K-MEDOIDS BASED ROBUST CLUSTERING USING STOCHASTIC FRONTIER ANALYSIS RESULT: A CASE STUDY OF OPERATIONAL PERFORMANCE IN THE INDONESIAN STATE ELECTRICITY COMPANY

Thesis

By

SHINTIA PUTRI SUNARDI
NPM. 2217031125

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
LAMPUNG UNIVERSITY
BANDAR LAMPUNG
2026

ABSTRACT

K-MEDOIDS BASED ROBUST CLUSTERING USING STOCHASTIC FRONTIER ANALYSIS RESULT: A CASE STUDY OF OPERATIONAL PERFORMANCE IN THE INDONESIAN STATE ELECTRICITY COMPANY

By

SHINTIA PUTRI SUNARDI

This study aims to analyze the technical efficiency of electricity distribution units in Indonesia and to identify patterns of their operational performance. The analysis employs *Stochastic Frontier Analysis* (SFA) with a Bayesian estimation approach to obtain more robust efficiency estimates in the presence of assumption violations and potential outliers. The data used in this study are secondary data derived from *PLN Statistics 2024 (Unaudited)* published by PT PLN (Persero), covering the operational performance of 33 electricity distribution units (UID/UIW) across Indonesia. The input variables include the length of medium-voltage distribution lines, the average electricity tariff, and the connected customer load, while the output variable is electricity sales. The results indicate that the Cobb–Douglas (CD) frontier with a Half-Normal (HN) inefficiency distribution provides the most appropriate model specification. Due to the violation of the normality assumption in the error term, Bayesian estimation with a Student's t disturbance distribution was applied to obtain more stable and robust parameter estimates. The estimated technical efficiency scores range from 0.6306 to 0.9622, with an average value of 0.8719, suggesting that most electricity distribution units operate at relatively high efficiency levels. The efficiency scores were subsequently incorporated into a K-Medoids clustering analysis together with the connected customer load variable. The clustering results reveal three clusters with good clustering quality, as indicated by a Silhouette coefficient of 0.657, a Davies–Bouldin Index (DBI) of 0.399, and an R^2 value of 0.815. Overall, the integration of Bayesian SFA and K-Medoids clustering provides a more comprehensive understanding of efficiency patterns in Indonesia's electricity distribution system.

Keywords: Stochastic Frontier Analysis, Bayesian estimation, technical efficiency, K-Medoids clustering, electricity distribution

ABSTRAK

KLASTERISASI ROBUST BERBASIS K-MEDOIDS MENGGUNAKAN HASIL STOCHASTIC FRONTIER ANALYSIS: STUDI KASUS KINERJA OPERASIONAL PADA PERUSAHAAN LISTRIK NEGARA DI INDONESIA

Oleh

SHINTIA PUTRI SUNARDI

Penelitian ini bertujuan untuk menganalisis efisiensi teknis unit distribusi listrik di Indonesia serta mengidentifikasi pola kinerja operasionalnya. Analisis dilakukan menggunakan *Stochastic Frontier Analysis* (SFA) dengan pendekatan estimasi Bayesian untuk memperoleh estimasi efisiensi yang lebih robust ketika terdapat pelanggaran asumsi dan potensi *outlier*. Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari *PLN Statistics 2024 (Unaudited)* yang diterbitkan oleh PT PLN (Persero), yang mencakup kinerja operasional dari 33 unit distribusi listrik (UID/UIW) di seluruh Indonesia. Variabel input yang digunakan meliputi panjang jaringan distribusi tegangan menengah, rata-rata tarif listrik, dan beban pelanggan tersambung, sedangkan variabel output adalah penjualan listrik. Hasil penelitian menunjukkan bahwa frontier Cobb–Douglas (CD) dengan distribusi inefisiensi Half-Normal (HN) merupakan spesifikasi model yang paling sesuai. Karena terjadi pelanggaran asumsi normalitas pada komponen galat, estimasi Bayesian dengan distribusi gangguan Student's t diterapkan untuk memperoleh estimasi parameter yang lebih stabil dan robust. Nilai efisiensi teknis yang diperoleh berada pada rentang 0,6306 hingga 0,9622 dengan rata-rata sebesar 0,8719, yang menunjukkan bahwa sebagian besar unit distribusi listrik beroperasi pada tingkat efisiensi yang relatif tinggi. Nilai efisiensi tersebut kemudian digunakan dalam analisis klaster K-Medoids bersama dengan variabel beban pelanggan tersambung. Hasil klasterisasi menunjukkan terbentuknya tiga klaster dengan kualitas klaster yang baik, yang ditunjukkan oleh nilai koefisien Silhouette sebesar 0,657, *Davies–Bouldin Index* (DBI) sebesar 0,399, dan nilai R^2 sebesar 0,815. Secara keseluruhan, integrasi metode Bayesian SFA dan klasterisasi K-Medoids memberikan pemahaman yang lebih komprehensif mengenai pola efisiensi dalam sistem distribusi listrik di Indonesia.

Kata Kunci: Analisis Frontier Stokastik, estimasi Bayesian, efisiensi teknis, klastering K-Medoids, distribusi listrik.

KLASTERISASI ROBUST BERBASIS K-MEDOIDS MENGGUNAKAN HASIL STOCHASTIC FRONTIER ANALYSIS: STUDI KASUS KINERJA OPERASIONAL PADA PERUSAHAAN LISTRIK NEGARA DI INDONESIA

SHINTIA PUTRI SUNARDI

Thesis

Submitted as a Partial Fulfilment of the Requrement for Degree of
DEGREE DEPARTMENT OF MATHEMATICS

At

Department Mathematics

Faculty of Mathematics and Natural Sciences

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
LAMPUNG UNIVERSITY
BANDAR LAMPUNG
2026

Thesis Title	:	**K-MEDOIDS BASED ROBUST CLUSTERING USING STOCHASTIC FRONTIER ANALYSIS RESULT: A CASE STUDY OF OPERATIONAL PERFORMANCE IN THE INDONESIAN STATE ELECTRICITY COMPANY**
Name of Student	:	Shintia Putri Sunardi
Student Identification Number	:	**2217031125**
Department	:	**Department Mathematics**
Faculty	:	**Mathematics and Natural Sciences**

APPROVED

1. Advisory Committee

Dr. Khoirin Nisa, S.Si., M.Si.
NIP 197407262000032001

Riza Sawitri, S.Si., M.Sc.
NIP 19890504202462001

2. Head of the Department Department Mathematics

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

VALIDATED BY

1. Examination Committee

Chair : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Secretary : **Riza Sawitri, S.Si., M.Sc.** .....

Examiner
(Non-Advisor) : **Widiarti, S.Si., M.Si.** .....

2. Dean of the Faculty of Mathematics and Natural Sciences

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Date of Thesis Examination: **April 15, 2026**

STATEMENT OF STUDENT DECLARATION

The undersigned:

Name	:	**SHINTIA PUTRI SUNARDI**
Student Identification Number	:	**2217031125**
Department	:	**Department Mathematics**
Thesis Title	:	**K-Medoids Based Robust Clustering Using Stochastic Frontier Analysis Result: A Case Study of Operational Performance In the Indonesian State Electricity Company**

Hereby declare that this thesis is my own work and has not been submitted previously for any degree at this or any other institution. If it is later proven that this thesis is the result of plagiarism or the work of others, I am willing to accept sanctions in accordance with the applicable academic regulations.

Bandar Lampung, April 15, 2026

Author

Shintia Putri Sunardi

AUTHOR BIOGRAPHY

The author's full name is Shintia Putri Sunardi, who was born in Panjang Bandar Lampung on January 14, 2003. She is the only daughter of Sunardi and Susanti.

The author began her formal education at Aisyiyah Kindergarten Panjang Bandar Lampung in 2008 and completed it in 2009. She then continued her education at Dwiwarna Elementary School from 2009 to 2015. After that, she studied at SMP Negeri 11 Bandar Lampung from 2015 to 2018, and later completed her senior high school education at SMA Negeri 6 Bandar Lampung in 2021.

In 2022, the author was admitted to the Undergraduate Program in Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Lampung through the Joint Selection for State University Admission (SBMPTN). At the end of 2024, the author carried out an internship (Internship) at Tarahan Bandar Lampung Port Unit, PT Bukit Asam Tbk for 40 days until January 2025. In addition, as a form of community service, the author also participated in the Community Service Program (KKN) for 40 days in Nusantara Permai Village, Sukabumi District, Bandar Lampung City, until August 2024.

During her academic journey, the author has shown perseverance and dedication in completing various academic tasks. The author hopes that the results of this research will contribute to the development of knowledge, particularly in the field of Statistics and its applications.

WORDS OF INSPIRATION

“For indeed, with hardship comes ease. Indeed, with hardship comes ease.”
(Qur’an 94:5–6)

“Whoever follows a path in pursuit of knowledge, Allah will make easy for him a path to Paradise.”
(Prophet Muhammad)

“Knowledge that does not benefit others has no goodness in it.”
(Imam Al-Shafi‘i)

DEDICATION

Alhadulillahirobbil”alamin
Praise and gratitude are expressed to Allah Subhanahu Wata’ala for His blessings and grace, which have enabled the author to complete this thesis properly and on time.
May peace and blessings always be bestowed upon our beloved Prophet Muhammad Shallallahu Alaihi Wasallam.
With great gratitude and joy, the author would like to express sincere thanks to:

My beloved father and mother

My sincere gratitude goes to my parents for all their sacrifices, motivation, prayers, blessings, and continuous support. Thank you for the valuable lessons you have given me about the true meaning of life’s journey, so that one day I may become someone who is beneficial to many people.

My thesis supervisors and discussants

I would like to express my sincere gratitude to my thesis supervisors and discussants for their guidance, motivation, direction, and valuable knowledge that have greatly contributed to the completion of this thesis.

My best friends

I would like to express my sincere gratitude to all the kind people who have given me valuable experiences, encouragement, motivation, and prayers, and who have continuously supported me in every aspect.

My beloved alma mater
University of Lampung

ACKNOWLEDGEMENTS

Alhamdulillah, all praise and gratitude are devoted to Allah Subhanahu Wata'ala for His blessings and mercy, which have enabled the author to complete this thesis entitled "K-Medoids Based Robust Clustering Using Stochastic Frontier Analysis Result: A Case Study of Operational Performance in the Indonesian State Electricity Company" properly, smoothly, and within the specified time. May peace and blessings always be upon the Prophet Muhammad Shallallahu Alaihi Wasallam.

During the process of writing this thesis, many parties have provided assistance in the form of guidance, support, direction, motivation, and valuable suggestions so that this thesis could be completed successfully. Therefore, on this occasion, the author would like to express sincere gratitude to:

1. Thank you to myself for holding on and making it this far, and for completing this thesis. Thank you for not giving up despite the exhaustion and doubts along the way. This journey and all the struggles have become valuable lessons for the next chapter of life.

2. I would like to express my sincere gratitude to Dr. Khoirin Nisa, M.Si., as my first supervisor, for her valuable guidance, time, and support throughout the preparation of this thesis. Her insightful suggestions, constructive feedback, and continuous encouragement have greatly helped the author in completing this research. The guidance and motivation she provided have been truly meaningful during the entire research process.

3. I would also like to express my sincere gratitude to Riza Sawitri, S.Pd., M.Sc., as my second supervisor, for her guidance, support, and prayers throughout the completion of this thesis. Her advice and encouragement have been very meaningful and have helped the author complete this research.

4. I would like to express my sincere appreciation to Widiarti, S.Si., M.Si., as the discussant, for her valuable suggestions, constructive criticism, and insightful feedback that have helped improve the quality of this thesis.

5. I would like to express my gratitude to Prof. Dr. Asmiati, S.Si., M.Si., as my academic advisor.
6. I would like to express my gratitude to Dr. Aang Nuryaman, S.Si., M.Si., as the Head of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung.
7. I would like to express my gratitude to all lecturers, staff, and employees of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung.
8. I would like to express my gratitude to all teachers from kindergarten, elementary school, junior high school, and senior high school who have taught and shared their knowledge, enabling the author to pursue higher education.
9. My father and mother, who have always given their love, financial support, moral encouragement, and endless prayers, enabling the author to complete this thesis.
10. My friends from the KKN Nusantara Permai group Rara, Opi, Nisa, Dinda, Doni, Heber, and Ridho for the support, togetherness, and many memorable stories and experiences shared during the KKN program.
11. I would also like to thank Gebrina, Anisah, Okta, Ainda, and Najwa for their support as friends and for sincerely helping the author in many ways since the beginning of university life until the completion of this thesis.
12. I would also like to thank Agustino, Apri, Ejia, and Khusni for their help, support, and companionship during our time in the C Statistics class until the author was able to complete this thesis.
13. I would also like to thank my fellow thesis supervision friends for the encouragement, support, and shared motivation during the supervision meetings and throughout the process of completing this thesis.

It is hoped that this thesis will be beneficial for all readers. The author realizes that this thesis is still far from perfect therefore, constructive criticism and suggestions are highly appreciated in order to improve this work in the future.

Bandar Lampung, April 15, 2026

Shintia Putri Sunardi

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

BAB I
INTRODUCTION

1.1 Background and Issue

The development of modern statistical methods has contributed significantly to the measurement of efficiency and performance analysis of complex systems. One widely used econometric approach is Stochastic Frontier Analysis (SFA), which aims to estimate the efficiency or inefficiency of a Decision Making Unit(DMU). This method models the frontier or optimal boundary that can only be achieved by fully efficient units (Makieła & Mazur, 2020).SFA was initially developed by Aigner, Lovell, and Schmidt in 1977, along with Meeusen and van den Broeck in the same year. This method provides a parametric approach to efficiency measurement by distinguishing between two unobserved errors, namely components statistical noise, which captures random fluctuations, and technical inefficiency, which indicates how far actual output falls short of the maximum attainable output (Khumbakar et al., , 2015). Based on the form of the frontier function, SFA is generally divided into two main types, namely Cobb Douglas (CD) and Trans Logarithmic (TL), where the TL model provides greater flexibility in capturing nonlinear relationships between the production factors (Pechrová & Šimpach 2020). In terms of error term distribution, classical SFA models typically use a combination of Normal–Half Normal (HN) and Normal–Truncated Normal (TN), which are design to model non-negative inefficiency terms (Papadopoulos, 2021; Sakouvogui et al., 2021).

In SFA, Maximum Likelihood Estimation (MLE) is widely used due to its efficiency and its ability to allow inference on inefficiency effects when the model assumptions are satisfied (Nguyen & O'Donnell, 2025). However, MLE has limitations, particularly when calculating non-linear functions of parameters or when facing parameter constraints. This can result in zero standard errors and invalid confidence intervals (Nguyen & O'Donnell, 2025). In SFA with asymmetric error, MLE often has difficulty converging and estimating the skewness parameter. The Bayesian approach, which uses prior distributions and Markov chain Monte Carlo

(MCMC), is an alternative that can produce robust and stable estimates and handle parameter uncertainty effectively (Wei et al., 2025). Simulations and empirical studies show that Bayesian SFA provides more accurate parameter estimates and better models than classical MLE.

Next, to gain a more detailed understanding of efficiency patterns, clustering is performed to group units with similar characteristics. Clustering is a multivariate analysis method that has several types, including hierarchical clustering and partitioning-based clustering (Pitafi et al., 2020). Partitioning-based clustering offers several advantages, such as simple principles and implementation, high convergence speed, flexibility in determining the number of clusters, the ability to handle various data types, and support for parallelization to improve computational efficiency (Mahdi et al., 2021; Zhang et al., 2023). In performing clustering analysis, there are several important steps. After running the clustering algorithm and obtaining the results, the next step is to evaluate the quality of the clusters (Wang et al., 2025). The quality of clustering results is commonly assessed through clustering evaluation indices, which measure the internal cohesion of objects within a cluster and the degree of separation among clusters, ensuring that the clustering structure reflects meaningful and well-defined groupings (Wang et al., 2025). Several commonly used indices include the Silhouette Index, Calinski–Harabasz, Davies–Bouldin (Awong & Zielinska, 2023).

One of the partitioning clustering analysis methods widely used in various fields of research, including electrical systems, is K-Medoids. This method is particularly suitable for small and low-dimensional data (Zhang et al., 2023). K-Medoids selects medoids that are real objects in the dataset, which makes it more resistant to outliers and abnormal data distributions, thus making it a robust clustering method (Budiaji & Leisch, 2019; Nahdliyah et al., 2019) . In addition, this method minimizes the total distance between objects in a cluster and its medoid, resulting in a more stable and interpretable clustering results compared to the K-Means method (AbdElSamea & Saif, 2024; Budiaji & Leisch, 2019).

Electricity is a vital necessity in the modern era, supporting almost all aspects of human life, from household activities to large-scale industries. According to Zhou & Mai (2021), global electricity demand is projected to increase at an average rate of 3.4% per year until 2026, in line with the growing digitalization and electrification across various economic sectors. Therefore, an efficient and sustainable distribution system is required to minimize disruptions and maintain service continuity, as explained by Seppälä, et al. (2024). In which electricity companies play a strategic

role in ensuring equitable and affordable energy availability across all regions (Hendrocahyo & Kurniawati, 2022). The efficiency of the electricity sector reflects the ability of power companies to manage operational factors such as generation capacity, distribution networks (Rüde et al., 2024), electrification rates (Motherway et al., 2024; Zhou & Mai , 2021), system reliability (Ryu et al., 2020), and tariff and energy sales structures (Susanty et al., 2022; Zaki & Hamdy, 2022) to optimally generate and distribute electricity. Improving efficiency contributes to reducing power losses (Khatiwada et al., 2024), lowering operating costs, and enhancing the reliability and overall performance of the electricity system through optimized network configurations (Agrawal et al., 2020). In Indonesia, this role is carried out by PT Perusahaan Listrik Negara (PLN) (Persero) through Regional Main Units and Distribution Main Units, which are responsible for managing and distributing electricity across regions.

Previous studies have widely used SFA to measure the efficiency of electricity providers and analyzed the impact of environmental factors on their performance. Based on the findings of Nascimento, et al. (2021), the proportion of renewable energy, power plant ownership, and geographic location significantly affect the operational efficiency of electricity distribution utilities in Brazil. Campos, et al. (2022) compared Data Envelopment Analysis (DEA) and SFA finding that SFA is more flexible in handling outliers, although if sometimes faces convergence issues that can be addressed through a Bayesian approach. However, these studies have not continued with further analysis in the form of clustering to examine patterns of efficiency between units. On the other hand, the K-Medoids clustering method has been widely used in the field of electrical power systems, such as by Sarnovsky & Bednar (2025) grouped customers of electricity distribution companies based on their annual consumption profiles from smart meter data. as well as by, who identified areas with electricity shortages and promoted energy efficiency improvements in remote areas of Colombia. These studies have not incorporated the efficiency dimension, even though efficiency-based segmentation can provide more relevant insights for improving electricity sector performance, even though efficiency remains a fundamental aspect of energy management (Kallel et al., 2025). In addition, Li, et al. (2022) used a two-stage approach with Bootstrapped DEA to measure efficiency and Medoid Clustering to map energy performance between regions in China. However, the DEA method does not consider random components as in SFA, which is capable of separating inefficiency from random disturbance factors.

Therefore, this study combines parametric efficiency measurements using SFA with a Bayesian approach, which is more resistant to outliers, effectively handles parameter

uncertainty, and is more flexible when facing parameter constraints, and then further analyzed the efficiency results using K-Medoids clustering. This cluster analysis is supplemented with additional variables, such as the proportion of renewable energy and non-renewable energy fuel usage, the proportion of installed capacity owned, leased, or project-based, and the proportion of energy sold to households, industry, and government, to comprehensively identity patterns of efficiency k operational differences across regions.

1.2 Issue Statement

1. How to measure the operational efficiency of PT PLN's Regional Main Units and Distribution Main Units using the SFA method?
2. How to group Regional Main Units and Distribution Main Units based on SFA efficiency levels and operational characteristics using the K-Medoids method?
3. How can the integration of the SFA and K-Medoids methods describe the efficiency patterns between Regional Main Units and Distribution Main Units in Indonesia's electricity distribution system?

1.3 Research Objective

1. Applying SFA to measure the operational efficiency of PT PLN's Regional Main Units and Distribution Main Units.
2. Using the K-Medoids method to cluster Regional Main Units and Distribution Main Units based on their efficiency and operational characteristics.
3. Identifying efficiency patterns between regions to provide an overview of electricity distribution performance in Indonesia.

1.4 Research Benefits

1. Contributing to the development of statistical methods in efficiency analysis by integrating SFA and K-Medoids clustering.

2. Providing empirical understanding of efficiency patterns and operational characteristics between Regional Main Units and Distribution Main Units PT PLN in Indonesia.
3. Providing information that can assist PT PLN and policymakers in formulating strategies to improve electricity distribution performance in an efficient and sustainable manner.

BAB II
LITERATURE REVIEW

2.1 Outliers in Multivariate Analysis

Multivariate analysis is a set of statistical techniques used to analyze more than two variables simultaneously. This approach helps identify and explore relationships among variables within one or more data structures simultaneously, thereby providing a more comprehensive understanding of the data and revealing complex patterns that cannot be detected through univariate or bivariate analysis alone (Jr et al., 2019). In this context, multivariate outlier detection is an important step because an observation may not appear extreme on a single variable, but may be an outlier when the relationships between several variables are considered simultaneously.

2.1.1 Definition of Outliers

Outliers are data points that significantly deviate from the overall pattern of a dataset. They may occur due to data inaccuracies, data entry mistakes, sampling variations, or genuine yet extreme observations. The presence of outliers can distort statistical analysis, bias model estimation, and reduce accuracy. Some outliers should be removed when caused by errors, while others should be retained if they represent meaningful or rare information. Proper detection and handling of outliers are essential to ensure reliable and valid analytical results (Smiti, 2020).

2.1.2 Detection of Multivariate Outliers

Outlier detection is important before analyzing data because extreme observations can affect the results (Ghorbani, 2019). Univariate detection only examines one

variable at a time, so it cannot capture extreme combinations across multiple variables simultaneously. Therefore, multivariate detection is more relevant.

According to Ghorbani (2019), the Mahalanobis Distance is commonly used to detect multivariate outliers. This distance takes into account the scale of each variable and the covariance between variables. Ghorbani (2019), the Mahalanobis distance is defined as:

$$D(\mathbf{X}, \mu) = \sqrt{(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu)} \tag{2.1}$$

where:

$\mathbf{X}$: the observation vector,
$\mu = E[X]$: the mean vector of all observations,
$\Sigma = \mathrm{cov}(X)$: the covariance matrix

The squared Mahalanobis distance (D^2) is compared with the critical value of the Chi-square distribution with p degrees of freedom. If $D^2 > \chi^2_{p,1-\alpha}$, the observation is considered an outlier. The Chi-square distribution is used as an approximate reference to determine the threshold.

2.2 Robust Statistic

2.2.1 Definition of Robustness in Statistics

The concept of robustness in statistics carries varying interpretations depending on the method being applied. In regression analysis, robustness refers to the ability of parameter estimates to remain reliable even in the presence of outliers (Khan et al., 2021). In the context of estimation, robustness emphasizes that an estimator continues to perform reliably despite departures from ideal distributional assumptions (Loh, 2025). Furthermore, in clustering methods, robustness denotes the stability of clustering results even when initial parameters or conditions are altered (Lu et al., 2019).

Therefore, a robust statistical method is one designed to maintain efficiency and optimality even when the data deviate from specific probabilistic assumptions. In other words, such methods are resistant to minor perturbations or outliers in the data, thereby ensuring reliable performance under diverse conditions (Lesosky, 2020).

2.2.2 Bayesian Robust Estimation of SFA Parameters

In SFA, the Bayesian approach, which uses prior distributions and the MCMC method, is a robust method for estimating frontier parameters. This approach can handle outliers, operational heterogeneity, and model assumptions violations because it generates a full posterior distribution of the parameters, allowing for effective consideration of parameter uncertainty (Wei et al., ,2025). show that Bayesian inference can overcome misspecification in the error component distribution through the use of skew-normal distributions, in addition, Bayesian SFA is flexible in handling asymmetric error distributions and parameter constraints, ensuring reliable and valid estimates.

2.2.3 Robust Clusterings

Robust clustering is an approach in cluster analysis specifically developed to handle the presence of outliers or anomalous data that may distort clustering results. Unlike conventional clustering methods, which are often highly sensitive to data irregularities, robust clustering produces more stable and reliable groupings even when the dataset contains deviations or noise (García-Escudero & Mayo-Iscar, 2024). This approach is grounded in the principles of robust statistics, which emphasize statistical techniques that remain consistent and dependable under small departures from the model's underlying assumptions (Maronna et al., 2019).

2.3 Stochastic Frontier Analysis (SFA)

2.3.1 Definition and Basic Model of SFA

SFA is a widely applied parametric technique for evaluating the efficiency of Decision Making Units (DMU). It assesses the ability of a unit to transform inputs into outputs, separating inefficiency from random noise. This method enables comparison of relative efficiency across DMU, which may include private firms, public organizations, or even nations, and provides valuable information for

enhancing operational performance (Wu, 2025).

SFA was originally introduced by Aigner, Lovell, and Schmidt in 1977, as well as by Meeusen and van den Broeck in the same year. At first, this method was developed to evaluate the technical efficiency of production units by distinguishing technical inefficiency from random errors that influence output (Khumbakar et al., 2015).

The basic functional form of the stochastic frontier model is given by (Aigner et al., 1977):

$$Y_i = f(X_i; \beta) \cdot \exp(v_i - u_i) \tag{2.2}$$

where:

Y_i : the output of the i^{th} unit,
X_i : the vector of input variables,
β : the technology parameters,
v_i : the random error component, and
$u_i \geq 0$: the technical inefficiency term.

Based on this formulation, the technical efficiency of the i-th unit is defined as (Aigner et al., 1977):

$$TE_i = \exp(-u_i) \tag{2.3}$$

where technical efficiency represents the efficiency level of a unit, and values closer to 1 indicate that the unit operates nearer to the production frontier, implying higher efficiency. This fundamental model has served as the basis for many subsequent extensions of SFA, including applications for estimating cost and profit efficiency, as well as incorporating exogenous variables to explain factors influencing inefficiency.

2.3.2 Types of SFA Models

1. Based on the Form of the Frontier Function
   According to Pechrová & Šimpach 2020 the production functions commonly applied in SFA are the Cobb Douglas (CD) and Trans Logarithmic (TL) functions.
   a) CD Function
      The CD production function is one of the most commonly applied forms in empirical studies due to its simplicity and ease of estimation. The

coefficients indicate the output elasticity with respect to each input, showing how changes in an input affect total output. One limitation of this model is the assumption of constant elasticity of substitution among inputs, which implies that the share of each input in total output remains fixed. Nevertheless, the CD function often provides a reasonable approximation of the production process even if some of its assumptions are not strictly satisfied. The CD production function can be expressed as (Aigner et al., 1977):

$$\ln y_{it} = \sum_{k=1}^{K} \beta_k \ln x_{k,it} + \varepsilon_{it} \tag{2.4}$$

$$\varepsilon_{it} = v_{it} - u_{it} \tag{2.5}$$

where:

y_{it}	: the output of unit i at time t,
β_k	: the estimated parameters for each input,
$x_{k,it}$	: the k-th input of unit i at time t,
ε_{it}	: the composite error term consisting of the stochastic noise v_{it} and technical inefficiency u_{it},
v_{it}	: random disturbances beyond the control of the unit,
u_{it}	: the technical inefficiency of the unit,
$i = 1, 2, \ldots, N$	: index of units, where N is the total number of units,
$t = 1, 2, \ldots, T$	: index of time periods, where T is the total number of observed periods,
$k = 1, 2, \ldots, K$	: index of inputs, where K is the total number of inputs.

b) TL Function

The TL production function relaxes the assumption of constant substitution elasticity. It is a more flexible specification that includes both second-order and cross-product terms among input variables. When these additional terms are restricted to zero, the Trans-Log function simplifies back to the CD form. Following Pechrová & Šimpach 2020 The functional form can be expressed as:

$$\ln y_{it} = \sum_{k=1}^{K} \beta_k \ln x_{k,it} + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{k,it} \ln x_{l,it} + \varepsilon_{it} \tag{2.6}$$

The form of the error component ε_{it} in equation (2.6) is the same as in equation (2.5).

where:

y_{it} : the output of unit i at time t,
β_k : the estimated parameters for each input,
$x_{k,it}$: the k-th input of unit i at time t,
ε_{it} : the composite error term consisting of the stochastic noise v_{it} and technical inefficiency u_{it},
v_{it} : random disturbances beyond the control of the unit,
u_{it} : the technical inefficiency of the unit,
$i = 1, 2, \ldots, N$: index of units, where N is the total number of units,
$t = 1, 2, \ldots, T$: index of time periods, where T is the total number of observed periods,
$k = 1, 2, \ldots, K$: index of inputs, where K is the total number of inputs.

In many SFA studies, researchers compare the performance of different functional forms, particularly the CD and TL models.

2. Based On the Form of the Error Function
In 1977, Aigner, Lovell, and Schmidt, as well as Meeusen and van den Broeck, introduced the SFA model, which initially employed the Half Normal (HN) and Exponential distributions for the inefficiency error term. Both distributions assume that inefficiency is positive and most likely occurs near zero. Later, in 1980, Stevenson proposed the Truncated Normal (TN) distribution as an alternative, allowing the mode of inefficiency to take a non-zero value and thus providing a more flexible representation of efficiency variations among production units (Papadopoulos, 2021). in The SFA, the inefficiency term u_i represents the shortfall of actual output from its maximum feasible value, while the random error v_i is normally distributed and independent of u_i (Sakouvogui et al., 2021). The Probability Density Function (PDF) of $\varepsilon_i = v_i - u_i$ can be derived by integrating out u_i as follows:

a) Normal-Half Normal (N-HN): $v \sim N(0, \sigma_v^2), ; u \sim HN(0, \sigma_u^2)$
The marginal PDF of $\varepsilon = v - u$ is

$$f(\varepsilon) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(-\frac{\lambda\varepsilon}{\sigma}\right) \tag{2.7}$$

$$\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}, \qquad \lambda = \frac{\sigma_u}{\sigma_v} \tag{2.8}$$

where:

ϕ : the standard normal PDF,
Φ : the standard normal cumulative distribution function (CDF),
ε : the composite error term consisting of the stochastic noise v and technical inefficiency u,

σ_u : the standard deviation (scale parameter) of component u,
σ_v : the standard deviation (scale parameter) of component v,
σ_v^2 : the variance of the symmetric noise term v,
σ_u^2 : the variance of the inefficiency term u in HN distributions.

b) Normal–Truncated Normal (N–TN): $v_i \sim N(0, \sigma_v^2), \quad u_i \sim TN(\mu, \sigma_u^2)$
The marginal PDF of $\varepsilon_i = v_i - u_i$ is:

$$f(\varepsilon) = \frac{1}{\sigma}\phi\left(\frac{\varepsilon+\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1} \tag{2.9}$$

The form of the error component ε_i and equation (2.9) is the same as in equation (2.8).

ϕ : the standard normal PDF,
Φ : the standard normal cumulative distribution CDF,
ε : the composite error term consisting of the stochastic noise v and technical inefficiency u,
σ_u : the standard deviation (scale parameter) of component u,
σ_u^2 : the variance of the inefficiency term u in HN distributions.
μ : the mean of the truncated normal inefficiency distribution, allowing non-zero mode of inefficiency.

2.3.3 Estimation of the Production Function Using OLS

Parameter estimation in a regression model can be carried out using the Ordinary Least Squares (OLS) method. This method determines parameter estimates by minimizing the sum of squared residuals, which represent the differences between the observed values and the predicted values from the regression model. In general, the simple linear regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n \tag{2.10}$$

where y_i denotes the response variable, x_i represents the predictor variable, β_0 and β_1 are the parameters to be estimated, and e_i is the error term. The OLS estimators

are obtained by minimizing the residual sum of squares (RSS), defined as

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{2.11}$$

This procedure produces the regression line that best fits the observed data according to the least squares criterion (Weisberg, 2014).

2.3.4 Estimation of SFA Parameters Using Maximum Likelihood

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model by selecting the parameter value θ that maximizes the likelihood of the observed data. For independent samples $x_1, x_2, \ldots, x_n$ with probability density function $f(x_i \mid \theta)$, the likelihood function can be expressed as (Hogg et al., 2019):

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{2.12}$$

where:

$L(\theta \mid x)$: likelihood function, representing the probability of observing the sample data given the parameter vector θ,
x_i : the i^{th} observed data point, assumed to be independently drawn from the underlying distribution,
θ : the vector of model parameters to be estimated.

To simplify the optimization process, the log-likelihood form is used,

$$\ell(\theta) = \ln L(\theta \mid x) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) \tag{2.13}$$

where $\ell(\theta)$ is the likelihood function expressed in logarithmic form, known as the log-likelihood.

$$\hat{\theta} = \arg\max_{\theta}, \ell(\theta) \tag{2.14}$$

The MLE estimator is obtained by choosing the parameter value that maximizes the log-likelihood function: Mathematically, the optimum is achieved by satisfying the first-order condition, where the derivative of the log-likelihood with respect to the parameter equals zero:

$$\frac{\partial}{\partial \theta}, \ell(\theta) = 0 \tag{2.15}$$

where $\frac{\partial}{\partial\theta}$ denotes the partial derivative operator with respect to the parameter θ.

In the context of SFA, the specification of the likelihood function depends on the form of the frontier function and the assumptions regarding the composite error term (ε_i), which consists of two components: the inefficiency term (u_i) and the random noise term (v_i).

Because the likelihood function in the SFA model involves the joint distribution of the error components v_i and u_i, its form becomes more complex than that of the classical regression model. Consequently, the maximization of the log-likelihood function generally cannot be solved analytically. Therefore, parameter estimation in the SFA model is typically performed using numerical optimization methods. One commonly used approach is the quasi-Newton algorithm, an iterative method designed to maximize the log-likelihood function by utilizing gradient information while constructing an approximation of the Hessian matrix without explicitly computing second-order derivatives. Several algorithms belonging to the quasi-Newton family include the Davidon–Fletcher–Powell (DFP) and Broyden–Fletcher–Goldfarb–Shanno (BFGS) methods, which are widely applied in Maximum Likelihood estimation. The stages and procedures of the quasi-Newton algorithm in optimization are described in detail in Nocedal & Wright 2006.

In practice, this optimization procedure has been implemented in various statistical software packages. In the *frontier* package in R, parameter estimation for the stochastic frontier model is carried out using MLE, where the log-likelihood function is maximized through a quasi-Newton optimization algorithm to obtain efficient parameter estimates.

2.3.5 SFA Model Assumptions and Tests

In classical studies, SFA was commonly developed under a set of core assumptions concerning the form of the frontier function and the distributional features of the inefficiency component. As the field of efficiency and production economics has advanced, numerous contemporary studies have revisited and refined these assumptions to improve the adaptability and robustness of efficiency estimation.

1. Multicollinearity
   A multicollinearity test was conducted prior to performing the SFA to ensure

that the explanatory variables were not highly correlated (Rauniyar & Kim ,2025). Multicollinearity among the explanatory variables was assessed using the Variance Inflation Factor (VIF), which is calculated as: Multicollinearity among the explanatory variables was assessed using the VIF. According to Gujarati & Porter (2009), the VIF for the $i-th$ explanatory variable is defined as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.16}$$

where R_i^2 is the coefficient of determination obtained from the regression of the i-th explanatory variable on all other explanatory variables. All VIF values were below 10, indicating that there was no serious multicollinearity among the explanatory variables (Gujarati & Porter, 2009).

2. Frontier Functional Form SFA

   The appropriateness of the stochastic frontier functional form has been extensively examined in the literature, with numerous studies comparing CD and TL specifications using the Likelihood Ratio (LR) test (Rauniyar & Kim ,2025; Sabar & Kamil, 2023; Tirkaso & Gren, 2023). The hypotheses for selecting the most suitable frontier functional form are formulated as follows:

   $H_0 : \beta_{jk} = 0,\ \forall j \leq k$

   (Cobb–Douglas provides an adequate representation of the data)

   $H_1 : \beta_{jk} \neq 0$ (Translog is more appropriate)

   The likelihood ratio (LR) test, originally developed by Wilks (1938), is defined as:

$$LR = -2 \ln \left(\frac{L(H_0)}{L(H_1)} \right) \tag{2.17}$$

   where $L(H_0)$ and $L(H_1)$ denote the log-likelihood values under the null and alternative hypotheses, respectively. The LR statistic is compared with the chi-square distribution with degrees of freedom equal to the number of restrictions. If $LR > \chi^2_{\text{critical}}$, the null hypothesis is rejected, indicating that the Translog specification provides a more appropriate functional form.

3. Existence Inefficiency

   A commonly used method to test for the existence of inefficiency in stochastic frontier analysis is the LR test. This test evaluates whether the variance of the inefficiency component (σ_u^2) is significantly different from zero. The hypotheses can be formulated as follows (Fenyves et al., 2022):

   $H_0 : \sigma_u^2 = 0$ (no inefficiency)

   $H_1 : \sigma_u^2 \neq 0$ (inefficiency is present)

   The test statistic follows the same formulation as Equation (2.17) and uses the

same decision criterion based on the chi-square distribution. If the computed LR statistic exceeds the critical chi-square value, the null hypothesis is rejected, indicating the presence of inefficiency in the model.

4. Distribution Inefficiency

   The inefficiency component (u_i) represents the difference between actual output and the maximum possible output. A common approach is to test the distribution of inefficiency using the HN and TN specifications, with the hypotheses evaluated through the LR test (Sabar & Kamil, 2023):

   $H_0 : \mu = 0$

   $H_1 : \mu \neq 0$

   The test statistic follows the same formulation as Equation (2.17) and uses the same decision criterion based on the chi-square distribution. If the calculated LR statistic exceeds the critical chi-square value, the null hypothesis is rejected, indicating that the truncated-normal distribution is more appropriate. Conversely, acceptance of H_0 implies that the inefficiency component u_i follows a half-normal distribution.

5. Endogeneity

   In SFA, endogeneity in input variables can lead to biased and inconsistent estimates of the production parameters (β). Recent studies highlight the importance of testing for endogeneity using the control function approach (Hou et al., 2025). In this approach, the residuals obtained from the first-stage regression are included in the frontier model as a bias-correction term $(\eta\hat{v}_{it})$, where η measures the correlation between the input variables and the composed error term. The significance of η is examined using the Wald chi-square test, formulated as follows:

   $H_0 : \eta = 0$ (no endogeneity)

   $H_1 : \eta \neq 0$ (presence of endogeneity)

   The Wald test statistic is defined as (Wald, 1943):

$$W = \hat{\eta}', [\mathrm{Var}(\hat{\eta})]^{-1}\hat{\eta} \sim \chi^2(q) \tag{2.18}$$

   where q represents the number of parameters tested for endogeneity. The decision rule is as follows: if $W > \chi^2_{(q,\alpha)}$, the null hypothesis is rejected, indicating the presence of endogeneity; otherwise, if $W \leq \chi^2_{(q,\alpha)}$, the null hypothesis cannot be rejected, suggesting that the explanatory variables are exogenous.

6. Specific Heteroscedasticity

   In the SFA model, the assumption of homoscedasticity requires the error variances to be constant across all observations. In practice, however, the variances of the two-sided random error (v_i) and the one-sided inefficiency error (u_i) may differ

across observations, resulting in heteroscedasticity, which can lead to biased and inefficient parameter estimates (Rauf et al., 2024). To detect heteroscedasticity, the Breusch–Pagan (BP) test can be applied with the following hypotheses (Rauf et al., 2024):
$H_0 : \sigma^2_{u_1} = \sigma^2_{u_2} = \cdots = \sigma^2_{u_n}$ and $\sigma^2_{v_1} = \sigma^2_{v_2} = \cdots = \sigma^2_{v_n}$
(no heteroscedasticity)
H_1 : at least one $\sigma^2_{u_i}$ or $\sigma^2_{v_i}$ differs across observations $i = 1, 2, \ldots, n$
(heteroscedasticity)
The BP test involves regressing the squared residuals from the SFA model (v_i) and (u_i) on the explanatory variables. Following Breusch & Pagan (1979), the Lagrange Multiplier statistic is defined as:

$$LM = n \cdot R^2_{\text{aux}} \tag{2.19}$$

where R^2_{aux} is the coefficient of determination obtained from the auxiliary regression of (v_i) and (u_i) on the independent variables, and n is the sample size.

The BP test statistic follows a χ^2 distribution with degrees of freedom equal to the number of regressors in the auxiliary regression. If the computed LM value exceeds the critical chi-square value χ^2_{critical}, the null hypothesis is rejected, indicating the presence of heteroscedasticity in one or both error components. Conversely, acceptance of H_0 suggests that the model assumes homoscedastic errors.

7. Distribution of the Noise Component
In the SFA model with a composite error term $\varepsilon_i = v_i - u_i$, a key assumption is that the noise component v_i follows a normal distribution, i.e., $v_i \sim N(0, \sigma^2_v)$. To verify this assumption, one commonly used method is the Shapiro–Wilk test, which evaluates the conformity between the sample distribution and the theoretical normal distribution using the following statistic:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.20}$$

where $x_{(i)}$ denotes the ordered sample values, $\bar{x}$ is the sample mean, and a_i are coefficients computed from the covariance matrix of a normally distributed sample (Shapiro & Wilk 1965). The hypotheses are as follows:
$H_0 : v_i \sim N(0, \sigma^2_v)$
(noise is normally distributed)
$H_1 : v_i \not\sim N(0, \sigma^2_v)$
(noise is not normally distributed)

The test statistic W reflects how closely the data follow a normal distribution; the closer W is to 1, the more normal the data appear. The $p - value$ is computed numerically based on the empirical distribution of W:

$$\text{p-value} = P\left(W \leq W_{\text{observed}} \mid H_0\right) \tag{2.21}$$

Decision rule: if the p-value is less than the significance level α, the null hypothesis is rejected, indicating that the noise component does not follow a normal distribution. Otherwise, the null hypothesis cannot be rejected, suggesting that the normality assumption is satisfied.

a) If $p - value > 0.05$: there is no strong evidence against the assumption that v_i follows a normal distribution.

b) If $p - value \leq 0.05$, the null hypothesis is rejected, indicating that the data deviate from normality.

2.4 Estimation of SFA Parameters Using Bayesian

2.4.1 Basic Concept of the Bayesian Approach

The Bayesian approach treats unknown parameters as random variables characterized by probability distributions, allowing parameter uncertainty to be explicitly modeled rather than assumed fixed as in the frequentist framework. A Bayesian model is composed of a likelihood function, which summarizes information contained in the observed data, and a prior distribution, which reflects prior beliefs about the parameters. After observing the data, prior information is updated into a posterior distribution through Bayes' theorem which is expressed as:

$$\pi(\theta \mid x) = \frac{L(x \mid \theta)\pi(\theta)}{\int L(x \mid \theta)\pi(\theta), d\theta} \tag{2.22}$$

where:

$L(x \mid \theta)$: the likelihood function, representing the probability of observing the data given the parameter θ,

$\pi(\theta)$: the prior distribution, expressing prior beliefs about the parameter before observing the data,

$\pi(\theta \mid x)$: the posterior distribution, representing the updated beliefs after incorporating the observed data.

The posterior distribution provides updated information about the parameters after considering the observed data. Thus, the Bayesian approach combines prior knowledge and empirical evidence to produce more informative parameter estimates (Karagiannis, ,2022).

2.4.2 Likelihood Function

In Bayesian inference, the likelihood function represents the probability of observing the data for a given parameter value within a model. The likelihood plays a central role because it is combined with the prior information to form the posterior distribution once the data are observed (Coventry & Bartlett, 2024). The mathematical structure of the likelihood used in the Bayesian framework is essentially the same as the likelihood employed in MLE, as shown in Equation (2.12), since both approaches model the probability of the observed data as a function of the parameter θ.

2.4.3 Prior Distributions

The prior distribution $\pi(\theta)$ represents the researcher's initial beliefs or assumptions about the uncertain parameter θ within its parameter space Θ. Selecting an appropriate prior is important because it helps produce meaningful posterior and predictive distributions, which form the basis of Bayesian analysis (Karagiannis, 2022). Priors can be classified as either informative or non-informative. An informative prior reflects the researcher's existing knowledge or assumptions about a parameter, whereas a non-informative prior is used when little or no prior information is available, thereby exerting minimal influence on the analysis results (Coventry & Bartlett, 2024).

In the context of Bayesian SFA, prior specifications are designed to incorporate prior knowledge while maintaining analytical tractability and computational stability. Consistent with the Bayesian SFA literature, particularly Griffin & Steel (2007, and

the recommendations of Wheat et al. (2019), the random disturbance component is modeled using a Student-t distribution, $v_i \sim t_\nu(0, \sigma_v)$. This specification allows extreme deviations in the data to be more appropriately captured as random noise rather than being misclassified as technical inefficiency.

The parameter σ is treated as a common scale parameter and does not directly represent the variance of the error term in the likelihood function. Accordingly, σ is assigned a weakly informative HN prior to maintain estimation stability and to prevent unrealistically extreme values, particularly in settings with limited sample sizes (Gelman, 2006). Meanwhile, the regression parameters, including the intercept and slope coefficients, are assigned weakly informative Student-t priors. These priors provide mild regularization through heavy-tailed distributions, thereby reducing sensitivity to outliers without unduly constraining the information contributed by the data (Gelman et al., 2008).

2.4.4 Model Evidence in Bayesian Inference

In Bayesian inference, the denominator of the posterior distribution, $\int L(x \mid \theta)\pi(\theta)\, d\theta$, is referred to as the model evidence (Coventry & Bartlett, 2024). The model evidence measures the probability of observing the data under a specific model of the data-generating process. Although its calculation may appear complex, the model evidence represents a weighted average of the likelihood over all possible parameter values, with weights determined by the prior distribution. Consequently, the model evidence also serves as a normalization factor, ensuring that the posterior $\pi(\theta \mid x)$ is a valid probability distribution.

The model evidence can also be used to compare the plausibility of competing models (Johnson et al., 2023). Historically, evaluating the model evidence was challenging due to the need to compute complex integrals. However, with the development of Markov Chain Monte Carlo (MCMC) methods and advances in computational capability, the evaluation of model evidence can now be performed more efficiently.

2.4.5 Markov Chain Monte Carlo (MCMC)

In this study, parameter estimation in the SFA model is performed using the Hamiltonian Monte Carlo (HMC) algorithm implemented in the Stan software. Unlike random-walk methods, HMC utilizes gradient information from the log-posterior distribution to guide the simulation toward high-density regions more efficiently (Gelman, 2013). Each iteration of the HMC algorithm consists of the following steps:

1. Momentum Initialization (ϕ)
   A new momentum variable is drawn from a multivariate normal distribution, $\phi \sim N(0, M)$, and used as an auxiliary variable to facilitate the movement of parameters in the sampling process.
2. Leapfrog Integration
   The position (ω) and momentum (ϕ) are updated simultaneously through L leapfrog integration steps with step size ϵ, using the gradient of the log-posterior distribution.
3. NUTS Adaptation (No-U-Turn Sampler)
   Stan automatically adjusts the number of steps L to prevent the trajectory from turning back, ensuring efficient exploration of the parameter space.
4. Metropolis Acceptance Criterion
   The proposed position (ω^*) is evaluated using an acceptance ratio. The proposed sample is accepted if the Hamiltonian energy remains approximately conserved.
5. Parameter Estimation Samples obtained after the warm-up period are used to estimate posterior values and credible intervals.

2.4.6 Bayesian Model Evaluation

Bayesian model evaluation is conducted using the Deviance Information Criterion (DIC) to balance model fit and model complexity. Given the deviance defined as $D(\theta) = -2 \log p(y \mid \theta)$, the DIC value is calculated using the following equation (Galan et al., 2014):

$$DIC = \bar{D} + p_D \tag{2.23}$$

where:

$\bar{D}$: the expected deviance, which measures the model fit,
p_D : the penalty for model complexity.

2.5 Cluster Analysis

2.5.1 Basic Concept of Clustering

Clustering is a technique used to group data into several clusters based on the level of similarity among their characteristics, without the need for predefined labels. Objects within the same cluster share greater similarity compared to those in other clusters, enabling this method to uncover hidden patterns and the natural structure within the data. It has been widely applied in various fields, including bioinformatics, image segmentation, anomaly detection, document analysis, and customer segmentation, making it an important approach in modern data analysis (Wani, 2024).

2.5.2 Types of Clustering

Clustering techniques are broadly divided into two main categories (Pitafi et al., 2020): hierarchical clustering and partitional clustering. These categories represent different approaches in how data objects are grouped either by forming a hierarchy of clusters or by directly partitioning data into a predetermined number of groups.

1. Hierarchical Clustering
   Hierarchical-based clustering algorithms are grouping methods that illustrate the relationships among clusters based on the degree of similarity or dissimilarity between objects. The clustering process is visualized in a hierarchical structure known as a dendrogram, allowing researchers to observe how objects or clusters are merged or separated at each level of the hierarchy (Benabdellah et al., 2019). Hierarchical method is further divided into several subtypes (Pitafi et al., 2020):
   a) Agglomerative Method is a bottom-up approach that starts by considering each object as an individual cluster and then gradually merges them based on their similarity level until a single large cluster is formed.
   b) Divisive Method is a top-down approach that begins with one large cluster containing all objects and then progressively splits it into smaller clusters according to their degree of dissimilarity.
2. Partitional Clustering
   This clustering method divides a dataset into a specific number of clusters, where

each pair of clusters may be either completely distinct or share some common members. The process begins by generating an initial partition of the data, which is then iteratively refined until a partition that is locally optimal is achieved (Oti et al., 2021). Partitioning-based clustering methods offer several advantages, including robustness, scalability, and ease of implementation. Furthermore, this approach is easy to understand and does not require specialized knowledge of the data domain. Its iterative nature also allows the cluster structure to be dynamically updated according to the computation of cluster centers, several algorithms fall under the category of partitioning-based clustering, which are commonly used in data analysis (Benabdellah et al., 2019):

a) K-Means is the most widely used method, based on the principle of minimizing the sum of squared distances between data points and their respective cluster centroids.

b) K-Modes was developed to handle categorical data by replacing the mean-based distance calculation with a mode-based one.

c) K-Medoids uses medoids as cluster center, making it more robust against outliers.

d) Partitioning Around Medoids (PAM) serves as the foundational algorithm for the K-Medoids approach.

e) For large-scale datasets, Clustering Large Applications (CLARA) is introduced as a variant of K-Medoids that employs a sampling strategy to improve computational efficiency.

2.5.3 Data Standardization in Clustering

Before applying the K-Medoids algorithm, all numerical variables were standardized to have a comparable scale. This process is important to prevent variables with large value ranges from dominating the clustering results. Standardization was performed using the Z-score method, which transforms each variable to have a mean of zero and a standard deviation of one (Wongoutong, 2024):

$$z_i = \frac{x_i - \bar{x}}{s} \tag{2.24}$$

where:

z_i : the standardized score (z-score) of the i-th observation, indicating how far x_i

is from the mean in terms of standard deviations,
x_i : the value of the i-th observation of the variable being analyzed,
$\bar{x}$: the mean of all observations x_i,
s : the standard deviation of all observations x_i.

Research Wongoutong (2024) shows that standardization improves the accuracy and quality of clustering results, especially when data has different units or scales.

2.5.4 Distance Measure in Clustering

Distance measures play a crucial role in clustering analysis, as they determine the degree of similarity between data points and directly influence the formation and quality of clusters. Several distance measures are commonly used, including Euclidean, Manhattan, Chebyshev, Mahalanobis, Canberra, and Cosine distance (Yaro et al., 2024).

Mathematically, the Manhattan distance between two objects f_1 and f_2 with N variables can be expressed as (Solikhun et al., 2024):

$$d_{\text{Manhattan}}(f_1, f_2) = \sum_{i=1}^{N} |f_{1i} - f_{2i}| \tag{2.25}$$

2.5.5 Clustering Validation Measures

According to Hassan, et al. (2024), clustering validation is a technique used to determine the optimal number of clusters that best represent the natural patterns or inherent structure within a dataset. This process is performed without relying on any predefined labels or classes. In general, clustering validation methods are categorized into three types: external validation, which uses known data labels as a reference; internal validation, which evaluates cluster quality based on distance and data density measures; and relative validation, which compares the results obtained from different parameters or methods to select the most suitable solution.

Several validation metrics are widely applied to assess the performance of partitioning clustering algorithms such as K-Means and K-Medoids.

1. The Silhouette Coefficient is used to assess how well each observation fits within its assigned cluster relative to other clusters. The coefficient ranges from -1 to 1, with higher values indicating more coherent and well separated cluster structures (Firmansyah et al., 2025). Based on empirical analyses of their data, Dalmaijer, et al. (2022) showed that silhouette values below 0 indicate potential misassignment of observations to clusters, values equal to 0 indicate observations located at cluster boundaries, values between 0.5 and 0.7 provide evidence of reasonably formed clusters, while values above 0.7 indicate strong evidence of a well established clustering structure
2. Davies–Bouldin Index (DBI) Index calculates the average ratio between the within-cluster distance and the distance between clusters for all cluster pairs. A lower DBI value indicates that the clusters are more distinct and internally homogeneous, reflecting better clustering performance (Ikotun et al., 2025).
3. The R-Squared Index (RS) is an internal clustering validation index that measures how much of the data variability is explained by the separation between clusters. The RS value ranges from 0 to 1, where higher values indicate better cluster separation. According to Hassan, et al. (2024) since the RS value generally increases as the number of clusters increases, the optimal number of clusters is commonly determined using the elbow method. As reported in Cluster validity indices for automatic clustering: A comprehensive review by Ikotun, et al. (2025), the Silhouette index exhibits relatively high computational complexity, limiting its practicality for large-scale or high-dimensional datasets.

2.6 K-Medoids Clustering Algorithm

The K-Medoids algorithm is a clustering technique that aims to group data into several clusters by minimizing the distance among data points within each cluster. Unlike K-Means, which uses centroids that may not correspond to actual data points, K-Medoids employs real data objects as medoids (Rahmawati & Fauzan, 2024). A medoid represents a true cluster center because of its resilience to outliers and noise (Sureja et al., 2022).

K-Medoids is considered a popular clustering algorithm due to its simplicity, efficiency, and robustness against noisy data and outliers. The algorithm functions by minimizing the total dissimilarity between each point and its corresponding medoid. However, the resulting cluster configuration may vary depending on the initial selection of medoids, which can lead to instability in clustering outcomes

(Fitriyanto & Syafiqoh, 2024).

According to Adriani, et al. (2023), K-Medoids models data in a way similar to K-Means, but instead of using centroids, it relies on k representative objects known as medoids that serve as cluster prototypes rather than arbitrarily defined centers. Medoids are selected based on their minimum average deviation from other objects within the same cluster, making them more resistant to the influence of outliers compared to centroids.

1. Initialization of initial medoids: Determine the number of clusters K and select initial medoid positions,
2. Assignment of data objects: Each data point is assigned to the nearest medoid based on a chosen distance metric,
3. Selection of new medoid candidates: Randomly select a potential new medoid from each existing cluster,
4. Distance computation: Calculate the distance of every data point to the newly selected medoid candidates,
5. Evaluation of total deviation: Compute the total deviation (S) between the new and previous configurations. If $S > 0$, swap the objects to form the new set of medoids,
6. Iteration until convergence: Repeat steps 3–5 until the medoids no longer change, indicating that the final cluster structure has been reached,
7. Evaluate the quality of the resulting clusters.

2.7 Overview of the Electric Power Industry

2.7.1 PT. Perusahaan Listrik Negara (PLN)

Perusahaan Listrik Negara (PLN) is Indonesia's state-owned electricity company responsible for supplying power to nearly all households and businesses across the nation. As the sole entity authorized to distribute electricity, PLN controls most of the generation, transmission, and distribution infrastructure (Aditya et al., 2023). While some power plants are owned by private companies, PLN remains the dominant player in ensuring affordable, reliable, and sustainable electricity supply (Hendrocahyo & Kurniawati, 2022). The company also plays a crucial role in expanding electrification, supporting national infrastructure projects, and integrating

renewable energy sources to promote sustainable energy development. Moreover, PLN's commitment to reliable and cost-effective operations aligns with national goals for asset optimization, maintenance efficiency, and continuous financial improvement through revenue growth, asset expansion, and risk management (Ayu & Yunusa-Kaltungo, 2020; Hendrocahyo & Kurniawati, 2022).

Organizationally, PT PLN (Persero) manages Indonesia's electricity system through an operational structure divided into several main units, including the Regional Main Unit and the Distribution Main Unit. Both units are responsible for electricity distribution within their respective regions, performing their functions according to operational duties while maintaining PLN's primary role as the national electricity provider.

2.7.2 Operational Efficiency and Performance Indicators

1. The length of medium-voltage distribution lines (JTM)
   Refers to the total length of the distribution network between the substation and the customer, which indicates the service area. A longer network length can result in greater power losses and higher operating costs, which negatively impact efficiency (Rüde et al., 2024).
2. Average Electricity Tariff
   Electricity tariffs determine the price consumers pay for energy use. Tariff structures affect the balance between demand and supply and reflect network efficiency and regulatory policies (Zaki & Hamdy, 2022).
3. Connected Load of Customers
   Connected load is the total rated capacity of all electrical equipment installed at a customer's premises and is directly linked to sold electricity as it defines the maximum potential energy consumption; thus, any increase in connected load through electrification inherently drives the growth of total energy sales (Gunkel et al., 2023).
4. Sold Energy
   Represents the total amount of electricity delivered and consumed by customers within a specific period. It reflects operational efficiency, as a higher volume of sold energy relative to inputs indicates better utilization of available resources (Susanty et al., 2022).

BAB III
RESEARCH METODOLOGY

3.1 Research Time and Location

This research was conducted in the odd semester of the 2025/2026 academic year at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung.

3.2 Research Data

This study uses secondary data from the 2024 PLN Statistics (Unaudited) published by PT PLN (Persero), which provide information on the operational conditions and performance of electricity distribution in Indonesia, particularly for UID and UIW units. Only selected variables relevant to the SFA and K-Medoids clustering analyses are used, as summarized in Table 3.1.

Table 3.1 Research Data

Method	Variable	Description
SFA	Input 1	Length of Medium-Voltage Lines (kilometers of circuit)
	Input 2	Average Electricity Tariff (Rupiah per kiloWatt-hour (Rp/kWh))
	Input 3	The connected load of customers (Mega Volt Ampere (MVA))
	Output	Average Electricity Sales (kWh)
K-Medoids	Variabele 1	Efficiency Score (Result of the SFA Method)
	Variable 2	The connected load of customers (MVA)

3.3 Research Method

The data analysis in this study employed statistical and computational tools to perform SFA estimation and clustering procedures efficiently. Steps of data analysis in this study are as follows:

1. Descriptive analysis of data for 2024 presents the operational characteristics of each Regional Main Units and Distribution Main Units based on variables from Statistik PLN 2024 (Unaudited),
2. Outlier detection is performed using Mahalanobis distance to identify extreme data and recommend robust methods for SFA and clustering analysis,
3. Building a SFA model,
    a) Estimate SFA parameters using MLE assuming normal noise,
    b) Model assumption testing based on MLE, including,
        i. Multicollinearity test using VIF.
        ii. The specification test of the frontier function (CD or TL) is conducted using the LR Test.
        iii. The existence of inefficiency is tested using the LR Test.
        iv. The distribution of the inefficiency component (HN or TN) is tested using the LR Test.
        v. Endogeneity among variables is tested using the Wald Chi-square Test.
        vi. Heteroskedasticity is tested using the BP Test.
        vii. The distribution of the noise component (v_i) is tested using the Shapiro–Wilk Test.
    c) Estimation of SFA Parameters Using the Bayesian Approach, including:
        i. Constructing the likelihood function based on the selected frontier form and the assumed inefficiency error distribution,
        ii. Defining the prior distribution for each model parameter,
        iii. Calculating model evidence to evaluate model fit and normalize the posterior distribution,
        iv. Estimating parameters numerically using the MCMC algorithm, such as Gibbs Sampling,
        v. Estimating Technical Efficiency (TE) based on the posterior distribution of the inefficiency component (u_i).
4. Clustering Analysis with K-Medoids,

a) Data Standardization: Normalize all variables to ensure equal contribution in the distance computation,

b) Initialize Medoids: Specify the number of clusters (C) and select the initial medoids,

c) Initial Clustering: Group each Regional Main Units or Distribution Main Units to the nearest medoid based on Euclidean distance,

d) Selection of New Medoid Candidates: Randomly select medoid candidates from each cluster of the initial grouping results,

e) Calculate the Return Distance – Recalculate the distance of each object to each medoid in each cluster,

f) Total Deviation Evaluation (S) – Compare the new configuration with the previous one; if $S > 0$, perform object exchange to update the medoid,

g) Iterate until Convergence – Repeat steps 3–5 until the medoid positions stabilize and the final clusters are formed,

h) Clustering Results Evaluation – Cluster quality values using Silhouette Coefficient, CHI, and DBI.

5. Analysis and Interpretation of Results.

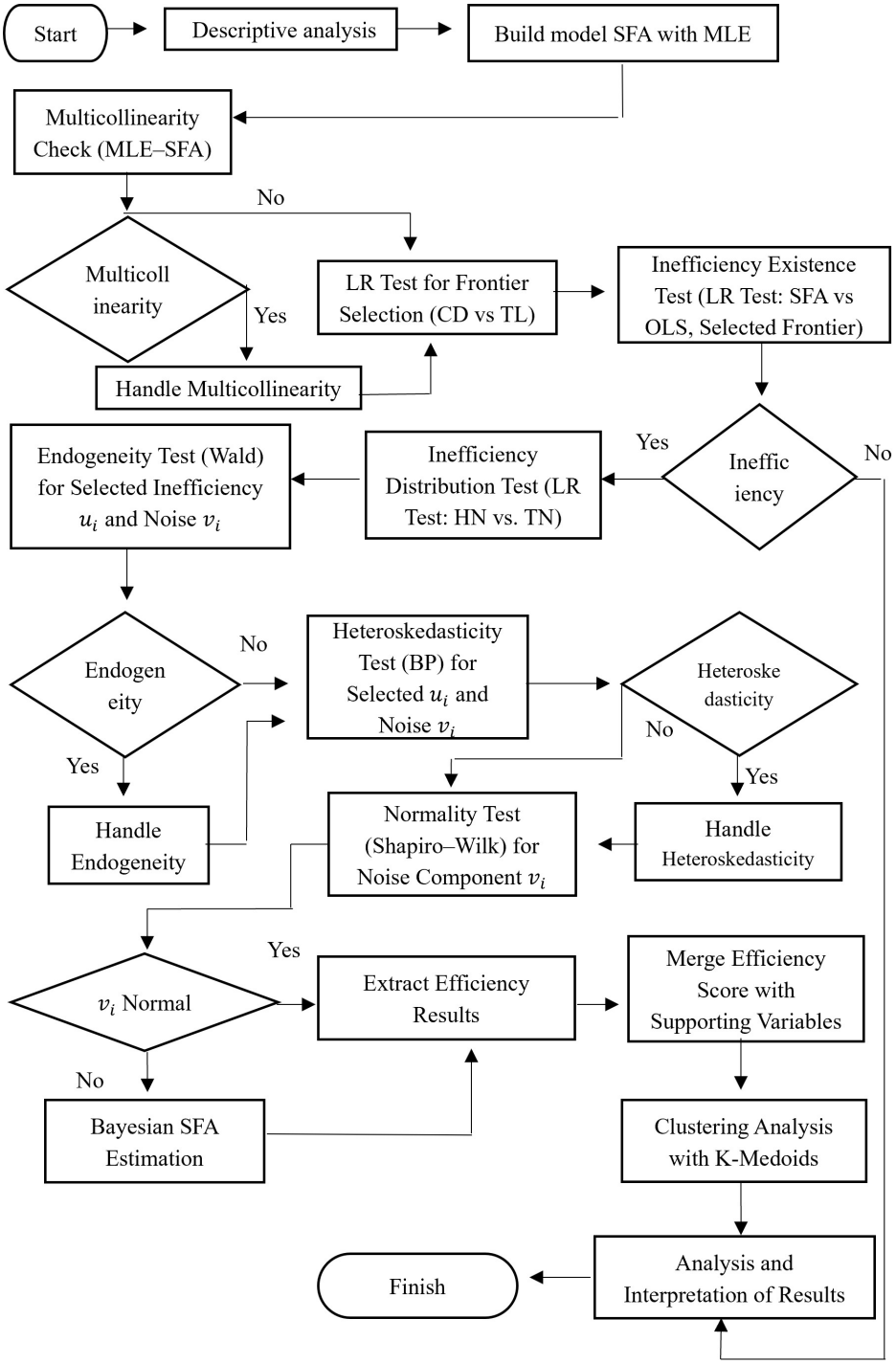

Figure 3.1 Research Flowchart.

BAB V
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study measures the technical efficiency of PLN's UID/UIW using the Stochastic Frontier Analysis (SFA) approach and clusters the efficiency results using the K-Medoids method. Based on the analysis, the following conclusions are drawn:

1. The SFA approach with a Cobb–Douglas frontier and a Half-Normal inefficiency distribution successfully estimates the technical efficiency of each unit using network length, average electricity tariff, and connected customer load as inputs, with energy sold as the output. Initial estimation using MLE was affected by violations of the normality assumption of the random error term, leading to results that were not fully robust. Therefore, a Bayesian MCMC approach with a Student-t disturbance specification was employed, yielding stable, convergent parameter estimates and more reliable inference.
2. Clustering of technical efficiency scores and selected technical variables, specifically technical efficiency (TE) and connected customer load (Input 3), was conducted using the K-Medoids method, as connected customer load shows a strong influence on electricity energy sold, reflecting the scale of demand served by each distribution unit. The analysis resulted in three well-defined clusters, supported by a Silhouette Coefficient of 0.657, a Davies–Bouldin Index of 0.4035, and an R-square value of 0.815, indicating good cluster separation and strong explanatory power.
3. The integration of Bayesian SFA and K-Medoids clustering enables the identification of efficiency patterns through two main stages, namely frontier estimation and clustering-based grouping. This approach allows technical efficiency and key explanatory variables particularly connected customer load to be directly analyzed and grouped while preserving the essential characteristics captured by the SFA model. Consequently, it provides clear and interpretable quantitative insights into performance variations among PLN's UID and UIW

units within the electricity distribution system.

5.2 Recommendation

To improve the accuracy and scope of analysis in future research, it is highly recommended to include a more diverse set of technical and operational variables as well as to increase the sample size. This aims to make the analysis results more representative and comprehensive. Additionally, it is important to explore and compare other clustering methods that are more robust against outliers and capable of handling non-normally distributed data. With this approach, future studies are expected to provide deeper insights and more accurate results regarding the technical efficiency of electricity distribution units.

REFERENCES

AbdElSamea, A., & Saif, S. M. 2024. K-Medoid Clustering Containerized Allocation Algorithm for Cloud Computing Environment. *Journal of Electrical Systems and Information Technology*. **11**(1): 35.

Aditya, I. A., Haryadi, F. N., Haryani, I., Rachmawati, I., Ramadhani, D. P., Tantra, T., & Alamsyah, A. 2023. Understanding Service Quality Concerns from Public Discourse in Indonesia State Electric Company. *Heliyon*. **9**(8): e18768.

Adriani, D., Dewi, R., Saleh, L., Heryadi, D. Y., Sarie, F., Sudipa, I. G. I., & Rahim, R. 2023. Using Distance Measure to Perform Optimal Mapping with the K-Medoids Method on Medicinal Plants, Aromatics, and Spices Export. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*. **14**(3): 103-111.

Agrawal, P., Kanwar, N., Gupta, N., Niazi, K. R., Swarnkar, A., Meena, N. K., & Yang, J. 2020. Reliability and Network Performance Enhancement by Reconfiguring Underground Distribution Systems. *Energies*. **13**(18): 4719.

Aigner, D., Lovell, C. A. K., & Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models.*Journal of Econometrics*. **6**(1): 21-37.

Awong, L. E. E., & Zielinska, T. 2023. Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification. *23*(28): 7925.

Ayu, K., & Yunusa-Kaltungo, A. 2020. A Holistic Framework for Supporting Maintenance and Asset Management Life Cycle Decisions for Power Systems. *Energies*. **13**(8): 1937.

Battese, G. E., & Corra, G. S. 1977. Estimation of a production frontier model: With application to the pastoral zone of Eastern Australia. *Australian Journal of Agricultural Economics*. **21**(3): 169-179.

Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. 2019. A survey of clustering algorithms for an industrial context. *Procedia Computer Science*. **148**: 291-302.

Breusch, T. S., & Pagan, A. R. 1979. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*. **47**(5): 1287-1294.

Budiaji, W., & Leisch, F. 2019. Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. *Algorithms*. **12**(9): 177.

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. *Decision Analytics Journal*. **3**: 100051.

Choi, K. 2023. Approach with the MCMC and the HMC as a Competitor of Classical Likelihood Statistics for Pharmacometricians. Translational and Clinical Pharmacology. *Translational and Clinical Pharmacology*. **31**(2): 69-84.

Coventry, B. S., & Bartlett, E. L. 2024. Practical Bayesian Inference in Neuroscience: Or How I Learned to Stop Worrying and Embrace the Distribution. *eNeuro*. **11**(7): ENEURO.0484-23.2024.

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. 2022. Statistical power for cluster analysis. *BMC Bioinformatics*. **23**: 205.

Fenyves, V., Tarnóczi, T., Bács, Z., Kerezsi, D., Bajnai, P., & Szoboszlai, M. 2022. Financial Efficiency Analysis of Hungarian Agriculture, Fisheries and Forestry Sector. *Agricultural Economics (Zemědělská Ekonomika)*. **68**(11): 413-426.

Firmansyah, M. I., Kustiyahningsih, Y., Rahmanita, E., Abidin, M. S., & Satoto, B. D. 2025. Optimization of MSMEs Clustering in Sampang District Using K-Medoids Method and Silhouette Coefficient Method. *Teknika*. **14**(1): 1-8.

Fitriyanto, R., & Syafiqoh, U. 2024. Multilevel Modal Value Analysis for Interpreting Categorical K-Medoids Clusters Data. *Jurnal Techno Nusa Mandiri*. **21**(2): 134-143.

Galán, J. E., Veiga, H., & Wiper, M. P. 2014. Bayesian Estimation of Inefficiency Heterogeneity in Stochastic Frontier Models. *Journal of Productivity Analysis*. 42(1): 85-101.

García-Escudero, L. A., & Mayo-Iscar, A. 2024. Robust Clustering Based on Trimming. *WIREs Computational Statistics*. **16**(4): e1658.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis*. **1**(3): 515-534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. 2013. *Bayesian Data Analysis* (3rd ed.). CRC Press. Boca Raton, FL, USA.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*. **2**(4): 1360-1383.

Ghorbani, H. 2019. Mahalanobis Distance and its Application for Detecting Multivariate Outliers. *Facta Universitatis, Series: Mathematics and Informatics*. **34**(3): 583-595.

Griffin, J. E., & Steel, M. F. J. 2007. Bayesian Stochastic Frontier Analysis Using WinBUGS. *Journal of Productivity Analysis*. **27**: 163-176.

Gujarati, D. N., & Porter, D. C. 2009. *Basic Econometrics*. 5th. Ed. McGraw-Hill Irwin. Boston, Mass.

Gunkel, P. A., Jacobsen, H. K., Bergaentzlé, C. M., Scheller, F., & Andersen, F. M. 2023. Variability in electricity consumption by category of consumer: The impact on electricity load profiles. *International Journal of Electrical Power & Energy Systems*. **147**: 108852.

Hadri, K. 1999. Estimation of a Doubly Heteroscedastic Stochastic Frontier Cost Function. *Journal of Business & Economic Statistics*. **17**(3): 359-363.

Hassan, B. A., Tayfor, N. B., Hassan, A. A., Ahmed, A. M., Rashid, T. A., & Abdalla, N. N. 2024. From A-To-Z Review of Clustering Validation Indices. *Neurocomputing*. **601**: 128198.

Hendrocahyo, H., & Kurniawati, L. 2022. Understanding the Financial Performance of PT PLN (Persero): A Narrative on State-Owned Enterprise (SOE) with a Mandate of Electricity in Indonesia. *Binus Business Review*. **13**(3): 242-258.

Hogg, R. V., McKean, J., & Craig, A. T. 2019. *Introduction to Mathematical Statistics*. (8th ed.). Boston, Pearson.

Hou, Z., Ramalho, J. J. S., & Roseta-Palma, C. 2025. Dealing With Endogeneity in Stochastic Frontier Models: A Comparative Assessment of Estimators.*Energy Economics*. **151**: 108922.

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. 2025. Benchmarking Validity Indices for Evolutionary K-Means Clustering Performance. *Scientific Reports*. **15**(1): 21842.

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. 2025. Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*. **11**: e41953.

Johnson, V. E., Pramanik, S., & Shudde, R. 2023. Bayes factor functions for reporting outcomes of hypothesis tests. *Statistics Psychological and Cognitive Sciences*. **120**(8): e2217331120.

Jr, J. F. H., Black, W. C., Babin, B. J., & Anderson, R. E. 2019. *Multivariate Data Analysis*. 8th. Ed. Annabel Ainscow. Andover, Hampshire, United Kingdom.

Kallel, S., Amayri, M., & Bouguila, N. 2025. Clustering and Interpretability of Residential Electricity Demand Profiles. *Sensors*. **25**(7): 2026.

Karagiannis, G. P. 2022. Introduction to Bayesian Statistical Inference. In L. J. M. Aslett, F. P. A. Coolen, & J. De Bock (Eds.), *Uncertainty in Engineering: Introduction to Methods and Applications*. 1–13.

Khan, D. M., Yaqoob, A., Zubair, S., Khan, M. A., Ahmad, Z., & Alamri, O. A. 2021. Applications of Robust Regression Techniques: An Econometric Approach. *Mathematical Problems in Engineering*. **2021**(1): 6525079.

Khatiwada, R., Bajracharya, T. R., & Khan, S. 2024. Improving the Energy Efficiency of a Power Distribution Network by Loss Reduction: A Case Study in Rural 11 Kv Feeder. *Journal of Advanced College of Engineering and Management*. **9**: 339–349.

Khumbakar, S. C., Wang, H.-J., & P. Horncastle, A. 2015. *A Practitioner's Guide to Stochastic FrontierAnalysis Using Stata*. Cambridge University Press. New York, NY, USA.

Lesosky, M. 2020. Robust Statistical Methods with R, 2nd Edition. *International Journal of Epidemiology*. **49**: 1056–1056.

Li, Y., Liu, A.-C., Wang, S.-M., Zhan, Y., Chen, J., & Hsiao, H.-F. 2022. A Study of Total-Factor Energy Efficiency for Regional Sustainable Development in China: An Application of Bootstrapped DEA and Clustering Approach. *Energies*. **15**(9): 3093.

Loh, P.-L. 2025. A Theoretical Review of Modern Robust Statistics. *Annual Review of Statistics and Its Application*. **12**: 477–496.

Lu, Y., Phillips, C. A., & Langston, M. A. 2019. A Robustness Metric for Biological Data Clustering Algorithms. *BMC Bioinformatics*. **20**(15): 503.

Mahdi, M. A., Hosny, K. M., & Elhenawy, I. 2021. Scalable Clustering Algorithms for Big Data: A Review. *IEEE Access*. **9**: 80015–80027.

Makieła, K., & Mazur, B. 2020. Bayesian model averaging and prior sensitivity in stochastic frontier analysis. *Econometrics*. **8**(2): 13.

Makieła, K., & Mazur, B. 2022. Model Uncertainty and Efficiency Measurement in Stochastic Frontier Analysis with Generalized Errors. *Journal of Productivity Analysis*. **58**(1): 35–54.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. 2019. *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons. Hoboken, New Jersey, USA.

Motherway, B., Bilodeau, J., Cohen, J., Mooney, E., Rozite, V., & Slade, M. 2024. *Energy Efficiency 2024*.

Nahdliyah, M. A., Widiharih, T., & Prahutama, A. 2019. Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018). *Jurnal Gaussian*. **8**(2): 161–170.

Nascimento, M. G. L., Silva, R. S., Mendonça, M. J., & Pereira Jr., A. O. 2021. Estimating the Efficiency of Brazilian Electricity Distribution Utilities. *Journal of Applied Statistics*. **49**(8): 2157–2166.

Nguyen, H. N., & O'Donnell, C. 2025. Using Stochastic Frontier Analysis to Assess the Performance of Public Service Providers in the Presence of Demand Uncertainty. *Journal of Productivity Analysis*. **64**(1): 61–79.

Nocedal, J., & Wright, S. J. 2006. *Numerical Optimization*. Springer. New York, NY, USA.

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. 2021. Comprehensive Review of K-Means Clustering Algorithms. *International Journal of Advances in Scientific Research and Engineering*. **7**(8): 64–69.

Papadopoulos, A. 2021. Stochastic Frontier Models Using the Generalized Exponential Distribution. *Journal of Productivity Analysis*. **55**(1): 15–29.

Pechrová, M. Š., & Šimpach, O. 2020. Cobb Douglass or Translog Production Function in Efficiency Analysis. page. 558–563. International Conference on Mathematical Methods in Economics (MME 2020), Prague, Czech Republic.

Pitafi, S., Anwar, T., & Sharif, Z. 2023. A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms. *Applied Sciences*. **13**(6): 3529.

Rahmawati, O., & Fauzan, A. 2024. Provincial Clustering Based on Education Indicators: K-Medoids Application and K-Medoids Outlier Handling. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*. **18**(2): 1167–1178.

Rauf, R. I., Ayinde, K., Hamidu, B. A., Kikelomo, B. O., & Olusegun, A. O. 2024. New Approach in Stochastic Frontier Analysis Estimation for Addressing Joint Assumption Violation of Heteroscedasticity and Multicollinearity. *SSRN Electronic Journal*. **26**(9): 9–26.

Rauniyar, P. B., & Kim, J. 2025. Assessing the Technical Efficiency of Rice Producers in the Parsa District of Nepal. *Agriculture*. **15**(3): 342.

Rüde, L., Wussow, M., Heleno, M., Gust, G., & Neumann, D. 2024. Estimating Electrical Distribution Network Length and Capital Investment Needs from Real-World Topologies and Land Cover Data. *Energy Policy*. **195**: 114368.

Ryu, H., Kim, Y., Jang, P., & Aldana, S. 2020. Restructuring and Reliability in the Electricity Industry of OECD Countries: Investigating Causal Relations between Market Reform and Power Supply. *Energies*. **13**(18): 4746.

Sabar, M., & Kamil, A. 2023. Estimating Technical Efficiency of Crude Palm Oil in Malaysia. *Journal of Numerical Optimization and Technology Management*. **1**(2): 52-58.

Sakouvogui, K., Shaik, S., Doetkott, C., & Magel, R. 2021. Sensitivity Analysis of Stochastic Frontier Analysis Models. *Monte Carlo Methods and Applications*. **27**(1): 71–90.

Sarnovsky, M., & Bednar, P. 2025. Segmentation of Electricity Consumers Using Clustering. *Acta Polytechnica Hungarica*. **22**(7): 285–298.

Seppälä, J., Kari, J., & Järventausta, P. 2024. The Total Cost of Reliable Electricity Distribution. *Electricity*. **5**(4): 916–930.

Shapiro, S. S., & Wilk, M. B. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. **52**(3/4): 591.

Smiti, A. 2020. A Critical Overview of Outlier Detection Methods. *Computer Science Review*. **38**: 100306.

Solikhun, S., Siregar, M. R., Pujiastuti, L., & Wahyudi, M. 2024. Manhattan distance-based K-Medoids clustering improvement for diagnosing diabetic disease. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. **8**(6): 710–718.

Sureja, N., Chawda, B., & Vasant, A. 2022. N Improved K-Medoids Clustering Approach Based on the Crow Search Algorithm. *Journal of Computational Mathematics and Data Science*. **3**: 100034.

Susanty, A., Purwanggono, B., & Faruq, C. A. 2022. Electricity Distribution Efficiency Analysis Using Data Envelopment Analysis (DEA) and Soft System Methodology. *Procedia Computer Science*. **203**: 342–349.

Tirkaso, W. T., & Gren, I.-M. 2023. Evaluation of Cost Efficiency in Hydropower-Related Biodiversity Restoration Projects in Sweden – A Stochastic Frontier Approach. *Journal of Environmental Planning and Management*. **66**(2): 221–240.

Van den Broeck, J., Koop, G., Osiewalski, J., & Steel, M. F. J. 1994. A Bayesian stochastic frontier analysis. *Journal of Econometrics*. **61**(2): 273–303.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*. **54**(3): 426–482.

Wang, J., Zhang, Z., & Yue, S. 2025. A Validity Index for Clustering Evaluation by Grid Structures. *Mathematics*. **13**(6): 1017.

Wani, A. A. 2024. Comprehensive Analysis of Clustering Algorithms: Exploring Limitations and Innovative Solutions. *PeerJ Computer Science*. **10**: e2286.

Wei, Z., Choy, S. T. B., Wang, T., & Zhu, X. 2025. Bayesian stochastic frontier models under the skew-normal half-normal settings. *Journal of Productivity Analysis*. **64**(1): 81–91.

Weisberg, S. 2005. *Applied Linear Regression*. Vol. 528. John Wiley & Sons. New York, NY, USA.

Wheat, P., Stead, A. D., & Greene, W. H. 2019. Robust stochastic frontier analysis: a Student'st-half normal model with application to highway maintenance costs in England. *Journal of Productivity Analysis*. **51**(1): 21-38.

Wilks, S. S. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*. **9**(1): 60–62.

Wongoutong, C. 2024. The Impact of Neglecting Feature Scaling in K-Means Clustering. *PLOS ONE*. **19**(12): e0310839.

Wu, J.-S. 2025. Assessing The Impact of Macroeconomic Factors on the Efficiency of Taiwan's Real Estate Firms Using Stochastic Frontier Analysis. *Journal of Asian Architecture and Building Engineering*. 1–13.

Yaro, A., Filip, M., Maly, K., & Prazak, P. 2024 Clustering Performance Analysis of the K-Medoids Algorithm for Improved Fingerprint-Based Localization. *Jordan Journal of Electrical Engineering*. **10**(3): 1.

Zaki, D. A., & Hamdy, M. 2022. A Review of Electricity Tariffs and Enabling Solutions for Optimal Energy Management. *Energies*. **15**(22): 8527.

Zhang, C., Huang, W., Niu, T., Liu, Z., Li, G., & Cao, D. 2023. Review of Clustering Technology and Its Application in Coordinating Vehicle Subsystems. *Automotive Innovation*. **6**: 89–115.

Zhou, E., & Mai, T. 2021. *Electrification Futures Study: Operational Analysis of U.S. Power Systems with Increased Electrification and Demand-Side Flexibility* (No. NREL/TP–6A20-79094, 1785329, MainId:33320).