Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 5, No. 2, September 2022, pp. 98 – 110
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Handling Outliers in The Stochastic Frontier Model Using Cauchy and Rayleigh Distributions to Measure Technical Efficiency of Rice Farming Bussiness

[1]Retna Nurwulan, [2]Anik Djuraidah, [3]Anwar Fitrianto
[1,2,3]Department of Statistics, IPB University
Email: [1]retnanurwulan@apps.ipb.ac.id, [2]anikdjuraidah@apps.ipb.ac.id, [3]anwarstat@gmail.com

Article Info

Article history:
Received Aug 18th, 2022
Revised Aug 30th, 2022
Accepted Sep 20th, 2022

Keyword:
Cauchy distribution
Fat-tailed
Production frontier
Rayleigh distribution
Translog

ABSTRACT

Technical Efficiency (TE) is one of the essential indicators used to evaluate the development of the agricultural sector. Generally, the statistical model used to measure TE is a stochastic frontier model with the noise being normally distributed and the inefficiency being half-normally distributed. The problem is that the model is not robust when outlier observations occur. The results of estimating technical efficiency will be inaccurate if there are outliers in the observed data. This study proposed a stochastic production frontier model with a fat-tailed distribution to overcome outlier observations. This study used two stochastic models with fat-tailed distribution used in this study: Cauchy-half normal and normal-Rayleigh stochastic models. The translog production function was selected as connecting the input and output. These two models were applied to estimate the technical efficiency of rice farming in Central Kalimantan. The results showed that the proposed model could reduce or eliminate outliers in the remaining inefficiencies. In addition, the range of technical efficiency values had also narrowed. The MAE of the Cauchy-half normal and normal-Rayleigh models are 0.84 and 1.14, respectively.

Corresponding Author:
Retna Nurwulan
Departement of Statistics,
IPB University,
Raya Dramaga Road, Babakan, Dramaga 16680, Bogor, Indonesia.
Email: retnanurwulan@apps.ipb.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v5i2.19597

1. INTRODUCTION

Technical efficiency is an indicator to measure the performance of a company. One method to estimate technical efficiency is stochastic frontier analysis (SFA). This method has been widely used in various economic sectors, such as agriculture [1], industry [2], banking [3], and many other sectors [4]. In the agricultural sector itself, this method has been used in various parts of the world, such as South Asia [5], America [6], Africa [7], and Europe [8].

SFA is a development of deterministic frontier analysis (DFA). In DFA, when maximum output is not achieved, presumably only caused by inefficiencies in processing existing inputs. However, in reality, many other factors can cause maximum output not to be performed. The SFA model exists to accommodate these other factors. This is one of the advantages of using SFA over other models covering all aspects, namely statistical noise, measurement errors, and external vibration outside the control of the production unit [9]. In addition, this model also produces good performance for models with single-output and multi-input [10]. The SFA model decomposes the residual into noise (v) and inefficiency (u). The commonly used SFA model has a normal distribution of noise and a half normal distribution of inefficiency. However, this model has limitations when there are outliers in the observed data [11]. This model is sensitive to outliers [12]. The presence of outliers can interfere with the model's performance both in terms of estimating the frontier function and the efficiency itself. The presence of the top outlier makes the frontier function turn higher; this makes the efficiency underestimate [13]. Outlier observations also widen the range of efficiency scores [14].

Handling outlier observations can be done by eliminating these observations. Outliers are removed along with other observations in the vicinity of outlier observations to produce a better estimate of the production frontier [15]. However, this step only sometimes produces an accurate production frontier [13]. Therefore, we need a method that can overcome the outliers but involves all observations. One of them is the SFA model with a fat-tailed distribution of the residual. The advantage of the fat-tailed distribution is having a heavier tail than the normal distribution, so the probability of covering extreme values is greater than the normal distribution. The SFA model with fat-tailed distribution can produce estimates of technical efficiency better than the conventional model when the observations contain outliers [14], [16], [17]. The SFA model is formed by changing the noise distribution from normal to Cauchy while the inefficiency distribution remains half normal [16]. The SFA model is constructed by changing the inefficiency distribution from half normal to Rayleigh while the noise distribution remains normal [18].

The presence of outliers in production data is a problem that often occurs. Therefore, estimating technical efficiency using a robust SFA model is essential to produce the proper technical efficiency estimate. Technical efficiency needs to be calculated to evaluate the extent to which a production unit can achieve its maximum output with the resources it has. This study examines the best SFA model for estimating technical efficiency when the observed data contains outliers. This model is applied to rice production data for Central Kalimantan Province to assess the best model. Rice commodity is the leading staple food for Indonesian people. Based on data [19], rice consumption in Indonesia reaches 6.75 kg per capita monthly. This amount of consumption exceeds the consumption of other staple foods such as corn, cassava, and sweet potatoes. However, data [20] shows that rice production in the last four years has stagnated in the range of 50-60 million tons per year. Food security will be threatened ff this continues. This phenomenon needs to be addressed, one of which is by developing rice productivity in provinces outside Java. Central Kalimantan Province is a province with potential for development. This province is included in the national Food Estate development program area.

2. RESEARCH METHOD

2.1 Literature View

2.1.1 Stokastik Production Frontier: Normal-Half Normal Model

The basic model of the stochastic production frontier, namely the normal-half normal model, in which the noise is assumed to be normally distributed ($v_i \sim iid\ N(0, \sigma_v^2)$) and the inefficiency is assumed to have a half normal distribution ($u_i \sim iid\ N^+(0, \sigma_u^2)$) [21]. This basic model estimates the parameters using the maximum likelihood method. The principle of this method is to determine an estimated parameter that maximizes the likelihood function. It is necessary to define the probability density functions according to the theoretical distribution assumptions on the u_i and v_i components. The probability density function of v is (normal distribution):

$$f_v(v) = \frac{1}{\sigma_v\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2\sigma_v^2}\right\}. \tag{1}$$

The probability density function of $u \geq 0$ is (half normal distribution):

$$f_u(u) = \frac{2}{\sigma_u\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2\sigma_u^2}\right\} \tag{2}$$

Assuming u and v are independent, then the joint density function is

$$f_{u,v}(u, v) = \frac{2}{2\pi\sigma_u\sigma_v} \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\}. \tag{3}$$

since the residual $\varepsilon = v - u$, then the joint density function for u and ε is as follows:

$$f_{u,\varepsilon}(u, \varepsilon) = \frac{2}{2\pi\sigma_u\sigma_v} \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon + u)^2}{2\sigma_v^2}\right\} \tag{4}$$

The marginal density function of ε is obtained by integrating u with $f_{u,\varepsilon}(u, \varepsilon)$. The marginal density function of ε is as follows:

$$f_\varepsilon(\varepsilon) = \frac{2}{\sigma\sqrt{2\pi}}\left[1 - \Phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right]\exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\} \quad (5)$$

where $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$, $\lambda = \frac{\sigma_u}{\sigma_v}$, and $\Phi(.)$ are standard normal cumulative distribution functions.

Using the marginal function $f_\varepsilon(\varepsilon)$ we get the log-likelihood function which will be maximized for the model parameters. Here is the log-likelihood function [22]:

$$\begin{aligned}\ln(L) &= \sum_{i=1}^{n}\left\{\frac{1}{2}ln\left(\frac{2}{\pi}\right) - ln(\sigma) + ln\left[\Phi\left(-\frac{\lambda\varepsilon_i}{\sigma}\right)\right] - \frac{\varepsilon_i^2}{2\sigma^2}\right\} \\ &= n\frac{1}{2}ln\left(\frac{2}{\pi}\right) - nln(\sigma) + \sum_{i=1}^{n} ln\left[\Phi\left(-\frac{\lambda\varepsilon_i}{\sigma}\right)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\end{aligned} \quad (6)$$

Next is the estimation of the technical efficiency value (TE_i). In this study, the value of technical efficiency is the exponential of u_i where u_i is the expected value of u_i, the conditional ε_i $E(u_i|\varepsilon_i)$. If $u_i \sim iid\ N^+(0, \sigma_u^2)$, then the conditional function is

$$f(u_i|\varepsilon_i) = \frac{f(u_i, \varepsilon_i)}{f(\varepsilon_i)} = \frac{\frac{1}{\sigma_*\sqrt{2\pi}}\exp\left\{-\frac{(u_i - \mu_*)^2}{2\sigma_*^2}\right\}}{1 - \Phi\left(-\frac{\mu_*}{\sigma_*}\right)} \quad (7)$$

so that the value of $E(u_i|\varepsilon_i)$ can be obtained from the following description

$$E(u_i|\varepsilon_i) = \int_0^\infty u_i f(u_i|\varepsilon_i)\, du \quad (8)$$

The value of Technical Efficiency is finally obtained with the following formula [22]

$$\begin{aligned}E(u_i|\varepsilon_i) &= \mu_{*i} + \sigma_*\left[\frac{\phi(-\mu_{*i}/\sigma_*)}{1 - \Phi(-\mu_{*i}/\sigma_*)}\right] \\ &= \sigma_*\left[\frac{\phi(\varepsilon_i\,\lambda/\sigma)}{1 - \Phi(\varepsilon_i\,\lambda/\sigma)} - \left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right]\end{aligned} \quad (9)$$

$$TE_i = \exp\{-E(u_i|\varepsilon_i)\} \quad (10)$$

where $\mu_* = -\varepsilon\,\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_u^2\,\sigma_v^2/\sigma^2$.

2.1.2 Stochastic Production Frontier: Cauchy-Half Normal Model

In the Cauchy-half normal model, the distribution of the v component is replaced by a fat distribution, namely Cauchy $\left(v_i \sim iid\ Ca(0, \sigma_v^2)\right)$, while u remains in half normal distribution $(u_i \sim iid\ N^+(0, \sigma_u^2))$. Estimation of parameters in this model using the simulated maximum likelihood method. The following is a description to obtain the likelihood function. The probability density function of v is (Cauchy distribution)

$$f_v(v) = \frac{1}{\pi\sigma_v}\left[1 + \left(\frac{v}{\sigma_v}\right)^2\right]^{-1} \quad (11)$$

The probability density function of $u \geq 0$ is (half normal distribution)

$$f_u(u) = \frac{2}{\sigma_u\sqrt{2\pi}}\exp\left(-\frac{u^2}{2\sigma_u^2}\right) \quad (12)$$

Assuming u and v are independent, then the joint density function is

$$f_{u,v}(u,v) = \frac{2}{\sigma_v \sigma_u \pi \sqrt{2\pi^2}} \left[1 + \left(\frac{v}{\sigma_v}\right)^2\right]^{-1} \exp\left(-\frac{u^2}{2\sigma_u^2}\right) \tag{13}$$

error component $\varepsilon = v - u$ so that $v = \varepsilon + u$ then the joint density function of u and ε is:

$$f_{u,\varepsilon}(u,\varepsilon) = \frac{2}{\sigma_v \sigma_u \pi \sqrt{2\pi^2}} \left[1 + \left(\frac{\varepsilon + u}{\sigma_v}\right)^2\right]^{-1} \exp\left(-\frac{u^2}{2\sigma_u^2}\right) \tag{14}$$

The marginal density function of ε is obtained by integrating the joint function u and ε ($f_{u,\varepsilon}(u,\varepsilon)$)

$$f_\varepsilon(\varepsilon) = \int_0^\infty \frac{2}{\sigma_v \sigma_u \pi \sqrt{2\pi^2}} \left[1 + \left(\frac{\varepsilon + u}{\sigma_v}\right)^2\right]^{-1} \exp\left(-\frac{u^2}{2\sigma_u^2}\right) du \tag{15}$$

However, the solution of $f_\varepsilon(\varepsilon)$ cannot be obtained using standard techniques because the integral of the equation is not closed form, so the approach used is the simulation technique [16]. The equation $f_\varepsilon(\varepsilon)$ is the expectation of $f_v(\varepsilon + u)$, where u comes from a half-normal distribution,

$$h(u) = E\{f_v(\varepsilon + u) | u \geq 0\}; \qquad u \sim N^+(0, \sigma_u^2) \tag{16}$$

The equation $h(u)$ is estimated by

$$\hat{h} = \frac{1}{Q} \sum_{q=1}^{Q} f_v(\varepsilon + u_q) \tag{17}$$

where u_q is generated from the half-normal distribution. This gives the simulation probability density function for ε [16]:

$$\widehat{f_\varepsilon(\varepsilon)} = \frac{1}{Q\pi\sigma_v} \sum_{q=1}^{Q} \left[1 + \left(\frac{\varepsilon + u_q}{\sigma_v}\right)^2\right]^{-1}. \tag{18}$$

and the simulated log-likelihood function:

$$\ln SL = -n \ln (Q\pi\sigma_v) + \sum_{i=1}^{n} ln \sum_{q=1}^{Q} \left[1 + \left(\frac{\varepsilon + u_{qi}}{\sigma_v}\right)^2\right]^{-1} \tag{19}$$

With the log-likelihood simulation equation, the model parameters can be obtained in the same way as the conventional probability function maximization method. The value of $E(u_i|\varepsilon_i)$ and the estimation of its individual technical efficiency are [16]:

$$E(u_i|\varepsilon_i) = \frac{\sum_{q=1}^{Q} u_q \left[1 + \left(\frac{\varepsilon_i + u_q}{\sigma_v}\right)^2\right]^{-1}}{\sum_{q=1}^{Q} \left[1 + \left(\frac{\varepsilon_i + u_q}{\sigma_v}\right)^2\right]^{-1}} \tag{20}$$

$$TE_i = \exp\{-E(u_i|\varepsilon_i)\} \tag{21}$$

where u_q is generated from a half-normal distribution.

2.1.3 Stochastic Production Frontier: Normal-Rayleigh Model

In the normal-Rayleigh model, the distribution of the u component is replaced by a fat distribution, namely Rayleigh ($u_i \sim iid\ Ra(0, \sigma_u^2)$), while v remains normally distributed ($v_i \sim iid\ N(0, \sigma_v^2)$). Estimation of parameters in this model with maximum likelihood method. The following is a description to obtain the likelihood function. The probability density function of v is (normal distribution)

$$f_v(v) = \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left\{-\frac{v^2}{2\sigma_v^2}\right\} \tag{22}$$

The probability density function of $u \geq 0$ is (Rayleigh distribution)

$$f_u(u) = \frac{u}{\sigma_u^2} exp\left\{-\frac{u^2}{2\sigma_u^2}\right\} \tag{23}$$

Assuming u and v are independent, then the joint density function is

$$f_{u,v}(u,v) = \frac{u}{\sigma_u^2\sqrt{2\pi\sigma_v^2}} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\} \tag{24}$$

since the residual $\varepsilon = v - u$, then the joint density function for u and ε is as follows:

$$f_{u,\varepsilon}(u,\varepsilon) = \frac{u}{\sigma_u^2\sqrt{2\pi\sigma_v^2}} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\} \tag{25}$$

The marginal density function of ε is obtained by integrating u with $f_{u,\varepsilon}(u,\varepsilon)$. The marginal density function of ε is as follows [18]:

$$f_\varepsilon(\varepsilon_i) = \frac{\sqrt{2\pi\sigma^2}}{\sigma_u^2\sqrt{2\pi\sigma_v^2}}\left[\sigma\phi\left(\frac{\mu_i}{\sigma}\right) + \mu_i\Phi\left(\frac{\mu_i}{\sigma}\right)\right] exp\left\{\frac{\mu_i^2}{2\sigma^2} - \frac{\varepsilon_i^2}{2\sigma_v^2}\right\} \tag{26}$$

where $\sigma^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma_v^2+\sigma_u^2}$, $\mu_i = \frac{\sigma_u^2\varepsilon_i}{\sigma_v^2+\sigma_u^2}$, $\phi(.)$ and $\Phi(.)$ is standard normal pdf and cdf.

By using the marginal function $f_\varepsilon(\varepsilon)$ we get the log-likelihood function which will be maximized for the model parameters. Here is the log-likelihood function [18]:

$$\ln(L) = n\frac{1}{2}ln\sigma^2 - n\frac{1}{2}ln\sigma_u^2 - n\frac{1}{2}ln\sigma_v^2 + \sum_{i=1}^{n} ln\left[\sigma\phi\left(\frac{\mu_i}{\sigma}\right) + \mu_i\Phi\left(\frac{\mu_i}{\sigma}\right)\right] + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mu_i^2 - \frac{1}{2\sigma_v^2}\sum_{i=1}^{n}\varepsilon_i^2 \tag{27}$$

Estimating of individual technical efficiency values is done by exponentiating the value of $E(u_i|\varepsilon_i)$. The $E(u_i|\varepsilon_i)$ value of and the estimation of the individual technical efficiency are [18]:

$$E(u_i|\varepsilon_i) = \frac{\mu_i\sigma\phi(\mu_i/\sigma) + (\mu_i^2+\sigma^2)\Phi(\mu_i/\sigma)}{\sigma\phi(\mu_i/\sigma) + \mu_i\Phi(\mu_i/\sigma)} \tag{28}$$

$$TE_i = exp\{-E(u_i|\varepsilon_i)\} \tag{29}$$

where $\sigma^2 = \frac{\sigma_u^2\sigma_v^2}{\sigma_v^2+\sigma_u^2}$, $\mu_i = \frac{\sigma_u^2\varepsilon_i}{\sigma_v^2+\sigma_u^2}$, $\phi(.)$ and $\Phi(.)$ are standard normal pdf and cdf.

2.2 Method

2.2.1 Data

The data that used in this paper were from Statistics Indonesia (BPS), the results of the 2017 Agricultural Cost Structure Survey (SOUT2017-SPD) of Central Kalimantan Province in 2017. The data used consists of 3.646 rice farming household data. The unit of observation was the household. The variables used in this study are presented in Table 1.

Table 1. Research variables

Variables	Description	Unit
Y	Rice production	Kg
X_1	Area	M^2
X_2	Seeds	Kg
X_3	Labours	Day of Work (DoW)
X_4	Fertilizers	Kg

Rice production is the amount of rice production harvested in the standard quality of Harvested Dry Grain; the area is harvested; the seed is the number of seeds used; the labours are paid and unpaid workers/ family workers, both male and female. The following formula obtains the DoW unit on the labours variable:

$$\text{DoW} = (\Sigma\,\text{labour} \times \text{working days} \times \text{working hours per day})/8$$

Fertilizers consist of urea, TSP/SP36, ZA, KCL, NPK, and organic fertilizers, both subsidized and non-subsidized fertilizers [23].

An overview of the value of each variable for the ten observed data (the first 5 data and the last 5 data) is presented in Table 2.

Table 2. Example of data visualization for ten observed data

No	Y	X_1	X_2	X_3	X_4
1	1350	10000	50	19.875	1
2	225	5000	45	8.375	1
3	750	10000	48	20.624	1
4	600	5000	45	14.750	1
5	990	5000	35	13.500	1
…	…	…	…	…	…
…	…	…	…	…	…
….	….	….	….	….	….
3642	1279	13000	40	83.500	1
3643	1418	13000	40	81.750	1
3644	1360	13000	40	95.875	10
3645	1336	20000	60	97.625	1
3646	1162	13000	40	75.750	1

2.2.2 Data Analysis

Data exploration

Data exploration is carried out using scatterplot and boxplot. Exploration was carried out to see the relationship between the Y and X variables and any outliers.

Production Function Selection

Before entering into the modeling, the production function is selected first, connecting the output variable (Y) with the input variable (X). There are two candidate production functions: the Cobb-Douglas production function and the transcendental logarithmic (translog) production function. The likelihood ratio test (LR-test) is carried out to determine which production function is better. The linear equation for the SFA model with the Cobb-Douglas production function as the connecting function is as follows [24]:

$$ln\,Y_i = \beta_0 + \beta_1 lnX_{i1} + \beta_2 lnX_{i2} + \beta_3 lnX_{i3} + \beta_4 lnX_{i4} + v_i - u_i \tag{30}$$

Meanwhile, if the translog production function is the connecting function, the SFA linear form is as follows [24]:

$$\ln Y_i = \beta_0 + \sum_{j=1}^{4} \beta_j\, lnX_{ij} + \frac{1}{2}\sum_{j=1}^{4} \beta_j\, lnX_{ij} lnX_{ij} + \sum_{j=1}^{4}\sum_{k=1}^{4} \beta_{jk}\big(lnX_{ij}\big)(lnX_{ik}) + v_i - u_i, \quad \beta_{jk} = \beta_{kj} \tag{31}$$

where Y_i is the output of the i-th observation, X_{ij} is the j-th input of the i-observation, v_i is the noise of the i-th observation, and u_i is the inefficiency of the i-th observation. The hypotheses for the LR test are as follows:

H_0: Cobb-Douglas is better than Translog

H_1: Translog is better than Cobb-Douglas

The test statistics of this LR test are:

$$\text{LR} - \text{stat} = -2(l_{CD} - l_{TR}) \tag{32}$$

where $\text{LR} - \text{stat}$ is the test statistic of the log-likelihood ratio test, l_{CD} is the log-likelihood value of the SFA model with the Cobb-Douglas link function, and l_{TR} is the log-likelihood value of the SFA model with the translog link function. The decision to reject H_0 if the $\text{LR} - \text{stat}$ value is greater than the table statistic, namely $\chi^2_{\alpha;df}$, the translog production function is better than Cobb-Douglas. However, if the $\text{LR} - \text{stat}$ is less than $\chi^2_{\alpha;df}$, there is not enough evidence to reject H_0, so the Cobb-Douglas production function is better.

Estimation of Parameters

The estimation of model parameters using the two proposed models is then compared with the conventional model. Akaike Information Criterion (AIC) and Mean Absolute Error (MAE) indicators are used for model evaluation.

$$AIC = 2p - 2l \tag{33}$$

where p is the number of parameters in the model and l is the log-likelihood value of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i| = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{34}$$

where ε_i is the residual of the i-th observation, $\hat{y}_i$ is the predicted maximum output of the i-th observation, and y_i is the i-th observation output. This study uses two software, the rfrontier package in STATA and the sfaR package in Rstudio. Technical efficiency scores were compared between the three models, then studied descriptively. The flow chart of this research methodology is presented in Figure 1.

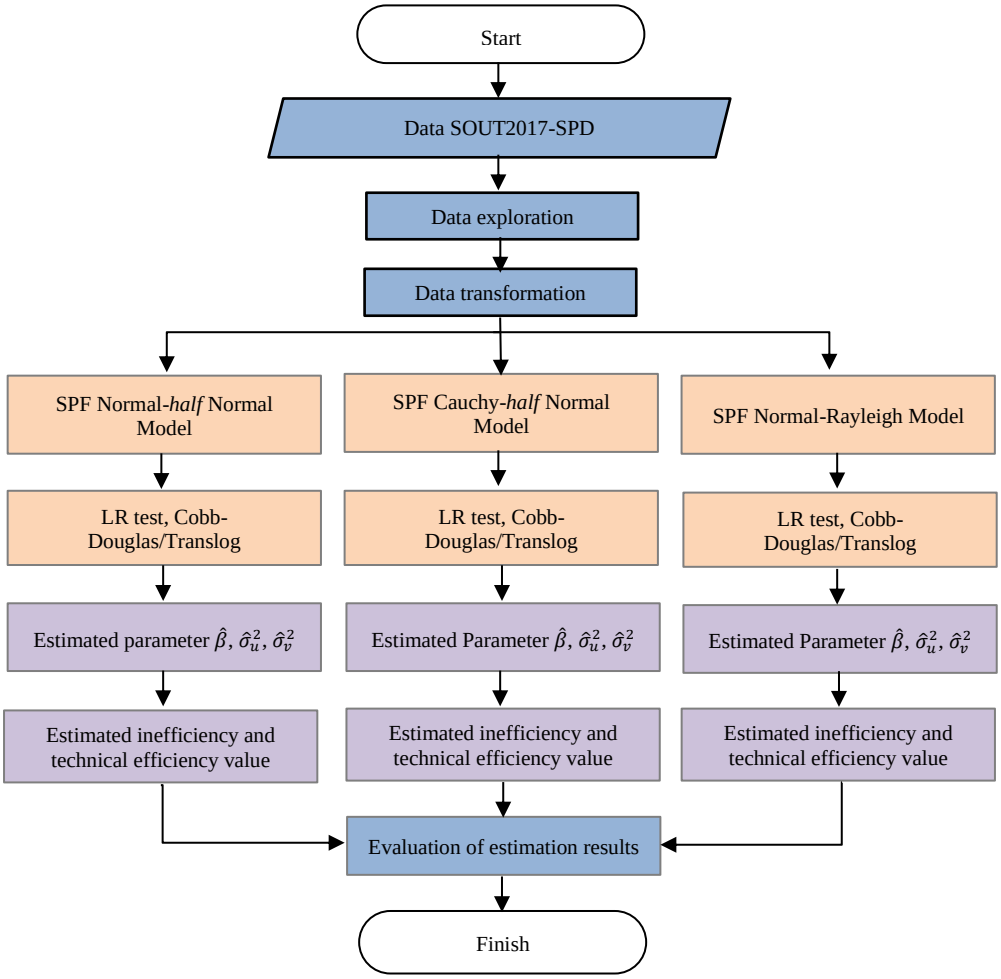


Figure 1. The flow chart of the research methodology

3. RESULTS AND ANALYSIS

The scatterplot of data points between the output variables, namely rice production, and the input variables, namely area, seeds, labor, and fertilizers, in pairs, can be seen in Figure 2. This scatterplot provides information on how the relationship between output and input variables. In Figure 2, it can be seen that the input variable has a positive relationship with the output variable. If the input variable has a large value, the output variable will have a large value and vice versa. If the input variable has a small value, the output variable will have a small value. Other information that can be given from this scatterplot is to provide an overview of the outliers. In Figure 2, it can be seen that quite a number of observations are far from the data set. This is further clarified by the boxplot presented in Figure 2. In the boxplot presentation, it can be seen that there are outliers on both the output variable and each input variable.

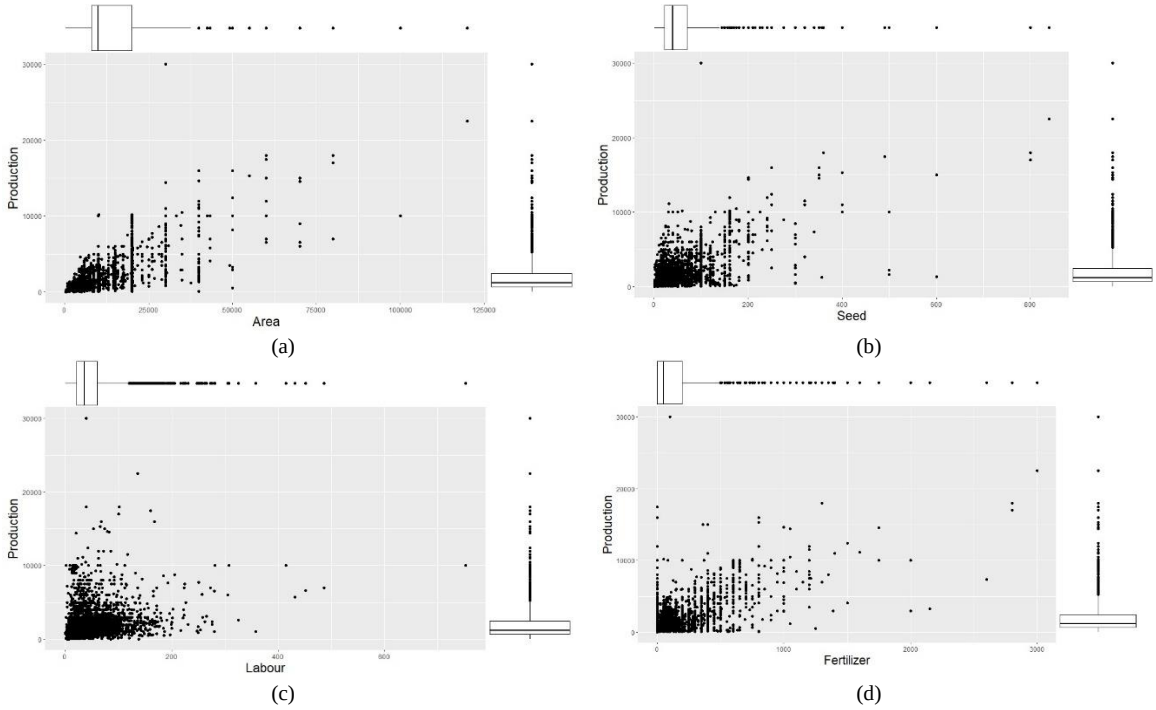


Figure 2. Scatterplot and boxplot between output and input
(a) land area, (b) seeds, (c) labors, and (d) fertilizer

The existence of outliers affects the pattern of point scattering, so the estimation of the production frontier is also affected. This can impact on an individual's estimate of his technical efficiency. Using the conventional SFA model sensitive to outliers can result in inaccurate estimates of technical efficiency. Therefore, it is necessary to estimate technical efficiency with an SFA model that is not sensitive to outliers, such as the SFA model, whose residual is a fat-tailed distribution.

A production function is needed in the SFA model as a liaison between output and input. Two production functions are commonly used, namely the Cobb-Douglas production function and the translog. The likelihood ratio (LR) test selects the best production function between the two. Table 3 shows the results of the LR test for each SFA model. The LR-stat value in the third model is greater than the critical value $\chi^2_{0.05;10}$, so it is decided to decide H_0. The translog production function is better than Cobb-Douglas as an input and output link for each SFA model.

Table 3. The results of the likelihood ratio test for the selection of the production function between Cobb-Douglas and translog

Model	Log Likelihood Value		LR-Stat	Critical value ($\chi^2_{\alpha;df}$)		Decision	Appropriate Model
	Cobb-Douglas	Translog		df	$\alpha = 0.05$		
Normal-Half Normal	-3768.13	-3629.71	276.839	10	18.307	Tolak H_0	Translog
Cauchy-Half Normal	-3875.97	-3713.38	325.175	10	18.307	Tolak H_0	Translog
Normal-Rayleigh	-3797.92	-3669.76	256.319	10	18.307	Tolak H_0	Translog

This study compares the conventional SFA model, namely normal-half normal, with the SFA development model, namely Cauchy-half normal and normal-Rayleigh. Table 4 shows the results of the estimated parameters of the SFA model using the three models. Generally, the parameter estimates for the three models give values that are not too different. Even the normal-Rayleigh model has the same sign as the normal-half-normal model for all the estimated coefficients of its parameters. While in the normal Cauchy-half model, there are several coefficients with different signs, namely on the constant variable, the square of the area variable, and the interaction of the seeding variable with labor.

In addition to the value of the parameter estimation coefficients, which are not too much different, the standard error values for each parameter estimate between models are also not much different. In the Cauchy-half normal model, the estimated parameter error for each variable is smaller than in the conventional model except for the constant variable. Parameter estimation using the Cauchy-half normal model is more precise than the conventional model. In addition, the significance level of the parameter estimates also increases. This is following previous studies where the SFA model with fat-tailed noise increases robustness in parameter estimation against outliers [14], [25]. Meanwhile, in the normal-Rayleigh model, there are several estimations of variable parameters whose standard error values are higher or lower than in conventional models.

Table 4 also presents the estimated variance for noise (σ_v^2) and inefficiency (σ_u^2). These two residual components are significant at the 0.001 level of significance for the three models. The significance of the two components shows that noise and inefficiency affect the prediction of frontier output. The estimated value of σ_u^2 is greater than σ_v^2 in each model. It shows that the deviation of the production unit to its maximum output (frontier) is more due to inefficiency.

Table 4. Estimating of modeling results parameters using normal-half normal, Cauchy-half normal, and normal-Rayleigh.

Variable	Parameter	Normal-Half Normal		Cauchy-Half Normal		Normal-Rayleigh	
		Coefficient	Standar Error	Coefficient	Standar Error	Coefficient	Standar Error
Constant	β_0	1.552	1.253	-0.170	1.287	1.858	1.243
$\ln X_1$	β_1	0.498	0.332	0.954**	0.304	0.522	0.327
$\ln X_2$	β_2	-1.293***	0.206	-1.304***	0.195	-1.284***	0.200
$\ln X_3$	β_3	1.376***	0.217	1.266***	0.212	1.354***	0.220
$\ln X_4$	β_4	0.150*	0.065	0.247***	0.064	0.124 ·	0.066
$(\ln X_1)^2$	β_{11}	0.015	0.024	-0.015	0.019	0.014	0.023
$(\ln X_2)^2$	β_{22}	0.035*	0.014	0.074***	0.010	0.031*	0.014
$(\ln X_3)^2$	β_{33}	0.055***	0.012	0.068***	0.015	0.055***	0.012
$(\ln X_4)^2$	β_{44}	0.032***	0.004	0.035***	0.003	0.064***	0.004
$\ln X_1 \ln X_2$	β_{12}	0.114***	0.029	0.108***	0.025	0.116***	0.028
$\ln X_1 \ln X_3$	β_{13}	-0.176***	0.029	-0.154***	0.025	-0.176***	0.029
$\ln X_1 \ln X_4$	β_{14}	-0.008	0.009	-0.019*	0.008	-0.006	0.009
$\ln X_2 \ln X_3$	β_{23}	0.042*	0.019	-0.012	0.015	0.047*	0.019
$\ln X_2 \ln X_4$	β_{24}	-0.020**	0.007	-0.013*	0.006·	-0.022**	0.007
$\ln X_3 \ln X_4$	β_{34}	-0.017**	0.006	-0.022***	0.005	-0.013*	0.006
	σ_v^2	0.137***	0.069	0.095***	0.072	0.094***	0.111
	σ_u^2	0.884**	0.046	0.941***	0.013	0.834***	0.042

Note: Signif. codes: <0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' Not significant ' '

The values of AIC and MAE measure the goodness of the three models. In Table 5, it can be seen that the AIC values of the three models are not too different, namely in the range of 7000. The smallest AIC value is the conventional model, which is normal-half normal at 7293.414. The AIC value only provide information about quality relative to other model, not about the absolute quality of a model. So, it is important to evaluate the residual of the model. Table 5 also presents the evaluation of the model using the MAE value to see how good the model is from the residual. The conventional model generates the smallest MAE value. The results are in line with the evaluation using AIC. However, the other two models give different results with AIC. The MAE value of the Cauchy-half normal model is smaller than the normal-Rayleigh model. Based on the values of AIC and MAE, the normal-half normal model provides a better prediction of frontier output than the other two models in the data of this study.

Table 5. Comparison of AIC and MAE

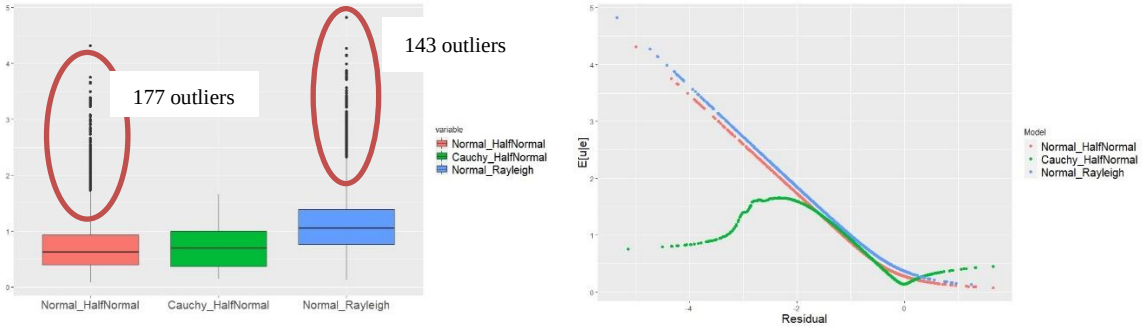
Model	AIC	MAE
Normal-Half Normal	7293.414	0.7903
Cauchy-Half Normal	7460.758	0.8412
Normal-Rayleigh	7373.524	1.1422

The main feature of the SFA model is the presence of a one-sided error that represents technical inefficiency. It is essential to examine the presence of this one-sided error in the SFA model [26]. The test was carried out with the LR test between the translog regression model and the SFA. Table 6 shows the results of the likelihood ratio test for the presence of a one-sided error. The LR-stat value is greater than the critical value, so it is concluded that there is a one-sided error in the model. The SFA analysis has been adequately carried out. The value of technical efficiency can be done with SFA modeling.

Table 6. The results of the likelihood ratio test between the SFA and the translog

Model	LR-Stat	Critical value		Decision	Remark
		df	$\alpha = 0.05$		
Normal-Half Normal	362.683	1	3.841	Tolak H_0	One side error existence
Cauchy-Half Normal	195.339	1	3.841	Tolak H_0	One side error existence
Normal-Rayleigh	282.574	1	3.841	Tolak H_0	One side error existence

In the stochastic frontier analysis, the main concern is the estimation of individual efficiency. After the model parameters are estimated, the next step is to estimate the technical efficiency. Estimating the technical efficiency depends on the value of $E(u_i|\varepsilon_i)$ (the expected value of the residual inefficiency). Figure 3 shows the diagnosis of the value of $E(u_i|\varepsilon_i)$ of the three models. In Figure 3(a), a boxplot of $E(u_i|\varepsilon_i)$, it can be seen that in the conventional model, there is an outlier in $E(u_i|\varepsilon_i)$. The SFA model with the fat-tailed distribution can reduce or even eliminate these outliers. In the normal-Rayleigh model, the outliers are only slightly reduced. Meanwhile, in the Cauchy-half normal model, there are absolutely no outliers. All observations fall within the range of distribution. If seen in Figure 3(b), which is a scatterplot between the residuals and $E(u_i|\varepsilon_i)$, it can be seen that in the Cauchy-half normal distribution, the observations in the upper tail are decreased downwards. In contrast, the observations in the lower tail are decreased upwards.



(a) (b)

Figure 3. Diagnosis of $E(u_i|\varepsilon_i)$: (a) Boxplot of $E(u_i|\varepsilon_i)$ and (b) Plot between residual and $E(u_i|\varepsilon_i)$

Table 7 shows the model residual's variance, which is recalculated based on the estimated variance results presented in Table 4. Based on the calculation results presented in Table 7, it can be seen that the estimated variance values given by the three models are not too different. In the residual inefficiency component, the normal-half normal model generates the smallest variance value. Meanwhile, the residual noise component of the fat-tailed model is smaller than the conventional model. Overall, the residual variance of the fat-tailed model is smaller than the conventional model. These results align with previous studies that used fat-tailed distributions when the data contained outliers can reduce the residual variance [14], [25].

Table 7. Model residual variance

	Normal-Half Normal	Cauchy-Half Normal	Normal-Rayleigh
Var (u)	0.321	0.342	0.358
Var (v)	0.137	0.095*	0.094
Var (e)	0.458	0.436	0.452

Note: * is taken from the value of σ_v^2 in Table 3 for the normal Cauchy-half model

A summary of the results of the estimated efficiency with the three models is shown in Table 8. The model with the fat-tailed distribution produces a narrower range of efficiency values than the standard model. The shrinkage in the fat-tailed distribution, as shown in Figure 3, narrows the efficiency range. This result aligns with previous research that the robust model will narrow the estimated technical efficiency scores. The maximum values of the estimated efficiency score of the fat-tailed distribution model are lower than the conventional model. The minimum value in the Cauchy-half normal model is higher than in the conventional model. At the same time, the normal-Rayleigh model is not the case. So, the Cauchy-half normal model has

the smallest range compared to the three models. It can be seen in Table 8 that the normal-Rayleigh model has a smaller median and mean value than the other two models. A comparison of the distribution of technical efficiency values for the three models is shown in Figure 4.

Table 8. Summary of technical efficiency score

Model	Min	Median	Mean	Max	Range
Normal-HalfNormal	0.013	0.539	0.528	0.927	0.914
Cauchy-HalfNormal	0.191	0.501	0.527	0.872	0.681
Normal-Rayleigh	0.008	0.349	0.364	0.878	0.871

Figure 4 is a density plot for the technical efficiency values generated by each model. The density plot for the model with the fat-tailed distribution looks different from the density plot for the normal-half normal model. If in the normal-half normal model, the distribution of technical efficiency values tends to be left skewed. In contrast, the other two models tend to deviate to the right. In the Cauchy-half normal model, it can be seen in Table 7 that the lower tail values of the efficiency score shifted to be greater than the standard model. Meanwhile, in the normal-Rayleigh model, the tail values of the shifting efficiency scores become smaller. The observations with high-efficiency scores in the normal-half normal model shifted to a lower technical efficiency value in the normal-Rayleigh model.

Figure 4. Density plot of technical efficiency score

Based on the results of estimating the model parameters and the results of estimating the value of technical efficiency, it can be said that the normal Cauchy-half model is a better model than other models in estimating the value of technical efficiency when the observed data contains outliers. Based on the normal Cauchy-half model, the average technical efficiency of rice farming households in Central Kalimantan in 2017 was only around 0.527. This means that rice production produced by rice farming households has only reached 52.70 percent of the total production, which should have been achieved by using broad inputs, seeds, labor, and fertilizers. With the existing inputs, rice farming households in Central Kalimantan still have the potential to increase their production.

4. CONCLUSION

Outliers in the data result in residual inefficiency. Frontier stochastic model with fat-tailed distribution can reduce or even eliminate outliers. In the normal-Rayleigh model, the outliers are only slightly reduced. In contrast, the normal Cauchy-half model can eliminate outliers. So, the estimation of technical efficiency with a narrower range is obtained. The Cauchy-half normal model best predicts technical efficiency when outliers exist. The MAE value of this model was 0.8421.

REFERENCES

[1] Sholikah S, Kadarmanto. The Analysis of Technical Efficiency of Inbred and Hybrid Lowland Rice Farming Business. *SOCA J.Sos. Ekon. Pertan*. 2020; 14(3): 381-397.
[2] Pratiwi AM, Bendesa IKG, Yuliarmi N. Analisis Efisiensi dan Produktivitas Industri Besar dan Sedang di Wilayah Provinsi Bali (Pendekatan Stochastic Frontier Analysis). *JEKT*. 2014;7(1):73-79.
[3] Linh Trinh Doan Tuan. Technical Efficiency of Vietnamese Commercial Banks. *American Journal of Theoretical and Applied Business*. 2020; 6(2): 17-22.
[4] Martín-Rivero R., Ledesma-Rodríguez FJ, Lorenzo-Alegría RM. Technical Efficiency and Agglomeration Economies in the Hotel Industry: Evidence from Canary Islands. *Applied Spatial Analysis and Policy.* 2021.
[5] Bibi Z, Khan D, Haq Iul. Technical and Environmental Efficiency of Agriculture Sector in South Asia: A Stochastic Frontier Analysis Approach. *Environmental, Development, and Sustainability*. 2020.

[6] Moreno-Moreno JJ, Morente FV, Diaz MTS. Assessment of The Operational and Environmental Efficiency of Agriculture in Latin America and the Caribbean. *Agricultural Economics*. 2018; 64(2): 74-88.
[7] Obianefo CA, Nwigwe CA, Meludu TN, Anyasie IC. Technical Efficiency of Rice Farmers in Anambra State Value Chain Development Programme. *Journal of Development and Agricultual Economics.* 2020; 12(2): 67-74.
[8] Kumbhakar, Subal C, Lien, Gudbrand, Hardaker JB. Technical Efficiency in Competing Panel Data Models: A Study of Norwegian Grain Farming. *Journal of Productivity Analysis*. 2014; 41.
[9] Jeewanthi DG, Shantha AA. The Technical Efficiency of Small-scale Tea Plantation in Sri Lanka. *Iranian Journal of Management Studies.* 2021; 01(01): 128-149.
[10] Hong NB, Yabe M. Resource Use Efficiency of Tea Production in Vietnam: Using Translog SFA Model. *International Journal of Biology*. 2015; 7(9): 160.
[11] Campos MS, Costa MA, Gontijo TS, Lopes-Ahn AL. Robust Stochastic Frontier Analysis Applied to Brazilian Electricity Distribution Benchmarking Method. *Decision Analytics Journal*. 2022; 3.
[12] Fusco E, Benedetti R, Vidoli F. Stochastic Frontier Estimation Through Parametric Modelling of Quantile Regression Coefficients. *Empircal Economics*. 2022.
[13] Khezrimotlagh D, Cook WD, Zhu J. A Nonparametric Framework to Detect Outliers in Estimating Production Frontiers. *European Journal of Operational Research 286*. 2020; 375-388.
[14] Wheat P, Stead AD, Greene WH. Robust Stochastic Frontier Analysis: A Student's t-half normal Model with Application to Highway Maintenance Costs in England. *J. Product. Anal*. 2019; 51(1): 21-38.
[15] Banker RD, Chang H, Zheng Z. On The Use of Super-efficiency Procedure for Ranking Efficient Units and Identifying Outliers. *Annals of Operations Research*. 2017; 250(1): 21-35.
[16] Zulkarnain R, Indahwati. Robust Stochastic Frontier Using Cauchy Distribution for Noise Component to Measure Efficiency of Rice Farming in East Java. *J.Phys.: Conf.* 2021; Ser.1863 012031.
[17] Zulkarnain R, Djuraidah A, Sumertajaya IM, Indahwati. Utilization of Student's t Distribution to Handle Outliers in Technical Efficiency Measurement. *Media Statistika*. 2021; 14(1): 56-67.
[18] Hajargasht G. Stochastics frontier with a Rayleigh distribution. *J. Product Anal*. 2015; 44(2):199-208.
[19] BPS (Badan Pusat Statistik). Ringkasan Eksekutif Pengeluaran dan Konsumsi Penduduk Indonesia berdasarkan Survei Sosial Eonomi Nasional (Susenas) September 2021. Jakarta: BPS. 2021.
[20] BPS (Badan Pusat Statistik). Statistik Indonesia 2022. Jakarta: BPS. 2022.
[21] Kumbhakar SC, Peresetsky S, Schetynin Y, Zayytsev A. Technical Efficiency and Ineffieciency: Reliability of Standar SFA Models and A Misspecification Problem. *Econometrics and Statistics*. DOI: https://doi.org/10.1016/j.ecosta.2021.12.006. 2021.
[22] Santha AA. Stochastic Frontier Analysis: Theory and Practise. LAMBERT Academic Publishing. 2019.
[23] Ardiansyah M, Kurnia A, Sadik K, Djuraidah A, and Wijayanto H. Numerical Prediction of Paddy Weight of Crop Cutting Survey using Generalized Geoadditive Linear Mixed Model. *Journal of Physics*: Conference Series. 1863 012024. 2021; 1-17.
[24] Umar HS, Girei AA, Yakubu D. Comparison of Cobb-Douglas and Translog Froniter Model in The Analysis of Technical Efficiency in Dry-season Tomato Production. 2017; 17(2): 67-77.
[25] Stead AD, Wheat P, Greene WH. *Estimating Efficiency in The Presence of Extreme Outlier: A Logistic-Half Normal Stochastic Frontier Model with Application to Highway Maintenance Costs in England*. Springer Proceedings in Business and Economics. Switzerland. 2018: 1-19.
[26] Kumbhakar SC, Wang H, Horncastle AP. A Practioner's Guide to Stochastic Frontier Analysis Using Stata. New York: Cambridge University Press. 2015.

BIBLIOGRAPHY OF AUTHORS

Retna Nurwulan, A Civil Servant at the Central Statistics Agency (BPS). Formal education in statistics is obtained from Sekolah Tinggi Ilmu Statistik (STIS) and IPB University. Work experience as section head of regional accounting and statistical analysis and field supervisor in various surveys in the fields of social, production, distribution, and regional accounting and statistical analysis. In addition, she is also involved as a data processing officer from field surveys, as well as an analyst and author of several publications published by BPS.

Anik Djuraidah, Lecturer at the Department of Statistics, IPB University. Author of scientific publications and books in the field of statistics. Formal education in statistics is obtained from IPB University. She is an expert in statistical theory and spatial data modeling and analysis.

Anwar Fitrianto, Lecturer at the Department of Statistics IPB University and Statistics Consultant. Formal education in statistics is obtained from the IPB University and University Putra Malaysia. Active in the International Statistical Engineering Association. He is an author of various scientific journals in the field of statistics.