International Journal on
Advanced Science
Engineering
Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Small Area Estimation with Measurement Error in t Distributed Covariate Variable

Soni Hariyanto[a,b], Khairil Anwar Notodiputro [a,1], Anang Kurnia[a,2], Kusman Sadik[a,3]

[a] *Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Bogor 16680, Indonesia*
E-mail: [1]khairil@apps.ipb.ac.id; [2]anangk@apps.ipb.ac.id; [3]kusmansadik@gmail.com

[b] *Statistics-Indonesia, Jalan Dr Sutomo 6-8 Jakarta Pusat 10710, Indonesia*
E-mail: soni.hariyanto@gmail.com

Abstract— **The large need for small area data and limited auxiliary information drive the development of small area estimation methods with auxiliary information comes from survey data. The consequence of the existence of the auxiliary variables from survey data is the development of measurement error models. Survey data is used as auxiliary variables that are taken randomly so that the data is considered to be stochastic. Thus, the measurement error model is assumed to be structural. Meanwhile, auxiliary information or covariates does not always have a normal distribution but sometimes contain outliers so the assumption of the t-distribution is considered to be more appropriate. Therefore, we use the moment-method to estimate the parameters and develop an empirical Bayes-EB predictor in a nested error regression model with measurement errors in the area-level covariates. In addition, the covariate in this model is assumed in the t-distribution which were previously always considered normal. Using simulation studies, we can report the performance of EB predictor under true covariates and measurement errors assumed to be t-distributions based on mean squared prediction errors (EMSPE). The results show that the model we developed leads to a significant increase in efficiency compared to EB predictors previously proposed. Furthermore, this approach is applied in National socio-economic survey (Susenas) data in Malang Regency with the aim of predicting mean years of schooling by districts using monthly per capita household expenditure data as the covariate variables that are considered to have the t-distribution.**

Keywords— **structural measurement error in covariate; empirical bayes predictor; nested error regression; t-distribution.**

I. INTRODUCTION

In modern era, the need of complete and up to date data and information are getting higher. Problems in developing countries such as Indonesia are the data and information that can satisfy those needs are not available. Data can be obtained from census or registration and surveys. The problem of registration data is they are incomplete and less recent information. Census data can be obtained once in every ten years. Meanwhile, the most recent and complete survey data are very limited since the sample was designed only for national or provincial level. The cost and time efficiency are the common reasons why the survey is not designed to estimate the smaller area [1]. One way to overcome this problem is small area estimation method [2]. The small area is not only limited to the administrative area but also to social demographic groups and so on. Principally, small area estimation methods use the small sample data (even no sample at all) to estimate the small areas by buying the related information robustness of the area as auxiliary information [3]. Since the basic assumption that must be satisfied is that there is an error in auxiliary information (also called the covariate variable in the model), the data used is census or registration data. Considering the census or registration data that are incomplete and less recent, a small area estimation method using survey data as auxiliary information was developed. The consequence of the use of survey data as the auxiliary information is that it should contains measurement errors. Measurement errors is defined as the difference in value between measurement value and the actual values in the survey or experiment [4]. Therefore, a small area estimation method with measurement errors on the covariate variables was developed. The development review of small area estimation model with measurement errors can be seen in previous study [5], [6]. This method is expected to answers the problem in Indonesia that is the big need of small area data and the limited availability of complete and up to date auxiliary information.

One of the models that is developed in the small area estimation method is the nested error regression model [7], [8]. The nested regression model, known as the unit level model requires the auxiliary variable in the unit level and does not contain measurement errors. There are 3 prediction methods in the small estimation area model: Best Linear Unbiased Prediction (BLUP), Empirical Bayes (EB), and Hierarchical Bayes (HB). The small area estimation model using the nested error regression model with measurement error on the auxiliary variable firstly introduced by Ghosh, Sinha, and Kim (GSK) [9]. The method used is EB and HB which is used to estimate the small area means where the covariate variable assumed random and normally distributed in the error measurement model and called structural measurement error model. Covariate variables are available in the area level, and measurement error variance is unknown. The EB estimation method used is based on, which applies to the finite population [10]. The deficiency of the GSK model is that the estimation model only considers the response variables without considering the covariate variables. Furthermore, the same model and EB method with a finite population by assuming the measurement error variance are known and covariate variables available at the unit level [11]. On the other hand, a study developed nested error regression model with measurement error which assume covariate variable in the area level but the estimation model considering the covariate variable and called Torabi, Datta, and Rao (TDR) model [12]. The method of moment is used in parameter estimation while the prediction uses EB. This model is more efficient than the GSK model.

Another development of the nested error regression model with measurement errors on the covariate variables with the HB method was carried out [13]. The difference lies in the prior determination of the estimation parameter model, which called GSK model by using prior Inverse Gamma (IG). This study uses Jeffrey's prior, which is considered more suitable for the model used, and the explanation is more acceptable for official statistics. The Bayes pseudo-Empirical method was developed in the nested regression model with measurement errors in the covariate variable [14] of the TDR base model, which develop using sample survey weighting. The results show that the method used is consistent with the increase in sample size. The nested error regression model with measurement errors in the covariate variables was also developed under the conditions of the binomial spread response variable [15], [16]. This model is also called the unit level logistics model with measurement errors on the auxiliary variables. The parameter estimation model used is the Laplace approach in the maximum likelihood method, and the prediction method used the Minimum Mean Square Error (MMSE) method.

In the previous model, the nested regression error model with measurement errors in the covariate variables still used one covariate. The previous study [17] investigated a model that uses more than one covariate variable, both containing or not containing measurement errors and their effects on predictions produced by the EB method using the method of moment as parameter estimation. Covariates variables which do not contain measurement errors can be derived from other survey data with larger sample sizes. The result is that the EB predictor is more efficient when the sample size of the auxiliary variable which is assumed contain no measurement error is greater.

All the studies that have been discussed assumed that the covariate variable is normally distributed but it is different in reality. One known distribution in modeling is t-distribution. This distribution is used when there is outlier's data or if the data has long tailed distribution [18], [19]. The subject of this paper is to examine the small area estimation under a unit-level model with measurement errors in the covariate variables which has t-distribution. This study is expected to solve the problem of auxiliary information using survey data as a covariate variable which contains outlier with a unit-level small area estimation model. The model is developed based on the TDR model. Furthermore, this paper is structured as follows: Chapter II Material and Method explains the TDR model and TDR model with auxiliary variables which have t-distribution or *t*-TDR model. Chapter III Results and Discussion explains the simulation and application of data to test the goodness of the basic and developed model. The general conclusion of the results of this study is explained in Chapter IV.

II. Material and Method

Consider a finite population, there are m area labelled $1, \dots, m$ and let N_d denote the known population size of the dth area. We denote by y_{dk} the response of the kth unit in the dth area $(k = 1, \dots, N_d; d = 1, \dots, m)$. A sample of size n_d is drawn from the dth area. Without loss of generality, we denote the sampled units by $1, \dots, n_d$ $(d = 1, \dots, m)$. Throughout, we will use the notations $y_d^{(s)} = \left(y_{d1}, \dots, y_{dn_d}\right)^T$, $y_d^{(r)} = \left(y_{dn_{d+1}}, \dots, y_{dN_d}\right)^T$, $y_d^T = \left(y_d^{(s)^T}, y_d^{(r)^T}\right)$, and $X_d^{(s)} = \left(X_{d1}, \dots, X_{dn_d}\right)^T$ with $y_d^{(s)}$ and $X_d^{(s)}$ corresponding to the sample unit and $y_d^{(r)}$ corresponding to the non-sample unit. The basic problem in finite population sampling is inference about $y_d^{(r)}$ conditional on $y_d^{(s)}$ and $X_d^{(s)}$ [10]. More specifically, we are interested in the prediction of finite population means

$$\gamma_d = \frac{1}{N_d} \sum_{k=1}^{N_d} y_{dk} \qquad (d = 1, \dots, m)$$

given the data.

A. Empirical Bayes in TDR Model

According to [12], we assume the superpopulation model

$$y_{dk} = b_0 + b_1 x_d + v_d + e_{dk} \ \ (k = 1, \dots, N_d) \qquad (1)$$

$$X_{dk} = x_d + \delta_{dk} \qquad (d = 1, \dots, m) \qquad (2)$$

where y_{dk}, $(d = 1, \dots, m; k = 1, \dots, N_d)$ is the response variable of concern of the *k*th unit in the *d*th area, X_{dk} is the covariate variable which is assumed to be linearly related and is the result of the survey, x_d is the actual area level covariate variable but it is unknown, δ_{dk} is the measurement error, (b_0, b_1) is the regression coefficient, v_d is the random effect area, and e_{dk} is the model error. It is assumed that x_d, v_d, e_{dk} and δ_{dk} are mutually independent with $x_d \overset{i.i.d}{\sim} N(\mu_x, \sigma_x^2)$, $v_d \overset{i.i.d}{\sim} N(0, \sigma_v^2)$, $e_{dk} \overset{i.i.d}{\sim} N(0, \sigma_e^2)$ and $\delta_{dk} \overset{i.i.d}{\sim} N(0, \sigma_\delta^2)$. The available data consist of $(y_{dk}, X_{dk}), (k = 1, \dots, n_d; d = 1, \dots, m)$. Also, we write $\phi = (b_0, b_1, \mu_x, \sigma_v^2, \sigma_e^2, \sigma_\delta^2, \sigma_x^2)^T$. An alternative way to express is

(i) $y_{dk} = \theta_d + e_{dk}$ $(d = 1, \dots, m; k = 1, \dots, n_d)$ where e_{dk} are i.i.d $N(0, \sigma_e^2)$

(ii) $\theta_d = b_0 + b_1 x_d + v_d \ (d = 1, \dots, m)$ where v_d are i.i.d $N(0, \sigma_v^2)$

(iii) $X_{dk} = x_d + \delta_{dk} \ (k = 1, \dots, n_d; d = 1, \dots, m)$ where δ_{dk} are i.i.d $N(0, \sigma_\delta^2)$ and $x_d \sim N(\mu_x, \sigma_x^2)$

In this way, it is possible to identify equation (1) and equation (2) as a Bayesian model.

From TDR model, the Bayes predictor of γ_d is linear function of $y_d = \left(y_d^{(s)^T}, y_d^{(r)^T}\right)^T$ and the independence sample of (y_d, X_d), $d = 1, \dots, m$ for known ϕ, we can get Bayes predictor $\hat{\gamma}_d^B = E\left(\gamma_d \middle| y_d^{(s)}, X_d^{(s)}, \phi\right)$ based on the sample data by first deriving the Bayes predictor of $y_d^{(r)}$ given $y_d^{(s)}, X_d^{(s)}$ and ϕ.

Under the nested error model given by equation (1) and equation (2) the Bayes predictor of γ_d is given by

$$\hat{\gamma}_d^B = (1 - h_d A_d)\bar{y}_d$$
$$+h_d A_d(b_0 + b_1\mu_x) + h_d A_d \left\{\frac{n_d \sigma_x^2}{\sigma_\delta^2 + n_d \sigma_x^2}\right\} b_1(\bar{X}_i - \mu_x) \quad (3)$$

Then the posterior variance of γ_d is

$$V(\hat{\gamma}_d^B) = h_d^2 A_d \left\{b_1^2 \sigma_x^2 + \sigma_v^2 - \frac{n_d b_1^2 \sigma_x^4}{\sigma_\delta^2 + n_d \sigma_x^2}\right\} + \frac{1}{N_d} h_d \sigma_e^2 \quad (4)$$

where

$$\bar{X}_d = n_d^{-1} \sum_{k-1}^{n_d} X_{dk}; \ A_d = \frac{\sigma_e^2(\sigma_\delta^2 + n_d \sigma_x^2)}{n_d b_1^2 \sigma_x^2 \sigma_\delta^2 + (n_d \sigma_v^2 + \sigma_e^2)(\sigma_\delta^2 + n_d \sigma_x^2)} \quad (5)$$

and $\bar{y}_d = n_d^{-1} \sum_{k-1}^{n_d} y_{dk}$, $h_d = (N_d - n_d)/N_d$ with h_d is the finite population correction fraction.

The EB predictor $\hat{\gamma}_d^{EB}$ γ_d is obtained by replacing ϕ in equation (4) by a consistent estimator $\hat{\phi}$. We use the method of moments estimators $\hat{\phi}$. A consistent estimator $\hat{A}_d$ A_d is obtained from formula (5) for A_d by replacing ϕ by $\hat{\phi}$. The EB predictor γ_d is given by

$$\hat{\gamma}_d^{EB} = \left(1 - h_d \hat{A}_d\right)\bar{y}_d$$
$$+h_d \hat{A}_d\left(\hat{b}_0 + \hat{b}_1 \bar{X}\right) + h_d \hat{A}_d \left\{\frac{n_d \hat{\sigma}_x^2}{\hat{\sigma}_\delta^2 + n_d \hat{\sigma}_x^2}\right\} b_1(\bar{X}_d - \bar{X}) \quad (6)$$

B. *Empirical Bayes in t-TDR Model*

From the model in equation (1) and equation (2), we developed a new approach for the TDR model by assuming covariate variables with a t-distribution measurement error or called the *t*-TDR model. The *t*-TDR model assumed that x_d , v_d e_{dk} and δ_{dk} are mutually independent with $x_d \overset{i.i.d}{\sim} t(\mu_x, \sigma_x^2, p)$ $v_d \overset{i.i.d}{\sim} N(0, \sigma_v^2)$, $e_{dk} \overset{i.i.d}{\sim} N(0, \sigma_e^2)$ and $\delta_{dk} \overset{i.i.d}{\sim} N(0, \sigma_\delta^2, q)$. The difference of the *t*-TDR model is the assumption of the actual covariate variable (x_d) and the measurement error variable that follows the t-distribution. Based on the nature of the variance in t-distribution, estimation of the variance multiplied by the degrees of freedom minus 2. Therefore, in this case, the t-distribution used in the model is minimum with 3 degrees of freedom [18]. Furthermore, the posterior distribution compiler, which was previously assumed to follow the normal distribution, in the *t*-TDR model, the distribution of *x* and δ is changed to the t-distribution. For known ϕ, the Bayes predictor of γ_d in *t*-TDR model given by:

$$\hat{\gamma}_d^B = E\left(\gamma_d \middle| y_d^{(s)}, X_d^{(s)}, \phi\right)$$
$$= (1 - h_d A_d)\bar{y}_d + h_d B_d(b_0 + b_1\mu_x)$$
$$+h_d B_d \left\{\frac{n_d\left(\frac{p}{p-2}\right)\sigma_x^2}{\left(\frac{q}{q-2}\right)\sigma_\delta^2 + n_d\left(\frac{p}{p-2}\right)\sigma_x^2}\right\} b_1(\bar{X}_i - \mu_x) \quad (7)$$

Then the posterior variance of γ_d is

$$V\left(\gamma_d \middle| y_d^{(s)}, X_d^{(s)}, \phi\right)$$
$$= h_d^2 B_d \left\{ b_1^2\left(\frac{p}{p-2}\right)\sigma_x^2 + \sigma_v^2 - \frac{n_d b_1^2\left(\frac{p}{p-2}\right)^2 \sigma_x^4}{\left(\frac{q}{q-2}\right)\sigma_\delta^2 + n_d\left(\frac{p}{p-2}\right)\sigma_x^2} \right\}$$
$$+\frac{1}{N_d} h_d \sigma_e^2 \quad (8)$$

where

$$B_d = \frac{\sigma_e^2\left(\left(\frac{q}{q-2}\right)\sigma_\delta^2 + n_d\left(\frac{p}{p-2}\right)\sigma_x^2\right)}{n_d b_1^2\left(\frac{p}{p-2}\right)\sigma_x^2\left(\frac{q}{q-2}\right)\sigma_\delta^2 + (n_d\sigma_v^2 + \sigma_e^2)\left(\left(\frac{q}{q-2}\right)\sigma_\delta^2 + n_d\left(\frac{p}{p-2}\right)\sigma_x^2\right)} \quad (9)$$

We use the method of moments estimators, $\hat{\phi}$, proposed by TDR. Let $SSW_X = \sum_{d=1}^{m} \sum_{k}^{n_d} (X_{dk} - \bar{X}_d)^2$ $SSW_y = \sum_{d=1}^{m} \sum_{k}^{n_d} (y_{dk} - \bar{y}_d)^2$, and $MSW_X = SSW_X/(n_T - m)$ $MSW_y = SSW_y/(n_T - m)$ $n_T = \sum_{d=1}^{m} n_d$ is the total sample size. Then σ_e^2 and σ_δ^2 consistently estimated by

$$\hat{\sigma}_\delta^2 = MSW_X; \ \ \hat{\sigma}_e^2 = MSW_y \quad (10)$$

Further, b_0, b_1, and μ_x are consistently estimated by

$$\hat{b}_1 = \frac{\sum_{d=1}^{m} n_d \bar{y}_d(\bar{X}_d - \bar{X})}{(MSB_X - MSW_X)(m-1)} \quad (11)$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{X}; \quad \mu_x = \bar{X} \quad (12)$$

where $\bar{X} = n_T^{-1} \sum_{d=1}^{m} n_d \bar{X}_d$; $\bar{y} = n_T^{-1} \sum_{d=1}^{m} n_d \bar{y}_d$ and $MSB_X = (m-1)^{-1} \sum_{d=1}^{m} n_d (\bar{X}_d - \bar{X})^2$.

The remaining parameters σ_x^2 and σ_v^2 are consistently estimated by

$$\hat{\sigma}_x^2 = \max\left\{0, \frac{(m-1)}{g_m}(MSB_X - MSW_x)\right\} \quad (13)$$

$$\hat{\sigma}_v^2 = \max\left\{0, \frac{(m-1)}{g_m}\left(MSB_y - MSW_y\right) - \hat{b}_1^2 \hat{\sigma}_x^2\right\} \quad (14)$$

where $MSB_y = (m-1)^{-1} \sum_{d=1}^{m} n_d (\bar{y}_d - \bar{y})^2$ and $g_m = n_T - \sum_{d=1}^{m} \frac{n_d^2}{n_T}$.

The consistent estimator $\hat{B}_d$ B_d is obtained from equation (9) for B_d by replacing ϕ by $\hat{\phi}$. The EB predictor γ_d is given by

$$\hat{\gamma}_d^{EB} = \left(1 - h_d \hat{B}_d\right)\bar{y}_d + h_d \hat{B}_d\left(\hat{b}_0 + \hat{b}_1 \bar{X}\right)$$
$$+h_d \hat{B}_d \left\{\frac{n_d\left(\frac{p}{p-2}\right)\hat{\sigma}_x^2}{\left(\frac{q}{q-2}\right)\hat{\sigma}_\delta^2 + n_d\left(\frac{p}{p-2}\right)\hat{\sigma}_x^2}\right\} b_1(\bar{X}_d - \bar{X}) \quad (15)$$

III. Result and Discussion

In this section, the simulation is conducted by assuming that x and δ have t-distribution with zero mean value, variance σ_x^2, σ_δ^2 and degree of freedom p and q. For the application, data of the National socio-economic survey (Susenas) in Malang Regency March 2015 is used to obtain data of mean years of schooling by sub-district based on the direct estimator, TDR model, and *t*-TDR model. The number of estimated sub-district is 33 sub-districts.

A. Simulation

In this paper, the simulations are conducted to prove the superior performance of the proposed model and compare it to the basic model. The EB predictor of the TDR model in equation (6) and the proposed model called the *t*-TDR model in equation (15). The comparisons are also associated with the sample size (n_d). Previously, a finite population are generated with population size of 2400 which are divided into 18 areas with sizes 50, 250, 50, 100, 200, 150, 50, 150, 100, 150, 100, 50, 250, 200, 150, 50, 200, and 150. First, set the parameter values b_0, b_1, σ_e^2, σ_v^2, σ_δ^2, σ_x^2. The variable response population y_{dk} is obtained from data generation with b_0=100, b_1=2, σ_e^2=10, σ_v^2=100, σ_δ^2=225, μ_x=10 dan σ_x^2=9, with degrees of freedom p=q=3. A sample of 2 percent of each population is taken with simple random sampling to generate pairs of sample data of 18 groups. Thus, for each group (area), there are sample sizes (n_d) of 1, 5, 1, 2, 4, 3, 1, 3, 2, 3, 2, 1, 5, 4, 3, 1, 3, 2. To find the effect of sample size, besides using this sample size (n_d), sampling is also done 5 times $(5n_d)$ for each group (area).

A data set of B=5000 that is mutually independent and has a normal distribution is generated for $\left\{v_d^{(b)}; d = 1, \dots, m\right\}$, $\left\{e_{dk}^{(b)};\ k = 1, \dots, N_d\right\}$ with zero mean value and determined variance is σ_v^2 and σ_e^2. Furthermore, also generated B=5000 mutually independent set for $\left\{x_d^{(b)}; d = 1, \dots, m\right\}$ which has t-distribution with determined mean value is μ_x, variance σ_x^2, and degree of freedom p and $\left\{\delta_{dk}^{(b)}; k = 1, \dots, N_d\right\}$ has t-distribution with zero mean value and variance σ_δ^2 and degree of freedom q. From the generated data, obtained B=5000 data sets $\left\{\left(y_{dk}^{(b)}, X_{dk}^{(b)}\right); d = 1, \dots, m; k = 1, \dots, N_d\right\}$, which are obtained from the equation: $y_{dk} = b_0 + b_1 x_d + v_d + e_{dk}$ and $X_{dk} = x_d + \delta_{dk}$. Next, the mean value of b the population obtained from

$$\gamma_d^{(b)} = \frac{1}{N_d} \sum_{k=1}^{N_d} y_{ij}^{(b)}$$

For each population, the sample $\left\{\left(y_{dk}^{(b)}, X_{dk}^{(b)}\right); d = 1, \dots, m; k = 1, \dots, n_d\right\}$ is taken with sample sizes n_d and $5n_d$ with simple random sampling. The model parameter ϕ is estimated using the method of the moment based on the formula (10)-(14) and for each b $\hat{\gamma}_i^{TDR}$ dan $\hat{\gamma}_i^{t-TDR}$ are estimated. To compare the performance TDR model and *t*-TDR model, empirical Mean Square Prediction Error (EMSPE) calculated by the formula:

$$Empirical\ MSPE(\hat{\gamma}_d^{TDR}) = \frac{1}{B} \sum_{b=1}^{5000} \left(\hat{\gamma}_d^{TDR(b)} - \gamma_d^{(b)}\right)^2$$

$$Empirical\ MSPE(\hat{\gamma}_d^{tTDR}) = \frac{1}{B} \sum_{b=1}^{5000} \left(\hat{\gamma}_d^{t-TDR(b)} - \gamma_d^{(b)}\right)^2$$

Table 1 shows that when variables x and δ have the smallest sample size (n_d) and assumed have a *t*-distribution with a degree of freedom 3, the empirical MSPE value of the *t*-TDR model is better than TDR model for all areas. Table 1 also shows that substantially the t-TDR estimator is more efficient than the TDR model estimator with relative efficiency value ranging from 100.38% to 107.57%. The efficiency value is calculated from EMSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$)/EMSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$).

TABLE I

EMPIRICAL MSPE OF $\hat{\gamma}_d^{TDR}$ AND $\hat{\gamma}_d^{t-TDR}$

Area *d*	**n_d**	**EMSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$)**	**EMSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$)**
1	1	10.0892	9.7823
2	5	2.0370	1.9764
3	1	10.2135	10.0437
4	2	5.0989	5.0467
5	4	2.4695	2.4545
6	3	3.3941	3.3434
7	1	10.5041	10.1479
8	3	3.1974	3.1690
9	2	5.0187	4.9882
10	3	3.3609	3.3123
11	2	4.9268	4.8804
12	1	10.0517	9.9514
13	5	1.9306	1.9209
14	4	2.4822	2.4599
15	3	3.3127	3.2793
16	1	11.1073	10.3254
17	3	2.4581	2.4488
18	2	3.3296	3.3029

Fig. 1 shows the effect of an increase in the *t*-TDR model with a degree of freedom 3 from the sample size n_d to $5n_d$. The additional sample size to the *t*-TDR model substantially more efficient when we use a larger sample size $(5n_d)$ than the same sample size (n_d). Simulation results show that the increase of the sample size can reduce the value of EMSPE between 73.16% to 79.93% or an average of 77.58%.

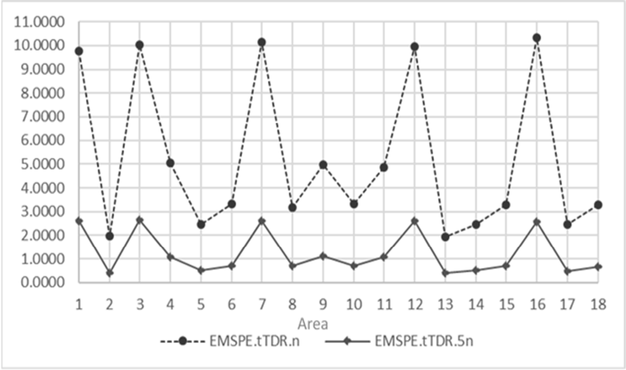
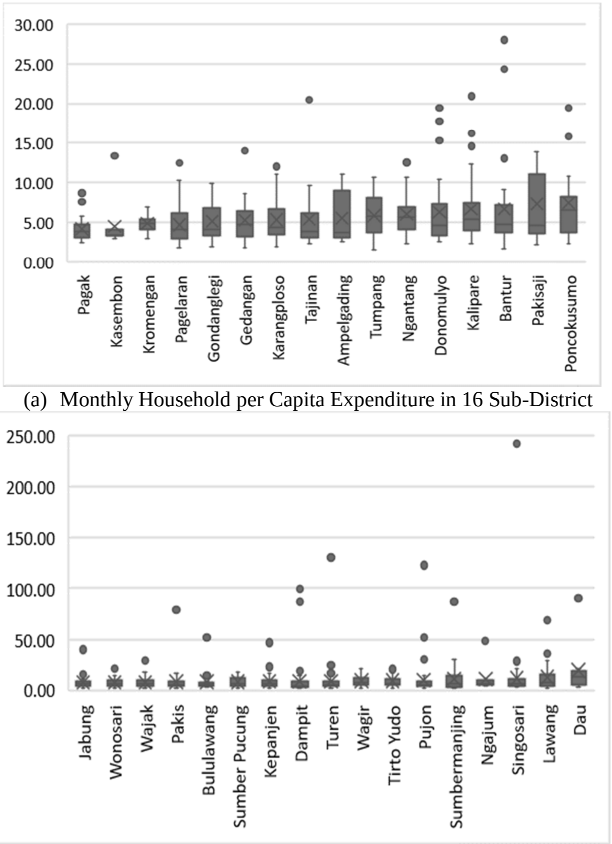


Fig. 1 Empirical MSPE of t-TDR Model with a degree of freedom 3 and sample size n_i and $5n_i$

B. Application

The variables used in the model are mean years of schooling (*y*) and monthly per capita expenditure for consumption (*X*). Both are used in March 2015 Susenas conducted by Badan Pusat Statistik (BPS). Susenas data collection is conducted 2 times a year in March and September. The survey sampling of March Susenas is designed to estimate the parameter at the regency/municipality level while the September Susenas is designed to estimate the parameter at the provinces level. The research observation unit is the household, and the area level to be estimated is the sub-district level. We select the March 2015 period of Susenas because of the availability of small samples to estimate sub-district levels for each sub-district and the availability of complete data of sub-district household populations required by the small area estimation model. Since the model relates to the measurement errors in the covariate variable (X) and the monthly per capita expenditure data, which becomes the covariate variable is obtained from the respondent interview, and it is not coming from the measurement results, it is assumed that there is a measurement error. In addition, since the monthly per capita expenditure for consumption is obtained from survey data by selecting a random sample, the measurement error model is assumed to be the structural model. This research covers all sub-districts (33 sub-districts) in Malang Regency, East Java Province.

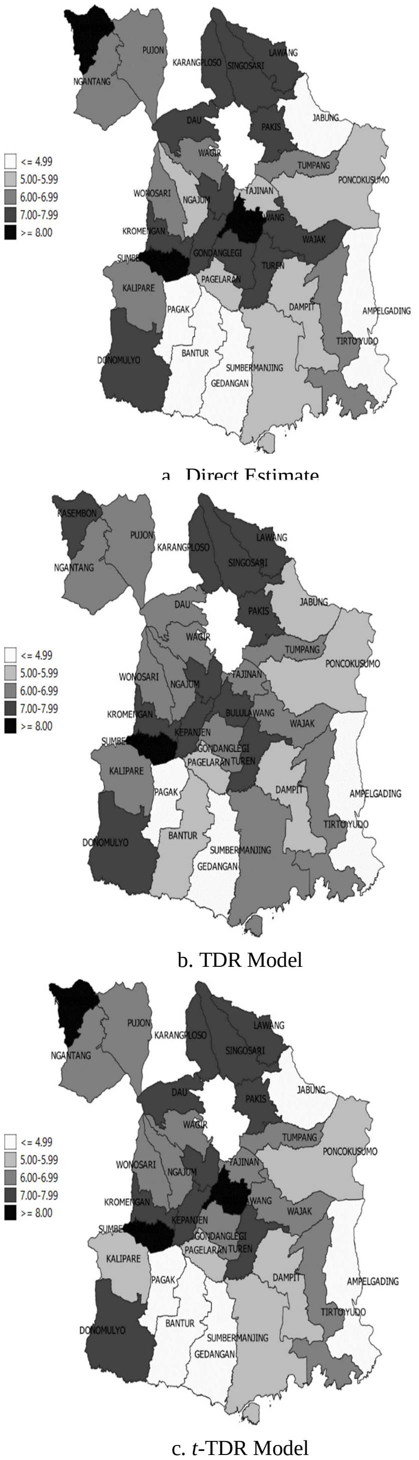


(a) Monthly Household per Capita Expenditure in 16 Sub-District

(b) Monthly Household per Capita Expenditure in 17 Sub-District

Fig. 2 Monthly Household Per capita Expenditure by Sub-District in Malang Regency, 2015.

Mean years of schooling of population aged 25 years and above is one indicator of the compilation of the Human Development Index (HDI), and it is also one of the indicators that are monitored by Sustainable Development Goals (SDGs). This indicator describes equitable development in the education sector of a region. The higher mean years of schooling shows the success of the development of education. Therefore, this indicator is mostly needed until the smallest administrative area since it plays a vital role in evaluating equitable development in education. The small area data can show the area where the education level is still low. Thus, the right target policies can be formulated to accelerate the distribution of education. The direct estimator value of mean years of schooling is obtained by selecting the population aged 25 years and above of a region and then converting the highest level of education and the highest grade ever completed to the duration of education (years). Furthermore, mean years of schooling of populations aged 25 years and above is obtained by summing mean years of schooling of populations aged 25 years and above and dividing it by the number of populations aged 25 years and above in that area.



a. Direct Estimate

b. TDR Model

c. *t*-TDR Model

Fig. 3 Mean Years Schooling Prediction based on Direct Estimate, TDR Model, and *t*-TDR Model by Sub-District in Malang Regency, 2015.

TABLE II
MEAN YEARS OF SCHOOLING PREDICTION BASED ON DIRECT ESTIMATE, TDR MODEL, AND T-TDR MODEL BY SUB-DISTRICT IN MALANG REGENCY, 2015

Sub-District	Sample size	Predicted		
		Direct	$\hat{\gamma}_d^{TDR}$	$\hat{\gamma}_d^{t-TDR}$
1. Donomulyo	30	7.4478	7.3934	7.3932
2. Kalipare	30	6.0758	6.0442	5.9453
3. Pagak	20	4.9375	4.9033	4.4540
4. Bantur	29	4.9552	5.0157	4.6952
5. Gedangan	17	4.2791	4.8885	4.3482
6. Sumbermanjing	40	5.9036	6.0190	5.8968
7. Dampit	59	5.6149	5.3335	5.1693
8. Tirto Yudo	29	6.6774	6.5699	6.5547
9. Ampelgading	8	4.7143	4.9958	3.9600
10. Poncokusumo	29	5.7059	5.7227	5.5446
11. Wajak	30	7.1714	6.7337	6.7669
12. Turen	69	7.5364	7.5058	7.5960
13. Bululawang	20	8.4615	7.9585	8.3572
14. Gondanglegi	18	7.1429	6.5112	6.5185
15. Pagelaran	39	5.8434	5.8330	5.7401
16. Kepanjen	56	7.5447	7.3392	7.4294
17. Sumber Pucung	29	8.8154	8.0344	8.3394
18. Kromengan	10	7.5000	7.1994	7.5710
19. Ngajum	10	6.0000	6.4182	6.3013
20. Wonosari	19	6.4595	6.3081	6.2288
21. Wagir	47	6.3364	6.3684	6.3252
22. Pakisaji	30	7.6557	7.3970	7.5708
23. Tajinan	28	5.8525	6.1820	6.1208
24. Tumpang	30	6.7015	6.1768	6.1144
25. Pakis	57	7.5360	7.1881	7.2626
26. Jabung	30	4.9577	5.1202	4.8219
27. Lawang	34	7.6104	7.1693	7.2280
28. Singosari	65	7.8562	7.4663	7.5283
29. Karangploso	40	7.1190	7.0323	7.1321
30. Dau	9	7.3846	6.9824	7.1097
31. Pujon	50	6.6535	6.7344	6.7350
32. Ngantang	19	6.1667	6.2757	6.2024
33. Kasembon	10	8.2500	7.6643	8.2932

In contrast to the direct estimator, mean years of schooling data calculated by the model based on individual data. So that the mean years of schooling, in this case, is the number of mean years of schooling of each person aged 25 years and above in dth sub-district divided by the number person aged 25 years and above in dth sub-district. The auxiliary variable is selected from the monthly per capita expenditure for food consumption and non-food consumption in a month. This data is obtained by summing the total household expenditure for food consumption and non-food consumption in a month divided by the number of household members. The selection of auxiliary variables based on the theory that the expenditure reflects the level of household welfare. Higher household expenditure shows a higher level of household welfare. Furthermore, higher welfare is positively correlated with the education improvement of household members. High education is reflected by the higher mean years of schooling of the household members. In addition, per capita, household expenditure tends to have a very spread distribution. In other words, this data has outliers data or long-tailed distribution. Per capita expenditure model assumes that per capita expenditure data has t-distribution [20]. Therefore, the auxiliary variables used in this study have t-distribution. Fig. 2 shows that variable X is assumed to have t-distribution with a degree of freedom 3 considering the outliers in almost all sub-districts. Thus, Malang Regency is chosen to apply the model proposed in equations (1) and (2).

Table 2 and Fig. 3 shows mean years of schooling prediction obtained from the direct estimator, the EB estimator with the TDR model, and the EB estimator with the *t*-TDR model. The estimator value of the TDR model and *t*-TDR model has the same pattern. It means that the estimation value of the TDR model is higher than the direct estimator value, likewise with the *t*-TRD model, although the difference is relatively smaller. The prediction result shows that the highest mean years of schooling in Malang Regency with a direct estimate and *t*-TDR model are Sub-district Bululawang. Using the TDR model, it is obtained that the highest mean years of schooling in Malang Regency is Sub-district Sumber Pucung. The prediction result shows that the lowest mean years of schooling in Malang Regency with the direct estimate and TDR model is Sub-district Gedangan. Using the *t*-TDR model, it is obtained that the lowest mean years of schooling in Malang Regency is the Sub-district Ampelgading.

TABLE III
MEAN SQUARED PREDICTION ERROR (MSPE) FROM TDR MODEL AND T-TDR MODEL BY SUB-DISTRICT IN MALANG REGENCY, 2015

Sub-District	MSPE		
	Direct	$\hat{\gamma}_d^{TDR}$	$\hat{\gamma}_d^{t-TDR}$
1. Donomulyo	10.3353	0.0147	0.0180
2. Kalipare	7.5172	2.1399	0.0180
3. Pagak	12.5279	2.8212	0.0256
4. Bantur	9.1040	2.1908	0.0185
5. Gedangan	10.1107	3.1489	0.0295
6. Sumbermanjing	16.1613	1.7626	0.0142
7. Dampit	16.1432	1.3694	0.0103
8. Tirto Yudo	10.3205	2.1899	0.0185
9. Ampelgading	18.5143	5.0350	0.0588
10. Poncokusumo	10.9868	2.1926	0.0186
11. Wajak	13.0137	2.1413	0.0180
12. Turen	18.0903	1.2427	0.0091
13. Bululawang	14.4495	2.8235	0.0256
14. Gondanglegi	23.8908	3.0328	0.0281
15. Pagelaran	15.5483	1.7904	0.0145
16. Kepanjen	18.6271	1.4144	0.0108
17. Sumber Pucung	24.3404	2.1881	0.0185
18. Kromengan	26.1667	4.4174	0.0478
19. Ngajum	5.3333	4.4190	0.0478
20. Wonosari	23.7553	2.9199	0.0267
21. Wagir	11.6014	1.5822	0.0124
22. Pakisaji	17.6295	2.1412	0.0180
23. Tajinan	15.7945	2.2415	0.0191
24. Tumpang	22.1520	2.1409	0.0180
25. Pakis	16.9120	1.4000	0.0106
26. Jabung	11.5553	2.1408	0.0180
27. Lawang	28.5041	1.9673	0.0162
28. Singosari	24.3999	1.2907	0.0095
29. Karangploso	15.4796	1.7609	0.0142
30. Dau	17.3662	4.7068	0.0527
31. Pujon	15.2917	1.5181	0.0118
32. Ngantang	10.0447	2.9226	0.0268
33. Kasembon	9.5870	4.4149	0.0478

Table 3 reports the values of MSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$) based on equation (4), the value of MSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$) based on equation (8), and MSE of direct estimation based on standard error. It is obviously presented in Table 1 that MSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$) is substantially smaller than MSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$) and MSE of direct estimation. The reduction in MSPE by using $\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$ over $\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$ the range from 8.82% to 13.51% in 32 sub-districts. The different condition occurs in Sub-district Donomulyo since MSPE($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{t-TDR}}$) is greater than MSPE ($\hat{\boldsymbol{\gamma}}_{\boldsymbol{d}}^{\boldsymbol{TDR}}$) which is increased to 18.74%. Hence, the use of the assumption of t-distribution that considering outlier in covariate leads to significant improvement in efficiency relatively than using the normal distribution. In other words, since variable *X* is the household per capita expenditure whose distribution is relatively spread, the t-assumption is appropriate, and the performance of the *t*-TDR model is better than the TDR model, which assumes the normal distribution of variable *X*.

IV. Conclusions

Outlier data can lead to violations of the normality assumption. Outlier data may appear on covariate variables that are based on survey data, which are assumed to contain measurement errors. Outliers in a small area estimation model can come from the unit level. This research is successful in developing a small area estimation model with measurement errors on the t-distribution based covariate variable that can overcome the problem of outliers. Simulation studies using the EB method show that when the covariate variable contains measurement errors and has t-distribution, EMSPE small area estimation models with normal distribution-based measurement errors (TDR models) are greater than small area estimation models with t-distribution based measurement errors (*t*-TDR model). Therefore, in general, it is shown that the *t*-TDR model is more efficient than the TDR model. The application data used to predict mean years of schooling with monthly per capita expenditure for consumption as the covariate variables which is assumed contain measurement errors with a t-distribution. It also shows that the performance of the *t*-TDR model is better than the TDR model.

Acknowledgment

The authors thank Statistics-Indonesia for providing access to the Susenas 2015 data in Malang Regency. We are grateful to referees for careful checking of the details and for helpful comments which improve this paper.

References

[1] W. G. Cochran, Sampling Technique, 3rd Ed, New York: Wiley, 1977.

[2] J. N. K. Rao and I. Molina, *Small Area Estimation*, 2nd ed., New York: John Wiley & Sons, Inc., 2015.

[3] D. Pfeffermann, "Small Area Estimation-New Development and Directions," *International Statistical Review*, vol. 70 pp. 125-143, 2002.

[4] W. A. Fuller, *Measurement Error Models*, New York: John Wiley & Sons, Inc., 1987.

[5] D. Pfeffermann, "New Important Developments in Small Area Estimation," *Statist Science*, vol. 28 pp. 40-68, 2013.

[6] S. Arima, G. S. Datta, and B. Liseo, "Models in Small Area Estimation when Covariates are Measured with Error," in M. Pratesi (ed.), *Analysis of Poverty Data by Small Area Estimation*, New York: John Wiley & Sons Ltd, 2016.

[7] G. E. Battese, R. M. Harter, W.A. Fuller, "An Error Component Model for Prediction of Country Crop Area using Survey and Sattelite Data," *J. Amer. Statist. Assoc.* vol. 83, pp. 28-36, 1988.

[8] N. G. N. Prasad and J. N. K. Rao, "The Estimation of Mean Square Error of Small Area Estimators," *J. Amer Statist. Assoc*., vol. 85, pp. 163-171, 1990.

[9] M. Ghosh, K. Sinha, and D. Kim, "Empirical and Hierarchical Bayesian Estimation in Finite Population Sampling under Structural Measurement Error Models," *Scand. J. Statist*, vol. 33, pp. 591-608, 2006.

[10] M. Ghosh and G. Meeden, "Empirical Bayes Estimation in Finite Population Sampling," *J. Amer. Statist. Assoc.*, vol. 81, pp. 1058-1062, 1986.

[11] M. Torabi, "Some Contribution to Small Area Estimation" PhD thesis, Carleton University, Ottawa, Ontario, Canada, 2006.

[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. Statist.*, vol. 36, pp. 355-368, 2009.

[13] S. Arima, G. S. Datta, and B. Liseo, "Objective Bayesian Analysis of a Measurement Error Small Area Model," *Bayesian Anal.*, vol. 7, pp. 363-384, 2012.

[14] M. Torabi, "Small Area Estimation using Survey Weights under a Nested Error Linear Regression Model with Structural Measurement Error", *J. Multivariate Anal.*, vol. 109, pp. 52-60, 2012.

[15] A. L. Erciulescu, "Small Area Prediction based on Unit Level Models when the Covariate Mean is Measured with Error ", PhD thesis, Iowa State University, Ames, US, 2015.

[16] A. L. Erciulescu and W.A. Fuller, "Small Area Prediction under Alternative Model Specification", *Statist in Transit New Series and Survey Methodology Joint Issue: Small Area Estimation 2014*, vo. 17, pp. 9-24, 2016.

[17] G. S. Datta, M. Torabi, J. N. K. Rao, B. Liu, "Small Area Estimation with Multiple Covariates Measured with Errors: A Nested Error Linear Regression Approach of Combining Multiple Surveys", *J. Multivariate Anal.,* vol. 167, pp. 49-59, 2018.

[18] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust Statistical Modeling Using the t Distribution" *J. Amer. Statist. Assoc*. vol. 408, pp. 881-896, 1989.

[19] W. R. Bell and E. T. Huang, "Using the *t*-distribution to Deal with Outliers in Small Area Estimation," in *Proc. of Statist.* Canada Symposium on Methodological Issues in Measuring Population Health, 2006.

[20] A. Ubaidillah, K. H. Notodiputro, A. Kurnia, I.W. Mangku, "A Comparative Study of Robust t Linear Mixed Models with Application to Household Consumption Per Capita Expenditure Data," *Applied Mathematical Science*, vol. 12, pp. 57-68, 2018.