Journal of Applied Probability and Statistics

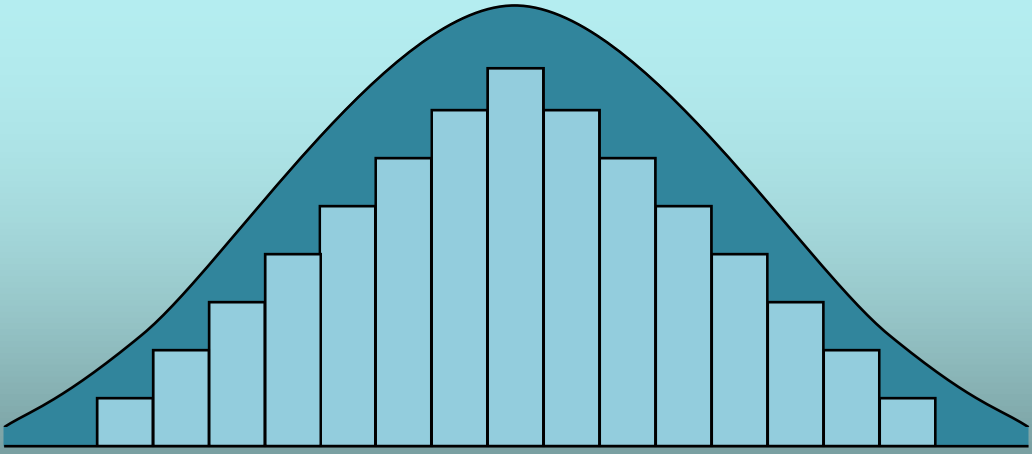
Volume 14 No. 3, December 2019

ISSN 1930-6792

Journal of Applied Probability & Statistics

Volume 14 Number 3 December 2019

Founding Chief Editor: Shahjahan Khan



ISOSS PUBLICATIONS
http://japs.isoss.net

JOURNAL OF APPLIED PROBABILITY & STATISTICS

Volume 14 Number 3 December, 2019

Contents

Journal of Applied Probability and Statistics
2019, Vol. 14, No. 3, pp. 75-93
Copyright ISOSS Publications 2019

ESTIMATION OF PER CAPITA HOUSEHOLD EXPENDITURE: A LIKELIHOOD APPROACH OF ROBUST EXTENSION OF SMALL AREA ESTIMATION

Cucu Sumarni[1,5], Kusman Sadik[2], Khairil Anwar Notodiputro[3], Bagus Sartono[4]
[1,2,3,4]*Departement of Statistics, Bogor Agricultural (IPB) University, Indonesia,*
[5]*BPS-Statistics, Indonesia*
Email: [1]*cucu_s@bps.go.id,*[2]*kusmansadik@gmail.com,*[3]*khairil@apps.ipb.ac.id,*
[4]*bagusco@gmail.com*

SUMMARY

Small area estimation (SAE) models have been widely used by statistician and policy maker to get small area statistics. The models based on normal distribu-tions (normal-SAE) may perform poorly in estimation when data contains out-liers. A robust method can be used to solve this, such as using Huber function or replacing normality assumption with t-distribution. The robust Huber can well overcome the area level outliers. However, the outliers arise from unit level or/and area level. Thus, we propose a robust SAE that handles these types of outliers by assuming t-distributions in both sampling errors and random effects (t-SAE). The inference was built using the likelihood approach. An expectation conditional maximization (ECM) algorithm was presented on getting its para-meters model estimation. The simulation study showed that the t-SAE model had better performance than normal-SAE and model based on Huber function when the data contained both unit and area level outliers. We have applied the proposed model for estimating per capita household expenditure at sub-districts in Bandung city, Indonesia. It results a better estimates. Thus, we recommend using the robust t-SAE that is proposed for handling the unit and area level outliers.

Keywords and phrases: area level model; t-distribution; unit and area level outliers; robust extension of SAE; adapted t-t linear mixed model; ECM algorithm.

2010 Mathematics Subject Classification: Primary 62F10, secondary 62J12, 62P20.

1 Introduction

The average of per capita expenditure as a measure of the well-being in a region is quite important in government policy making in developing countries. The official statistics Indonesia used data from National Social Economic Survey (SUSENAS) to produce this welfare indicator for national, provincial or district levels. Until now, it does not produce for sub-districts because of expensively. However, such data is very important for local policy. In this case, direct estimators from survey data might have large sampling errors. Therefore, the estimations become inefficient because of the small samples sizes. This is known as a *small area.*

The small area estimation models (SAE) are widely used to provide small area statistics because of its efficiency. These models use survey data from related areas through a linking model, thus increasing the "effective" sample size [18]. It means, without having to increase sample size, SAE models can produce estimators with a higher precision than direct estimators. These models use survey data as a response variable and borrow strength the census or administrative data as auxiliary variables to get parameter estimators in small areas.

Based on the availability of auxiliary variables, there are two types of small area models: unit and area level. A unit level model relates the unit values of the study variable to unit-specific auxiliary variables, but such data are rarely available. While an area level model relates the small area means to area-specific auxiliary variables (aggregated data), (see [18] for more detail). We focus here on the area level models, because the auxiliary variables in aggregate are more accessible.

The basic area level model was introduced by Fay and Herriot [9] and can be written as,

$$y_i = \theta_i + e_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \quad i = 1, 2, \dots, m, \tag{1.1}$$

where y_i is a direct estimate of i-th small area parameter θ_i with m small areas, obtained from survey and θ_i is assumed to have a linear relation to some p auxiliary variables $\mathbf{x}_i^T = (x_{1i}, x_{2i}, \dots, x_{pi})$, random effects u_i and sampling errors e_i are commonly assumed have normal distributions $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$, $e_i \overset{iid}{\sim} N(0, \sigma_{ei}^2)$, and σ_{ei}^2 are assumed known from sampling variances. Next, we call this model (1.1) as normal-SAE models.

The normal distribution is very sensitive to outliers since the influence function of the estimators under normal distribution is a trend line and unbounded [14, 17, 23, 25]. A simple way to solve this problem is to remove these outliers from the data, but this kind of practice is not recommended because the cost of data collection is very expensive and we can also lose some information. Another technique is to use the log-transformation method [22], but sometimes can be quite problematic, such as it can generate inaccurate estimates [10]. Other methods that can handle outliers and also provide accurate estimate is robust models.

Broadly, in the existing literatures, there are three approaches in the robust SAE models, for instance, firstly by using influence functions i.e Hubers function (we call this as robust-SAE Huber). This model have been studied by Sinha and Rao [21] in context unit level models. Warnholz [26] adapted this model for area level models and showed that this model effectively overcomes area level outliers. Secondly, by using M-quantile regression see for examples [5]. The third, by replacing the normal distribution with heavy tailed distributions. In context Hierarchical Bayesian framework, Datta and Lahiri [7] recommended a scale mixture of normal distributions specifically Cauchy distribution for random effects, Xie *et al.* [29] used a t-distribution with an unknown degrees of freedom parameter and Chakraborty *et al.*[4] used a two-component mixture normal model.

Bell and Huang [3] said that there are two types of outliers in area level models, such as, unit level outliers (that one or some units within an area are outlying) and area level outliers (that all units within an area are outlying). However, they realized that it was difficult to specify which type of outliers were observed. Therefore, they assumed a t-distribution on sampling error or on random effects in Hierarchical Bayes framework with known degrees of freedom. Pinheiro *et al.*[17] in the context of linear mixed-effects models for longitudinal data, also said that it may not be potential to separate these two types of outliers, so they proposed a model in which the random effects and the within subject-errors (sampling error in terms of SAE model) have multivariate t-distributions. In this paper, we propose the adapted t-t linear mixed model by Pinheiro *et al.*[17] for small area level models.

The t-distribution is widely used in robust models because the outlying observations are down-weighted on parameters estimation [12] and it has a bounded influence function [14, 17, 23, 25]. Moreover, on getting parameters estimation in robust t model, Lange *et al.* [12] suggested expectation maximization (EM) algorithm because of the simplicity. Liu and Rubin [13] showed that the expectation conditional maximization (ECM) was a simple and stable algorithm. It also has a faster rate convergence, when the degree of freedom is fixed. So did Bai *et al.* [2] also suggested to be chosen of its degree of freedom. If it is estimated, the estimate is not very accurate. Nevertheless, the use of ECM algorithm, according to authors knowledge, has not been applied on SAE models. Thus, this study aims to apply the ECM algorithm in parameter estimation of SAE model based on t distribution and to analyze the robustness of this model compared to SAE models based on normal distribution and robust Huber in term of bias and the efficiency. An empirical study of estimating the average of per capita household expenditure is conducted to show the performance of proposed model.

In recent years, several papers on the application of SAE method in estimating the average per capita household expenditure have been published. Susianto *et al.* [24] used the SUSENAS data for estimating in district levels, while Salma *et al.* [19] for sub-districts level. They have taken data from Village Potential (PODES) as auxiliary variables. Salma *et al.* [19] used unit level model and the results showed that the household expenditure data contained outliers. They applied the robust SAE model based on Huber function to deal with this problem. Nonetheless, the SAE Huber model is more suitable for dealing with area level outliers [26], but such data may contains unit and area level outliers. Thus, the proposed model is also applied to estimate the average of per capita household expenditure for sub-district level. Then, compared it with direct estimation and other SAE models, such as normal-SAE and its log-transformation, and robust SAE-Huber.

2 Small Area Model Based on t-Distribution

Here the t-t linear mixed model by Pinheiro *et al.*[17] is adapted for area level model,

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{iid}{\sim} t(0, \sigma_u^2, \nu), e_i \overset{iid}{\sim} t(0, \sigma_{ei}^2, \nu). \tag{2.1}$$

We call this (2.1) as t-SAE model. According to Lucas [14], the model based on the t distribution can be robust if the degrees of freedom is fixed. So, in this paper, the degrees of freedom ν of the model are assumed to be known.

The model based on t-distribution can be constructed from normal - Chi-squared hierarchical model [12] or generally, normal-gamma model [17] and the t-SAE model (2.1) can be written as,

$$\begin{aligned} y_i|u_i, \tau_i &\sim N(\mathbf{x}_i^T\boldsymbol{\beta} + u_i, \sigma_{ei}^2/\tau_i) \\ u_i|\tau_i &\sim N(0, \sigma_u^2/\tau_i) \\ \tau_i &\sim Gamma(\nu/2, \nu/2), \end{aligned} \tag{2.2}$$

where τ_i here is a latent variable from gamma distribution. From this model (2.2) and according to Bayes theorem, we get some properties as follows:

- The conditional distribution of u_i given y_i and τ_i:

$$u_i|y_i, \tau_i \sim N\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta})\left(\frac{\sigma_u^2}{\sigma_u^2 + \sigma_{ei}^2}\right), \frac{1}{\tau_i}\left(\frac{1}{\sigma_u^2} + \frac{1}{\sigma_{ei}^2}\right)^{-1}\right). \tag{2.3}$$

- The conditional distribution of y_i given τ_i:

$$y_i|\tau_i \sim N\left(\mathbf{x}_i^T\boldsymbol{\beta}, \frac{1}{\tau_i}\left(\sigma_u^2 + \sigma_{ei}^2\right)\right). \tag{2.4}$$

- The conditional distribution of τ_i given y_i :

$$\tau_i|y_i \sim Gamma\left(\frac{\nu+1}{2}, \frac{\nu + z_i^2}{2}\right), \tag{2.5}$$

  where $z_i^2 = (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2/(\sigma_u^2 + \sigma_{ei}^2)$.

- The marginal distribution of y_i is $y_i \sim t(\mathbf{x}_i^T\boldsymbol{\beta}, \mathbf{V}_i = \sigma_u^2 + \sigma_{ei}^2, \nu)$ with log-likelihood:

$$l = \sum_{i=1}^{m} -\frac{1}{2}\log(\sigma_u^2 + \sigma_{ei}^2) - \frac{1}{2}(\nu+1)\log(1 + \frac{z_i^2}{\nu}). \tag{2.6}$$

2.1 Parameters Estimation

In small area estimations, our focus is to predict the small area characteristics θ_i as in equation (1.1) which is a function of $\boldsymbol{\beta}$ and u_i, so the latent variables of model (2.2) are u_i and τ_i. If the degrees of freedom ν is fixed, then the vector of parameter models is $\omega = (\boldsymbol{\beta}^T, \sigma_u^2)$. The log-likelihood of completed data (2.2) is

$$l^c(\omega) = \sum_{i=1}^{m} l_i^c(\omega), \tag{2.7}$$

where

$$\begin{aligned} l_i^c(\omega) &= ln(f(y_i|u_i,\tau_i)f(u_i|\tau_i)f(\tau_i)) \\ &= ln(f(y_i|u_i,\tau_i)) + ln(f(u_i|\tau_i)) + ln(f(\tau_i)) \\ &= -\frac{\tau_i}{2\sigma_{ei}^2}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - u_i\right)^2 \\ &- \frac{1}{2}\ln(\sigma_u^2) - \frac{\tau_i}{2\sigma_u^2}u_i^2 \\ &+ \frac{\nu}{2}\ln(\tau_i) - \frac{\nu}{2}\tau_i + constant. \end{aligned} \tag{2.8}$$

The ECM algorithm can be computed as follows.
E-step: compute the expected of complete data log-likelihood (2.8).

$$\begin{aligned} Q\left(\omega|\hat{\omega}^{(h)}\right) &= \mathrm{E}\left(l^c(\omega)|y_i,\hat{\omega}^{(h)}\right) \\ &= \sum_{i=1}^{m}\mathrm{E}\left(-\frac{\tau_i}{2\sigma_{ei}^2}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - u_i\right)^2|y_i,\hat{\omega}^{(h)}\right) \\ &+ \sum_{i=1}^{m}\mathrm{E}\left(-\frac{1}{2}\ln(\sigma_u^2) - \frac{\tau_i}{2\sigma_u^2}u_i^2|y_i,\hat{\omega}^{(h)}\right) \\ &+ \sum_{i=1}^{m}\frac{\nu}{2}\mathrm{E}\left(\ln(\tau_i) - \tau_i|y_i,\hat{\omega}^{(h)}\right). \end{aligned} \tag{2.9}$$

We can see that $\sum_{i=1}^{m}\frac{\nu}{2}\mathrm{E}\left(\ln(\tau_i) - \tau_i|y_i,\hat{\omega}^{(h)}\right)$, the third term in the right side of (2.9) is a function of ν. When ν is known, the target is only $\boldsymbol{\beta}$ and σ_u^2 parameters, so this term becomes a constant and can be ignored. This is in line with Lange *et al.* [12] and Pawitan [16] which stated that $\mathrm{E}(\ln(\tau_i)|y_i,\hat{\omega}^{(h)})$ is sought when the degree of freedom ν is estimated. Thus, when the ν is fixed, at this stage only $\mathrm{E}(u_i^2|y_i,\hat{\omega}^{(h)})$ and $\mathrm{E}(\tau_i|y_i,\hat{\omega}^{(h)})$ is calculated.

Given $\omega = \hat{\omega}^{(h)}$ at h-th iteration, according to properties in equations (2.5) and (2.3) respectively we get:

$$\hat{\tau}_i^{(h)} = \mathrm{E}(\tau_i|y_i,\hat{\omega}^{(h)}) = (\nu+1)/\left(\nu + \hat{z}_i^{2(h)}\right), \tag{2.10}$$

where,

$$\hat{z}_i^{2(h)} = (y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(h)})^2/(\hat{\sigma}_u^{2(h)} + \sigma_{ei}^2). \tag{2.11}$$

The expectation of $\mathrm{u_i^2}$ given $\mathrm{y_i}, \hat{\omega}^{(\mathrm{h})}$ is:

$$\begin{aligned} \mathrm{E}(u_i^2|y_i,\hat{\omega}^{(h)}) &= \mathrm{E}\left[\mathrm{E}(u_i^2|y_i,\hat{\tau}_i^{(h)},\hat{\omega}^{(h)})\right] \\ &= var(u_i|y_i,\hat{\omega}^{(h)}) + \left\{\mathrm{E}(u_i|y_i,\hat{\omega}^{(h)})\right\}^2 \\ &= \left(\frac{1}{\hat{\tau}_i^{(h)}}\right)\hat{\eta}_i^{(h)} + \left(\hat{u}_i^{(h)}\right)^2, \end{aligned} \tag{2.12}$$

where,

$$\hat{\eta}_i^{(h)} = \left(\frac{1}{\hat{\sigma}_u^{2(h)}} + \frac{1}{\sigma_{ei}^2} \right)^{-1}, \tag{2.13}$$

and

$$\begin{aligned} \hat{u}_i^{(h)} &= \mathrm{E}(u_i | y_i, \hat{\omega}^{(h)}) = \mathrm{E}\left[\mathrm{E}(u_i | y_i, \hat{\tau}_i^{(h)}, \hat{\omega}^{(h)}) \right] \\ &= (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(h)}) \left(\frac{\hat{\sigma_u}^{2(h)}}{\hat{\sigma_u}^{2(h)} + \sigma_{ei}^2} \right). \end{aligned} \tag{2.14}$$

CM-step: update the estimated value of $\omega = (\boldsymbol{\beta}^T, \sigma_u^2)$ by maximizing the expected of completed data log-likelihood (2.9). It can be shown that,

$$\hat{\boldsymbol{\beta}}^{(h+1)} = \left(\sum_{i=1}^{m} \frac{\hat{\tau}_i^{(h)}}{\sigma_{ei}^2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{m} \left(\frac{\hat{\tau}_i^{(h)}}{\sigma_{ei}^2} \mathbf{x}_i \left(y_i - \hat{u}_i^{(h)} \right) \right), \tag{2.15}$$

and

$$\hat{\sigma}_u^{2(h+1)} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\tau}_i^{(h)} \left(\hat{u}_i^{(h)} \right)^2 + \hat{\eta}_i^{(h)} \right). \tag{2.16}$$

CM-1: fix $\sigma_u^2 = \hat{\sigma_u}^{2(h)}$ and update $\hat{\boldsymbol{\beta}}^{(h+1)}$ (2.15).
CM-2: fix $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(h)}$ and update $\hat{\sigma}_u^{2(h+1)}$ (2.16).
Do these *E-step* and *CM-step* iteratively until the log-likelihood of marginal distribution y_i (2.6) converges to some value, then we obtain the estimate of $\boldsymbol{\beta}$ and σ_u^2.

2.2 EBLUP Estimation

If σ_u^2 is known, the best linear unbiased prediction (BLUP) of θ_i without normality assumptions is defined by [8],

$$\tilde{\theta}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + \tilde{u_i} = (1 - \mathrm{B}_i) y_i + \mathrm{B}_i (\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}), \tag{2.17}$$

where $\mathrm{B}_i = \sigma_{ei}^2 / (\sigma_u^2 + \sigma_{ei}^2)$ and $\tilde{\boldsymbol{\beta}}$ is the estimate for $\boldsymbol{\beta}$ when σ_u^2 is known. In practice, σ_u^2 is unknown, then it will be substituted by its estimator, then we get the empirical BLUP (EBLUP) of θ_i:

$$\hat{\theta}_i = (1 - \hat{\mathrm{B}}_i) y_i + \hat{\mathrm{B}}_i (\mathbf{x}_i^T \hat{\boldsymbol{\beta}}), \tag{2.18}$$

where $\hat{\mathrm{B}}_i = \sigma_{ei}^2 / (\hat{\sigma_u}^2 + \sigma_{ei}^2)$.

The difference between EBLUP under normal-SAE model and t-SAE is the model parameters estimation. Under normal-SAE model (1.1), $\hat{\boldsymbol{\beta}}$ can be obtained by weighted least squared,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \frac{1}{\sigma_{ei}^2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{m} \left(\frac{1}{\sigma_{ei}^2} \mathbf{x}_i \left(y_i - \hat{u}_i \right) \right), \tag{2.19}$$

and the random effect variance σ_u^2 can be estimated by maximum likelihood (ML), or restricted (REML) or method of moment (see [18]).

Meanwhile, the proposed t-SAE model estimators ($\hat{\boldsymbol{\beta}}$ and $\hat{\sigma_u}^2$) can be obtained by using the ECM algorithm (as described in Section 2.1). For the t-SAE model, let the values in (2.14) and (2.15) converge respectively to,

$$\hat{u}_i = (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \left(\frac{\hat{\sigma_u}^2}{\hat{\sigma_u}^2 + \sigma_{ei}^2} \right), \tag{2.20}$$

and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \frac{\hat{\tau}_i}{\sigma_{ei}^2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{m} \left(\frac{\hat{\tau}_i}{\sigma_{ei}^2} \mathbf{x}_i \left(y_i - \hat{u}_i \right) \right), \tag{2.21}$$

where,

$$\hat{\tau}_i = (\nu + 1) \, / \left(\nu + \hat{z}_i^2 \right), \tag{2.22}$$

and

$$\hat{z}_i^2 = (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2 / (\hat{\sigma}_u^2 + \sigma_{ei}^2). \tag{2.23}$$

To see the robustness of EBLUP under t-SAE model, for example let us compare between (2.19) and (2.21). We can see that $\hat{\boldsymbol{\beta}}$ under t-SAE model is a robust estimator, since the outlying cases are down-weighted by (2.22).

3 Simulation Study

The purpose of this simulation is to investigate the robustness of EBLUP under robust t-SAE model compared to EBLUP under normal-SAE and robust-SAE based on Huber function. We used absolute relative biases (ARB) and asymptotic relative efficiency (ARE) for measure the robustness. The computation is carried out by using some packages in R software. We used *saeSim* package for data generating [28], *sae* package to get EBLUP under normal-SAE model [15] and *saeRobust* package to get EBLUP under robust-SAE Huber [27]. The method for getting EBLUP under the proposed t-SAE model can be seen in Section 2. There is no package yet for this. However, it can be implemented easily in standard software, such as R.

The simulated data were generated from basic area level models (1.1) and (2.1) with several scenarios:

1. both u_i and e_i were generated from normal distributions,
2. both u_i and e_i were generated from t-distributions with 5 degree of freedom ($\nu = 5$),
3. both u_i and e_i were generated from t-distributions with $\nu = 4$,
4. both u_i and e_i were generated from t-distributions with $\nu = 3$,
5. both u_i and e_i were generated from t-distributions with $\nu = 2$.

Set total of small area $m = \{15, 30\}$; $\mathbf{x}_i^T = (1, x_{1i}, x_{2i})$ with $\{x_{1i} = i : i = 1, \ldots, m\}$, and $x_{2i} \sim U(2, 5)$; $\boldsymbol{\beta} = (5, 1, -1)^T$; $\sigma_u^2 = \{1, 4\}$ and the pattern of σ_{ei}^2 is $\{1, 3, 5, 9, 19\}$. Every value in this set was taken 3 areas for $m = 15$ or 6 areas for $m = 30$.

The steps of simulation are:

1. Generate for i-th area and r-th replication, the area-specific random effects u_{ir} and errors e_{ir}. We get samples of θ_{ir} and y_{ir} without outliers by equation (1.1) when u_{ir} and e_{ir} from normal distribution, and ones that have unit and area level outliers by (2.1) when u_{ir} and e_{ir} from t-distribution.

2. Compute the EBLUP of θ_{ir} under normal-SAE, robust Huber and t-SAE models for several degrees of freedom $\nu = 3, 4, 5$ (here we used REML method for normal-SAE model, because it produced a consistent estimator of random effects variance even if normality is violated [11]).

3. Repeat the step (1)-(2) as many $R = 1000$ times.

4. Compute the percentage ARB of EBLUP averaged over areas for each model,

$$\%\overline{ARB} = \frac{1}{m} \sum_{i=1}^{m} ARB\left(\hat{\theta}_i\right) \times 100,$$

   where

$$ARB\left(\hat{\theta}_i\right) = \left(\frac{1}{R} \sum_{r=1}^{R} \theta_{ir}\right)^{-1} \left| \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_{ir} - \theta_{ir}\right) \right|.$$

5. Compute the mean squared error (MSE) of EBLUP averaged over areas for each model,

$$\overline{MSE} = \frac{1}{m} \sum_{i=1}^{m} MSE\left(\hat{\theta}_i\right),$$

   where

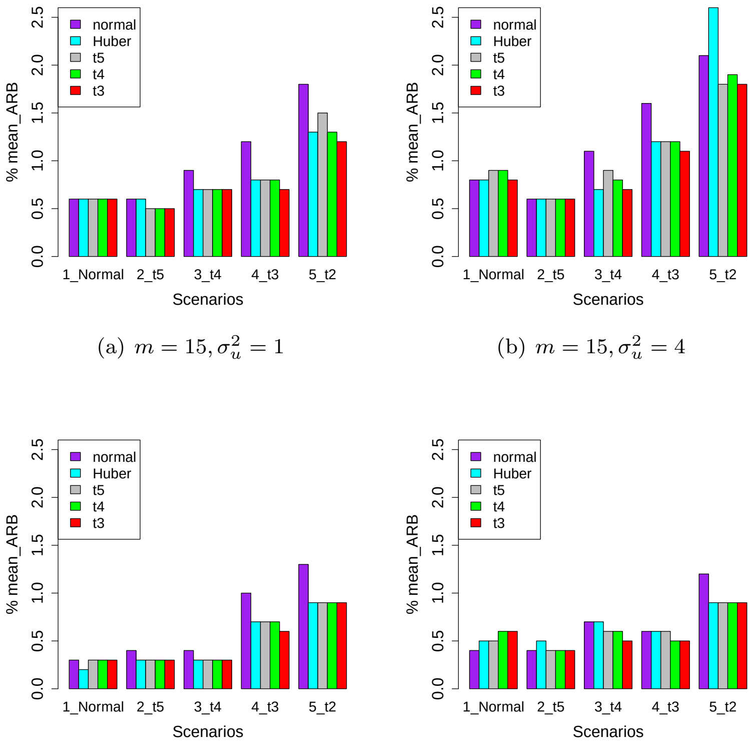
$$MSE\left(\hat{\theta}_i\right) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_{ir} - \theta_{ir}\right)^2.$$

6. Compute ARE of EBLUP under robust-SAE Huber and robust t-SAE models with respect to the normal-SAE,

$$ARE_k = \overline{MSE}_{normal} / \overline{MSE}_k,$$

   with k is for robust Huber or t-SAE models.

The results of simulation are presented in Figure 1 and 2. Figure 1 shows the percentage ARB of EBLUP averaged over areas under normal-SAE, robust SAE Huber and t-SAE models with several degrees of freedoms $\nu = \{3, 4, 5\}$. It shows generally that increasing the



(a) $m = 15, \sigma_u^2 = 1$

(b) $m = 15, \sigma_u^2 = 4$

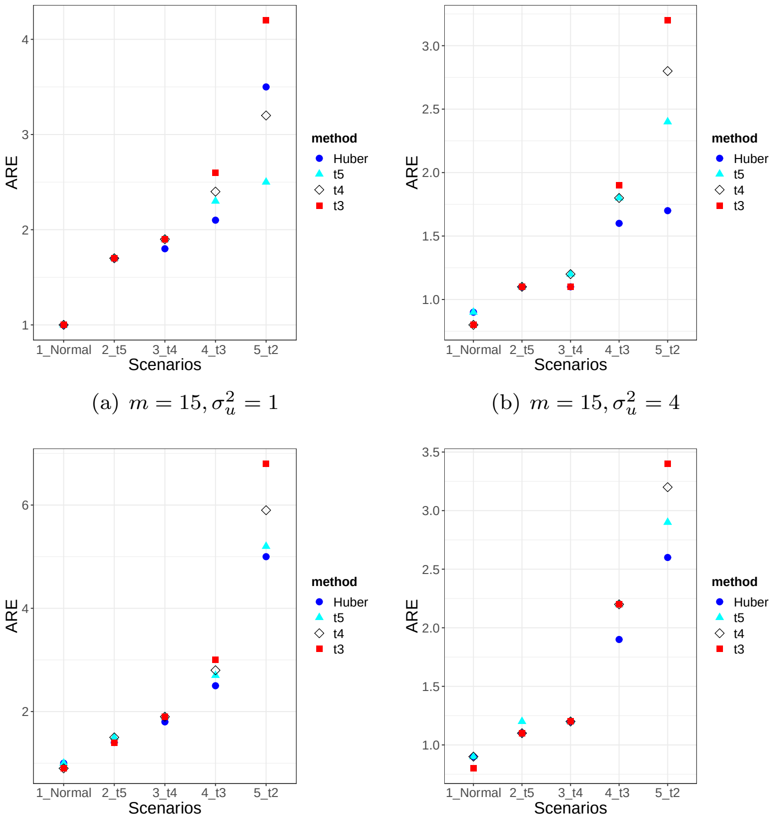
(c) $m = 30, \sigma_u^2 = 1$

(d) $m = 30, \sigma_u^2 = 4$

Figure 1: Simulated values of percentage absolute relative biased ($\%\overline{ARB}$) of EBLUP for normal-SAE, Huber, t-SAE with several degrees of freedom ($\nu = 3, 4, 5$)

number of areas m may reduce the ARB of EBLUP for all of the SAE models, but otherwise the greater random effect variance will enlarge the ARB. It also shows that the robust (Huber and t-SAE) also normal-SAE models give similar results in term of percent ARB when data has no outliers (normal scenario that both u_i and e_i from normal distribution). When there are unit and area level outliers (both u_i and e_i from t-distribution), the robust t-SAE model produces the smallest ARB, especially for t3 (t-SAE with 3 degree of freedom).

Figure 2 depicts the comparison of asymptotic relative efficiency (ARE) of EBLUP under the robust models (based on Huber function and t-distribution) with respect to normal-SAE model. It shows that ARE value of EBLUP under robust method is higher than 1 as its MSE decreases. It means, both robust models are more efficient than the normal-SAE when data has unit and area level outliers (both u_i and e_i from t-distribution). Its efficiency can be up to three times or more when u_i and e_i are from t-distribution with 2 degrees of freedom



(a) $m = 15, \sigma_u^2 = 1$

(b) $m = 15, \sigma_u^2 = 4$

(c) $m = 30, \sigma_u^2 = 1$

(d) $m = 30, \sigma_u^2 = 4$

Figure 2: Simulated values of asymptotic relative efficiency (ARE) of robust Huber and t-SAE with several degrees of freedom ($\nu = 3, 4, 5$) w.r.t normal-SAE model

and two times or more when data from t-distribution with 3 degrees of freedom. In this condition, ARE of EBLUP under all robust t-SAE models (3, 4 and 5 degrees of freedoms) are higher than the ARE under robust Huber with respect to normal-SAE models, and the t-SAE with 3 degree of freedom is highest.

For more detail, we can see Table 1. It presents the comparison of MSE of EBLUP averaged over areas under normal, robust Huber and t-SAE models. It is interesting to analyze the MSE of EBLUP. Table 1 shows that the MSE can decrease as the number of areas m increases, but when the random effect variance σ_u^2 gets bigger, the MSE becomes larger. Over all, we can see that all t-SAE models have smaller MSE than normal-SAE and Huber models when data contains unit and area level outliers (u_i and e_i from t-distributions), and t-SAE model with 3 degree of freedom has smallest of MSE, mainly for data with smaller random effect variance. So, there is no doubt based on the MSE and ARE value, the t-SAE

Table 1: Simulated values of MSE of EBLUP and asymptotic relative efficiency (ARE) of robust Huber and robust t-SAE models w.r.t normal-SAE model (averaged over areas)

Simulation scenarios			$\overline{MSE}$					ARE			
m	σ_u^2	dist. of u_i & e_i	normal	Huber	t3	t4	t5	Huber	t3	t4	t5
15	1	normal	1.9	1.9	2.0	1.9	1.9	1.0	1.0	1.0	1.0
		t5	5.0	3.0	2.9	2.9	2.9	1.7	1.7	1.7	1.7
		t4	6.7	3.8	3.5	3.5	3.6	1.8	1.9	1.9	1.9
		t3	12.9	6.1	5.0	5.3	5.6	2.1	2.6	2.4	2.3
		t2	83.5	23.9	20.1	26.1	33.4	3.5	4.2	3.2	2.5
	4	normal	3.3	3.7	4.2	4.0	3.8	0.9	0.8	0.8	0.9
		t5	6.5	5.8	6.1	5.9	5.7	1.1	1.1	1.1	1.1
		t4	8.3	7.3	7.3	6.9	6.8	1.1	1.1	1.2	1.2
		t3	17.3	10.7	9.3	9.4	9.6	1.6	1.9	1.8	1.8
		t2	103.7	59.5	32.8	36.9	43.0	1.7	3.2	2.8	2.4
30	1	normal	1.4	1.4	1.5	1.5	1.4	1.0	0.9	0.9	1.0
		t5	3.2	2.3	2.2	2.2	2.2	1.4	1.4	1.5	1.5
		t4	5.0	2.8	2.6	2.6	2.6	1.8	1.9	1.9	1.9
		t3	10.7	4.2	3.6	3.8	3.9	2.5	3.0	2.8	2.7
		t2	74.9	15.1	11.0	12.8	14.3	5.0	6.8	5.9	5.2
	4	normal	2.7	2.9	3.3	3.1	3.0	0.9	0.8	0.9	0.9
		t5	5.4	5.0	4.9	4.7	4.7	1.1	1.1	1.1	1.2
		t4	6.6	5.7	5.5	5.4	5.4	1.2	1.2	1.2	1.2
		t3	17.2	9.3	7.8	7.8	8.0	1.9	2.2	2.2	2.2
		t2	82.5	32.2	24.3	26.2	28.0	2.6	3.4	3.2	2.9

models are more efficient than normal-SAE even robust-SAE Huber since data contains both unit and area level outliers.

4 Application

The average per capita household consumption expenditure is one of the welfare indicators in a region. Statistics Indonesia (BPS) has been collecting the data of welfare through the

National Social Economic Surveys (SUSENAS), and one of the product is to estimate the average of per capita household expenditure (PCHE). SUSENAS March 2015 that conducted to estimate for districts level, by SAE model, the same survey data can be used to get the estimation for sub-districts. In this paper we used SUSENAS March 2015 to estimate the average of per capita household expenditure for $m = 29$ sub-districts in Bandung city with sample sizes as much 1040 households.

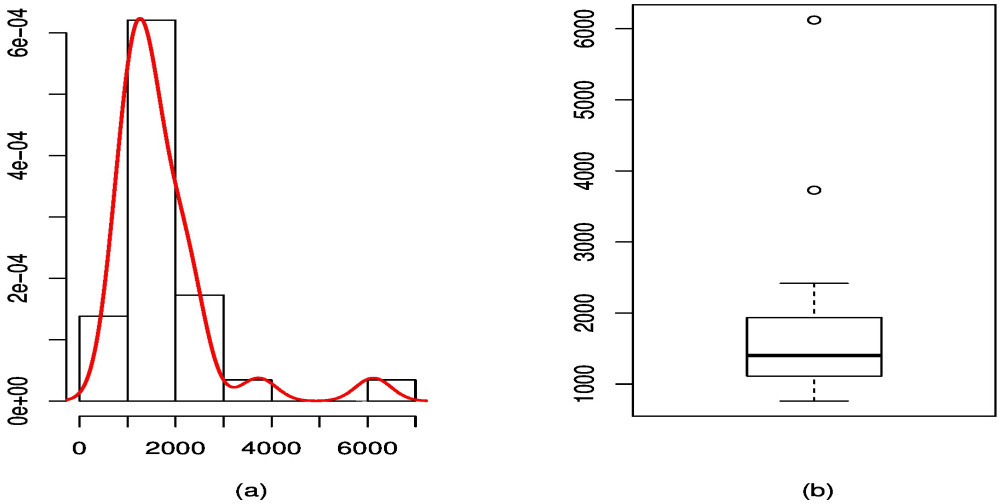


Figure 3: Histogram (a) and boxplot (b) of the direct estimates of PCHE at sub-districts in Bandung city based on SUSENAS March 2015

Before calculating the EBLUP estimator, let us look at Figure 3 first. It displays the histogram and box-plot of the direct estimates of PCHE at sub-districts in Bandung city based on SUSENAS March 2015. This figure shows that the distribution of PCHE is not normal, its tail is stretched to the right. It indicates the existence of some outliers, but not explain the type of outliers. Generally logarithmic transformation is done to overcome this. So in this paper, we compare the direct estimates of PCHE with EBLUP from several SAE models based on normal distribution, logarithmic transformation method, Huber model and t-distribution based model. Explanation in how to get the log-transformed EBLUP see [22], and [26] for Huber's EBLUP.

4.1 Variables selection

In theory, household income directly influences its expenditure. However, the data is difficult to obtain, especially for the sub-district level until now it is not yet available. According to empirical studies, demographic social factors can affect the level of household expenditure. Sekhampu and Niyimbanira [20] showed employment status and education have a positive effect on household expenditure. Coskun *et al.* [6] stated that housing wealth is significantly and positively related to household expenditure. Arias-Granada *et al.* [1] in his research at Dhakka Bangladesh showed that there were differences in water and sanitation services based on the level of household welfare. This showed that the higher level of income, the ability to buy drinking water with high quality is higher too. Therefore, a household income is approximated by the housing condition, the main source of drinking water is used, education, employment, and other social or cultural status. These variables can be found in Village Potential (PODES) data.

To get the EBLUP estimators of PCHE at sub-districts, we developed SAE model by modeling its direct estimates from SUSENAS March 2015 with the auxiliary variables from PODES 2014. The variables which have significant effect were:

- the proportion of villages where most households use bottled water as the primary source of drinking water (x_{1i}),
- the proportion of villages with the availability of communal library (x_{2i})
- the proportion of villages with settlements below extra high voltage air ducts (x_{3i}).

Variable (x_{2i}) is an approximation for education, while (x_{3i}) is for housing conditions. The coefficients estimates based on normal-SAE and t-SAE models are presented in Table 2. Only the third variable has a negative effect, while the first two have positive effects. It is interesting that in urban areas, people reading interest can affect the level of welfare, which in this case is reflected in household expenditure. On the contrary, a large number of illegal settlements can cause an increase in urban poverty.

Table 2: Regression parameter estimates and standard errors (in parentheses) for normal-SAE model and t-SAE with 3 degree of freedom (t3)

Variables	normal	t3
(Intercept)	989.20 (222.3)	976.24 (50.8)
x_{1i}	16.25 (3.9)	9.21 (1.7)
x_{2i}	10.34 (3.7)	7.34 (1.6)
x_{3i}	-14.33 (5.7)	-7.18 (2.1)
log-likelihood	-231.06	-193.23
AIC	472.13	396.47

Figure 4 depicts the comparison of estimated PCHE at Bandung city 2015, between the direct estimates and the EBLUPs. EBLUPs under robust t-SAE model give similar with the Hubers results. We can see that the EBLUPs are around the median of direct estimates.

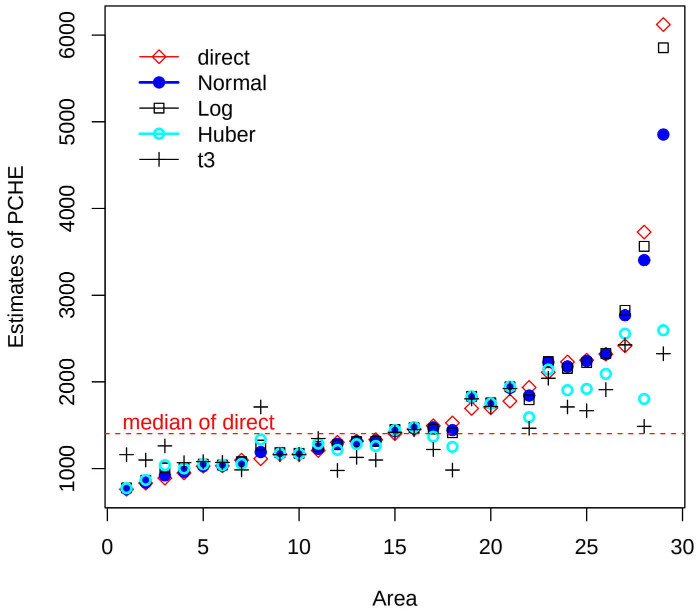


Figure 4: Comparison between direct estimates of PCHE from SUSENAS March 2015 and EBLUPs under normal-SAE, log-transformation, Huber and t-SAE ($\nu = 3$)

Meanwhile, the EBLUPs under normal-SAE and the logarithmic transformation method are still too close to its direct estimates, and we know that they are unreliable because the survey is not designed to sub-districts level. Therefore, the SAE model based on normal distribution and even its logarithmic transformation are not good estimates in this time.

Because of that, we should investigate the violation of normality assumption. Here we present the qq-norm plot of standardized residuals $\hat{e}_i^*$ and random effect $\hat{u}_i$ for normal-SAE model (Figure 5) and log-transformed normal-SAE model (Figure 6). The standardized residuals are computed by $\hat{e}_i^* = \hat{e}_i \sigma_{ei}$, where $\hat{e}_i = y_i - \hat{\theta}_i$ [26]. Figure 5 indicates the violation of normality in both sampling errors and random effects. Thus, we can understand that the data contains both unit and area level outliers and these affected the calculation of EBLUP under normal-SAE model that is close to the direct estimates. Sometimes a logarithmic transformation can overcome the violation of normality, but Figure 6 shows us that this method just overcome the normality violation on sampling error term, but not on the random effects term. So the robust models are more advantage in this condition.

Which robust models must be choose. Back to Table 2, it shows that the estimated values

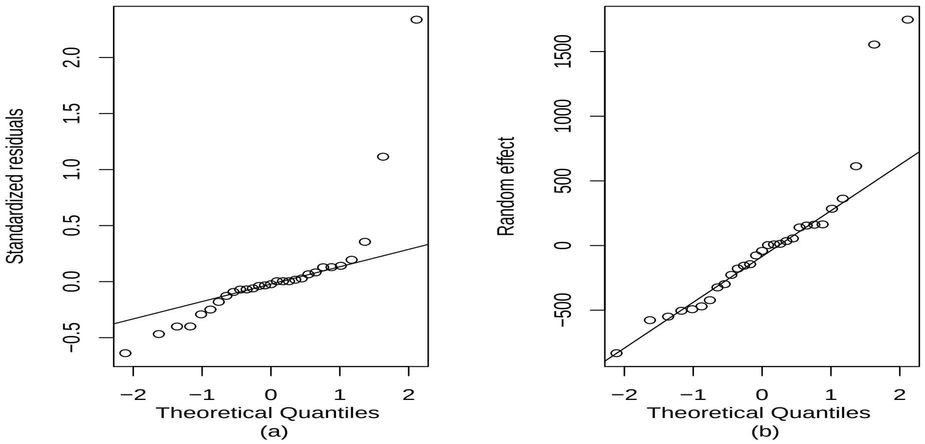


Figure 5: Q-Q norm plot for standardize residuals (a) and random effects (b) based on normal-SAE model

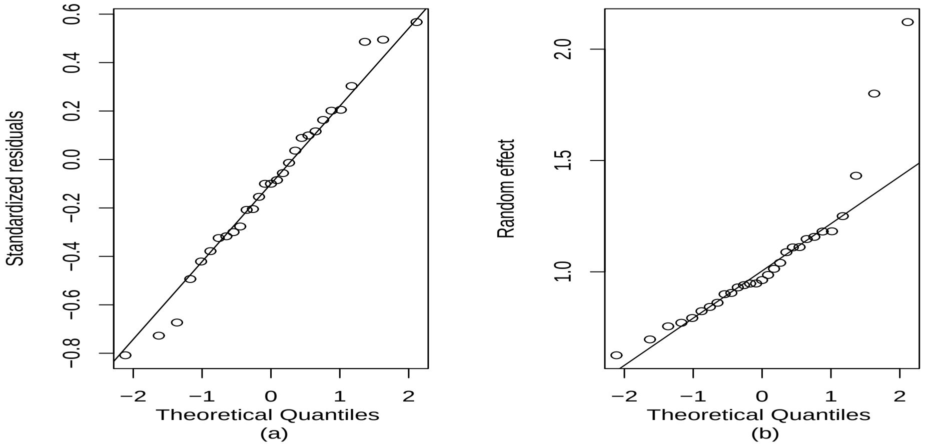


Figure 6: Q-Q norm plot for standardize residuals (a) and random effects (b) based on log-transformed normal-SAE model

of the normal-SAE model has the larger value than the robust t-SAE model. This indicates that the outliers influence estimation of parameters. we also can see that the robust t-SAE model produces larger log-likelihood and smaller AIC compare to normal model. In other words, the robust t-SAE model is better than normal model. Now let us look at Table 3 given in Appendix. It shows the results of PCHE estimation and standard errors of each direct method and SAE models such as normal, Huber and t-distribution. It can be seen that the standard errors of the normal-SAE model tends to be large and still close to the value of direct estimation. Meanwhile, the robust models (Huber and t-SAE) produce standard error of smaller values. We know that a good model produces an estimator with a small standard error. The table shows that the robust t-SAE (with 3 degrees of freedom, t3) model produces the smallest standard errors value. Thus, in this case, the SAE model based on the t-distribution produces a best estimation.

Table 3: Estimates of PCHE and standard errors between direct, normal, Huber and t-SAE ($\nu = 3$) methods (sorted ascending by direct estimates)

Area	Estimates				Standard errors			
	direct	normal	Huber	t3	direct	normal	Huber	t3
01	761.73	764.46	773.51	1160.24	*44.86*	*44.78*	*42.72*	*53.15*
02	831.55	837.89	869.02	1098.56	*89.08*	*88.45*	*86.45*	*54.11*
03	889.91	921.31	1039.45	1261.46	*173.56*	*169.02*	*125.15*	*79.40*
04	949.70	962.31	1001.55	1068.01	*180.89*	*175.73*	*154.75*	*54.31*
05	1028.22	1027.75	1049.84	1080.67	*146.43*	*144.05*	*143.49*	*107.85*
06	1032.41	1032.10	1040.44	1072.12	*97.75*	*96.92*	*96.84*	*66.56*
07	1098.01	1074.37	1054.10	984.47	*183.59*	*178.84*	*153.82*	*99.68*
08	1112.88	1190.83	1335.06	1710.40	*194.80*	*189.38*	*170.50*	*142.70*
09	1165.39	1171.36	1172.45	1159.78	*177.98*	*172.98*	*145.64*	*55.65*
10	1165.53	1169.76	1170.35	1160.46	*109.00*	*107.88*	*98.95*	*52.51*
11	1209.24	1228.81	1288.34	1347.44	*211.15*	*203.31*	*180.47*	*93.28*
12	1303.65	1285.39	1214.99	979.00	*142.90*	*140.32*	*125.48*	*61.14*
13	1317.08	1315.04	1281.63	1129.77	*122.76*	*121.15*	*115.64*	*52.27*
14	1333.41	1322.71	1260.71	1098.61	*164.06*	*160.17*	*142.19*	*53.97*
15	1402.76	1431.62	1443.43	1418.86	*226.88*	*219.00*	*177.67*	*134.46*
16	1455.84	1461.15	1477.25	1449.14	*227.20*	*218.89*	*176.55*	*119.76*
17	1493.54	1474.04	1367.82	1220.97	*237.56*	*225.83*	*198.72*	*60.22*
18	1526.60	1443.41	1250.53	982.83	*234.46*	*223.94*	*162.05*	*85.15*
19	1693.88	1826.05	1830.30	1803.88	*329.77*	*300.90*	*210.70*	*113.22*
20	1701.42	1751.34	1750.88	1717.91	*200.28*	*194.56*	*185.09*	*127.92*
21	1777.72	1931.59	1944.92	1922.12	*328.76*	*301.09*	*216.06*	*130.89*
22	1936.38	1842.39	1592.39	1465.69	*482.50*	*400.41*	*269.22*	*96.01*
23	2111.14	2227.02	2142.36	2042.39	*396.54*	*353.21*	*245.06*	*156.55*
24	2230.29	2177.97	1903.99	1710.42	*370.53*	*334.57*	*226.21*	*142.69*
25	2251.47	2242.22	1918.81	1667.07	*329.73*	*303.97*	*221.96*	*111.66*
26	2320.17	2318.82	2091.89	1909.69	*391.55*	*352.02*	*294.20*	*153.34*
27	2416.67	2768.44	2555.82	2426.46	*550.51*	*454.86*	*330.78*	*205.48*
28	3728.03	3403.95	1803.00	1487.67	*290.72*	*271.18*	*213.47*	*92.73*
29	6121.52	4853.46	2594.96	2324.51	*542.46*	*449.09*	*295.54*	*198.64*

5 Conclusions

This study has proven that the proposed SAE model based on t-distribution is more robust than SAE models based on normal distribution and Huber function. The proposed model is more efficient to be used when the data have both unit and area level outliers. In addition, the study has also found that the smaller degrees of freedom is more effective for longer tail data.

The application of SUSENAS data to predict the average of per capita household expenditure in some sub-districts at Bandung city showed that the logarithmic transformation is not able to overcome the violation of normality in SAE models, since the data have outliers in both sampling errors (unit level) and random effects (area level). This transformation can only fix the assumption of normality violation in sampling errors but not on the random effects. On the other hand, the robust t-SAE models which assumes t-distribution in both sampling errors and random effects can handle this type of problem better.

Finding MSE of EBLUP and its estimates in small area estimation model is one of the challenges for SAE researchers. However, this study has not been conducted. Next research, we will discuss how to build analytically the MSE of EBLUP in the t-SAE model that is proposed, and find the estimates.

Acknowledgments

The authors would like to thank the Head of Statistics Indonesia for giving us permission to use SUSENAS data for this paper and Directorate General Reinforcement of Research and Development (Ministry of Research Technology and Higher Education, Indonesia) for funding this article publication. We are also thankful to the associate editors and referees for spending the significant amount of time in processing this manuscript for JAPS.

References

[1] Arias-Granada, Y., Haque, S., Joseph, G., and Yanez-Pagans, M. (2018). Water and sanitation in Dhaka slums: Access, quality, and informality in service provision. *Policy Research Working Paper, 8552.*

[2] Bai, X., Chen, K., Yao, W. (2016). Mixture of linear mixed models using multivariate t distribution. *Journal of Statistical Computation and Simulation, 86(4)*, 771-787.

[3] Bell, W.R., and Huang, E.T. (2006). Using the t-distribution to deal with outliers in small area estimation. *Proceedings of Statistics Canada symposium.* Canada.

[4] Chakraborty, A., Datta, G.S., and Mandal, A. (2014). A two-component normal mixture alternative to the Fay-Herriot model. *Statistics in Transtition new series and Survey Methodology, 17(Small Area Estimation)*, 67-90.

[5] Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *76*, 47-69.

[6] Coskun, Y., Atasoy, B., Morri, G., and Alp, E. (2018). Wealth effects on household final consumption: Stock and housing market channels.*International Journal of Financial Studies*, *6*.

[7] Datta, G.S., and Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariate and outliers. *Journal of Multivariate Analysis*, *54*, 310-328.

[8] Datta, G.S., and Ghosh M. (2012). Small area shrinkage estimation. *Statistical Science*, *27*, 95-114.

[9] Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269-277.

[10] Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., and Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*, 105-109.

[11] Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, *24*(1), 255-286.

[12] Lange, K.L., Little, R.J., and Taylor, J.M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, *84*(408), 881-896.

[13] Liu, C., and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, *5*, 19-39.

[14] Lucas, A. (1997). Robustness of the student t based M-estimator. *Communications in Statistics - Theory and Methods*, *26*(5), 1165-1182.

[15] Molina, I., and Marhuenda, Y. (2015). *sae: Small Area Estimation.* Retrieved from R package version 1.0-5: https://CRAN.R-poject.org/package=sae

[16] Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood.* New York: Oxford University Press Inc.

[17] Pinheiro, J., Liu, C., and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects model using the multivariate t distribution. *Journal of Computational, Graphics and Statistics*,*10*, 249-276.

[18] Rao, J.N.K, and Molina, I. (2015). *Small area estimation second edition.* New Jersey: John Wiley and Sons Inc.

[19] Salma, A., Sadik, K., and Notodiputro, K. A. (2017). Small area estimation of per capita expenditures using robust empirical best linear unbiased prediction (REBLUP). *AIP Conference Proceedings.*

[20] Sekhampu, T., and Niyimbanira, F. (2013). Analysis of the factors influencing household expenditure. *International Business & Economics Research Journal*, *12*, 279-284.

[21] Sinha, S.K., and Rao, J. N. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, *37*, 381-399.

[22] Slud, E.V., and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot model. *Journal Royal Statistics B*, *68*, 239-257.

[23] Sumarni, C., Sadik, K., Notodiputro, K.A., and Sartono, B. (2017). Robustness of location estimators under t-distributions: A literature review. *IOP Conf. Series: Earth and Environmental Science*, *58*.

[24] Susianto, Y., Notodiputro, K.A., Kurnia, A., and Wijayanto, H. (2017). Small area estimation models with time factor effects for repeated measurement data. *Applied Mathematical Sciences*, *11*, 1995 - 2010.

[25] Ubaidillah, A., Notodiputro, K.A., Kurnia, A., Fitrianto, A., and Mangku, I. W. (2017). A robustness study of student-t distributions in regression models with application to infant birth weight data in Indonesia. *IOP Conf. Series: Earth and Environmental Science*, *58*.

[26] Warnholz, S. (2016a). Small area estimation using robust extentions to area level models. Retrieved from Refubium-Freire Universitat Berlin Repository: http://refubium.fu-berlin.de/handle/fub188/9706.

[27] Warnholz, S. (2016b, May). *saeRobust: Robust small area estimation.* Retrieved from R package version 0.1.0: https://CRAN.R-project.org/package=saeRobust

[28] Warnholz, S., and Schmid, T. (2016). Simulation tools for small area estimation: introducing the R package saeSim. *Austrian Journal of Statistics*, *45*, 55-69.

[29] Xie, D., Raghunathan, T.E., and Lepkowski, J.M. (2007). Estimation of the proportion of overweight individuals in small areas - a robust extension of the FayHerriot model. *Statistics in Medicine*, *26*, 2699-2715.