STATISTICS IN TRANSITION new series, September 2025
Vol. 26, No. 3, pp. 117–133, https://doi.org/10.59139/stattrans-2025-030
Received – 04.09.2024; accepted – 02.04.2025

Exploring the stochastic production frontier in the presence of outliers: a simulation study

Anik Djuraidah[1], **Ismail Pranata**[2]

Abstract

Technical efficiency measures the performance of an observation unit in generating outputs effectively. The stochastic production frontier (SPF), which is a commonly used method for this purpose, determines how close a unit is to achieving maximum output based on its inputs. However, the outliers in the data can distort the accuracy of SPF models. To address this issue, various error distribution modifications like gamma, Student's-t, Weibull and Rayleigh distributions have been proposed. However, there is limited research comparing these distributions in handling outliers. This study describes a simulation conducted to compare five SPF models: Normal-half Normal, Normal-Gamma, Normal-Weibull, Normal-Rayleigh, and Student's-t-half Normal. Applying simulated data across nine scenarios with varying data amounts and outlier percentages, the findings demonstrate that the SPF Student's t-half Normal model provides the most accurate prediction of technical efficiency. Using a heavy-tailed distribution, such as the Student's t distribution, for the disturbance component is more effective in handling outliers in the response variable than modifying the inefficiency of the component distribution.

Key words: robust, outlier, simulation, stochastic production frontier, technical efficiency.

1. Introduction

Stochastic Frontier Analysis (SFA) is one of the analytical methods used to measure the performance of individual units by estimating the technical efficiency level in transforming input into output. SFA modelling in performance measurement assumes that the outputs generated by individual units are not always efficient or do not reach the maximum output limit (Coelli et al. 2005). Additionally, the SFA model assumes that inefficiency effects do not solely cause the deviations occurring in SFA modelling but can also be attributed to factors of statistical errors/disturbances (noise) (Mariyono

[1] School of Data Science, Mathematics, and Informatics, IPB University, Indonesia. E-mail: anikdjuraidah@apps.ipb.ac.id. ORCID: https://orcid.org/0000-0002-3163-4343.
[2] BPS-Statistics Indonesia, Indonesia. E-mail: pranata@bps.go.id ORCID: https://orcid.org/0009-0000-7973-7133.

2006). The error component of SFA is a combination of inefficiency and disturbance components. Both components are assumed to have specific parametric distributions. The disturbance component has a symmetric distribution that is assumed to be independently and identically distributed, while the inefficiency component is assumed to have a one-sided distribution independent of the disturbance component (Aigner et al. 2023).

SFA modelling is supported by only one assumption: the production approach that can utilize distance, income, cost, or profit functions (O'Donnell 2018). In general, the SFA model uses a distance approach to calculate the efficiency level of individual units, where the model measures the distance between the observed production and the estimated maximum production (frontier) at a given input. The SFA model using the distance function approach is called the Stochastic Production Frontier (SPF). The SPF model will estimate the maximum production by incorporating disturbance components into its estimation process, thus yielding a stochastic estimation of the maximum production. The estimation of the performance level of an individual unit in the SPF model begins with estimating the inefficiency. Estimating inefficiency in the SPF model can be measured through two approaches, namely the output-oriented and the input-oriented approach (Kumbhakar et al. 2015). The SPF model in measuring inefficiency estimation using an output-oriented approach is the difference between the actual production and the maximum possible production by a certain amount of input. The production function is theoretically common in various literature, namely the Cobb-Douglas (CB) and the Transcendental Logarithmic (Translog) (Zulkarnain et al. 2021).

The distribution of the standard SPF model assumes that the disturbance component follows a Normal distribution ($N[0, \sigma_v^2]$) and the inefficiency component follows a half-normal distribution ($N^+[0, \sigma_u^2]$) (Meeusen et al. 1977; Aigner et al. 2023). The standard SPF model, in its application, has limitations in estimating technical efficiency values when there are outliers present in a dataset (Campos et al. 2022). The presence of outliers in the SPF model can cause parameter estimation to become inaccurate and lead to an excessive spread in the estimated values of technical efficiency (Wheat et al. 2019). One approach to deal with outliers in a dataset when forming an SPF model is modifying the distribution assumptions with a more flexible parameterization to describe the data distribution (Greene 1990).

SPF models that use a more flexible distribution have been widely implemented to improve robustness against outliers. The SPF Normal-Gamma model replaces the standard half-normal inefficiency distribution with a Gamma distribution, enhancing the accuracy and distribution of technical efficiency estimates (Greene 1990). The SPF Normal-Gamma model can also provide smaller standard deviations and ranges of estimated technical efficiency compared to the standard SPF model. Similarly, the

Normal-Weibull SPF model modifies the inefficiency distribution to follow a Weibull distribution, which has a longer tail and higher kurtosis than the Gamma distribution, helping to mitigate outlier effects (Tsionas 2007). The SPF Normal-Rayleigh model assumes a two-sided normal distribution for disturbances and a one-sided Rayleigh distribution for inefficiency, allowing for non-zero inefficiency estimates and reducing the risk of underestimation (Hajargasht 2015). Meanwhile, the SPF Student's t-half Normal model modifies the disturbance component to follow a Student's t-distribution while keeping the half-normal inefficiency distribution. With heavier tails in the disturbance component, this model improves inefficiency estimation, resulting in smaller standard deviations, a narrower range, and a more favorable distribution pattern (Wheat et al. 2019).

This research compares various SPF models with different parameter distributions for disturbance and inefficiency components to identify the most accurate model for estimating technical efficiency in the presence of outliers. The optimal SPF model is one that effectively handles outliers, assessed through its application to both real and simulated data. Previous studies on SPF model development using real data have examined the SPF Normal-half Normal, SPF Normal-Gamma, and SPF Student's t-half Normal models (Pranata et al. 2023). This study expands the analysis by incorporating simulation data to evaluate additional models, including SPF Normal-Weibull and SPF Normal-Rayleigh. The simulation data consists of varying sample sizes with different outlier proportions.

2. Stochastic Production Frontier

The Stochastic Production Frontier (SPF) is a Stochastic Frontier Analysis (SFA) model that measures the performance of individual units by measuring the distance between actual production and the estimated production frontier, given the same inputs. The production frontier function in the model is a production function that incorporates disturbance effects in its estimates, making it inherently stochastic. The disturbance component describes the possibility of random shocks that affect the production process and cannot be controlled, such as weather, economy, etc. (Coelli et al. 2005). In the SPF model, there is also an inefficiency component (Aigner et al. 1968). The measurement of estimation inefficiency in the SPF model can be categorized into output-oriented inefficiency and input-oriented inefficiency (Kumbhakar et al. 2015). Output-oriented inefficiency calculates the distance between the produced output and the estimated output frontier (achievable maximum production) for a given input. On the other hand, input-oriented inefficiency calculates the distance between the actual input and the estimated input frontier (minimum required) to achieve a certain output. In this study, the approach for measuring inefficiency in the model uses the output-oriented inefficiency measurement. The form of the SPF model using the Cobb-Douglas production function is as follows:

$$\ln y_i = \beta_0 + \sum_{j=1}^{n} \beta_j \ln x_{ij} + \varepsilon_i, \quad \varepsilon_i = v_i - u_i \tag{1}$$

where ε_i represents the error model for observation i. The error model consists of two components, v_i is the disturbance component following symmetric distribution for observation i and u_i is the inefficiency component following one-sided distribution that is always positive for observation i.

One of the parameter estimation methods in the SPF model is maximizing the log-likelihood function based on the marginal density function of the error. When the marginal density function of the error is not in a closed form, the estimates of the parameter cannot directly maximize the log-likelihood function, but use the simulated maximum likelihood method (Train 2009). The equation for the log-likelihood function is as follows:

$$\ln \mathcal{L} = \ln \prod_{i=1}^{n} f(\varepsilon_i) \tag{2}$$

where $\mathcal{L}$ represents the likelihood function and $f(\varepsilon)$ represents the marginal density function of the error SPF model. The probability density function of the error in the SPF model is formed by composing the probability density functions of the disturbance and the inefficiency component. The probability density function for the error in the SPF model begins with the formation of a joint probability density function between the disturbance and the inefficiency component. Since both components are independent, their joint probability density function is the product of their density functions. The joint probability density function of the two components is as follows:

$$f(v, u) = f(v) \times f(u) \tag{3}$$

where $f(v)$ represents the probability density function of the disturbance component and $f(u)$ represents the probability density function of the inefficiency component. The joint probability density function $f(\varepsilon, u)$ is constructed by transforming the disturbance component using $v = \varepsilon + u$. The error's probability density function is then obtained by integrating the joint density over the inefficiency component. The equation is given as follows:

$$f(\varepsilon) = \int_0^\infty f(\varepsilon, u)\, du \tag{4}$$

Estimating inefficiency in the SPF model begins by forming the probability density function of inefficiency values where the error values are known $(f(u|\varepsilon))$. This function can be determined as follows:

$$f(u|\varepsilon) = \frac{f(u, \varepsilon)}{f(\varepsilon)} \tag{5}$$

In the SPF model, inefficiency is estimated by calculating the expected value of inefficiency given the error $\mathrm{E}(\mathrm{u}|\varepsilon)$. The form of this expected value is as follows:

$$\mathrm{E}(\mathrm{u}|\varepsilon) = \int_0^\infty \mathrm{u}\, \mathrm{f}(\mathrm{u}|\varepsilon)\, \mathrm{du} \tag{6}$$

The estimation of technical efficiency ($\widehat{\mathrm{TE}}$) in the SPF model follows the equation introduced by Jondrow et al. (1982), where technical efficiency is computed as the exponent of negative estimated inefficiency. The equation as follows:

$$\widehat{\mathrm{TE}} = \exp(-\mathrm{E}(\mathrm{u}|\varepsilon)) \tag{7}$$

Several types of SPF models include the following:

1. **The SPF Normal-half Normal Model:** The SPF Normal-half Normal model is a standard SPF model in which the disturbance component follows a symmetric Normal distribution ($\mathrm{N}(0, \sigma_v^2)$), and the inefficiency component follows a half Normal distribution ($\mathrm{N}^+(0, \sigma_u^2)$) Meeusen & Van Den Broeck (1977) and (Aigner et al., 2023). The marginal density function of the error in the SPF Normal-half Normal model is closed-form so that the estimation of the parameter maximizes the log-likelihood function.
2. **The SPF Normal-Gamma Model**: The SPF Normal-Gamma model is a SPF model where the disturbance component follows a symmetric Normal distribution ($\mathrm{N}(0, \sigma_v^2)$), and the inefficiency component follows a Gamma distribution ($\mathrm{G}(\alpha, \theta)$) as introduced by (Greene, 1990). Since the marginal density function of the error lacks a closed-form solution, its log-likelihood function cannot be directly maximized. Instead, parameter estimation can be performed using the maximum simulated likelihood method, as suggested (Greene 2003).
3. **The SPF Normal-Weibull Model**: The SPF Normal-Weibull model assumes that the disturbance component follows a symmetric Normal distribution ($\mathrm{N}(0, \sigma_v^2)$), and the inefficiency component follows a Weibull distribution. This model was first introduced by (Tsionas 2007). The Weibull distribution has a long-tailed curve shape, classified as heavy-tailed. The characteristic of a long-tailed curve shape allows the Weibull distribution to handle outliers or extreme observations in data distributions. The formed marginal density function of the error in the SPF Normal-Weibull model does not have a closed-form expression, so parameter estimation using simulated maximum likelihood (Nurwulan et al. 2023).
4. **The SPF Normal-Rayleigh Model**: The SPF Normal-Rayleigh model assumes that the disturbance component follows a symmetric Normal distribution ($\mathrm{N}(0, \sigma_v^2)$), and the inefficiency component follows a Rayleigh distribution $\mathrm{Ra}(0, \sigma_u^2)$, as introduced by Hajargasht (2015). The characteristics of the Rayleigh distribution in the SPF model are different from the half-normal and exponential distributions.

The half-normal and exponential distributions tend to produce inefficiency estimates close to zero, raising concerns about potential underestimation. In contrast, the Rayleigh distribution exhibits a broader spread of data, reducing concentration near zero and potentially providing more accurate inefficiency estimates. The marginal density function of the error in the SPF Normal-Rayleigh model has a closed-form so that the estimation of the parameter maximizes the log-likelihood function.

5. **The SPF Student's t-half Normal Model**: The SPF Student's t-half Normal model assumes that the disturbance component follows a symmetric Student's t distribution ($T(a)$) with degrees of freedom (a), while the inefficiency component follows a half-normal distribution ($N^+(0, \sigma_u^2)$) (Wheat et al., 2019). The Student's t distribution is a two-tailed distribution characterized by a long-tailed curve on both sides of the distribution. Like the Weibull distribution, the Student's t distribution is also categorized as heavy-tailed. The marginal density function of the error in the SPF Student's t-half Normal model is not a closed form. As a result, the parameter estimation process for this SPF model utilizes simulated maximum likelihood.

3. Method

This study compares the estimated coefficients of model parameters, inefficiency values, and technical efficiency among five Stochastic Production Frontier (SPF) models. The five SPF models include the SPF Normal-half Normal model, SPF Normal-Gamma model, SPF Normal-Weibull model, SPF Normal-Rayleigh model, and SPF Student's t-half Normal model. Additionally, this study will also identify the best-performing SPF model or the SPF model robust to outliers in the response variable observations. The data used in this research is generated through simulation, where simulated data is created for nine dataset scenarios. These scenarios are formed based on combinations of predefined data quantities and outlier observation percentages. The process of modelling SPF in this study utilizes the "sfaR" package (Dakpo et al. 2022) and the "rfrontier" package (Wheat et al. 2019).

3.1. Generating Simulated Data

This study uses data simulation with Monte Carlo simulations. Monte Carlo simulation data repeatedly generates random numbers or values of a variable with a specific distribution (Sartono 2005). The number of generated data in the simulation consists of three different data quantities: 200 data points as the small data category, 500 points data as the medium data category, and 1000 points data as the large data category. The varying sample data sizes aim to observe the influence of data quantity on enhancing estimation accuracy. The simulation outlier percentages are divided into three levels: 1%, 3%, and 5% of the response variable. It aims to assess the model's consistency in handling data containing outlier observations.

The production function used in the SPF model is the Cobb-Douglas production function, where the model's parameter coefficients are determined as follows: $\beta_0 = 0{,}1$, $\beta_1 = 2$, and $\beta_2 = 2$. Thus, the form of the simulated data SPF model is as follows:

$$\ln y_i = 0{,}1 + 2 \ln x_{1i} + 2 \ln x_{2i} + v_i - u_i \tag{8}$$

Generating simulated data starts by creating independent variable data, disturbance components, and inefficiency components. The model uses two independent variables, where both are mutually independent and are generated using a Uniform distribution (10, 20). In generating independent variable data, using a variance-covariance matrix is an identity matrix to ensure that the generated data for both variables are mutually independent. The disturbance components are assumed to follow a Normal distribution. The disturbance components will be generated with a mean of 0 and a variance of 0.25 ($N(0, 0{,}25)$). Meanwhile, the inefficiency components are assumed to follow a half-normal distribution. The inefficiency components will be generated with a mean of 0 and a variance of 1 ($N^+(0{,}1)$). The results of generated data from independent variables, disturbance component, and inefficiency component will be input into the model to obtain the generated data for the response variable. The process will be repeated one hundred times for each set of generated data.

The outlier data in the simulated data is a set of data significantly different from the generated data. The outlier will replace a randomly selected portion of the response variable data, and the predetermined outlier percentage will determine the amount of these replaced data.

3.2. Modelling

The SPF model has a characteristic that the residual distribution is skewed to the left (Kumbhakar et al., 2015). Therefore, an error distribution test is performed on the generated data before conducting the SPF modelling to ensure the error distribution is skewed to the left. The testing can be done by calculating the statistically significant Coelli Test (M3T) (Mutz et al. 2017). The M3T test counts the standardized third moment of the error from the least squares model and has an asymptotic distribution of a standard normal random variable. The equation for the M3T test is as follows: (Coelli 1995)

$$M3T = {m_3}\Big/{\sqrt{6m_2^3/N}} \tag{11}$$

where m_2 and m_3 are the second and third moments of the least square error, forming a distribution $N(0, 6m_2^3/N)$. The Hypothesis is as follows:

H_0: The error has a symmetric distribution.

H_1: The error has a skewed distribution.

The generated data can be used for SPF modelling if the M3T test results reject the null hypothesis (the M3T value falls outside the critical range of ±1.96) at a significance level of α=5% and the M3T value is negative. This condition indicates that the distribution of least squares errors from the generated data has a left-skewed curve.

Once it is confirmed that the generated data for SPF modelling has a left-skewed distribution, the next step is to proceed with SPF modelling using the five different SPF model forms to obtain parameter estimation values, inefficiency values, and technical efficiency values from the models.

3.3. Estimation Performance

The criterion for selecting the best SPF model is its ability to provide estimated technical efficiency close to the true values. It can be determined by the mean square error (MSE) of the estimated inefficiency and technical efficiency compared to the true values. The SPF model that yields the smallest MSE can be considered the best SPF model. The MSE for estimating inefficiency is the average of the squared differences between the estimated inefficiency and the true inefficiency across all observations and repetitions. The calculation of the MSE for inefficiency is as follows:

$$\mathrm{MSE(u)} = \frac{1}{\mathrm{NR}}\sum_{r=1}^{R}\sum_{i=1}^{N}(\hat{u}_{i,r} - u_{i,r})^2 \tag{9}$$

where N is the number of generated dataset scenarios, R is the number of repetitions, $\hat{u}_{i,r}$ is the estimated inefficiency for observation i and repetition r, $u_{i,r}$ is the true inefficiency (generated) for observation i and repetition r. Meanwhile, the MSE for technical efficiency is the average of the squared differences between the estimated technical efficiency and the true technical efficiency across all observations and repetitions. The MSE for technical efficiency is as follows (Sakouvogui et al. 2021):

$$\mathrm{MSE(TE)} = \frac{1}{\mathrm{nR}}\sum_{r=1}^{R}\sum_{i=1}^{N}(\widehat{\mathrm{TE}}_{i,r} - \mathrm{TE}_{i,r})^2 \tag{10}$$

where $\widehat{\mathrm{TE}}_{i,r}$ is the estimated technical efficiency for observation i and repetition r and $\mathrm{TE}_{i,r}$ is the true value (generated) technical efficiency for observation i and repetition r.

4. Results and Discussion

The simulation of data of the SPF model begins by generating data for nine scenarios. These nine scenarios consist of three sets of generated data, each with a sample size of 200, 500, and 1000. Within each set there are three variations of outlier percentages: 1%, 3%, and 5%. The distribution of generated data in nine scenarios can be observed in Figure 1. The data distribution plots in Figure 1 represent a sample from one dataset of the independent variable ($\ln x_1$) and response variable ($\ln y$) generated from a hundred repetitions for each scenario.

The generated data from the nine scenarios is tested using the M3T test to ensure that the generated data has the error distribution criteria of the SPF model. The results of the minimum and maximum M3T values across a hundred repetitions are presented in Table 1. The M3T test uses a significance level of $\alpha = 5\%$, and the critical values for rejecting H_0 a are when the M3T value is less than -1.96 or greater than 1.96. The range of the M3T values from the nine scenarios with a hundred repetitions for each simulation yields values smaller than -1.96. Hence, all the generated datasets reject the null hypothesis with the negative M3T values. In other words, all the generated data distributions have a left-skewed error distribution with a 95% confidence level.

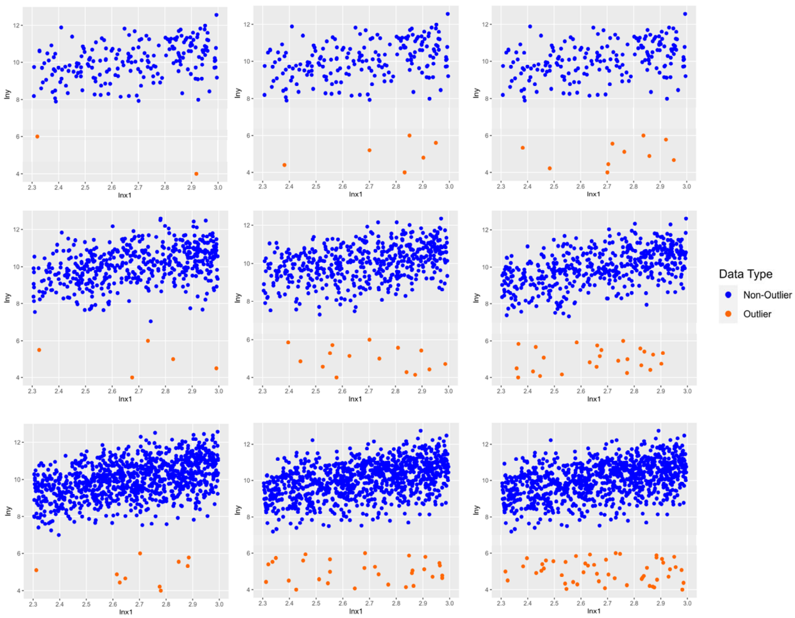


Figure 1: Plot the distribution of generated data between the response variable ($\ln y$) and the independent variable ($\ln x_1$) with outlier observation percentages 1%; 3%; and 5%, (a) a sample size of 200, (b) a sample size of 500, (c) a sample size of 1000

Source: own study.

Table 1. The range of M3T values across the nine simulation scenarios with 100 repetitions

Simulation Scenarios	M3T Score	
	Minimum	**Maximum**
Sample size=200		
Outlier 1%	-17.99	-5.63
Outlier 3%	-16.31	-10.15
Outlier 5%	-14.89	-10.01

Table 1: The range of M3T values across the nine simulation scenarios with 100 repetitions (cont.)

Simulation Scenarios	M3T Score	
	Minimum	**Maximum**
Sample size=500		
Outlier 1%	-22.07	-13.99
Outlier 3%	-24.67	-19.05
Outlier 5%	-23.07	-18.88
Sample size=1000		
Outlier 1%	-28.63	-20.44
Outlier 3%	-34.28	-27.39
Outlier 5%	-33.01	-28.12

Source: own study.

4.1. SPF Model Parameter Estimates

The distribution of estimated intercept (b_0) from the SPF Normal-half Normal model, SPF Normal-Gamma model, SPF Normal-Weibull model, SPF Normal-Rayleigh model, and SPF Student's t-half Normal model is presented in Figure 2. From the generated simulation data, the SPF Student's t-half Normal model is the model that can estimate the parameter value β_0 close to the true value. This can be seen from the distribution of estimated b_0, which are around the true parameter, and the median of b_0 has the smallest bias to the true parameter compared to the other four SPF models across the nine simulation scenarios.

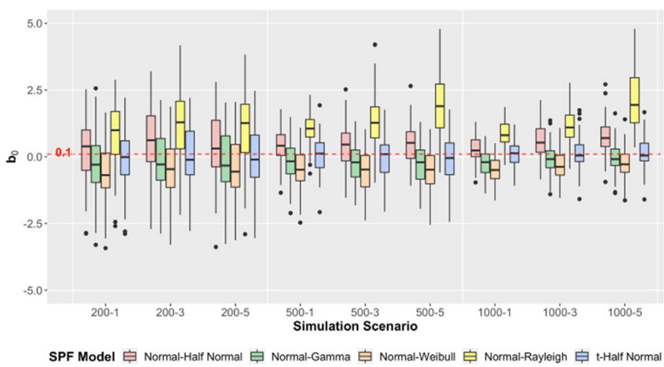
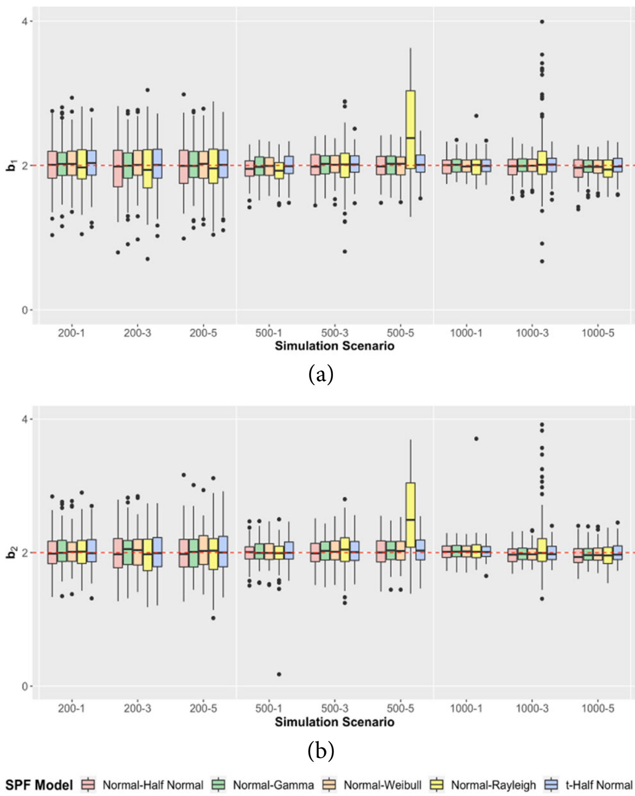


Figure 2: The estimation distribution of b_0 across nine simulation data scenarios using the SPF Normal-half Normal, SPF Normal-Gamma, SPF Normal-Weibull, SPF Normal-Rayleigh, and SPF Student's t-half Normal models

Source: own study.

Meanwhile, the SPF Normal-Rayleigh model exhibits significant estimation bias, with the distribution of b_0 estimates deviating considerably from the true parameter value and the estimated median of b_0 being far from the actual value. The median of the b_0 in the Normal-half Normal SPF model is greater than that in the SPF Student's t-half Normal model. Furthermore, in the Normal-Half Normal SPF model, the median of the b_0 increases with higher percentages of outliers. This indicates that the accuracy of the standard SPF model in estimating intercept coefficients is greatly affected when outlier observations are present in the data. The SPF Normal-Gamma model and SPF Normal-Weibull model have similar characteristics in estimating intercept parameter coefficients. Both SPF models have an estimated distribution of b_0 that tends to be lower than the true intercept. Despite having a heavier tail distribution than the SPF Normal-Gamma model, the SPF Normal-Weibull model does not yield better intercept estimations than the SPF Normal-Gamma model. A box plot in Figure 2 shows that as the sample size used in the SPF model increases, the variance of b_0 estimation becomes smaller. It is indicated by the decreasing length of the box plot as the simulation data size grows.



(a)

(b)

Figure 3: Distribution of estimated slope parameter across nine simulation data scenarios using the SPF Normal-half Normal, SPF Normal-Gamma, SPF Normal-Weibull, SPF Normal-Rayleigh, and SPF Student's t-half Normal models, (a) estimated of b_1 (b) estimated of b_2.

Source: own study.

The distribution of the estimated slope from nine simulation scenarios, composed of two coefficient parameters, namely the estimated b_1 and b_2, is depicted in the box plot diagram presented in Figure 3. The five SPF models were used to estimate the coefficients b_1 and b_2, which generally had a median parameter estimate close to the true values. However, in simulations with 500 observations and a 5% outlier percentage, the SPF Normal-Rayleigh model yields median slope estimates that deviate from the true values.

4.2. Inefficiency and Technical Efficiency Prediction

Figure 4 compares the estimated inefficiency distributions of the SPF Normal-Half Normal, SPF Normal-Gamma, SPF Normal-Weibull, SPF Normal-Rayleigh, and SPF Student's t-Half Normal models across nine simulation scenarios. The SPF Normal-Half Normal and SPF Normal-Rayleigh models tend to overestimate inefficiency, while the SPF Normal-Gamma and SPF Normal-Weibull models generally underestimate it. Among these models, the SPF Student's t-Half Normal model provides the most accurate inefficiency estimates, with its distribution closely aligning with the true inefficiency values. This is evident from the median of the estimated inefficiency distribution being close to the true inefficiency median, with the overall estimated distribution falling within the range of the true inefficiency values. In the SPF Student's t-Half Normal model, the number of estimated inefficiency outliers closely matches the true inefficiency outliers. In contrast, the other four SPF models tend to produce significantly more outliers than the true inefficiency distribution.

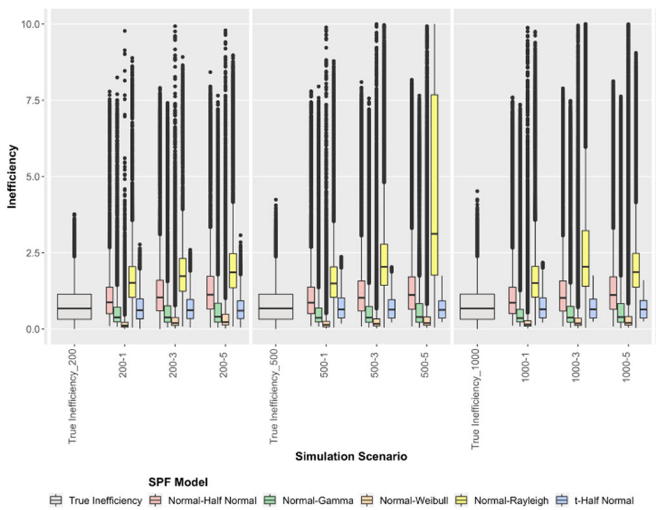


Figure 4: The distribution of true inefficiency and the estimated distributions of inefficiency using the SPF Normal-half Normal, SPF Normal-Gamma, SPF Normal-Weibull, SPF Normal-Rayleigh, and SPF Student's t-half Normal models across nine simulation scenarios

Source: own study.

In addition to comparing the estimated inefficiency scores of the five SPF models using box-and-whisker plots, the mean squared error (MSE) was calculated to assess how accurately each model predicts the true inefficiency. The MSE results, presented in Table 2, compare the estimated inefficiency with the true values across nine simulation scenarios using the five SPF models. The SPF Student's t-Half Normal model consistently produces lower MSE values than the other four SPF models across all nine simulation scenarios. In contrast, the SPF Normal-Weibull model performs the worst in predicting inefficiency, as indicated by its significantly higher MSE compared to the other models. Increasing the outlier percentage in the response variable generally leads to higher MSE values. However, this trend does not apply to the SPF Normal-Weibull model.

Table 2: Mean square error of the estimated inefficiency of five SPF models across nine simulation scenarios

Simulation scenario	MSE Scores of SPF				
	Normal-half Normal	Normal-Gamma	Normal-Weibull	Normal-Rayleigh	Student's t-half Normal
Sample size=200					
Outlier 1%	0.4619	0.4890	85.5771	1.1897	0.1941
Outlier 3%	1.1106	0.8847	3.1932e+05	2.2624	0.2100
Outlier 5%	1.7314	1.2923	6.4129	3.2150	0.2313
Sample size=500					
Outlier 1%	0.4524	0.4668	1.5906e+10	12.6826	0.1666
Outlier 3%	1.0874	0.8626	1.1455e+05	4.6814	0.1897
Outlier 5%	1.7247	1.2943	3.6382	39.2181	0.2081
Sample size=1000					
Outlier 1%	0.4326	0.4251	8.7324e+08	1.5060	0.1635
Outlier 3%	1.0787	0.8547	21.1244e+05	7.2978	0.1806
Outlier 5%	1.7409	1.3054	34.9856	290.9171	0.1982

Source: own study.

The distribution of estimated technical efficiency from the five SPF models is presented in Figure 5, with each model yielding a distinct distribution. Among the nine simulation scenarios, the SPF Student's t-half Normal model produces estimated technical efficiency distributions that closely align with the true distribution. This is evident from the box-and-whisker plot, which closely resembles the true technical efficiency plot, with the estimated median exhibiting minimal bias relative to the true median. The estimated technical efficiency of the Student's t-half Normal and SPF Normal-half Normal models does not contain outliers, whereas the other three SPF models do.

The accuracy of the model in predicting technical efficiency values is measured using the Mean Squared Error (MSE) between the estimated and true technical efficiency values. The MSE results for nine simulation scenarios are presented in Table 3. The SPF Student's t-half Normal model has the smallest MSE among all SPF models across all nine simulated data scenarios. In contrast, the SPF Normal-Weibull and SPF Normal-Rayleigh models have the highest MSE among the five SPF models.

These results indicate that handling outliers in the response variable by replacing the distribution of inefficiency components with a heavy-tailed distribution does not impact the SPF model's ability to improve the accuracy of predicting true technical efficiency. Conversely, modifying the distribution of disturbance components to a heavier-tailed distribution can enhance the model's accuracy in predicting true technical efficiency.

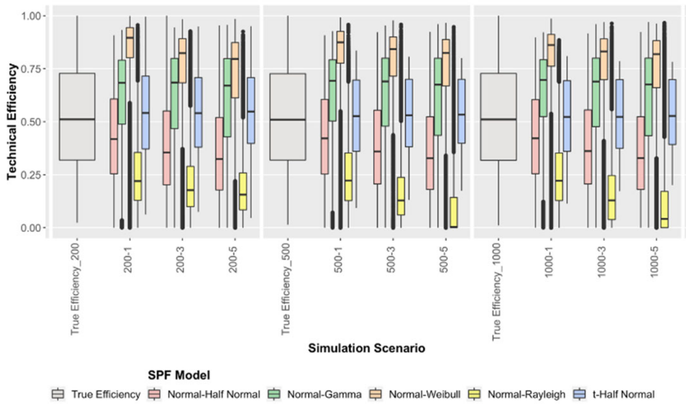


Figure 5: The distribution of true technical efficiency and the estimated technical efficiency using the SPF Normal-half Normal, SPF Normal-Gamma, SPF Normal-Weibull, SPF Normal-Rayleigh, and SPF Student's t-half Normal models across nine simulation scenarios

Source: own study.

Table 3: Mean square error of the estimated technical efficiency of five SPF models across nine simulation scenarios

Simulation scenario	MSE Scores				
	SPF Normal-half Normal	**SPF Normal-Gamma**	**SPF Normal-Weibull**	**SPF Normal-Rayleigh**	**SPF Student's t-half Normal**
Sample size=200					
Outlier 1%	0.0429	0.0520	0.1467	0.1058	0.0376
Outlier 3%	0.0578	0.0531	0.1044	0.1331	0.0371
Outlier 5%	0.0682	0.0564	0.0946	0.1491	0.0391
Sample size=500					
Outlier 1%	0.0419	0.0497	0.1312	0.1069	0.0314
Outlier 3%	0.0573	0.0523	0.1118	0.1715	0.0329
Outlier 5%	0.0683	0.0563	0.1039	0.2501	0.0344
Sample size=1000					
Outlier 1%	0.0419	0.0470	0.1205	0.1072	0.0308
Outlier 3%	0.0564	0.0517	0.1044	0.1768	0.0318
Outlier 5%	0.0676	0.0559	0.1005	0.2336	0.0330

Source: own study.

5. Conclusions

The findings suggest that the SPF Student's t-half Normal model produces more accurate intercept and slope estimates than the SPF Normal-half Normal, SPF Normal-Gamma, SPF Normal-Weibull, and SPF Normal-Rayleigh models. Additionally, it provides the most accurate inefficiency and technical efficiency predictions among the SPF models. Changing the assumption of the distribution in the disturbance component to a distribution with heavy tails, such as the Student's t distribution, is more effective in handling outliers in the response variable than altering the assumption of the distribution in the inefficiency component. The SPF Student's t-half Normal model is robust in dealing with outliers in the response variable when predicting technical efficiency.

Acknowledgement

The authors express gratitude to IPB University and the Statistics of Indonesia for their support, which has enabled the completion of this research.

References

Aigner, D. J., Chu, S. F., (1968). On estimating the industry production function. *American Economic Review*, 58(4), pp. 826–839.

Aigner, D., Lovell, C. A. K. and Schmidt, P., (2023). Reprint of: Formulation and estimation of stochastic frontier production function models. *J. Econometrics*, 234(S), pp. 15–24. doi: 10.1016/j.jeconom.2023.01.

Behr, A., (2015). Production and efficiency analysis with R. *Springer,* Cham.

Campos, M. S., Costa, M. A., Gontijo, T. S., Lopes-Ahn, A. L., (2022). Robust stochastic frontier analysis applied to the Brazilian electricity distribution benchmarking method. *Decis. Anal.*, *3*, 100051. doi:/10.1016/j.dajour.2022.100051.

Coelli, T., (1995). Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. *Journal of Productivity Analysis*, 6(3), pp. 247–268. doi: 10.1007/BF01076978.

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J. and Battese, G.E., (2005). An introduction to efficiency and productivity analysis. *Springer*, Boston, MA, USA.

Dakpo, K. H., Desjeux, Y. and Latruffe, L., (2022). {sfaR}: Stochastic frontier analysis using R.

Greene, W. H., (1990). A gamma-distributed stochastic frontier model. *J. Econometrics*, 46, pp. 141–163.

Greene, W. H., (2003). Simulated likelihood estimation of the normal-gamma stochastic frontier function. *Journal of Productivity Analysis*, 19, pp. 179–190.

Hajargasht, G., (2015). Stochastic frontiers with a Rayleigh distribution. *Journal of Productivity Analysis,* 44(2), pp 199-208. doi: 10.1007/s11123-014-0417-8.

Jondrow, J., Knox Lovell, C., Materov, I. S., Schmidt, P., (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *J. Econometrics*., 19(2-3), 233-238. Doi: 10.1016/0304-4076(82)90004-5.

Kumbhakar, C. S., Wang, J. H. and Horncastle, P. A., (2015). A practitioner's guide to stochastic frontier analysis using Stata. *Cambridge University Press*, New York.

Mariyono, J., (2006). Geographical distribution of technical efficiency in Indonesian rice production during the period of 1979-1994. *Ekon. Pembang*., 11, pp. 33–48.

Meeusen, W., Van Den Broeck, J., (1977). Efficiency estimation from Cobb-Douglas production function with composed error. *International Economic Review (Philadelphia)*, 18.

Mutz, R., Bornmann, L. and Daniel, H.-D., (2017). Are there any frontiers of research performance? Efficiency measurement of funded research projects with the Bayesian stochastic frontier analysis for count data. *JOI*, 11(3), pp. 613–628. doi: 10.1016/j.joi.2017.04.009.

Nurwulan, R., Djuraidah, A. and Fitrianto, A., (2022). Handling outliers in the stochastic frontier model using Cauchy and Rayleigh distributions to measure technical efficiency of rice farming business. *Indonesian Journal of Artificial Intelligence and Data Mining*, 5(2), pp. 98–110.

O'Donnell, C. J., (2018). Productivity and efficiency analysis. *Springer*, Singapore.

Pranata, I., Djuraidah, A. and Aidi, M., (2023). Robust stochastic production frontier to estimate technical efficiency of rice farming in Sulawesi Selatan. *Barekeng Journal of Mathematics and Applications*, 17(1). doi: 10.30598/barekengvol17iss1pp0391-0400.

Sakouvogui, K., Shaik, S., Doetkott, C. and Magel, R., (2021). Sensitivity analysis of stochastic frontier analysis models. *Monte Carlo Methods and Applications*, 27 (1), pp. 71-90. doi: 10.1515/mcma-2021-2083.

Sartono, B. , (2005). Pembangkitan bilangan acak untuk simulasi Monte Carlo non-parametrik. *Forum Statistika dan Komputasi*, 10, pp. 8–11.

Train, K., (2009). *Discrete choice methods with simulation*. 2nd ed. *Cambridge University Press*, Cambridge.

Tsionas, G. E., (2007). Efficiency measurement with the Weibull stochastic frontier. *Oxford Bulletin of Economics and Statistics*, 69(5), pp. 693–706. doi: 10.1111/j.1468-0084.2007.

Wheat, P., Stead, A. D. and Greene, W. H., (2019). Robust stochastic frontier analysis: A Student's t-half normal model with application to highway maintenance costs in England. *Journal of Productivity Analysis*, 51(1), pp. 21–38. doi: 10.1007/s11123-018-0541-y.

Zulkarnain, R., Djuraidah, A., Sumertajaya, I. M. and Indahwati, (2021). Utilization of Student's-T distribution to handle outliers in technical efficiency measurement. *Media Statistika*, 14(1), pp. 56–67.