Online version at https://journal.lenterailmu.com/index.php/jafotik

JAFOTIK

E-ISSN: 3031-2698, DOI prefix: 10.70356

Analysis of oil palm production patterns using the DBSCAN method based on historical data at PT. Suryabumi Agrolanggeng

Agnes Kristina Sipayung [1,*], ***Syahril Rizal*** [2]

[1,2] *Program Studi Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia*

ARTICLE INFO

Article history:
Received August 2026
Accepted 12 September 2026
Published 13 September 2026

Keywords:
Oil Palm
DBSCAN
Clustering
Data Mining
Silhouette Score

ABSTRACT

Oil palm is one of the important plantation commodities that plays a significant role in Indonesia's economy. PT. Suryabumi Agrolanggeng has historical oil palm production data covering production, rainfall, and fertilizer usage variables for the period 2019–2025. The main challenges are production fluctuations and the suboptimal utilization of historical data to identify production patterns and periods with different characteristics. This study aims to apply the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to group data based on the similarity of their characteristics. The research stages include data collection, preprocessing, standardization using the Z-Score method, parameter determination using a K-Distance Graph, DBSCAN clustering, and evaluation using the Silhouette Score. The results show that the parameters Eps = 0.95 and MinPts = 6 produce three main clusters, with 58 data points grouped into clusters and 26 data points identified as noise. The resulting Silhouette Score of 0.376 indicates a cluster structure with a moderate level of separation. The analysis results are further presented through a web-based dashboard to facilitate the monitoring and interpretation of production patterns. This study demonstrates that DBSCAN can assist in identifying production patterns and data points with uncommon characteristics, thereby supporting data-driven decision-making.

1. Introduction

Oil palm is one of the most important plantation commodities and plays a significant role in the Indonesian economy [1]. Effective management of oil palm production is essential for enabling companies to understand production variations and utilize historical data as a basis for planning and decision-making [2].

The development of information technology has further encouraged the utilization of data as a valuable source of information for supporting data-driven decisions. Data mining techniques can be used to discover hidden patterns and meaningful information from large datasets, allowing historical records to be transformed into valuable knowledge for organizational decision-making [3].

* *Corresponding author:* Agnes Kristina Sipayung
E-mail address: agneskristina11@gmail.com

PT. Suryabumi Agrolanggeng possesses historical oil palm production data consisting of production volume, rainfall, and fertilizer usage. These data were recorded monthly from 2019 to 2025. Oil palm production can fluctuate over time due to various factors, including environmental conditions and fertilizer application [4]. Variations in production across different periods create challenges in identifying groups of periods with similar characteristics and detecting observations that exhibit atypical patterns. Conventional or manual analysis may be less effective for identifying such patterns, particularly when multiple variables must be considered simultaneously.

Clustering is one of the data mining techniques used to group data based on similarities in their characteristics [5]. This study employs the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm because it forms clusters based on the density of data points and can identify observations that do not belong to any cluster without requiring the number of clusters to be predetermined. DBSCAN primarily uses the Eps and MinPts parameters to determine neighborhood proximity and data density. The DBSCAN algorithm was introduced by Ester et al. in 1996 [6], while the quality of the resulting clusters in this study is evaluated using the Silhouette Score, a clustering evaluation measure introduced by Rousseeuw in 1987 [7].

Previous studies have demonstrated the application of data mining and DBSCAN techniques to agricultural and numerical datasets. Ningsih et al. (2023) examined factors influencing oil palm production, while Hafid and Arisandi (2024) applied DBSCAN to data related to millennial farmers. However, these studies did not specifically investigate the clustering of historical oil palm production periods by simultaneously considering production volume, rainfall, and fertilizer usage. Therefore, this study addresses this gap by applying DBSCAN to historical production data from PT. Suryabumi Agrolanggeng to identify groups of periods with similar production characteristics and observations that exhibit atypical characteristics. The resulting clusters are subsequently presented through an interactive web-based dashboard to facilitate data interpretation and monitoring.

Based on these issues, this study aims to implement the DBSCAN method on historical oil palm production data, analyze the resulting cluster patterns and identified noise observations, and present the analysis results through a web-based dashboard. In addition, the dashboard provides production forecast information for 2026 as a complementary feature to support the interpretation of production conditions and provide additional information for data-driven decision-making.

2. Method

This study employs a quantitative approach using data mining techniques to analyze historical oil palm production data at PT. Suryabumi Agrolanggeng. Data mining techniques enable the extraction of meaningful patterns and knowledge from large datasets to support data-driven decision-making [8]. The research focuses on identifying patterns and grouping production periods based on similarities in production volume, rainfall, and fertilizer usage. The clustering process is performed using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, which is widely used for identifying dense groups and detecting noise within datasets [9].

A. Data Collection

The data used in this study consist of secondary data obtained from PT. Suryabumi Agrolanggeng. The dataset contains historical monthly records from 2019 to 2025. The variables used include oil palm production, rainfall, and fertilizer usage. These variables were selected because environmental conditions and agricultural inputs can influence oil palm production performance [10].

B. Data Preprocessing

Before the clustering process, the collected data undergo preprocessing to ensure that the dataset is suitable for analysis. Data preprocessing is an important stage in data mining because data quality can significantly affect the performance and reliability of the analytical results [8]. The preprocessing stage includes data cleaning, checking for missing values, removing duplicate data, and selecting relevant variables. Since the variables have different measurement scales, data standardization is applied before the clustering process. Standardization transforms the variables into a comparable scale and prevents variables with larger numerical ranges from having a dominant influence on distance-based calculations [11].

C. DBSCAN Clustering

The clustering process is conducted using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN groups data points based on the density of their distribution and does not require the number of clusters to be determined in advance [9]. The algorithm uses two main parameters, namely Eps (epsilon), which determines the maximum distance between neighboring data points, and MinPts, which determines the minimum number of data points required to form a dense region [6]. Based on these parameters, DBSCAN classifies data points into three categories: core points, border points, and noise points. Core points are located in dense regions, border points are located near core points but do not meet the density requirements to become core points, while noise points do not belong to any cluster.

D. Cluster Evaluation

The quality of the resulting clusters is evaluated using the Silhouette Score. This evaluation method measures the similarity of a data point to other points within the same cluster compared to points in different clusters [7]. The Silhouette Score ranges from -1 to 1, where values closer to 1 indicate that data points are well matched with their own cluster and clearly separated from other clusters.

The evaluation process is conducted to assess the effectiveness of the selected DBSCAN parameters and determine the clustering configuration that produces meaningful groups from the historical oil palm production data.

E. Data Visualization and Dashboard Development

The results of the clustering analysis are visualized through graphs and an interactive web-based dashboard. Data visualization plays an important role in transforming analytical results into understandable information that can support decision-making [12]. The dashboard is designed to provide users with an easier way to interpret the patterns identified from historical production data. The visualizations include cluster distribution, production trends, rainfall patterns, fertilizer usage, and identified noise data. By presenting the results interactively, users can explore production patterns and differences between clusters more effectively.

In addition to presenting clustering results, the dashboard also provides oil palm production forecast information for the year 2026 as an additional analytical feature. Forecasting utilizes historical data to estimate potential future conditions and can support planning and decision-making processes [13]. Previous research has shown that the quality and presentation of information within dashboard visualizations can influence users' information satisfaction, perceived task complexity, and decision-making quality [14]. Therefore, the dashboard developed in this study is designed not only as a reporting interface but also as an analytical tool that facilitates the interpretation of clustering results.

The dashboard presents several visualization components, including cluster distribution, production trends, rainfall patterns, fertilizer usage, and observations identified as noise by the DBSCAN algorithm. These visualizations enable users to examine differences in production characteristics among clusters and to identify periods with unusual combinations of production, rainfall, and fertilizer usage. Interactive dashboard features can further improve users' ability to explore analytical information by allowing them to examine different data views and relationships according to their analytical needs [15]. In this context, the interactive dashboard provides a more flexible means of exploring the historical production data than static reports or tabular presentations. The visualization of cluster characteristics also helps users interpret the practical meaning of the resulting groups and identify production periods that may require further investigation.

In addition to presenting the clustering results, the dashboard provides oil palm production forecast information for 2026 as a complementary analytical feature. Forecasting historical production data can provide estimates of potential future production conditions and support operational planning and decision-making. Previous studies have demonstrated the application of time-series and statistical forecasting methods to oil palm production, including approaches that incorporate rainfall and other agroecological variables [16].

3. Result and Discussion

3.1. DBSCAN Parameter Determination

The first stage in implementing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was determining the appropriate parameter values. DBSCAN requires two main parameters, namely Eps (epsilon) and MinPts (minimum points). The Eps parameter determines the maximum radius used to identify neighboring data points, while MinPts specifies the minimum number of points required to form a dense region.

In this study, parameter selection was performed using the K-Distance Graph. Several MinPts values, including 4, 5, 6, and 7, were tested to observe differences in the resulting curves. The optimal parameter was selected by identifying the point where a significant change in the slope of the curve occurred.
Based on the comparison of the K-Distance curves, MinPts = 6 was selected because it showed a clearer change in the slope compared with the other tested values. The elbow point was observed at approximately Eps = 0.95. Therefore, Eps = 0.95 and MinPts = 6 were selected as the final parameters for the DBSCAN clustering process.

The selected parameters were subsequently applied to the entire dataset consisting of historical monthly oil palm production data from 2019 to 2025 (Table 1).

Table 1 - DBSCAN Parameters Used in the Study

Parameter	Value	Description
Eps	0.95	Maximum radius used to determine the proximity between data points
MinPts	6	Minimum number of points required within the Eps neighborhood
Distance Metric	Euclidean Distance	Measures the distance between data points based on the research variables

The selected parameters determine how the algorithm identifies dense regions within the dataset. Data points located within the Eps radius are considered neighbors, while the MinPts value determines whether a data point can be classified as part of a dense region.

3.2. Manual DBSCAN Calculation

Manual calculations were performed to demonstrate the process of DBSCAN clustering using the research data. The calculation begins with data standardization, followed by the formation of multidimensional data points, Euclidean Distance calculation, comparison with the Eps parameter, and the

identification of core points, border points, and noise. The dataset consists of three variables: oil palm production, rainfall, and fertilizer usage. Since these variables have different measurement scales, data standardization was required before calculating the distance between data points. The parameters used in the manual calculation were Eps = 0.95 and MinPts = 6.

a. Z-Score Standardization

Before clustering, the production, rainfall, and fertilizer variables were standardized using the Z-Score method. Standardization was necessary because the variables have different numerical ranges. Without standardization, variables with larger numerical values could have a greater influence on the distance calculation. The Z-Score formula is expressed as follows (Eq. 1):

$$Z = (X - \mu) / \sigma \quad (1)$$

where Z represents the standardized value, X represents the original data value, μ represents the mean value, and σ represents the standard deviation.

Example 1: Oil Palm Production in January 2019

The oil palm production value in January 2019 was 5,875.52. The average production value was 3,965.13, while the standard deviation was 1,234.28. The calculation is expressed as:

$$Z = (5{,}875.52 - 3{,}965.13) / 1{,}234.28$$
$$Z = 1.548$$

Therefore, the standardized production value for January 2019 was 1.548. The positive value indicates that production during this period was above the average production value of the dataset.

Example 2: Rainfall in January 2019

The rainfall value in January 2019 was 200.00. The average rainfall value was 190.12, with a standard deviation of 49.49. The calculation is:

$$Z = (200.00 - 190.12) / 49.49$$
$$Z = 0.200$$

Therefore, the standardized rainfall value for January 2019 was 0.200.

Example 3: Fertilizer Usage in January 2019

The fertilizer usage value in January 2019 was 50.00. The average fertilizer usage was 46.31, with a standard deviation of 6.32. The calculation is:

$$Z = (50.00 - 46.31) / 6.32$$
$$Z = 0.584$$

Therefore, the standardized fertilizer usage value for January 2019 was 0.584.

b. Data Point Formation

After the standardization process, the three standardized variables were combined into a multidimensional data point. Each monthly observation was represented by a point consisting of production, rainfall, and fertilizer usage. The data point for January 2019 (P1) was represented as:

$$P1 = (1.548, 0.200, 0.584)$$

This representation describes the condition of oil palm production in January 2019 based on the three standardized research variables. The same standardization process was applied to the February 2019 data to form the second data point (P2). This procedure was repeated for all 84 monthly observations from 2019 to 2025.

c. Euclidean Distance Calculation

After forming the data points, the distance between observations was calculated using Euclidean Distance. This calculation was used to measure the similarity between different periods based on the three research variables. The Euclidean Distance formula for three variables is expressed as (Eq. 2):

$$d(P1,P2) = \sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]} \quad (2)$$

where x represents the standardized production value, y represents the standardized rainfall value, and z represents the standardized fertilizer usage value. Based on the calculation between January 2019 (P1) and February 2019 (P2), the Euclidean Distance obtained was:

$$d(P1,P2) = 1.258$$

This result indicates the level of difference between the characteristics of the two periods based on production, rainfall, and fertilizer usage.

d. Comparison with the Eps Parameter

The calculated distance was subsequently compared with the selected Eps value. In this study, the Eps parameter was set to 0.95. The distance between P1 and P2 was:

$$d(P1,P2) = 1.258$$

Meanwhile:

$$Eps = 0.95$$

Since:

$$1.258 > 0.95$$

P2 was not included in the ε-neighborhood of P1. This indicates that the two observations were not sufficiently close to be considered direct neighbors according to the selected DBSCAN parameter. The same comparison process was performed for the remaining observations in the dataset.

e. Identification of Core Points, Border Points, and Noise

After determining the neighborhood relationships between data points, DBSCAN classified each observation based on data density. Using Eps = 0.95 and MinPts = 6, the observations were categorized as core points, border points, or

noise. A core point is a data point that has at least the minimum required number of neighboring points within the Eps radius. Core points represent the central areas of dense clusters.

A border point does not meet the density requirement to become a core point but is located within the Eps neighborhood of a core point. Therefore, it can still be assigned to a cluster.

Noise refers to data points that do not satisfy the requirements of either core points or border points. These observations are located in low-density regions and are not assigned to any cluster.

The clustering results identified 50 core points, 8 border points, and 26 noise points. Therefore, a total of 58 observations were successfully assigned to clusters, while 26 observations were identified as noise.

The presence of noise does not necessarily indicate incorrect data. In the context of oil palm production, noise may represent periods with unusual combinations of production, rainfall, and fertilizer usage compared with the general pattern of the dataset.

3.3. Clustering Results

The implementation of DBSCAN using Eps = 0.95 and MinPts = 6 resulted in three main clusters and one noise category. Each cluster represents a group of periods with similar characteristics based on the selected variables.

Table 2 - Summary of DBSCAN Clustering Results

Group	Number of Data	Percentage
Cluster 1 – High	14	16.67%
Cluster 2 – Medium	29	34.52%
Cluster 3 – Relatively Low	15	17.86%
Noise – Requires Further Review	26	30.95%
Total	84	100%

Based on Table 2, Cluster 2 contained the largest number of observations, consisting of 29 data points or 34.52% of the total dataset. This result indicates that most production periods exhibited relatively moderate characteristics compared with the other groups.

Cluster 1 consisted of 14 observations or 16.67% of the dataset. This cluster represents periods with relatively high characteristics based on the combination of production, rainfall, and fertilizer usage.

Cluster 3 contained 15 observations or 17.86% of the total data. This group represents periods with relatively lower characteristics compared with the other clusters.

In addition to the three clusters, 26 observations or 30.95% of the dataset were classified as noise. The relatively large proportion of noise indicates that several periods had characteristics that differed significantly from the dense regions identified by DBSCAN. The visualization provides a clearer representation of the distribution of observations across the resulting clusters. Data points located close to one another tend to form dense groups, while observations positioned outside the main cluster regions are identified as noise.

3.4. Discussion

The results demonstrate that the DBSCAN algorithm can be applied to historical oil palm production data to identify groups of periods with similar characteristics. Using Eps = 0.95 and MinPts = 6, the algorithm successfully formed three clusters and identified observations that did not belong to dense regions as noise (Figure 1).

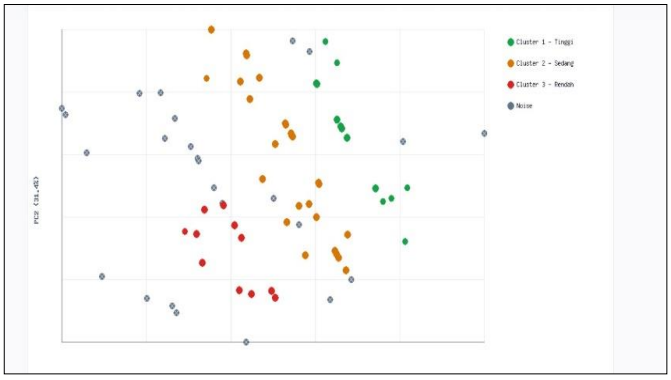

Figure 1 - Visualization of DBSCAN Clustering Results.

A total of 58 observations were assigned to clusters, while 26 observations were classified as noise. Cluster 2 was the largest group, containing 29 observations or 34.52% of the dataset. This indicates that the majority of historical periods exhibited relatively similar characteristics.

The identification of 26 noise observations is an important finding of this study. Rather than forcing all observations into predefined groups, DBSCAN separates observations with unusual characteristics. In the context of oil palm production, these unusual periods may provide useful information for further investigation.

The Silhouette Score of 0.376 indicates that the clustering structure has a moderate level of separation. The result suggests that the clusters are meaningful but could potentially be improved through further parameter optimization or the inclusion of additional variables.

Overall, the implementation of DBSCAN transforms historical production data into meaningful groups that can support the identification of production patterns. The integration of the clustering results into an interactive dashboard further improves data accessibility and supports the interpretation of complex historical information.

4. Conclusion

This study successfully applied the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to historical oil palm production data from PT. Suryabumi Agrolanggeng covering the period from 2019 to 2025, using production volume, rainfall, and fertilizer usage as the main variables. Based on the K-Distance Graph analysis, Eps = 0.95 and MinPts = 6 were selected as the DBSCAN parameters, resulting in three clusters consisting of 14 high-production observations, 29 medium-production observations, and 15 relatively low-production observations, while 26 observations

were identified as noise. The clustering evaluation yielded a Silhouette Score of 0.376, indicating a moderate clustering structure with a reasonable degree of separation among the resulting clusters. These results demonstrate that DBSCAN can identify groups of observations with similar characteristics while also detecting observations that exhibit atypical characteristics. The clustering results were subsequently implemented in an interactive web-based dashboard to facilitate the visualization and interpretation of production patterns, rainfall, fertilizer usage, and noise observations. The dashboard also incorporates oil palm production forecasting for 2026 as an additional analytical feature to provide information on potential future production conditions. Overall, the study demonstrates that DBSCAN can transform historical oil palm production data into meaningful information that supports production monitoring, pattern identification, and data-driven decision-making at PT. Suryabumi Agrolanggeng.

Acknowledgements

The authors would like to express their sincere gratitude to PT. Suryabumi Agrolanggeng for providing the data and supporting this research. The authors also gratefully acknowledge Universitas Bina Darma for providing academic support and guidance throughout the research process. Finally, the authors extend their appreciation to all individuals who contributed to the completion of this study.

REFERENCES

[1] Badan Pusat Statistik, Statistik Kelapa Sawit Indonesia 2023. Jakarta, Indonesia: Badan Pusat Statistik, 2024.

[2] Direktorat Jenderal Perkebunan, *Statistik Perkebunan Unggulan Nasional 2021–2023: Kelapa Sawit*. Jakarta, Indonesia: Kementerian Pertanian Republik Indonesia, 2023.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[4] T. Ningsih, I. O. Y. Sitompul, and S. D. Siahaan, "Analisa faktor-faktor yang mempengaruhi produksi kelapa sawit di Kebun Tanah Raja PT. Bakrie Sumatera Plantations," *JASc (Journal of Agribusiness Sciences)*, vol. 7, no. 2, 2023. doi: https://doi.org/10.30596/jasc.v7i2.16628.

[5] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA, USA: Pearson Education, 2006.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 226–231. https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[9] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1–21, 2017, doi: https://doi.org/10.1145/3068335.

[10] M. P. Corley and P. B. Tinker, *The Oil Palm*, 5th ed. Oxford, U.K.: Blackwell Science, 2016.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019. https://bcebakhtiyarpur.ac.in/wp-content/uploads/2020/03/file_5e748501c810d.pdf

[12] S. Wexler, J. Shaffer, and A. Cotgreave, *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. Hoboken, NJ, USA: John Wiley & Sons, 2017.

[13] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.

[14] S. Hjelle, P. Mikalef, N. Altwaijry, and V. Parida, "Organizational decision making and analytics: An experimental study on dashboard visualizations," *Information & Management*, vol. 61, no. 6, Art. no. 104011, 2024, doi: https://doi.org/10.1016/j.im.2024.104011.

[15] M. Nadj, A. Maedche, and C. Schieder, "The effect of interactive analytical dashboard features on situation awareness and task performance," *Decision Support Systems*, vol. 135, Art. no. 113322, 2020, doi: https://doi.org/10.1016/j.dss.2020.113322.

[16] E. Santosa, "Peramalan produksi kelapa sawit menggunakan peubah agroekologi di Kalimantan Selatan," *Jurnal Agronomi Indonesia (Indonesian Journal of Agronomy)*, vol. 39, no. 3, 2011, doi: https://doi.org/10.24831/jai.v39i3.14963.