\bar{x}$ is an unbiased estimator for μ. If the sampling process is repeated many times and each sample is calculated the value of $\bar{x}$, the average of $\bar{x}$ is equal to μ. Table 1 shows the average relative bias generated under conditions of small, medium and large autoregressive coefficients. The average relative bias value in the SAE-AR1-ME model is seen always smaller than the SAE-AR1 model.

In three types of autoregressive coefficient values (ρ), for the SAE-AR1-ME model on the value of variance of the auxiliary variable (σ_w^2), small and medium, the area random effect variance value (σ_v^2) tends to be smaller as well. Slightly different things occur under the condition of the large value of variance of the auxiliary variable (σ_w^2) which tends to be larger in value along with the greater value of the area random effect variance (σ_v^2). While, the SAE-AR1 model tends to be stable for the three conditions of area random effect variance (σ_v^2).

The Small Area Estimation model is known influenced by the magnitude of the random effect value, in this model the area effect (v_i) and the time-area random effect (u_{it}). Based on the equation $\hat{v}_i = \hat{\gamma}_{i(v)}(\bar{y}_i^* - \bar{\mathrm{x}}_i\hat{\beta} - w_i\hat{\alpha})$ where $\hat{\gamma}_{i(v)} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2+\hat{\alpha}_t^2\hat{\sigma}_w^2+\hat{\sigma}_e^2/n_i+\hat{\sigma}_\varepsilon^2}$ and the equation $\hat{u}_{it} = \hat{\gamma}_{i(u)}(\bar{y}_i^* - \bar{\mathrm{x}}_i\hat{\beta} - w_i\hat{\alpha})$ with $\hat{\gamma}_{i(u)} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_v^2+\hat{\alpha}_t^2\hat{\sigma}_w^2+\hat{\sigma}_e^2/n_i+\hat{\sigma}_\varepsilon^2}$, it appears that v_i and u_{it} are affected by relative unexplained inter-area variability to the total variability (γ).

Table 1. The Average Relative Bias of the Whole Area in the Population with the Variance of the Covariates of Small, Medium and Large Values According to the Type of Autoregressive Coefficient Value and the Type of Area Random Effect Variance Value in the SAE-AR1-ME and SAE-AR1 Models.

Auto-regressive Coefficient Value (ρ)	Area Random Effect Variance Value (σ_v^2)	Estimation Method		Estimation Method		Estimation Method	
		SAE-AR1-ME (σ_w^2 small)	SAE-AR1	SAE-AR1-ME (σ_w^2 medium)	SAE-AR1	SAE-AR1-ME (σ_w^2 large)	SAE-AR1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small	Small	-0.0563	-0.0627	-0.0428	-0.0556	-0.0155	-0.0667
	Medium	-0.0594	-0.0670	-0.0374	-0.0608	-0.0250	-0.0627
	Large	0.0115	-0.0627	0.0130	-0.0565	0.0195	-0.0622
Medium	Small	-0.0263	-0.0508	-0.0195	-0.0560	0.0010	-0.0638
	Medium	-0.0086	-0.0598	-0.0349	-0.0632	0.0174	-0.0678
	Large	-0.0001	-0.0621	-0.0155	-0.0570	-0.0041	-0.0605
Large	Small	-0.0560	-0.0627	-0.0505	-0.0649	0.0053	-0.0601
	Medium	-0.0207	-0.0603	-0.0224	-0.0571	0.0025	-0.0595
	Large	-0.0423	-0.0641	-0.0448	-0.0615	0.0271	-0.0594

b. Root Mean Square Error

One measure of error in estimation is the root mean square error or RMSE [16]. The RMSE is a value that is useful for evaluating the estimation technique of a model. The RMSE is the average value of the number of squares of errors, it can also state the size of the error generated by an estimator model. A low value of RMSE indicates that the diversity of values produced by an estimator model is close to the diversity of the observed values. Table 2 presents the average value of RMSE ($\overline{\text{RMSE}}$) generated from the simulation process.

The SAE-AR1-ME model produces a value $\overline{RMSE}$ which is always greater than the SAE-AR1 model. However, there is a condition that the value $\overline{RMSE}$ of the SAE-AR1-ME model is almost the same as the SAE-AR1 model. This happens when the population is derived from a variance of auxiliary variables with a small value and when the variance of the area of random effect is of small value. The greater value of the variance of the random effect area tends to result in the greater value of $\overline{RMSE}$ generated. The SAE-AR1 model also shows the same thing, that the greater value of the variance of the random effect area tends to result in the greater value of $\overline{RMSE}$ generated.

Table 2. The Mean of the Root Mean Square Error Overall Area in the Population with Small, Medium and Large Variance of Auxiliary Variable Values According to the Type of Autoregressive Coefficient Value and the Type of Area Random Effect Variance Value in the SAE-AR1-ME and SAE-AR1 Models.

Auto-regressive Coefficient value (ρ)	Area Random Effect Variance Value (σ_v^2)	Estimation Method		Estimation Method		Estimation Method	
		SAE-AR1-ME (σ_w^2 small)	SAE-AR1	SAE-AR1-ME (σ_w^2 medium)	SAE-AR1	SAE-AR1-ME (σ_w^2 large)	SAE-AR1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small	Small	95009	84499	108027	80986	236381	96667
	Medium	100180	84356	135703	81205	182437	85629
	Large	200403	128720	228235	133100	277504	108773
Medium	Small	99362	80044	150478	89350	209128	81196
	Medium	130330	83973	131557	86308	277234	95511
	Large	215714	124627	127669	77683	232365	111535
Large	Small	100063	90823	138001	92800	266387	113249
	Medium	118841	84742	124011	85239	243401	102338
	Large	155732	102346	108905	85612	270184	98703

c. Coefficient of Variance

The Coefficient of Variance (CV) is a measure of diversity that can be used to compare a distribution of data that has different units. The CV is a comparison between the standard deviation and the average of a data distribution and is expressed as a percentage. The magnitude of the CV value will affect the quality of the data distribution. The smaller the CV value means the more uniform or homogeneous a group of data is, and vice versa. Table 3 presents the mean value of CV ($\overline{\text{CV}}$) generated from the simulation process.

Table 3. The Average Coefficient of Variance of the Entire Area in the Population under the Conditions of Small, Medium and Large Variance of the Auxiliary Variable According to the Type of Autoregressive Coefficient Value and the Type of Area Random Effect Variance Value in the SAE-AR1-ME and SAE-AR1 Models

Auto-regressive Coefficient Value (ρ)	Area Random Effect Variance Value (σ_v^2)	Estimation Method		Estimation Method		Estimation Method	
		SAE-AR1-ME (σ_w^2 small)	SAE-AR1	SAE-AR1-ME (σ_w^2 medium)	SAE-AR1	SAE-AR1-ME (σ_w^2 large)	SAE-AR1
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Small	Small	12.216	10.865	14.165	10.619	27.527	11.257
	Medium	13.130	11.056	17.820	10.663	23.667	11.108
	Large	17.672	11.351	16.690	9.733	28.519	11.179
Medium	Small	12.680	10.214	17.736	10.531	28.388	11.022
	Medium	16.681	10.747	16.433	10.781	33.361	11.494
	Large	19.452	11.238	16.797	10.221	22.984	11.032
Large	Small	11.945	10.842	16.550	11.129	25.402	10.799
	Medium	14.931	10.647	15.243	10.477	25.848	10.868
	Large	17.352	11.404	13.499	10.612	27.765	10.143

The SAE-AR1-ME model produces a value $\overline{CV}$ which is always greater than the SAE-AR1 model. When the population comes from a variety of covariates with small and medium measurement errors, the value of $\overline{CV}$ for both models is almost the same. In the SAE-AR1-ME model, the greater value of the auxiliary variable with measurement error tends to result in the greater value of $\overline{CV}$ produced. The greater value of the variance of the random effect area tends to result in the greater value $\overline{CV}$ generated. The SAE-AR1 model shows things that tend to be stable for the $\overline{CV}$ value generated by changes in the value of the area random effect variance. The SAE-AR1-ME and the SAE-AR1 are models with homogeneous data groups for the population with small and medium measurement errors in the covariates.

6. Application of the Development Model

This section presents the results of the application of the developed method by using actual data. The data used is the 2015 *Susenas* data and the 2014 Village Potential data for West Java Province. The application of this development model should be carried out on longer time series data, not only using two points. The use of two points, namely March and September, is not satisfactory enough to provide an illustration that corresponds to the actual conditions. From this point of view, due to limited data available, this can be perceived as an early initiative to use time series data in Small Area Estimation for autoregressive models with measurement errors in the auxiliary variables.

The application of the development model is carried out with several variables derived from the results of data collection in West Java Province, which is divided into two periods with the same size of samples (n_{it}). For the application to the actual data, an estimator model is used according to equation (3), as follows:

$$\mathrm{y}^*_{itj} = \mathrm{X}'_{itj}\beta + \mathrm{w}_{it}\alpha + v_i + u_{it} + \mathrm{e}_{itj},$$

with: $\mathrm{W}_{it} = \mathrm{w}_{it} + \eta_{it}$, $\mathrm{y}^*_{itj} = log(y_{itj})$.

The variable of concern to y_{itj} is the average household consumption per capita at the sub-district level, from the 2015 *Susenas* for two periods of data collection, namely March and September. Index i is the area level which in this study, is the district/city. Index j is the unit level which in this study, is the sub-district. The t index is time which, in this study, consists of two periods, namely March and September. The auxiliary variable $\mathrm{X1}$ is the proportion of the number of restaurants to the total population in each sub-district for two periods. The auxiliary variable $\mathrm{X2}$

is the proportion of the number of food and beverage stalls/shops with the total population in each sub-district for two periods. The two auxiliary variables are proportions to the total population in 2014 and 2015. The auxiliary variable containing measurement error (w) in this study is the log of the average consumption per capita of the district/city a month ago from information on the results of the March 2015 *Susenas* data collection, which was used in both periods of the study. Based on the parameter estimation method presented previously, from the results of data processing using R software, the estimation results for each parameter are; $\hat{\sigma}_v^2 = 0.059$, $\hat{\sigma}_e^2 = 0.115$, $\hat{\sigma}_w^2 = 0.024$, $\hat{\sigma}_\varepsilon^2 = 0.012$, $\hat{\rho} = 0.499$, $\hat{\sigma}_u^2 = 0.016$, $\bar{y} = 13.66$, $\hat{\mu}_w = 13.64$, $\hat{\propto} = 0.924$, $\hat{\beta}_0 = 0.957$, $\hat{\beta}_1 = -9.336$, $\hat{\beta}_2 = 4186.775$, $\hat{\sigma}_y^2 = 0.211$, and $\hat{\sigma}_{y|v_i,u_i}^2 = 0.136$.

The estimated average consumption per capita at the provincial level is 905168 rupiahs by the SAE-AR1-ME method and 966767 rupiahs by the SAE-AR1 method. Previously, the results of the March data collection of the 2015 *Susenas* had been submitted in the amount of 896895 rupiahs and a direct estimator of the province as a result of the September data collection of the 2015 *Susenas* at 981968 rupiahs. These results indicate that by using an alternative method of estimating a small area of the autoregressive model, which examines the measurement error of auxiliary variable, an estimator at the provincial level is possibly obtained by first estimating the district/city level, presented in Figure 2 and Figure 3.

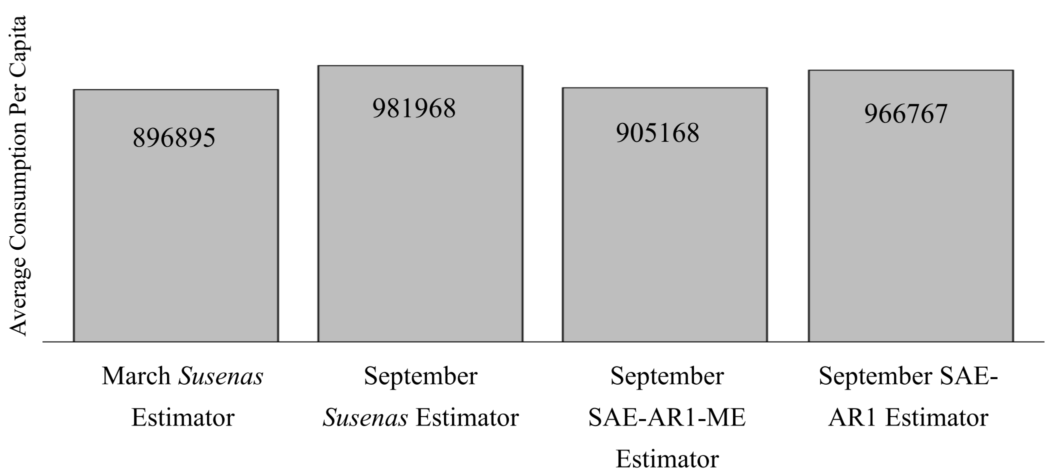

Figure 2. Estimator of the Average Consumption Per Capita of the Province According to the Estimation Method and the Period of Data Collection in West Java Province in 2015

Furthermore, the more detailed results of the estimation are presented in Table 4. In Table 4, an oddity appears, that is the estimate for September which is lower than the direct estimate for March for several districts/cities. In fact, in September, many specific household expenditures/consumptions should have made the September average consumption per capita relatively greater.

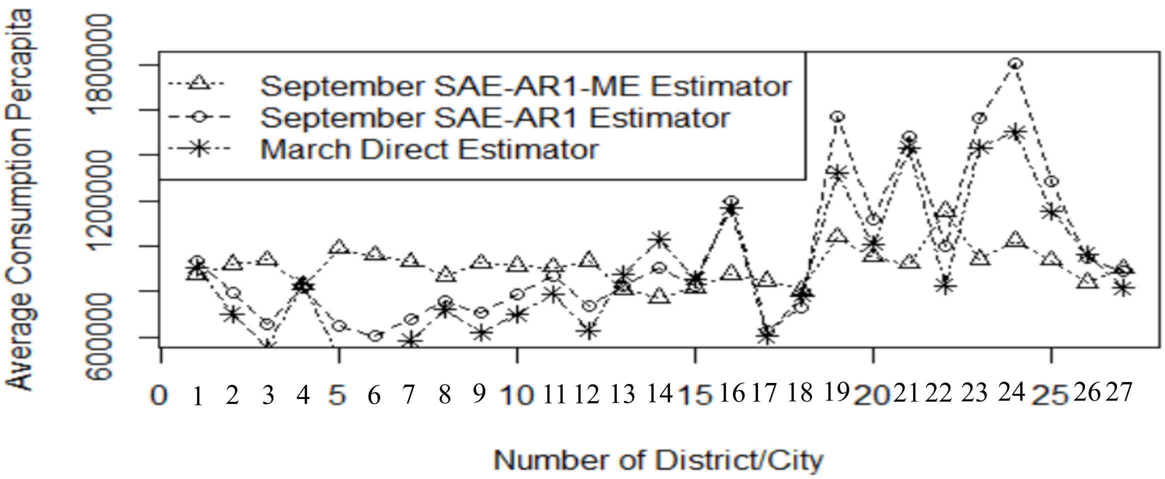

Figure 3. Estimating the Average Consumption Per Capita at the District/City Level According to the Estimation Method and the Period of Data Collection in West Java Province in 2015

Table 4. Estimation of the Average Consumption Per Capita at the District/City Level According to the Estimation Method and the Period of Data Collection in West Java Province in 2015

Number	Name of District/City	March 2015 *Susenas* Direct Estimator (Rupiahs)	September 2015 SAE-AR1-ME Estimator (Rupiahs)	September 2015 SAE-AR1 Estimator (Rupiahs)
(1)	(2)	(3)	(4)	(5)
1	Bogor District	907682	872930	938374
2	Sukabumi District	700506	922168	797681
3	Cianjur District	553869	939178	659246
4	Bandung District	834803	828551	817024
5	Garut District	513366	988430	648836
6	Tasikmalaya District	489726	959818	605723
7	Ciamis District	587214	933834	682479
8	Kuningan District	721786	868323	757879
9	Cirebon District	619552	924567	710528
10	Majalengka District	698224	914675	786194
11	Sumedang District	789992	907778	866075
12	Indramayu District	629355	935140	734590
13	Subang District	873718	808375	825099
14	Purwakarta District	1030583	775037	904521
15	Karawang District	855416	819626	828081
16	Bekasi District	1168767	877138	1196513
17	West Bandung District	605302	847035	624451
18	Pangandaran District	783266	809734	726469
19	Bogor City	1324986	1040251	1569321
20	Sukabumi City	1010902	952071	1118228
21	Bandung City	1433908	923639	1486214
22	Cirebon City	828197	1154793	997037
23	Bekasi City	1434648	943669	1567067
24	Depok City	1503423	1022586	1806383
25	Cimahi City	1153348	937919	1286326
26	Tasikmalaya City	964434	839330	946833
27	Banjar City	817072	904503	890024
	West Java Province	981968	905168	966767

The results in Table 4 show that the AR(1) model with 2 observation points still gives unsatisfactory results. It is necessary to add observation points to make the estimation results better.

Tabel 5. The Value of the Goodness of the Estimation Model by District/City, Estimation Method and the Period of Data Collection in West Java Province in 2015

Name of District/City	Percentage of Sub-district Sample Recorded	Direct Estimator		September 2015 SAE-AR1-ME Estimator		September 2015 SAE-AR1 Estimator	
		Standard Deviation	Coefficient of Variation	Standard Deviation	Coefficient of Variation	Standard Deviation	Coefficient of Variation
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bogor District	57.50	454226	50.04	213300	23.92	274138	28.22
Sukabumi District	42.22	248962	35.54	208018	23.32	191772	19.74
Cianjur District	53.13	130710	23.60	236329	26.50	186933	19.24
Bandung District	64.52	336764	40.34	225946	25.33	241003	24.81
Garut District	52.78	153603	29.92	294155	32.98	157768	16.24
Tasikmalaya District	51.28	134527	27.47	220865	24.76	224089	23.06
Ciamis District	65.38	140736	23.97	233305	26.16	195370	20.11
Kuningan District	51.61	163370	22.63	235464	26.40	210765	21.69
Cirebon District	55.00	191339	30.88	340289	38.15	441753	45.47
Majalengka District	65.38	211120	30.24	283368	31.77	318086	32.74
Sumedang District	53.85	197164	24.96	331468	37.17	322818	33.23
Indramayu District	64.52	126045	20.03	202516	22.71	207142	21.32
Subang District	53.33	255076	29.19	288419	32.34	254547	26.20
Purwakarta District	75.00	334324	32.44	264565	29.66	459989	47.35
Karawang District	68.97	285775	33.41	244085	27.37	233161	24.00
Bekasi District	65.22	492691	42.15	232800	26.10	528705	54.42
West Bandung District	81.25	206333	34.09	240959	27.02	189838	19.54
Pangandaran District	80.00	201413	25.71	290843	32.61	274896	28.29
Bogor City	83.33	151570	11.44	236956	26.57	577822	59.47
Sukabumi City	100.00	282980	27.99	300168	33.66	461592	47.51
Bandung City	62.07	1058837	73.84	298394	33.46	810775	83.45
Cirebon City	100.00	134414	16.23	323394	36.26	285615	29.40
Bekasi City	100.00	307722	21.45	372212	41.73	916790	94.36
Depok City	81.82	434811	28.92	224152	25.13	648399	66.74
Cimahi City	100.00	176360	15.29	258562	28.99	557129	57.34
Tasikmalaya City	100.00	366189	37.97	279177	31.30	429757	44.23
Banjar City	100.00	189586	23.20	276830	31.04	299573	30.83
West Java Province	62.05	272839	30.11	265057	29.72	366675	37.74

In Table 5, the value of the goodness of the estimation model is given, in the form of the value of the standard deviation/ RMSE and the CV. The SAE-AR1-ME estimator produces an estimator model looks better as a measure of the goodness of the model. The estimator of the results of the SAE-AR1-ME model as a whole has an average standard deviation/RMSE and an average CV that is smaller than the estimator of the SAE-AR1 model and its direct estimator have. The SAE-AR1-ME model produces more efficient and more homogeneous estimates than the SAE-AR1 model and its direct estimator have.

It was previously stated that the September *Susenas* was only able to produce estimates at the provincial level due to the small sample sizes. By using the Small Area Estimation method that examines measurement errors in the auxiliary variables (the results of the March *Susenas*), it is possible to produce estimates at the provincial and district/city levels with a measure of the tested goodness of the estimation model. Therefore, the use of survey data with limited samples can be wider in scope, one of which is by using a Small Area Estimation method with measurement errors in the auxiliary variables.

7. CONCLUSION

The Small Area Estimation for the autoregressive model with measurement error in the auxiliary variable (SAE-AR1-ME) has a likely smaller mean $\overline{\text{RB}}$ compared to the Small Area Estimation for the autoregressive model (SAE AR1). The SAE-AR1-ME is also an estimator as efficient as the SAE-AR1 estimator with the mean of the root mean square error ($\overline{\text{RMSE}}$) as low as the estimator small area for the autoregressive model (SAE-AR1) when each variance of the auxiliary variable and the variance of area random effect is small. The SAE-AR1-ME is an estimator as reliable as the SAE-AR1 estimator, which is indicated by the average value of the coefficient of variance ($\overline{\text{CV}}$) as low as the SAE-AR1 in the population originating from a variance of auxiliary variables with small and medium values of measurement errors.

The development model application to the 2015 *Susenas* and 2014 Village Potential data shows that the SAE-AR1-ME produces an estimate better than the SAE-AR1 does. The SAE-AR1-

ME estimator model has an RMSE smaller than the SAE-AR1 estimator model's. This shows that the SAE-AR1-ME estimator has a variety of estimator values closer to the diversity of the actual observed values than the SAE-AR1 model's.

The SAE-AR1-ME model has a mean CV smaller than the SAE-AR1's, which shows that the SAE-AR1-ME model is a more reliable estimator compared to SAE-AR1 model. The results of the application of two development models to the actual data, indicate that the use of the SAE-AR1-ME, is better than the use of the SAE-AR1.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] D. Thorburn, Model based estimation in survey sampling of lognormal distribution. In: G. Rosenqvist, K. Juselius, K. Nordström, J. Palmgren, (eds). A spectrum of statistical thought, Essays in statistical theory, economics and population genetics in honor of Johan Fellman. Helsinki: Swedish School of Economics and Business Administration; 1991. p. 228–243.

[2] E. Tanur, A. Kurnia, K.A. Notodiputro, et al. Estimation of small area means for subsample repeated measurement data, IOP Conf. Ser.: Earth Environ. Sci. 187 (2018) 012033. https://doi.org/10.1088/1755-1315/187/1/012033.

[3] E. Tanur, A. Kurnia, K.A. Notodiputro, et al. Log-normal small area estimation with measurement error – an application for estimating household consumption expenditure, ICIC Express Lett. 15 (2021), 449-456. https://doi.org/10.24507/icicel.15.05.449.

[4] F. Karlberg, Population total prediction under a lognormal superpopulation model, Metron, LVIII (2000), 53–80.

[5] G.E. Battese, R.M. Harter, W.A. Fuller, An error component model for prediction of county crop areas using survei and satellite data, J. Amer. Stat. Assoc. 83 (1988), 28–36.

[6] G.S. Datta, P. Lahiri, T. Maiti, Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data, J. Stat. Plan. Inference. 102 (2002), 83–97.

https://doi.org/10.1016/s0378-3758(01)00173-2.

[7] J.N.K Rao, Small area estimation, John Wiley and Sons, New York, (2003).

[8] J.N.K. Rao, M. Yu, Small-area estimation by combining time-series and cross-sectional data, Can. J. Stat. 22 (1994), 511–528. https://doi.org/10.2307/3315407.

[9] L.M.R. Ybarra, S.L. Lohr, Small area estimation when auxiliary information is measured with error, Biometrika. 95 (2008), 919–931. https://doi.org/10.1093/biomet/asn048.

[10] M.D. Esteban, D. Morales, A. Pérez, et al. Small area estimation of poverty proportions under area-level time models, Comput. Stat. Data Anal. 56 (2012), 2840–2855. https://doi.org/10.1016/j.csda.2011.10.015.

[11] M. Ghosh, K. Sinha, D. Kim, Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models, Scand. J. Stat. 33 (2006), 591–608. https://doi.org/10.1111/j.1467-9469.2006.00492.x.

[12] M. Torabi, G.S. Datta, J.N.K. Rao, Empirical bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates, Scand. J. Stat. 36 (2009) 355–369. https://doi.org/10.1111/j.1467-9469.2008.00623.x.

[13] N.G.N. Prasad, J.N.K. Rao, The estimation of mean squared errors of small area estimators, J. Amer. Stat. Assoc. 85 (1990), 163-171.

[14] P.P. Biemer, R.M. Groves, L.E. Lyberg, et al. Measurement error in surveys, John Wiley & Sons (2004), Hoboken.

[15] R.J. Carroll, D. Ruppert, L.A. Stefanski, et al. Measurement error in nonlinear models: A modern perspective. 2nd edition. Chapman & Hall, New York, (2006).

[16] S. Makridakis, A. Andersen, R. Carbone, et al. The accuracy of extrapolative (time series methods): Result of a forecasting competition, J. Forecast. 1 (1982), 111-153.

[17] S. Muchlisoh, A. Kurnia, K.A. Notodiputro, et al. Small area estimation of unemployment rate based on unit level model with first order autoregressive time effects, J. Appl. Probab. Stat. 12 (2017), 65-77.

[18] Statistics Indonesia, Book 1 Susenas September 2015, Jakarta, BPS (2015).

[19] W.A. Fuller, R.M. Harter, The multivariate components of variance model for small area estimation. In: R. Platek, J.N.K. Rao, C.E. Sarndal, M.P. Singh, (eds). Small Area Statistics, 103-123. Wiley, New York, (1987).