ISSN 2355-0082

Triangulation Approach Using K-Means, Hierarchical Clustering, and DBSCAN for Beef Production Analysis

Nurfia Oktaviani Syamsiah[1], Indah Purwandani[2], Mia Rosmiati[3], Siti Nurwahyuni[4]
[1, 2, 3, 4] Faculty of Engineering and Informatics, Bina Sarana Informatika University, Indonesia
[1] nurfia.nos@bsi.ac.id, [2] indah@bsi.ac.id, [3] mia.mrm@bsi.ac.id, [4] siti.swu@bsi.ac.id

Accepted 16 November 2025
Approved 12 January 2026

***Abstract*— This study implements a methodological triangulation approach for clustering highly skewed data using three algorithms with distinct paradigms: K-Means (partitional-based), Agglomerative Hierarchical Clustering with Ward Linkage (hierarchical-based), and DBSCAN (density-based). Applied to beef production data from 38 Indonesian provinces in 2024, the dataset exhibited extreme characteristics with a coefficient of variation of 171.89%, skewness of 2.87, and a maximum-minimum ratio of 664:1. Data were standardised using Z-score transformation to address scale dominance. Evaluation using the Silhouette Score for K-Means and Hierarchical Clustering, alongside qualitative outlier detection with DBSCAN, revealed high consistency across all algorithms in identifying k=2 as the optimal structure (Agreement: 99.7%). The algorithms consistently isolated three provinces (East Java, West Java, and Central Java) as a high-production cluster, distinctly separated from the remaining 35 provinces. Bootstrap resampling (B=100) confirmed the stability of this structure with a standard deviation of 0.0089. These findings demonstrate that relying on a single algorithm for skewed data is methodologically risky, whereas triangulation provides robust validation for policy formulation.**

***Index Terms*— DBSCAN; Hierarchical Clustering; Outlier Detection; Silhouette Score; Triangulation Algorithm**

I. INTRODUCTION

As an archipelagic nation with a population exceeding 270 million, Indonesia faces considerable challenges in ensuring food security, particularly regarding the availability of animal protein, a vital component of public health. The domestic demand for beef continues to rise alongside population growth and shifting consumption patterns increasingly oriented toward high-quality protein intake. Pressure to enhance livestock sector productivity has intensified, yet efforts to achieve sustainable beef self-sufficiency remain constrained by production disparities across regions [1]–[3].

The 2024 beef production data reveal a pattern in which a small number of provinces on Java Island, historically established as livestock centres, continue to dominate the national supply. As shown in Table I, the three major provinces (East Java, West Java, and Central Java) contribute significantly to the national output, while the majority of other provinces scattered across various islands contribute only marginally. This dominance pattern has persisted for several decades [4], [5].

Production inequality results from the accumulation of various interacting factors, ranging from differences in cattle genetic quality and the availability of modern slaughterhouse infrastructure to variations in the availability of adequate pastureland and regional-level resource allocation policies [6], [7]

In computational analysis, uneven data distribution poses methodological challenges when Euclidean distance-based clustering algorithms like K-Means are applied to real-world data. K-Means' sensitivity to initial centroids and the presence of outliers can easily distort clustering quality, indicating that these challenges are consistently encountered across different data analysis contexts [8]–[10].

Research Gap and Significance Although the value of triangulation and ensemble methods in clustering analysis is increasingly recognised for mitigating algorithmic bias [11], [12], no prior study has applied a robust cross-validation approach specifically to Indonesian beef production data. This data demonstrates extreme imbalance (CV > 170%) and unprecedented regional disparity (ratio 664:1), characteristics that often lead to convergence failures in standard algorithms [13]. Existing studies have been limited to single-algorithm applications, such as Ningsih [14] who utilized K-Means on raw data, or have failed to systematically validate findings through cross-paradigm triangulation [15], [16]. This represents a critical methodological gap given the proven sensitivity of clustering results to method selection [17]. Furthermore, previous research has not addressed the challenge of parameter optimisation for density-based algorithms in the context of agricultural data with extreme outliers. Recent literature emphasises the need for careful parameter adaptation to avoid misidentifying structural noise [18]. This research gap is crucial because policy interventions based on unvalidated clustering structures may lead to resource misallocation or a failure to address genuine production disparities.

TABLE I. BEEF PRODUCTION DATA PER PROVINCE IN INDONESIA, YEAR 2024

Province	Production (Tons)	Province	Production (Tons)
ACEH	11,006.40	SOUTH KALIMANTAN	5,272.95
NORTH SUMATRA	18,245.02	EAST KALIMANTAN	6,466.78
WEST SUMATRA	14,901.15	NORTH KALIMANTAN	632.91
RIAU	13,457.10	NORTH SULAWESI	1,840.64
JAMBI	3,571.72	CENTRAL SULAWESI	3,848.72
SOUTH SUMATRA	11,810.70	SOUTH SULAWESI	13,722.50
BENGKULU	1,762.84	SOUTHEAST SULAWESI	5,985.22
LAMPUNG	18,625.00	GORONTALO	1,900.72
BANGKA BELITUNG ISLANDS	2,490.59	WEST SULAWESI	1,174.18
RIAU ISLANDS	2,202.23	MALUKU	1,279.24
DKI JAKARTA	14,925.20	NORTH MALUKU	1,530.33
WEST JAVA	85,241.70	WEST PAPUA	742.54
CENTRAL JAVA	83,275.69	SOUTHWEST PAPUA	325.41
DIY	6,700.69	PAPUA	683.22
EAST JAVA	96,907.31	SOUTH PAPUA	439.03
BANTEN	19,259.70	CENTRAL PAPUA	667.16
BALI	4,882.25	HIGHLAND PAPUA	145.81
WEST NUSA TENGGARA	11,356.76	WEST KALIMANTAN	3,890.08
EAST NUSA TENGGARA	6,234.53	CENTRAL KALIMANTAN	1,448.19

Our study addresses this gap by implementing systematic triangulation across three clustering paradigms (partitional, hierarchical, and density-based). The selection of K-Means, Hierarchical Clustering, and DBSCAN is grounded in their fundamental differences. The theoretical complementarity of these three paradigms—partition-based optimisation, hierarchical structure discovery, and density-based outlier detection—provides a robust cross-validation unattainable by single-paradigm approaches [11], [12].

II. METHODOLOGY

The research followed a standard data mining methodology framework, encompassing data collection, preprocessing, algorithm implementation, and comparative evaluation, as illustrated in Figure 1.

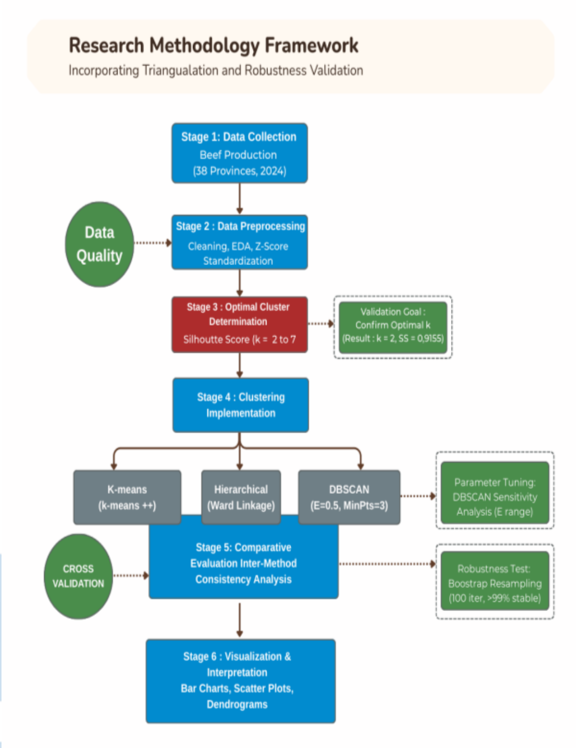

Fig. 1. Research Stages

A. Data Source and Preprocessing

The study utilised beef production data (in tons) from 38 Indonesian provinces in 2024, sourced from the Ministry of Agriculture. Given the extreme skewness (Skewness = 2.87), data preprocessing included Z-score standardisation to transform the data into a standard normal distribution**.** This step is critical to prevent provinces with large production volumes from dominating the Euclidean distance calculations in K-Means and Hierarchical clustering [19].

B. Clustering Algorithms Implementation

Three algorithms were implemented with specific configurations to ensure robustness:

1. K-Means: Implemented with k-means++ initialisation to select optimal initial centroids, accelerating convergence and reducing the probability of falling into local optima [20]. The optimal number of clusters (k) was determined using the Silhouette Score.

2. Agglomerative Hierarchical Clustering: Utilised Euclidean distance and Ward's linkage method, which minimises the total within-cluster variance. The cut-off point for the dendrogram was determined based on the largest vertical distance between merges.

3. DBSCAN: Selected for its ability to handle noise. Parameter selection was conducted systematically.

C. *Algorithm Behavior on Highly Skewed Data*

Each algorithm exhibits distinct sensitivities when applied to data with extreme outliers and high skewness. The Euclidean distance-based objective function in K-Means makes it inherently sensitive to outliers, as extreme values disproportionately influence centroid calculation and cluster assignment [13]. To mitigate this, k-means++ initialisation was employed. Hierarchical Clustering with Ward linkage minimises within-cluster variance, making it relatively robust compared to single or complete linkage. However, hierarchical methods are deterministic; once an outlier is merged, it cannot be reassigned. The density-based paradigm of DBSCAN differs fundamentally by not forcing every observation into a cluster. It defines clusters as dense regions and explicitly labels low-density observations as noise. This characteristic makes DBSCAN methodologically superior for highly skewed data, where outliers represent distinct production regimes rather than measurement errors [18], [21].

D. *DBSCAN Parameter Selection*

For the DBSCAN implementation, parameter selection was conducted systematically through exploratory analysis and sensitivity testing. The Epsilon parameter was initially estimated using a heuristic k-distance plot, plotting the distance to the k-th nearest neighbour (MinPts=3) for all observations sorted in ascending order [22]. The "elbow" in this plot suggested an initial epsilon range of 0.4–0.6. We selected MinPts=3 based on the rule of thumb MinPts dimensionality + 1 [21]; for univariate data (d=1), MinPts=3 provides sufficient density estimation while avoiding excessive noise labelling. The final parameters were validated through systematic sensitivity analysis.

E. *Validation Stability via Bootstrap Resampling*

To assess clustering stability against sampling variation, we implemented bootstrap resampling with 100 iterations. In each iteration, we generated a bootstrap sample by randomly sampling 38 observations with replacement from the original dataset. This resampling approach simulates the variability that would arise from repeated sampling from the population [20]. For each bootstrap sample, we applied K-Means and Hierarchical Clustering, recording Silhouette Scores and cluster membership consistency. A membership consistency near 100% indicates a highly stable clustering structure robust to sampling variations.

F. *Software Environment*

All computational analyses were implemented in Python 3.8.10 running on Windows 10 Pro (64-bit) with 16GB RAM. Data manipulation utilised Pandas 2.0.3 for structured data operations and NumPy 1.24.3 for high-performance numerical array computing. Clustering algorithms were implemented using Scikit-learn 1.3.0, specifically the KMeans (with k-means++), Agglomerative Clustering (Ward linkage), and DBSCAN classes. Statistical analysis utilised SciPy 1.11.1, particularly for dendrogram generation. Data visualizations were created using Matplotlib 3.7.2 for publication-quality figures and Seaborn 0.12.2 for enhanced statistical graphics. All analyses were executed within a Jupyter Notebook 6.5.4 environment to ensure full reproducibility.

III. RESULT

A. *Descriptive Analysis*

The descriptive statistics of the beef production data, summarised in Table II, reveal a fundamental structural imbalance in the national supply chain. The mean production stands at 12,195.26 tons, a figure that is mathematically pulled upward by extreme outliers, whereas the median is significantly lower at 5,629.58 tons. This substantial divergence between the mean and median confirms a heavy right-skewed distribution, indicating that the "average" province does not represent the typical production capacity. Furthermore, the Coefficient of Variation (CV) reached an extreme 171.89%, suggesting that the disparity among provinces is not merely a variation but a sign of high heterogeneity. The maximum-minimum ratio of 664:1 provides the clearest picture of the production gap's magnitude, necessitating differentiated policy interventions.

This structural gap is vividly illustrated in Figure 2. The distribution plot displays a distinct "long-tail" characteristic, where the three leading provinces form a high-production plateau that sharply drops off to a flat consolidation line for the remaining 35 provinces. This visual evidence supports the statistical indication of a dualistic production structure.

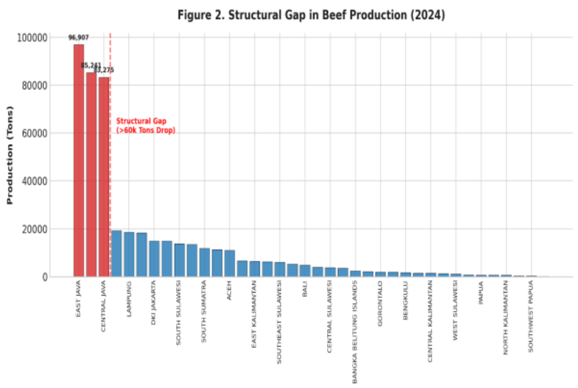

Fig 2. Visualisation of Structural Gap in Beef Production

TABLE II. DESCRIPTIVE STATISTICS OF 2024 BEEF PRODUCTION DATA

Statistical Metric	Value (Tons)	Interpretation
Mean (μ)	12,195.26	Average production per province
Median	5,629.58	Distribution midpoint
Standard Deviation (σ)	20,967.83	Very high variability level
Minimum	145.81	Highland Papua (lowest)
Maximum	96,907.31	East Java (highest)
Range	96,761.50	Huge max-min difference
Skewness	2.87	Positively skewed distribution
Kurtosis	8.45	Heavy-tailed distribution
Coefficient of Variation	171.89%	Very high heterogeneity
Max/Min Ratio	664:01:00	Extremely high disparity

B. *Optimal Cluster Number Determination and Parameter Sensitivity*

The determination of the optimal number of clusters for K-Means and Hierarchical Clustering was rigorously guided by the Silhouette Score validation. As detailed in Table III, the analysis produced a remarkably high score of 0.9155 at k=2. This value is significantly higher than the scores for k=3 (0.7842) or k=4 (0.7123), providing empirical evidence that the natural structure of the data partitions most cleanly into two distinct groups. A score exceeding 0.7 typically denotes a "strong" structure; achieving > 0.9 suggests that the separation between the production centres and the rest of the country is nearly absolute in the feature space. This distinct peak at k=2 is visually demonstrated in Figure 4, which charts the Silhouette Scores across different cluster numbers, highlighting the sharp drop in validation quality for k > 2.

To validate this partition through a density-based paradigm, we performed a sensitivity analysis on the DBSCAN algorithm. The critical challenge in DBSCAN is parameter selection, specifically Epsilon. The results in Table IV reveal a stable detection window at epsilon=0.5-0.55. Within this specific range, the algorithm consistently identified the three super-producer provinces as outliers while keeping the remaining provinces in a coherent cluster. At lower epsilon values (epsilon < 0.5), the algorithm became overly restrictive, fragmenting the main cluster into noise, whereas at higher values (epsilon > 0.6), the distinction collapsed as outliers were merged into the main group.

TABLE III. SILHOUETTE SCORE EVALUATION FOR VARIOUS K VALUES

K	K-Means SS	Hierarchical SS	Average	Category
2	0.9155	0.9155	0.9155	Very Strong
3	0.7842	0.7839	0.7841	Strong
4	0.7123	0.7118	0.7121	Strong
5	0.6845	0.6841	0.6843	Adequate
6	0.6492	0.6488	0.6490	Adequate
7	0.6201	0.6197	0.6199	Adequate

TABLE IV. DBSCAN PARAMETER SENSITIVITY ANALYSIS RESULTS

ε (Epsilon)	MinPts	Number of Clusters	Number of Noise	Interpretation
0.3	3	0	38	Too tight, all noise
0.4	3	0	38	Still too tight
0.5	3	1	3	Optimal: Clear outlier isolation
0.6	3	1	2	One outlier enters the central cluster
0.7	3	1	0	Too loose, no outliers

C. *Clustering Results and Inter-Method Consistency*

The clustering results demonstrate a complete consensus among the three paradigms. A comparative performance summary is provided in Table V, highlighting the structural agreement across methods. While Hierarchical Clustering required slightly more computational time due to dendrogram construction, all methods demonstrated high efficiency.

Table VI details the specific membership of the identified clusters. K-Means and Hierarchical Clustering identified identical partitions: Cluster 0 consists of the three major producers (East Java, West Java, Central Java), while Cluster 1 comprises the remaining 35 provinces. DBSCAN provided a complementary validation by identifying the same three provinces as "Noise" (Outliers) and the remaining 35 as the core cluster.

TABLE V. CLUSTERING RESULTS AND PRODUCTION CENTRE IDENTIFICATION

Method	k / Number of Clusters	Silhouette Score	Outlier / Noise	Time (ms)
K-Means	2	0.9155	N/A	12.4
Hierarchical	2	0.9155	N/A	45.8
DBSCAN	1	N/A	3 Provinces	8.7

TABLE VI. CLUSTERING RESULTS AND PRODUCTION CENTRE IDENTIFICATION 2024

Method & Cluster	Number of Provinces	Average (Tons)	Std Dev (Tons)	Cluster Members
K-Means Cluster 0	3 (7.89%)	88,474.90	6,979.82	East Java, Central Java, West Java
K-Means Cluster 1	35 (92.11%)	7,026.14	5,408.77	35 other provinces
Hierarchical Cluster 0	3 (7.89%)	88,474.90	6,979.82	East Java, Central Java, West Java
Hierarchical Cluster 1	35 (92.11%)	7,026.14	5,408.77	35 other provinces
DBSCAN Noise (-1)	3 (7.89%)	88,474.90	6,979.82	East Java, Central Java, West Java
DBSCAN Cluster 0	35 (92.11%)	7,026.14	5,408.77	35 other provinces

The hierarchical structure of this partition is illustrated in the dendrogram in Figure 3. The dendrogram shows a massive vertical distance before the first split, visually confirming that the data naturally divides into two distinct branches (production centers vs. others) before further granular sub-divisions occur.

The separation of these clusters is further confirmed in Figure 4. The scatter plot maps the provinces in the standardized Z-score space. The visual gap is striking: the top three provinces are located distinctly beyond the 3 sigma threshold (Z-score > 3.0), isolating them from the main consolidation group which is tightly clustered below Z-score < 1.0. This physical distance in the plot validates the mathematical separation found by the algorithms.

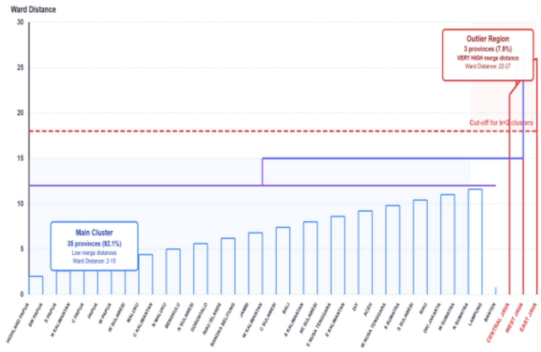

Fig. 3. Hierarchical Clustering Dendrogram (Ward Linkage)

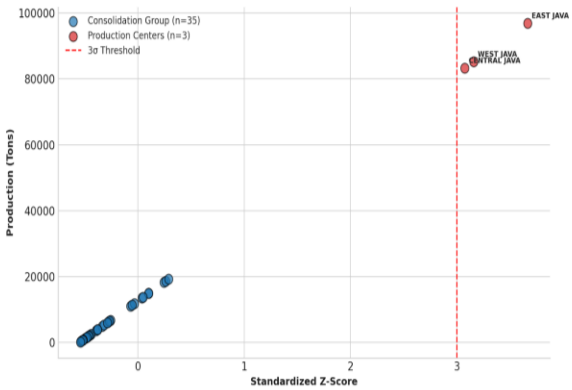

Fig 4. Scatter Plot of Cluster Distribution in Feature Space

D. *Cluster Profiling and Sub-segmentation*

A comprehensive profile of the identified clusters is presented in Table VII. The data underscores the depth of the disparity: Cluster 0 (Production Centres), despite containing only 7.89% of the provinces, commands a staggering 57.3% of the national beef production. The production ratio between the average province in Cluster 0 and Cluster 1 is approximately 12.6:1, highlighting a massive productivity divide that separates the industrial-scale producers in Java from the developing regions.

However, treating the 35 provinces in Cluster 1 as a monolith would be an oversimplification. To provide granular insights for policy targeting, we conducted a sub-segmentation analysis based on production ranges. Table VII breaks down this cluster into four sub-tiers (Upper-Mid to Low). This analysis reveals that even within the "developing" group, significant variation exists; the "Upper-Mid" tier (e.g., Lampung, Banten) shows potential to transition into higher production levels, whereas the "Low" tier requires fundamental capacity-building interventions.

TABLE VII. COMPREHENSIVE PROFILE OF BOTH CLUSTERS

Metric	Cluster 0 (Centre)	Cluster 1 (Consolidation)	Ratio (0:1)
Number of Provinces	3 (7.89%)	35 (92.11%)	0.09:1
Total Production (Tons)	265,424.70	197,995.18	1.34:1
National Contribution	57.3%	42.7%	1.34:1
Mean (Tons)	88,474.90	7,026.14	12.59:1
Median (Tons)	85,241.70	3,848.72	22.15:1
Std Dev (Tons)	6,979.82	5,408.77	1.29:1
CV (%)	7.89%	76.98%	0.10:1
Min (Tons)	83,275.69	145.81	571.04:1
Max (Tons)	96,907.31	19,259.70	5.03:1

TABLE VIII. SUB-SEGMENTATION OF CLUSTER 1 (CONSOLIDATION)

Sub-Segment	Number of Provinces	Production Range (Tons)	Average (Tons)
Upper-Mid	5	13,000 - 19,260	16,049
Mid	10	6,000 - 13,000	8,827
Lower-Mid	12	2,000 - 6,000	3,982
Low	8	145 - 2,000	993

E. *Stability Validation*

Finally, the reliability of these findings was stress-tested via bootstrap resampling. As shown in Table VIII, the results from 100 iterations demonstrated a mean Silhouette Score of 0.9142 with a negligible standard deviation of 0.0089. Furthermore, the membership consistency reached 99.7%, meaning that in almost every resampling scenario, the algorithms consistently assigned the same provinces to the same clusters. This level of stability is exceptionally high and confirms that the identified dualistic structure is a robust economic reality, resilient to sampling variations or minor data fluctuations.

TABLE IX. BOOTSTRAP RESAMPLING ANALYSIS RESULTS (100 ITERATIONS)

Metric	K-Means	Hierarchical	Interpretation
Mean Silhouette Score	0.9142	0.9148	High and consistent
Std Dev SS	0.0089	0.0076	Stable, minimal variation
Min SS	0.8973	0.9012	Remains in a strong category
Max SS	0.9278	0.9301	Not excessive, realistic

IV. DISCUSSION

A. *Theoretical Interpretation of Convergence*

The remarkable consistency among the three algorithmically distinct methods (99.7–100% agreement) reveals important theoretical insights into the data structure. In modern clustering theory, high ensemble agreement is recognized as the strongest indicator of natural structure, ensuring that results are not merely artifacts of algorithmic bias [23], [24]. The Silhouette Score of 0.9155 for k=2 substantially exceeds the 0.7 threshold categorized as "strong structure" in recent literature [25], approaching the theoretical maximum, which indicates nearly perfect linear separation.

This convergence can be mathematically explained by the extreme separation in the standardized feature space shown in Figure 3. The three super-producer provinces occupy Z-score positions > 3.0, creating a gap of approximately 2 standard deviations from the rest. In multivariate statistics, observations beyond 3 standard deviations represent the tail (<0.3%), effectively constituting a distinct population. This mathematical separation explains why algorithms with different optimisation criteria (variance minimization vs. connectivity vs. density) converged on identical solutions. The fact that DBSCAN independently identified the same three provinces as outliers provides non-circular validation that these observations are fundamentally different in density structure [18].

Furthermore, the stability evidence provided by the bootstrap results (see Table VIII) exceeds typical standards. While other ensemble studies typically report 85–90% agreement for outlier detection [23], our triangulation achieved near-perfect consistency. This confirms that the production dichotomy in Indonesia is a robust economic reality, not a statistical coincidence.

B. Comparison with Previous Research

Comparing our findings with previous research reveals critical methodological implications that extend beyond simple structural differences. Ningsih [14]identified k=3 as the optimal cluster number using K-Means on raw production data (2017–2022). In contrast, our triangulation approach consistently identified k=2. This discrepancy is not merely a difference in results but highlights the critical role of data preprocessing. Our analysis suggests that the third cluster identified in Ningsih's study likely emerged as an artifact of scale variance rather than a distinct production regime. Without Z-score standardization, the Euclidean distance function is disproportionately influenced by variables with large variances [19], potentially fragmenting naturally cohesive clusters. By standardizing the data, our study successfully mitigated this bias, revealing a more fundamental dualistic structure (Production Centres vs. Consolidation Group).

Our results align more closely with Indah [15], who utilized hierarchical methods and found a similar separation between major and minor producers. However, our study advances beyond Indah's findings by integrating DBSCAN for explicit outlier detection. While Indah's hierarchical approach effectively captured the global structure, it lacked a mechanism to distinguish between "extreme values within a cluster" and "true structural outliers." Our application of DBSCAN filled this gap by explicitly labeling the three super-producer provinces as "Noise," thereby providing a stronger, non-circular validation that these provinces constitute a structurally distinct entity [18].

Furthermore, compared to Ais et al. [16], who employed Fuzzy C-Means to analyze livestock meat production, our crisp clustering approach (K-Means and DBSCAN) offers a more definitive categorization necessary for clear policy formulation. While fuzzy clustering provides valuable insights into transitional memberships, policy interventions often require clear-cut segmentation to allocate resources effectively. The near-perfect stability of our results (99.7% consistency via bootstrapping) suggests that the ambiguity typically handled by fuzzy methods is minimal in this specific dataset, making our hard clustering approach both methodologically robust and practically actionable.

V. Conclusion

This study demonstrates that methodological triangulation across partitional, hierarchical, and density-based paradigms provides validation that is significantly more robust than single-algorithm approaches, particularly for data with extreme disparities like Indonesian beef production (CV=171.89%; ratio 664:1). All three algorithms consistently converged on the k=2 solution with 99.7–100% agreement and a Silhouette Score of 0.9155, statistically confirming that the dominance of the three Java provinces is a natural structure rather than an algorithmic artifact. The primary contribution of this study lies in demonstrating that integrating density-based outlier detection and bootstrap stability testing can mitigate the distortions often present in conventional methods. This provides a solid empirical foundation for policymakers to implement differentiated strategies between major production centers and developing regions, while recommending the adoption of this triangulation framework for future complex agricultural datasets.

References

[1] F. Andriansyah, Asifa, L. Isnaini, and Sahara, "Price Variability and Beef Market Integration in Java," *AGRISOCIONOMICS J. Socioecon. Agric. Policy*, vol. 9, no. November, pp. 612–630, 2025.

[2] A. B. Heatubun and M. J. Matatula, "PRODUKSI DAGING SAPI DI INDONESIA DAN SKENARIO PENINGKATAN : SEBUAH ANALISIS DAMPAK UNTUK PENGAMBILAN KEBIJAKAN," *Agrinimal J. Ilmu Ternak dan Tanam.*, vol. 11, no. 2, pp. 92–100, 2023.

[3] L. O. Nazaruddin and Q. Aulia, "The lacked access to beef in the low-income region : An evidence from the eastern part of Indonesia," *Open Agric.*, vol. 10, no. 1, pp. 1–16, 2025, doi: https://doi.org/10.1515/opag-2025-0411.

[4] A. K. Wardana, T. Rizal, and D. Adi, "Efisiensi saluran pemasaran sapi potong peternakan rakyat di kecamatan soko kabupaten tuban," *JIA (Jurnal Ilm. Agribisnis) J. Agribisnis dan Ilmu Sos. Ekon. Pertan.*, vol. 8, no. 2, pp. 123–134, 2023.

[5] I. Trisman, A. Firman, and L. Herlina, "Penentuan Wilayah Pengembangan Ternak Ruminansia Besar Di Provinsi Jawa Timur Determination of Superior Areas of Big Ruminant Development in East Java," *J. Ilmu Ternak Univ. Padjadjaran*, vol. 22, no. 2, pp. 125–136, 2022, doi: 10.24198/jit.v22i2.41717.

[6] N. Widyas *et al.*, "Promoting Sustainable Utilization and Genetic Improvement of Indonesian Local Beef Cattle Breeds : A Review," *Agriculture*, vol. 12, no. 10, pp. 3–25, 2022.

[7] N. Hilmiati *et al.*, "Smallholder Cattle Development in Indonesia : Learning from the Past for an Outcome-Oriented Development Model," *Int. J. Des. Nat. Ecodynamics*, vol. 19, no. 1, pp. 169–184, 2024.

[8] G. Gan, "A k -Means Algorithm with Automatic Outlier Detection," *Electronics*, vol. 14, no. 9, pp. 1–17, 2025.

[9] F. Khalish, N. M. Piranti, and O. Martadireja, "Implementasi Data Mining Menggunakan Teknik Clustering dengan Metode K-Means," *JIIP (Jurnal Ilm. Ilmu Pendidikan)*, vol. 8, no. 5, pp. 5392–5397, 2025.

[10] N. Nugroho and F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means ++ Sebagai Clustering Data Covid-19 di Pulau Jawa," *TEKNIKA*, vol. 11, no. 3, pp. 170–179, 2022, doi: 10.34148/teknika.v11i3.502.

[11] I. K. Khan *et al.*, "Addressing limitations of the K-means clustering algorithm : outliers , non-spherical data , and optimal cluster selection," *AIMS Math.*, vol. 9, no. 9, pp. 25070–25097, 2024, doi: 10.3934/math.20241222.

[12] C. X. Gao *et al.*, "An overview of clustering methods with guidelines for application in mental health research," *Psychiatry Res.*, vol. 327, pp. 1–28, 2023, doi: 10.1016/j.psychres.2023.115265.

[13] N. H. M. M. Shrifan, M. F. Akbar, N. Ashidi, and M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6365–6376, 2022, doi: 10.1016/j.jksuci.2021.07.003.

[14] D. R. Ningsih, "Pengelompokan Produksi Daging Sapi Menurut Provinsi di Indonesia Tahun 2017-2022 dengan Menggunakan Metode K-Means," *ESTIMASI J. Stat. Its Appl.*, vol. 5, no. 1, pp. 113–125, 2024.

[15] Y. M. Indah, "Clustering Analysis for Regional Variations in Beef Production : A Comparative Study of Hierarchical and Non-Hierarchical Article Info :," *Statistika*, vol. 25,

no. 1, pp. 35–49, 2025.

[16] C. Ais, A. Hamid, D. Candra, and R. Novitasari, "Analysis of Livestock Meat Production in Indonesia Using Fuzzy C-Means Clustering," *J. Ilmu Komput. dan Inf.*, vol. 15, no. 1, pp. 1–8, 2022.

[17] B. V Sánchez, V. Erich, and S. Arthur, *A comparative evaluation of clustering - based outlier detection*, vol. 39, no. 13. Springer US, 2025. doi: 10.1007/s10618-024-01086-z.

[18] E. S. Dalmaijer, C. L. Nord, and D. E. Astle, "Statistical power for cluster analysis," *BMC Bioinformatics*, vol. 23, no. 205, pp. 1–28, 2022, doi: 10.1186/s12859-022-04675-1.

[19] T. Li, Y. Ma, and T. Endoh, "Normalization-Based Validity Index of Adaptive K-Means Clustering for Multi-Solution Application," *IEEE Access*, vol. 8, pp. 9403–9419, 2020.

[20] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.

[21] M. R. Zuhdi, H. S. Al Jauhar, A. A. R. Fernandes, and N. W. S. Wardani, "COMPARISON OF DBSCAN AND K-MEANS CLUSTER ANALYSIS WITH PATH-ANOVA IN CLUSTERING WASTE MANAGEMENT BEHAVIOUR PATTERNS," *J. Tek. Inform.*, vol. 6, no. 1, pp. 105–112, 2025.

[22] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "Intelligent mapping of geochemical anomalies: Adaptation of DBSCAN and mean-shift clustering approaches," *J. Geochemical Explor.*, vol. 258, p. 107393, 2024, doi: https://doi.org/10.1016/j.gexplo.2024.107393.

[23] B. Ray, S. Ghosh, S. Ahmed, R. Sarkar, and M. Nasipuri, "Outlier detection using an ensemble of clustering algorithms," *Multimed. Tools Appl.*, vol. 81, no. 2, pp. 2681–2709, 2022, doi: 10.1007/s11042-021-11671-9.

[24] A. Saha, A. Chatterjee, S. Ghosh, N. Kumar, and R. Sarkar, "An ensemble approach to outlier detection using some conventional clustering algorithms," *Multimed. Tools Appl.*, vol. 80, no. 28, pp. 35145–35169, 2021, doi: 10.1007/s11042-020-09628-5.

[25] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.