Jurnal Natural
Vol. 23, (2) 2023
DOI 10.24815/jn.v23i2.31095
pISSN 1411-8513
eISSN 2541-4062
ORIGINAL RESEARCH

Performance of copula and nested error regression models in estimating per capita expenditure of sub-district in Pidie Regency

NUR HASANAH[1,2], KHAIRIL ANWAR NOTODIPUTRO[2*], BAGUS SARTONO[2]

[1]BPS-Statistics of Aceh Province, Banda Aceh, Indonesia
[2]Departement of Statistics, IPB University, Bogor, Indonesia

Abstract. In unit-level small area estimation (SAE), the commonly used nested error regression (NER) model assumes normality which is not always the case. To handle non-normal data, researchers in statistics have developed a novel approach using exchangeable and extendible copula called the multivariate exchangeable copula (MEC) model. This study compares the performance of parametric MEC and NER models in estimating the sub-district average of per capita expenditure (PCE) in Pidie Regency, Aceh Province. This study presents PCE, which has a skewed distribution of the three-parameter skew-normal. The parametric MEC model uses a Gaussian copula from the Elliptical family and an empirical best unbiased prediction (EBUP) estimator. Meanwhile, the NER model uses an empirical best linear unbiased prediction (EBLUP) estimator. The results reveal that at a 95% confidence level, the parametric MEC model outperforms the NER model with a smaller root of mean squared error (RMSE) and provides a more precise estimate of the sub-district average of PCE. This study highlights the importance of considering the parametric MEC model as an alternative method for skewed data in unit-level SAE. The results of this study have the potential to support the achievement of Goal 1 (to end poverty) and Goal 10 (to reduce inequality) of the sustainable development goals (SDGs) by providing average PCE estimates at the sub-district level.

Keywords: BHF, EBLUP and EBUP, extendible copula, unit-level SAE, SDGs

INTRODUCTION

The rising demand for reliable small area statistics for evidence-based policymaking in many fields has led to the vast development and implementation of small area estimation (SAE). SAE provides a statistical procedure to produce statistics with adequate precision for areas or domains with small or even zero sample sizes [1]. A small area may be a geographic area or a subpopulation of specific socio-economic or demographic characteristics in which statistics are of interest. However, the sample size is relatively small, so the direct estimates based solely on survey samples for this specific scale area come up with inadequate precision. Moreover, it is impossible to provide direct estimates for non-sampled areas.

SAE is intended to increase the efficiency of survey samples by borrowing strengths of additional information from related areas or relevant data sources with the help of modeling. Rao & Molina [1] recommend using a linear mixed model (LMM) in SAE. LMM combines the target or response variable of interest from survey samples with the fixed effects of auxiliary variables or covariates, random area effects, and sample error linearly. In SAE, a unit-level small area model is known as the nested error regression (NER) or Battese-Harter-Fuller (BHF) model. The model is called unit-level as it uses the response variable and the relevant auxiliary variables data available for each unit sampling in a small area.

The NER model uses hierarchically structured data with units located in the area, so the sample errors are nested in the area. The NER model assumes normality of random area effects and errors, independence among random area effects and errors, and independence between random area effects and errors [2]. The random area effect is specific to each small area and thus induces dependence among units within a small area. The units within the same small area are likely to be more similar than units in other small areas. The NER model accommodates the

*Corresponding Author:
khairil@apps.ipb.ac.id

Received: March 2023| Revised: May 2023 | Accepted: May 2023

dependence between units by assuming a joint distribution function of the errors within a small area following multivariate normal distribution [3].

In the application of unit-level SAE, the continuous response variable data is often non-normal or asymmetrically distributed, primarily for economic data such as income or consumption, which causes the deviation of errors within a small area from the multivariate normal distribution. Rivest et al. [3], referred to as RVB, proposed a copula-based unit-level small area model, the so-called multivariate exchangeable copulas (MEC) model, to model asymmetric continuous response variables. The model is based on a general linear model (LM). It utilizes the exchangeable and extendible copula to construct a joint cumulative distribution function (CDF) and capture the dependence structure between the errors within a small area.

The copula is a multivariate distribution function in which marginals are uniform at unit intervals [0,1] and can describe the dependence between random variables [4;5]. The copula in the MEC model satisfies the exchangeable and extendible conditions. The MEC model involves errors in a small area which are assumed to have identical marginal distributions and symmetric dependence, so they are exchangeable. The copula in the MEC model is also extendible to be extended to a large number of errors in a small area. Copulas that meet these conditions come from the Elliptical family, such as Gaussian and Student-t, and the one-parameter Archimedean family, such as Clayton, Frank, Gumbel, and Joe [6].

The MEC model provides the empirical best non-linear and linear unbiased estimator (predictor) for the small area parameter, respectively, through the empirical best unbiased prediction (EBUP) and empirical best linear unbiased prediction (EBLUP) method. Grover et al. [7], referred to as GAT, propose the MEC modeling method, which produces a smaller relative bias of the EBUP estimator of the small area mean and the mean squared prediction error (MSPE) estimator than RVB. From here, we use the terminology mean squared error (MSE) instead of MSPE. There are four contributions of GAT in the MEC model. First, the ordinary least squares (OLS) method replaces the maximum normal log-likelihood in estimating fixed effect parameters. Second, the maximum pseudo-copula log-likelihood replaces Kendall's Tau in estimating the copula parameter. Third, the resampling technique with the single-phase bootstrap replaces the jackknife in estimating the MSE. Fourth, the GAT method accommodates a parametric approach besides semi-parametric.

Based on the description above, the MEC model is a novel approach in unit-level SAE proposed by [3], developed by [7], and has yet to be widely applied. This study examines the performance of the MEC and NER models in estimating the sub-district average of per capita expenditure (PCE) in Pidie Regency, Aceh Province. The MEC model uses a GAT method, parametric approach, and Gaussian copula to provide the EBUP estimator of the sub-district average PCE. Meanwhile, the NER model provides the EBLUP estimator of the sub-district average PCE. The performance criterion for both models is assessed using RMSE and evaluated with independent samples T-test.

The direct estimates of the average PCE from the results of the March round National Socio-economic Survey (Susenas) by the Badan Pusat Statistik (BPS-Statistics Indonesia) are only presented (estimated) up to the district level due to the limited number of samples. PCE data generally has a right-skewed distribution that is in line with this focus study and is widely used in SAE studies, among others, by [8], [9], and [10] for area-level SAE and [2] for unit-level SAE, but none of those applied copula. In this study, PCE data is assumed to follow the skewed distribution of the three-parameter skew-normal (SN).

PCE is an essential indicator for measuring a decent living standard. PCE is the amount derived from the total household expenditure for food and non-food consumption over a month to the number of household members [11]. PCE represents a household's ability to fulfill its members' basic needs. PCE is essential to monitor progress towards the first and tenth goals of the sustainable development goals (SDGs). Related to the first goal to end poverty, the indicator accounts for the percentage of people whose PCE falls under one purchasing power parity [12]. The tenth goal is to reduce inequality by aiming to empower and promote the economic inclusion of all. It accounts for the proportion of the population whose PCE is still below half of the median PCE [13].

METHODOLOGY

Data Description

The data consists of one response variable, PCE, denoted as Y, and four auxiliary variables, X_1 to X_4 (see Table 1). All variables are numeric. The ascending hierarchical structure related to data is presented as individual, household unit, village area, sub-district area, district area

Table 1. Research variables

Notation	Description	Unit	Type	Source
Y	The monthly household per capita expenditure for food and non-food consumption.	a hundred thousand IDR	continuous	Susenas March 2021
X_1	The number of cooperatives other than village cooperative units (KUD), small industry and citizen handicraft cooperatives (Kopinra), and savings and loan cooperatives (Kospin).	unit	discrete	Podes 2021
X_2	The number of governments, private, and rural (BPR) banks.	unit	discrete	Podes 2021
X_3	The proportion of inability letters/certificates (SKTM) issued by the village government to the number of families in the village.	letters per family	continuous	Podes 2021
X_4	The number of years the head of household attends school (converted from the highest graduation certificate owned by the head of household).	year	discrete	SP 2010

called regency or city, province area, and national. This study defines a sub-district as a small area and a household as a unit. The PCE data is available at the household level, while the auxiliary variables are available or aggregated at the village area level; so-called the intermediate level lies between the small area (sub-district) and the unit level (household). Thus, the households within a village correspond to the same values of auxiliary variables.

The PCE data are from the March round of Susenas in 2021. Susenas was conducted by BPS-Statistics Indonesia biannually in March and September. The March round was designed to provide estimates up to the district level. The area of interest is Pidie Regency, Aceh Province, which consists of 23 sub-districts and includes samples of Susenas 2021 varying from 10 to 60 households per sub-district with a total of 695 samples from 124,015 household populations spread over 70 of 731 villages. Each sub-district has household samples, so there is no non-sampled sub-district.

The four auxiliary variables are from census results: the 2010 Population Census (SP 2010) and the 2021 Village Potential Data Collection (Podes 2021). Using auxiliary variables from census results is essential as they are assumed to be collected without measurement errors, particularly sampling errors. BPS-Statistics Indonesia conducts Podes thrice a decade, every two years before the primary census (population, agriculture, and economic census). Meanwhile, SP is conducted decennial every year ends with zero, i.e., 2010 and 2020. Podes 2021 was conducted in June 2022 and covered all 84.096 administrative areas aligned with villages. SP 2020 is the latest population census held in Indonesia, but the dataset is not yet accessible to the public. This study aggregated the individual SP 2010 data at the village area level. To do so, we need to align the village master file used in SP 2010 with the ones used in Susenas and Podes 2021.

The four auxiliary variables are selected from 17 related auxiliary variables using stepwise regression with a backward elimination method. The selection was based on the smallest Akaike's Information Criterion (AIC) value obtained by comparing LMM and LM fits. The 17 auxiliary variables included in the selection process are chosen from 51 variables based on their correlation with the PCE variable assessed by the Pearson correlation coefficient. The original ordinal variable of X_4 is converted to numeric, from the highest graduation certificate owned by the head of the household to the number of years the head of the household attends school. Hence, the population mean of numeric X_4 becomes meaningful.

Data Notation

The data used are (1) household-level response variable, (2) household-level auxiliary variables, and (3) sub-district area-level auxiliary variables. The auxiliary variables data are available at the village area level, so the household-level auxiliary variables data are prepared to correspond to the sample data of the household-level response variable. The sample data is denoted as $\{(y_{ij}, x_{ij}); i=1, \cdots, m; j=1, \cdots, n_i\}$ with y_{ij} as a response variable vector for sample household j in the sub-district i, $x_{ij} = (1, x_{ij1}, \ldots, x_{ijp})^T$ as a $(1 + p)$-dimensional random vector of the auxiliary variables for sample household j in the sub-district i, $p = 4$ as the number of auxiliary variables, $m = 23$ as the number of sub-districts, and n_i as the sample size of households in the sub-district i. The sub-district area-level auxiliary variables are population means of auxiliary variables in the sub-district i, defined as $\bar{X}_i$.

Nested Error Regression Model

The NER model was carried out to provide the estimates of (1) the fixed effect parameters, (2) the variance components, (3) the random area effects, (4) the EBLUP of the sub-district means, and (5) the parametric bootstrap MSE of the EBLUP. The parametric bootstrap in estimating the MSE of EBLUP was conducted by assuming that the marginal distribution of the random area effects and errors in the sub-district still follow the normal distribution. The procedures of the NER model are given as follows:

1. Specify the sample data with the NER model. The relationship between y_{ij} and x_{ij} is expressed in a linear mixed model:

$$y_{ij} = x_{ij}^{T}\beta + v_i + \xi_{ij};\ i = 1, \dots, m;\ j = 1, \dots, n_i \qquad (1)$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is a vector of the unknown fixed effect; v_i is a random area effect in the sub-district *i*, following normal distribution $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$, and independent of each other; ξ_{ij} is a sampling error term for sample household *j* in the sub-district *i*, following normal distribution $\xi_{ij} \overset{iid}{\sim} N(0, \sigma_\xi^2)$, and independent of each other; v_i is independent of ξ_{ij}; and both σ_v^2 and σ_ξ^2 are unknown variance components [14].

2. Estimate the fixed effect parameters (β) and the variance components (σ_v^2 and σ_ξ^2) using REML. The fixed effects and variance components are estimated using the residual or restricted maximum likelihood (REML). REML is a widely used estimation method to reduce the bias of the maximum likelihood estimators of the variance components [15]. The REML estimates the fixed effect parameters and the variance components separately. It gives $\hat{\beta}$ as an empirical best linear unbiased estimator (EBLUE) for β [15], can be estimated using the weighted least squares (WLS) method [16], and $\hat{\sigma}_v^2$ and $\hat{\sigma}_\xi^2$ as the variance component estimators, respectively, called between-variation and within-variation components [14]. From the estimated variance components, we can calculate the intra-cluster correlation (ICC), defined as $\hat{\rho} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_\xi^2)$, the so-called within-area dependence [3]. Here, the cluster is a sub-district.

3. Estimate the random area effects (v_i). The estimate of the random area effect in the sub-district *i*, $\hat{v}_i$, is the EBLUP estimator of v_i and defined as $\hat{v}_i = \hat{\gamma}_i \left(\bar{y}_{i,s_i} - \bar{x}_{i,s_i}^{T}\hat{\beta}\right)$ with $\bar{y}_{i,s_i} = n_i^{-1} \sum_{j\in s_i} y_{ij}$ as the sample mean of the response variable in the sub-district *i*, s_i as a random sample of size n_i in the sub-district *i*, $\bar{x}_{i,s_i} = n_i^{-1} \sum_{j\in s_i} x_{ij}$ as the sample mean of auxiliary variables in the sub-district *i*, and $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_\xi^2 / n_i)$ as the scaling factor [17].

4. Estimate the EBLUP of the sub-district mean. The EBLUP estimator of the mean in sub-district *i* is given by equation:

$$\hat{\mu}_i^{EBLUP} = f_i \bar{y}_{i,s_i} + \left(\bar{X}_i - f_i \bar{x}_{i,s_i}\right)^{T} \hat{\beta} + (1 - f_i)\, \hat{v}_i \qquad (2)$$

where $f_i = n_i / N_i$ is the sampling fraction in the sub-district *i* and N_i is the number of population households in the sub-district *i* defined as $N_i = \sum_{j=1}^{n_i} w_{ij}$, the sum of the sampling weights for the sub-district *i*, where w_{ij} is the official sampling weight for household *j* in the sub-district *i* [17].

5. Estimate the MSE of the EBLUP. The MSE measures the precision of the estimator. The MSE of the EBLUP in the sub-district *i* can be expressed by:

$$MSE(\hat{\mu}_i^{EBLUP}) = E\left\{\left(\hat{\mu}_i^{EBLUP} - \mu_i\right)^2\right\} \qquad (3)$$

where μ_i is the true mean value in the sub-district *i* [15]. In practical settings, the estimation of MSE of the EBLUP is carried out using sample data and can be handled using parametric bootstrap resampling. This method is generally carried out in *B* replicates with the following steps in each replicate: use the estimated model parameters to generate the bootstrap sample, then use the bootstrap sample to estimate the model parameters and the EBLUP of the sub-district means.

Using *B* bootstrap replicates, the estimate of MSE of the EBLUP in the sub-district *i* is calculated as:

$$\widehat{MSE}(\hat{\mu}_i^{EBLUP}) = B^{-1} \sum_{b=1}^{B} \left(\hat{\mu}_i^{EBLUP(b)} - \mu_i^{(b)}\right)^2 \qquad (4)$$

where $b = 1, \dots, B$ is the number of bootstrap replicates and $\mu_i^{(b)}$ is the true mean of population bootstrap *b* in the sub-district *i*, calculated as $\mu_i^{(b)} = \bar{X}_i^{T}\hat{\beta} + v_i^{(b)} + \bar{\xi}_i^{(b)}$. Here, $v_i^{(b)}$ and $\bar{\xi}_i^{(b)}$ are the random area effect and the mean error of population bootstrap *b* in the sub-district *i* that follow the normal distribution. For more details on the steps of parametric bootstrap in the NER model, see [16].

6. Calculate the estimate of RMSE of the EBLUP. The performance criterion of the NER model is assessed using the RMSE.

The estimate of RMSE of the EBUP in the sub-district i is defined as:

$$\widehat{\mathrm{RMSE}} = \left(\widehat{\mathrm{MSE}}\left(\hat{\mu}_{\mathrm{i}}^{\mathrm{EBLUP}}\right)\right)^{1/2} \quad (5)$$

From equation (5), we can see that the MSE is measured in the squared unit of the response variable, so by taking the square root of MSE, we get the estimate of RMSE in the same unit as the response variable.

Multivariate Exchangeable Copula Model

The parametric MEC model was carried out to provide the estimates of (1) the fixed effect parameters, (2) the marginal error distribution, (3) the dependency parameter of Gaussian copula, (4) the EBUP of the sub-district means, and (5) the parametric bootstrap MSE of the EBUP. The parametric approach in the MEC model, referred to as the parametric MEC model, means that the marginal error distribution is assumed to follow the parametric family, and the copula model comes from the one-parameter exchangeable and extendible copula family [3;7]. This study's marginal error distribution follows the three-parameter SN distribution, and the copula is Gaussian. The procedures to perform the parametric MEC model in [7] are given as follows:

1. Specify the sample data with the MEC model. The relationship between y_{ij} and x_{ij} is expressed in a linear model:

$$y_{\mathrm{ij}} = x_{\mathrm{ij}}^{\mathrm{T}}\beta + \varepsilon_{\mathrm{ij}};\ i = 1, \ldots, m;\ j = 1, \ldots, n_{\mathrm{i}} \quad (6)$$

Here, $\varepsilon_{\mathrm{ij}}$ is the error term with the marginal distribution F_{ε} which has zero mean and finite variance σ^2. For simplicity, F_{ε} is parameterized by δ. The MEC model is used to model the joint CDF of the errors in each sub-district using the continuous distribution function of the exchangeable and extendible family [3]. The joint distribution of errors for each sub-district is expressed by the exchangeable and extendible copula model $C_{1:\mathrm{n_i}}(\cdot;\theta)$ as:

$$\begin{aligned} &F_{1:\mathrm{n_i}}\left(\varepsilon_{\mathrm{i},1}, \ldots, \varepsilon_{\mathrm{i,n_i}}; \delta; \theta\right) = \\ &C_{1:\mathrm{n_i}}\{F_{\varepsilon}\left(\varepsilon_{\mathrm{i},1}; \delta\right), \ldots, F_{\varepsilon}\left(\varepsilon_{\mathrm{i,n_i}}; \delta\right); \theta\} \end{aligned} \quad (7)$$

where θ is the dependency parameter of the copula, representing the within-area dependence that is the intra-cluster dependence.

2. Estimate the fixed effect parameters (β). A linear model's fixed effect parameter (β) in equation (6) is estimated using OLS to the sample data. OLS finds the fixed effect estimates, which minimizes the sum of the squared residuals. The fixed effect parameter estimator is denoted by $\hat{\beta}$.

3. Estimate the marginal error distribution (F_{ε}). The residuals $e_{\mathrm{ij}} = y_{\mathrm{ij}} - x_{\mathrm{ij}}^{\mathrm{T}}\hat{\beta}$ are obtained to estimate the marginal error distribution. The parametric approach is applied if the parametric model of the marginal error distribution $F_{\varepsilon}\left(\cdot\,;\hat{\delta}\right)$ is known. In the parametric approach, it is essential to specify the marginal error distribution correctly. The marginal error distribution is estimated using $\widehat{F}_{\varepsilon} \equiv \widehat{F}_{\varepsilon}\left(e_{\mathrm{ij}};\hat{\delta}\right)$ with the maximum likelihood estimation. The maximum likelihood estimator of marginal error distribution parameters is denoted by $\hat{\delta}$.

4. Estimate the dependency parameter of the Gaussian copula (θ). The marginal error distribution is used to form pseudo-observations which are copula data:

$$\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}\right) = \left(\widehat{F}_{\varepsilon}\left(e_{\mathrm{i},1}\right), \ldots, \widehat{F}_{\varepsilon}\left(e_{\mathrm{i,n_i}}\right)\right) \quad (8)$$

The copula data is used to estimate the copula dependency parameter (θ) using the maximum pseudo-copula log-likelihood. The copula parameter estimator ($\hat{\theta}$) is obtained by maximizing the following function of pseudo-copula log-likelihood:

$$\ell\left(\theta; \{\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}\}_{\mathrm{i=1}}^{\mathrm{m}}\right) = \textstyle\sum_{\mathrm{i=1}}^{\mathrm{m}} \ln\left[c_{1:\mathrm{n_i}}\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}; \theta\right)\right] \quad (9)$$

Here, $c_{1:\mathrm{n_i}}\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}; \theta\right)$ is the copula density function in the sub-district i, defined as:

$$c_{1:\mathrm{n_i}}\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}; \theta\right) = \frac{\partial^{\mathrm{n_i}}}{\partial_{\hat{u}_{\mathrm{i},1}} \cdots \partial_{\hat{u}_{\mathrm{i,n_i}}}} C_{1:\mathrm{n_i}}\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}; \theta\right) \quad (10)$$

This study uses a Gaussian copula. The Gaussian copula in the sub-district i has the following form:

$$\begin{aligned} &C_{1:\mathrm{n_i}}\left(\hat{u}_{\mathrm{i},1}, \ldots, \hat{u}_{\mathrm{i,n_i}}; \theta\right) = \\ &\int_{-\infty}^{\Phi^{-1}(\hat{u}_{\mathrm{i},1})} \cdots \int_{-\infty}^{\Phi^{-1}(\hat{u}_{\mathrm{i,n_i}})} \frac{\exp\left(e_{\mathrm{i}}^{\mathrm{T}}(\Sigma(\theta,n_i))^{-1}e_{\mathrm{i}}/2\right)}{(2\pi)^{\mathrm{n_i}/2}|\Sigma(\theta,n_i)|^{1/2}}\, de_{\mathrm{i},1} \cdots de_{\mathrm{i,n_i}} \end{aligned} \quad (11)$$

where $\sum(\theta,n_{\mathrm{i}})$ is the correlation matrix of size $n_{\mathrm{i}} \times n_{\mathrm{i}}$ with the determinant $|\sum(\theta,n_{\mathrm{i}})| = \{1+(n_{\mathrm{i}}-1)\theta\}(1-\theta)^{\mathrm{n_i}-1}$, Φ^{-1} is the inverse CDF of Gaussian single variable (univariate) such that $\Phi^{-1}(\hat{u}_{\mathrm{i},1}) = e_{\mathrm{i},1}$ and $\Phi^{-1}(\hat{u}_{\mathrm{i,n_i}}) = e_{\mathrm{i,n_i}}$, and $e_{\mathrm{i}} = (e_{\mathrm{i},1}, \ldots, e_{\mathrm{i,n_i}})^{\mathrm{T}}$ [18].

5. Estimate the EBUP of the sub-district mean. The EBUP estimator of the mean in sub-district i is given by the following equation:

$$\hat{\mu}_{\mathrm{i}}^{\mathrm{EBUP}} = \overline{X}_{\mathrm{i}}^{\mathrm{T}}\hat{\beta} + \frac{n_{\mathrm{i}}}{N_{\mathrm{i}}}\bar{e}_{\mathrm{i,s_i}} + \frac{N_{\mathrm{i}}-n_{\mathrm{i}}}{N_{\mathrm{i}}}\hat{e}_{\mathrm{i}} \quad (12)$$

where $\bar{e}_{\mathrm{i,s_i}}$ is the sample mean of errors in the sub-district i and calculated as $\bar{e}_{\mathrm{i,s_i}} = n_{\mathrm{i}}^{-1}\sum_{\mathrm{j}\in\mathrm{s_i}} e_{\mathrm{ij}}$, and $\hat{e}_{\mathrm{i}}$ is the empirical estimator of the non-sampled error term in the sub-district *i,* which is obtained by:

$$\hat{e}_{\mathrm{i}} = \frac{\sum_{\mathrm{i_1=1}}^{\mathrm{m}} \sum_{\mathrm{j_1 \in s_{i_1}}} e_{\mathrm{i_1,j_1}} w_{\mathrm{i1}} \left\{ \widehat{F}_{\varepsilon}\left(e_{\mathrm{i_1,j_1}}\right), \widehat{F}_{\varepsilon}\left(e_{\mathrm{ij}}\right); \mathrm{j} \in \mathrm{s_i}; \hat{\theta} \right\}}{\sum_{\mathrm{i_1=1}}^{\mathrm{m}} \sum_{\mathrm{j_1 \in s_{i_1}}} w_{\mathrm{i1}} \left\{ \widehat{F}_{\varepsilon}\left(e_{\mathrm{i_1,j_1}}\right), \widehat{F}_{\varepsilon}\left(e_{\mathrm{ij}}\right); \mathrm{j} \in \mathrm{s_i}; \hat{\theta} \right\}} \tag{13}$$

where $e_{\mathrm{i_1,j_1}}$ is the non-sampled error term with a marginal distribution of $\widehat{F}_{\varepsilon}\left(e_{\mathrm{i_1,j_1}}\right)$, $\widehat{F}_{\varepsilon}\left(e_{\mathrm{ij}}\right)$ is the marginal error distribution for household *j* in the sub-district *i*, and w_{i1} is a weighting function. Here, w_{i1} is the conditional density of the non-sampled error term given the sampled error term in the sub-district *i* and defined as:

$$w_{\mathrm{i1}} \left\{ \widehat{F}_{\varepsilon}\left(e_{\mathrm{i_1,j_1}}\right), \widehat{F}_{\varepsilon}\left(e_{\mathrm{ij}}\right); \mathrm{j} \in \mathrm{s_i}; \hat{\theta} \right\} = \frac{c_{1:\mathrm{n_i}+1}\left(\widehat{F}_{\varepsilon}\left(e_{\mathrm{i_1,j_1}}\right), \widehat{F}_{\varepsilon}\left(e_{\mathrm{i,j_1}}\right), \ldots, \widehat{F}_{\varepsilon}\left(e_{\mathrm{i,j_{n_i}}}\right); \hat{\theta}\right)}{c_{1:\mathrm{n_i}}\left(\widehat{F}_{\varepsilon}\left(e_{\mathrm{i,j_1}}\right), \ldots, \widehat{F}_{\varepsilon}\left(e_{\mathrm{i,j_{n_i}}}\right); \hat{\theta}\right)} \tag{14}$$

6. Estimate the MSE of the EBUP. The MSE of the EBUP in the sub-district *i* following equation (3) by replacing the superscript EBLUP with EBUP. The estimation of MSE of the EBUP using sample data was carried out using parametric bootstrap resampling. This method is generally carried out with *B* replicates. Each replicate includes generating the bootstrap sample using the estimated model parameters, then estimating the model parameters and calculating the EBUP of the sub-district mean using the bootstrap sample. The estimate of the MSE of the EBUP is obtained as follows:
   (1). Given the parameter estimates $\hat{\beta}$, $\hat{\delta}$, and $\hat{\theta}$ (as the results of steps 1 to 4 above), the household-level of auxiliary variables $x_{\mathrm{ij}}^{\mathrm{T}}$, and the sub-district area-level population means of auxiliary variables $\overline{X}_{\mathrm{i}}^{\mathrm{T}}$, perform steps (2) to (6) below for $b = 1, \ldots, B$ replicate:
   (2). Generate copula data ($\hat{u}_{\mathrm{ij}}^{(\mathrm{b})}$) using fitted Gaussian copula of sample data $C_{1:\mathrm{n_i}}(\cdot\,; \hat{\theta})$.
   (3). Use the inverse CDF of the fitted marginal error distribution to the copula data to obtain the error term $\varepsilon_{\mathrm{ij}}^{(\mathrm{b})} = \widehat{F}_{\varepsilon}^{-1}\left(\hat{u}_{\mathrm{ij}}^{(\mathrm{b})}; \hat{\delta}\right)$.
   (4). Compute the true mean of population bootstrap *b* in the sub-district *i*, $\mu_{\mathrm{i}}^{(\mathrm{b})} = \overline{X}_{\mathrm{i}}^{\mathrm{T}} \hat{\beta} + \bar{\varepsilon}_{\mathrm{i}}^{(\mathrm{b})}$, with $\bar{\varepsilon}_{\mathrm{i}}^{(\mathrm{b})}$ as the error mean of sample bootstrap *b* in the sub-district *i*. The error mean of sample bootstrap is used instead of population bootstrap to efficiently the execution time.
   (5). Obtain the response data for the bootstrap sample, $y_{\mathrm{ij}}^{(\mathrm{b})} = x_{\mathrm{ij}}^{\mathrm{T}} \hat{\beta} + \varepsilon_{\mathrm{ij}}^{(\mathrm{b})}$.
   (6). Perform the MEC model to bootstrap sample $(y_{\mathrm{ij}}^{(\mathrm{b})}, x_{\mathrm{ij}}^{\mathrm{T}})$ and obtain bootstrap parameter estimates $\hat{\beta}^{(\mathrm{b})}, \hat{\delta}^{(\mathrm{b})}$, and $\hat{\theta}^{(\mathrm{b})}$, and bootstrap EBUP of sub-district means $\hat{\mu}_{\mathrm{i}}^{\mathrm{EBUP(b)}}$.
   (7). Using *B* bootstrap replicates results, the bootstrap MSE of the EBUP estimator is calculated as in equation (4) by replacing the superscript EBLUP with EBUP and $\mu_{\mathrm{i}}^{(\mathrm{b})}$ obtained in step (4). For more details on the steps of parametric bootstrap in the MEC model for sample data, see the algorithm R function available from the author [7].
7. Calculate the estimate of RMSE of the EBUP. The estimates of RMSE of the EBUP are calculated following equation (5) by replacing the superscript EBLUP with EBUP.

Skew-Normal Distribution

The MEC model's marginal error distribution follows the three-parameter SN distribution. The SN is the development of the normal distribution, which has a slant or skewness [19]. The three parameters of the skew-normal distribution are location, scale, and skewness/shape. The location parameter determines the center of the data, i.e., the center value (mean) or the peak of the shape of the distribution. The scale parameter states the dispersion or spread of data around the mean. Skewness measures the symmetry or asymmetry of the data. Positive skewness indicates a distribution shape skewed or slanted to the right, while negative skewness indicates a distribution shape skewed to the left.

A continuous random variable *Y* with a three-parameter skew-normal distribution of location ζ, scale ω, and skewness λ, denoted as $Y \sim \mathrm{SN}(\zeta, \omega^2, \lambda)$. The equation gives the probability density function (PDF) of *Y*:

$$\mathrm{F}(y; \zeta, \omega^2, \lambda) = 2\omega^{-1} \phi((y - \zeta)/\omega)\, \Phi\big(\lambda(y - \zeta)/\omega\big) = \omega^{-1}\phi(y - \zeta)/\omega\,;\, \lambda) \tag{15}$$

where ϕ and Φ, respectively, PDF and CDF of standard normal distribution N(0,1), and the value ω^2 correspond with $\mathrm{N}(\mu, \sigma^2)$. Random variable *Y* has the mean $\mathbb{E}\{Y\} = \zeta + \omega\, \kappa (2/\pi)^{1/2}$, variance $\mathrm{V}\{Y\} = \omega^2 (1 - 2\kappa^2/\pi)$, skewness $\Upsilon_1 = [(4 - \pi)\mathbb{E}\{Y\}^2]/\big(2\,\mathrm{V}\{Y\}^{3/2}\big)$, and kurtosis $\Upsilon_2 = 2\,(\pi - 3)\mathbb{E}\{Y\}^4/\mathrm{V}\{Y\}^2$ with $\kappa = \lambda/\big(1 + \lambda^2\big)^{1/2}$. The equation gives the CDF of *Y*:

$$F(y;\omega^2) = \Phi((y-\zeta)/\omega) - 2T((y-\zeta)/\omega;\lambda) \quad (16)$$

where T is the T Owen function.

Direct Estimation

Direct estimation is an estimation method based on the sample survey data collected on a certain area to make inferences about the population characteristics of that area. The estimation incorporates the sole response variable and the survey weights. The unbiased direct Horvitz-Thompson estimator of the mean for the sub-district i in [16] is defined by:

$$\hat{\mu}_i^{DE} = N_i^{-1} \sum_{j\in s_i} w_{ij} y_{ij} \quad (17)$$

where w_{ij} is the sampling weight for household j in the sub-district i and $N_i = \sum_{j\in s_i} w_{ij}$ is a direct estimator of the population size in the sub-district i. The precision for the mean estimator is estimated by an unbiased estimator of the sampling variance defined as:

$$\hat{V}(\hat{\mu}_i^{DE}) = N_i^{-2} \sum_{j\in s_i} w_{ij}(w_{ij}-1)\, y_{ij}^2 \quad (18)$$

The standard deviation can be obtained by taking the square root of the sampling variance.

Independent Samples T-test

The performance of RMSE from the parametric MEC and NER models was evaluated using independent samples T-test. An independent samples T-test is a parametric test to compare whether two independent samples from different populations have the same mean. When conducting an independent samples t-test, the data should fulfill the normality assumption. Here are the procedures to conduct the independent samples T-test:

1. Check the normality assumption and perform the transformation method to the data if needed.
2. State the null and alternative hypotheses. The null hypothesis $H_0: \mu_{MEC} = \mu_{NER}$, the same as $H_0: \mu_{MEC} - \mu_{NER} = 0$, is that the RMSE of the parametric MEC and NER models have the same population mean. For a one-tailed T-test, the alternative hypothesis is $H_1: \mu_{MEC} < \mu_{NER}$ that the population mean of RMSE of the parametric MEC model is smaller than the NER model, or $H_1: \mu_{MEC} > \mu_{NER}$ that the population mean of RMSE of the parametric MEC model is greater than the NER model.
3. Determine the significance level (α) for 5% (0.05), equivalent to determining the confidence level $(1-\alpha)$ for 95%.
4. Determine the T-test statistics. The T-test statistic is defined in [20] as follows:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{s/(n_1^{-1}+n_2^{-1})^{1/2}} \text{ where } s = \left(\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}\right)^{1/2} \quad (19)$$

   The indexes of 1 and 2 belong to the parametric MEC and NER models, respectively. $\bar{X}_1$ and $\bar{X}_2$ are the averages of the RMSE data, $\delta_0 = 0$ is the difference in means under the null hypothesis, n_1 and n_2 are the sample sizes, s_1 and s_2 are the standard deviations of the RMSE data, and s is the standard deviation of the difference in means.
5. Determine the criterion of rejection. The null hypothesis is rejected if $T < -t_{\alpha(n_1+n_2-2)}$ for $H_1: \mu_{MEC} < \mu_{NER}$, or $T > t_{\alpha(n_1+n_2-2)}$ for $H_1: \mu_{MEC} > \mu_{NER}$, or if the calculated p-value is less than the significance level. The $t_{\alpha(n_1+n_2-2)}$ is the critical value for T-test with a significance level of α and degree of freedom $n_1 + n_2 - 2$.
6. Calculate the test statistics and p-value.
7. Make conclusions.

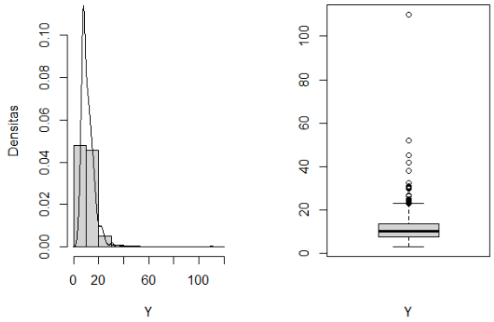
Procedures of Data Analysis

This study examines the performance of the parametric MEC and NER models in estimating average PCE at the sub-district level in Pidie Regency, Aceh Province. The study was conducted in R version 4.1.2 and R Studio version 2022.02.3+492 using various R packages. The MEC modeling uses the R functions from [7], mainly the copula [21] and sn [22] packages, while NER modeling and direct estimation use the sae package [17]. This study follows the given procedures:

1. Prepare and explore the data.
2. Specify the data with the parametric MEC and NER models.
3. Estimate the EBUP of the parametric MEC model and the EBLUP of the NER model for the sub-district average PCE.
4. Estimate the MSE of the EBUP/EBLUP using parametric bootstrap with $B = 1000$ replicates and calculate the RMSE estimates for each model.
5. Estimate the sub-district average PCE and its standard deviation using direct estimation.
6. Perform means difference test using independent samples T-test to the RMSE of parametric MEC and NER models.
7. Evaluate and interpret the results.

Table 2. Data summary

Summary	Sample data					Population data			
	Y	***X*₁**	***X*₂**	***X*₃**	***X*₄**	***X*₁**	***X*₂**	***X*₃**	***X*₄**
Observation	695.000	70.000	70.000	70.000	70.000	731.000	731.000	731.000	731.000
Minimum	3.104	0.000	0.000	0.028	3.000	0.000	0.000	0.000	3.000
Mean	11.579	0.014	0.014	0.219	6.557	0.007	0.041	0.220	7.124
SD	6.484	0.120	0.120	0.208	2.062	0.082	0.328	0.202	2.369
Maximum	110.103	1.000	1.000	1.045	12.000	1.000	6.000	1.700	12.000
Skewness	6.239	8.186	8.186	2.310	1.367	11.967	11.689	2.532	1.084
Kurtosis	82.496	68.014	68.014	8.498	5.115	144.207	174.556	12.651	3.201



(a) (b)

Figure 1. Per capita expenditure (a) histogram and density (b) boxplot.

RESULTS AND DISCUSSION

Data Exploration

In this section, we explore the data. Pidie Regency is one of 23 districts in Aceh Province. In March 2021, the direct estimate of the average PCE in Pidie was 1,049,186 IDR. The PCE data has an average of 11 (1,157,900 IDR) with a standard deviation of 6 (648,400 IDR) (see Table 2). The PCE data in Pidie has a moderate skewness of 6.239 among districts in Aceh Province, which varies from 1.657 to 11.219. The positive skewness in PCE data indicates that the distribution has skewed to the right (see Figure 1(a)).

If we convert the mean, variance, and skewness of PCE sample data to location, scale, and skewness of SN parameters (see Skew-Normal Distribution sub-section), we get the PCE data follows SN($\hat{\zeta}$ = 3.285; $\hat{\omega}$ = 10.528; $\hat{\lambda}$ = 6.239). Based on the boxplot (Figure 1(b)), there are 22 outliers, and 6 of them are categorized as extreme values in the PCE data. This study did not set aside the outliers since we intentionally presented the skewed distribution data, which outliers may also occur. Moreover, we have investigated that the outliers are not coming from measurement error as their reasonable values align with the households' profile.

Although there is a limitation in using Pearson correlation for skewed data, it still provides a useful and easy calculation to measure the linear relationship between variables. The Pearson correlation coefficients (and p-values) between the PCE and the auxiliary variables of the number of other cooperatives (X_1), the number of banks (X_2), the proportion of inability letters to number of families (X_3), and the number of years the head of household attends school (X_4) are -0.059 (0.118), 0.255 (0.000), -0.097 (0.011), and 0.125 (0.001). In a regression model with multiple auxiliary variables, it is necessary to detect the multicollinearity effect. Using the Pearson correlation coefficient among auxiliary variables and the variance inflation factor (VIF), we get a small correlation between -0.097 and 0.322 and a VIF of less than five, indicating no harm from multicollinearity.

Performance of the Parametric MEC and NER Models

We examine the performance of parametric MEC and NER models in estimating the sub-district average PCE using RMSE. The copula used is Gaussian. Selecting a suitable copula that fits the data can consider using theoretical and empirical normal score boxplots on residuals to visualize the within-area dependence; the details can be seen in the Supplementary Material of [3]. The results of the parametric MEC and NER models, which applied the indirect estimation methods, were also compared with those of direct estimation (DE). In this matter, the unbiased direct estimates of standard deviation (SD) of average

Table 3. The estimates of fixed effects of the parametric MEC and NER models

Result	Parametric MEC model					NER model				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	10.353	-4.818	12.714	-2.387	0.252	10.040	-5.034	12.536	-2.687	0.298
SE	0.909	2.121	2.223	1.156	0.130	1.090	2.538	2.367	1.263	0.151
T-Statistics	11.390	-2.271	5.718	-2.065	1.948	9.208	-1.984	5.295	-2.127	1.970
p-value	0.000*	0.023*	0.000*	0.039*	0.052	0.000*	0.051	0.000*	0.034*	0.050
95% confidence interval:										
Lower	8.568	-8.984	8.348	-4.657	-0.002	7.899	-10.311	7.888	-5.167	0.001
Upper	12.138	-0.653	17.079	-0.117	0.507	12.181	0.243	17.185	-0.207	0.595

Note: *Significant at confidence level 95%

Table 4. The estimates of variance components of the NER model

Estimator	Estimate	95% confidence interval	
		Lower	Upper
$\hat{\sigma}_v$	1.346	0.782	2.317
$\hat{\sigma}_\xi$	6.115	5.796	6.452

PCE are comparable to the RMSE estimates of the parametric MEC and NER models.

Given the estimates of fixed effects or regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{\beta}_4$, in Table 3 and the area effect for each sub-district $\hat{v}_i$ in Table 5, the parametric MEC and NER models have formed as follows:

MEC model: $$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \hat{\beta}_2 x_{ij2} + \hat{\beta}_3 x_{ij3} + \hat{\beta}_4 x_{ij4} \quad (20)$$

NER model: $$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \hat{\beta}_2 x_{ij2} + \hat{\beta}_3 x_{ij3} + \hat{\beta}_4 x_{ij4} + \hat{v}_i \quad (21)$$

which *i* is the sub-district, *j* is the household in the sub-district *i,* and x_{ij1} to x_{ij4} are the auxiliary variables of the number of other cooperatives, the number of banks, the proportion of inability letters to number of families, and the number of years the head of household attends school for household *j* in the sub-district *i.*

Both models yield estimated fixed effects that are relatively close in value. The sign of the estimated fixed effects for the auxiliary variables aligns with the sign of the Pearson correlation coefficient. As seen in Table 3, the auxiliary variables, such as the number of other cooperatives and the proportion of inability letters to the number of families, have a negative sign of estimated fixed effects. A negative sign of the estimated fixed effect indicates that an increase in the corresponding auxiliary variable, while holding all other auxiliary variables constant, is expected to be associated with a decrease in PCE. Similarly, a decrease in the corresponding auxiliary variable, while holding all other auxiliary variables constant, is expected to be associated with increased PCE.

Conversely, the number of banks and the number of years the head of household attends school have a positive sign of estimated fixed effects. A positive sign of the estimated fixed effect suggests that an increase in the corresponding auxiliary variable, while holding all other auxiliary variables constant, is associated with an increase in PCE. On the other hand, a decrease in the corresponding auxiliary variable, while holding all other auxiliary variables constant, is associated with a decrease in PCE.

From Table 3, we can also see that not all fixed effect is significant. The significant fixed effect comes with a p-value less than 0.05. We are not particularly concerned about the significance or insignificance of the auxiliary variables or which model produces better individual predictions of PCE. Our primary interest lies in the accuracy of the sub-district average of PCE estimate with RMSE as the performance criterion to validate the models.

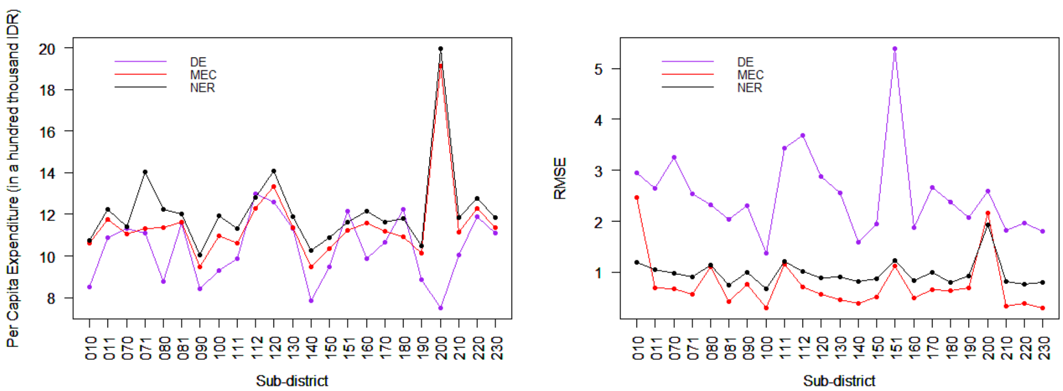
For the parametric MEC model, we found that residuals in all sub-districts exhibit a positive skewness range from 0.05 to 4.58. The estimates of skew-normal parameters $\hat{\zeta}$ = -6.071, $\hat{\omega}$ = 8.689, and $\hat{\lambda}$= 6.111 indicate residuals have skewed distribution. The estimate of the Gaussian copula dependency parameter is $\hat{\theta}$ = 0.166. Meanwhile, the NER model's estimate of the standard deviation of the sub-district area effect is $\hat{\sigma}_v$ =1.346 (see Table 4). It shows the variability of the PCE across different sub-districts. In Table 5, the estimate of the random area effect for sub-district 010 are presented with 0.000 due to the rounding, but the value is not zero. The estimate of intra-cluster correlation (ICC) is $\hat{\rho}$ = 0.052.

Even though the copula dependency parameter is said to quantify the within-area dependence, it is different from the ICC. The Gaussian copula has a dependence range from -1 to 1, but the extendible version has positive dependence. Zero indicates no dependence, and one for perfect dependence. The estimate of Gaussian

Table 5. Population and sample size and estimates of average per capita expenditure (in a hundred thousand IDR), RMSE, and area effect for each sub-district in Pidie Regency on March 2021.

	Sub-district	Sample size (n_i)	Population size (N_i)	DE		Parametric MEC		NER		
				DE	SD	EBUP	RMSE	EBLUP	RMSE	Area effect ($\hat{v}_i$)
010	Geumpang	10	2,694	8.497	2.953	10.625	2.474	10.745	1.193	0.000*
111	Keumala	10	1,575	9.885	3.429	10.617	1.163	11.306	1.209	-0.593
151	Grong Grong	10	2,352	12.151	5.392	11.218	1.128	11.625	1.232	-0.084
200	Kota Sigli	10	2,607	7.495	2.589	19.122	2.162	19.978	1.934	-0.847
112	Titeue	19	3,493	13.002	3.686	12.294	0.720	12.815	1.013	1.025
011	Mane	20	3,873	10.872	2.655	11.737	0.691	12.220	1.047	0.176
070	Glumpang Tiga	20	3,043	11.318	3.251	11.076	0.671	11.424	0.988	-0.151
080	Mutiara	20	2,219	8.795	2.316	11.377	1.106	12.221	1.141	-1.195
090	Tiro/truseb	20	3,727	8.412	2.314	9.477	0.767	10.040	0.997	-1.167
170	Peukan Baro	20	3,696	10.651	2.661	11.207	0.665	11.637	1.004	-0.136
120	Sakti	29	5,594	12.610	2.874	13.336	0.563	14.095	0.898	1.290
190	Simpang Tiga	29	4,930	8.876	2.076	10.115	0.687	10.506	0.935	-0.969
071	Glumpang Baro	30	5,415	11.091	2.544	11.32	0.575	14.017	0.919	2.802
130	Mila	30	5,087	11.324	2.563	11.364	0.455	11.876	0.915	0.329
150	Delima	30	4,998	9.469	1.951	10.35	0.520	10.865	0.868	-0.175
140	Padang Tiji	40	6,675	7.850	1.599	9.465	0.391	10.261	0.818	-1.625
160	Indrajaya	40	6,462	9.876	1.875	11.572	0.505	12.144	0.843	-0.145
210	Pidie	40	6,574	10.034	1.818	11.161	0.329	11.827	0.827	-0.382
180	Kembang Tanjong	48	8,131	12.230	2.380	10.924	0.643	11.782	0.799	-0.046
220	Batee	50	9,611	11.893	1.968	12.278	0.399	12.780	0.775	1.471
230	Muara Tiga	50	8,708	11.082	1.801	11.355	0.300	11.831	0.802	0.505
081	Mutiara Timur	60	10,085	11.562	2.038	11.607	0.434	12.013	0.740	0.202
100	Tangse	60	12,466	9.292	1.377	10.971	0.298	11.916	0.686	-0.286

Note: *The number appears as zero due to the rounding.



(a) (b)

Figure 2. Estimates of DE, parametric MEC, and NER, (a) average of per capita expenditure, (b) RMSE.

copula dependency $\hat{\theta} = 0.166$ means a weak positive dependence between the errors within the sub-district. However, it also shows that a linear model does not account for the dependence structure among errors. The ICC ranges from 0 to 1, with zero indicating no between-variation and one as no within-variation. The ICC $\hat{\rho} = 0.052$ means only 5% of the variability PCE is attributed to the differences between the sub-districts. The remaining 95% of the variability is due to differences within the sub-district.

We present Table 5 in ascending sample sizes, while the line plot in Figure 2 visualizes the estimates in ascending sub-district code. Generally, the estimates of average PCE at the sub-district level in Pidie Regency tend to be smaller for the DE results (Figure 2(a)). It is also shown that the estimates of average PCE of the parametric MEC and NER models have a wide gap with DE at sub-district 200. A possible explanation for this is the high mean population for the number of banks and the number of years the head of household attends school, respectively 0.6 and 10.4, which correspond to the positive fixed effects. As seen in equations (2) and (12), the estimates of EBLUP/EBUP incorporated the population means of auxiliary variables.

Compared to the DE method, the RMSE estimates of the parametric MEC and NER models tend to be smaller, indicating that the two models are more precise than DE (Figure 2(b)). As shown in Table 5, the RMSE estimates of parametric MEC and NER models tend to decrease with increased sample sizes. The RMSE estimates of the parametric MEC model tend to be smaller than the NER model, except for sub-districts 010 and 200. Both sub-districts cover only one village, so households within a village have the same values of auxiliary variables. The parametric MEC model has failed to provide more precise estimates than the NER model when the auxiliary variables provide less variation since the MEC model accounts only for the fixed effects of auxiliary variables. When the auxiliary variables do not vary much, it may lead the model to have larger residuals. As explained in [3], besides the sample size, the EBUP of the MEC model depends on the multivariate probability integral transform of the errors in which residuals are used to estimate the error distribution. The MSE estimation to finally obtain the estimates of RMSE accounts for this condition in each replicate since the same auxiliary variables data are used.

The RMSE estimates between the parametric MEC and NER models were evaluated using independent samples T-test. Since the RMSE data of the two models are not normal and positively skewed, the inverse transformation is applied to the RMSE data to meet the normality assumption. After the inverse transformation, the Shapiro-Wilk test results of both models have W statistic (and p-value) 0.959 (0.448) and 0.973 (0.781), respectively, in which insignificant p-value shows no evidence to conclude that the inverse RMSE data is not normal. At first, we want to evaluate the mean of RMSE of the parametric MEC model smaller than the NER model, but the inverse transformation changes the direction of the difference between the means. Thus, an independent samples T-test was conducted to evaluate if the mean of inverse RMSE of the parametric MEC model was greater than the NER model.

The independent samples T-test provides a T-statistic (and p-value) of 3.877 (0.000). The critical value of the T-test, $t_{\alpha(n_1+n_2-2)}$, with a significance level of 0.05 and a degree of freedom, $n_1 + n_2 - 2 = 44$, was 1.680. Since the T-statistic was greater than the critical value or the p-value was less than 0.05, the null hypothesis that the population means of inverse RMSE of the two models were equal was rejected. The results prove that the mean of inverse RMSE of the parametric MEC model is greater than the NER model at a confidence level of 95%. In conclusion to the original RMSE data, we are 95% confident that the true mean value of RMSE of the parametric MEC model is smaller than the NER model.

CONCLUSION

This study attempts to model the per capita expenditure data, which presents a skew-normal distribution, with the parametric MEC and NER models. In estimating the average per capita expenditure at the sub-district level, the parametric MEC model used the Gaussian copula and non-linear estimator of EBUP. Meanwhile, the NER model used a linear estimator of EBLUP. From the independent samples T-test results at a 95% confidence level, the parametric MEC model using Gaussian Copula performs better with a smaller RMSE than the NER model in estimating the sub-district average of per capita expenditure in Pidie Regency, Aceh Province. Thus, the parametric MEC model gives a more precise and reliable estimate for the sub-district average of per capita expenditure.

The estimates of the parametric MEC model for the sub-district average of per capita expenditure in Pidie Regency for the March 2021 period provide valuable insights into the economic landscape of the district. The estimated values range from 9.465 (946,500 IDR) to 19.122 (1,912,200 IDR), indicating variation in the living standard or economic well-being of different sub-districts. Among the sub-districts in Pidie Regency, the Padang Tiji sub-district emerges with the lowest estimate of average per capita expenditure, while the Kota Sigli sub-district exhibits the highest estimate of average per capita expenditure.

By highlighting the lowest and highest estimates of average per capita expenditure in specific sub-districts, we draw attention to the economic disparities and opportunities present within Pidie Regency. This analysis not only contributes to the understanding of local economic dynamics but also aligns with the broader goals of sustainable development, such as Goal 1 (to end poverty) and Goal 10 (to reduce inequality) of the Sustainable Development Goals (SDGs).

In addition to providing valuable insights into modeling per capita expenditure data with a skew-normal distribution, this study also addresses the practical implications of the estimates derived from the parametric MEC model. Specifically, considering the execution time required for analyzing 695 observations spread among 23 sub-districts, we find that the parametric MEC model remains viable despite its longer runtime of approximately 14 minutes.

Meanwhile, the NER model only requires approximately one minute (66 seconds).

The exclusion of outliers and using other copulas are also the subjects of further discussion. In our findings, excluding outliers drops the skewness of per capita expenditure data to 0.766 from 6.239. In that case, the drop in the RMSE estimate is more pronounced for the NER model than the parametric MEC model for the sub-district with the outliers. It is predictable as the data is close to normal but still skewed. However, in that case, the parametric MEC model still performs better than the NER model. Interestingly, the Gaussian copula also has a smaller RMSE estimate than other copulas from the one-parameter Archimedean family, such as Clayton, Frank, and Gumbel.

ACKNOWLEDGMENT

The first author expresses gratitude for the guidance and correction from the supervisory committees, Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S. and Dr. Bagus Sartono, S.Si., M.Sc. This research is part of the thesis of Nur Hasanah on the applied statistics master program. The scholarship funding from BPS-Statistics Indonesia and the data support from BPS-Statistics of Aceh Province are gratefully acknowledged.

REFERENCE

[1] Rao, J. N. K.; Molina, I. 2015. *Small area estimation* 2nd ed (Hoboken: John Wiley & Sons).

[2] Handayani, D.; Notodiputro, K. A.; Saefuddin, A.; Mangku, I. W.; Kurnia, A. 2018. Empirical best predictor for nested error regression small area models. *IOP Conf. Ser.: Earth Environ. Sci.* **187** (1) 012036. *DOI: 10.1088/1755-1315/187/1/012036*.

[3] Rivest, L. P.; Verret, F.; Baillargeon, S. 2016. Unit level small area estimation with copulas. *Can. J. Stat.* **44** (4) 397-415. *DOI: 10.1002/CJS.11296*.

[4] Nelsen, R. B. 2006. *An introduction to copulas* 2nd ed (Portland: Springer).

[5] Flores, M. Ú.; Artero, E. de A.; Durante, F. 2017. *Copulas and dependence models with applications: contributions in honor of Roger B. Nelsen.* Sánchez, J. F. (Ed.) (Switzerland: Springer).

[6] Mai, J. F.; Scherer, M. 2012. *Simulating copulas: stochastic models, sampling algorithms, and applications* 2nd ed (London: Imperial College Press).

[7] Grover, K.; Acar, E. F.; Torabi, M. 2020. Copula-based predictions in small area estimation. *Can. J. Stat.* **48** (4) 685-711. *DOI: 10.1002/CJS.11558*.

[8] Susianto, Y.; Notodiputro, K. A.; Kurnia, A.; Wijayanto, H. 2017. Small area estimation models with time factor effects for repeated measurement data. *Appl. Math. Sci.* **11** (41) 1995-2010. *DOI: 10.12988/ams.2017.74142*.

[9] Ubaidillah, A.; Notodiputro, K. A.; Kurnia, A.; Mangku, I. W. 2019. Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. *J. Appl. Stat.* **46** (15) 2845-2861. *DOI: 10.1080/02664763.2019.1615420*.

[10] Pusponegoro, N. H.; Kurnia, A.; Notodiputro, K. A.; Soleh, A. M.; Astuti, E. T. 2021. Small area estimation of sub-district's per capita expenditure through area effects selection using lasso method. *Procedia Comput. Sci.* **179** 754-761. *DOI: 10.1016/J.PROCS.2021.01.064*.

[11] [BPS] Badan Pusat Statistik. 2021. *Pengeluaran untuk konsumsi penduduk Indonesia, Maret 2021*. Badan Pusat Statistik.

[12] [Bappenas] Kementerian Perencanaan Pembangunan Nasional/Badan Perencanaan Pembangunan Nasional. 2020. *Metadata indikator pilar pembangunan sosial pelaksanaan pencapaian Tujuan Pembangunan Berkelanjutan/Sustainable Development Goals (TPB/SDGs)*. Bappenas.

[13] [Bappenas] Kementerian Perencanaan Pembangunan Nasional/Badan Perencanaan Pembangunan Nasional. 2020. *Metadata indikator pilar pembangunan ekonomi pelaksanaan pencapaian Tujuan Pembangunan Berkelanjutan/Sustainable Development Goals (TPB/SDGs)*. Bappenas.

[14] Sugasawa, S.; Kubokawa, T. 2020. Small area estimation with mixed models: a review. *Jpn. J. Stat. Data Sci*. **3** (2) 693-720. *DOI: 10.1007/S42081-020-00076-X/METRICS*.

[15] Morales, D.; Esteban, M. D.; Pérez, A.; Hobza, T. 2021. *A Course on small area estimation and mixed models* (Cham: Springer Nature Switzerland AG).

[16] Molina, I.; Marhuenda, Y. 2015. R package sae: methodology. [online] Available at: https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.pdf [accessed 1 Dec. 2022].

[17] Molina, I.; Marhuenda, Y. 2015. SAE: an R package for small area estimation. *The*

R Journal. **7** (1), 81-98. *DOI: 10.32614/rj-2015-007*.

[18] Rivest, L. P.; Verret, F.; Baillargeon, S. 2015. Estimation of the parameters in copula models for small areas. Proceedings of the Survey Methods Section; Canada: Statistical Society of Canada; p. 1-7.

[19] Azzalini, A.; Capitanio, A. 2014. *The skew-normal and related families* (New York: Cambridge University Press).

[20] Saefuddin, A.; Notodiputro, K. A.; Alamudi, A.; Sadik, K. 2009. *Statistika dasar* (Jakarta: Grasindo).

[21] Hofert, M.; Kojadinovic, I.; Maechler, M.; Yan, J.; Nešlehová, J. G.; Morger, R. 2023. Multivariate dependence with copulas [R package copula version 1.1-2]. [online] Available at: https://cran.r-project.org/package=copula [accessed 20 Jan. 2023].

[22] Azzalini, A. 2022. The R package "sn": the skew-normal and related distributions such as the skew-t and the SUN (version 2.1.0). [online] Available at: http://azzalini.stat.unipd.it/SN/, https://cran.r-project.org/package=sn [accessed 1 Dec. 2022].