Robust Cluster Approach in Handling Extreme Data on Segmentation of Lecturer Publication Productivity Based on SINTA Data

Feby Wulandari Sembiring, Eko Hariyanto

Abstract

Lecturer publication data on the SINTA portal are typically right-skewed and contain extreme values (outliers) that can distort clustering results, particularly in centroid-based algorithms such as K-Means. This study aims to segment lecturer publication productivity while handling outliers using a robust clustering approach. The data consist of 99 lecturers from Royal University, retrieved from the SINTA portal, with annual publication counts from 2020 to 2025 as features. The research stages include data cleaning, Z-Score normalization, exploratory analysis using box plots, outlier handling with the IQR method, determination of the number of clusters using the elbow method and silhouette score, K-Means segmentation, and evaluation using the Davies-Bouldin Index (DBI), silhouette score, and Calinski-Harabasz Index. The results show that the optimal number of clusters is three (k = 3). Outlier handling improved cluster quality, with DBI decreasing from 1.0837 to 0.6636 and the silhouette score increasing from 0.3877 to 0.4032. The Calinski-Harabasz Index decreased slightly from 66.7273 to 61.5589, which is attributed to reduced between-cluster variance after extreme values were removed. Cluster distribution by study program shows that computing-related programs dominate one cluster, while non-computing programs concentrate in another. These findings can support institutional policy, such as targeted publication training for low-productivity groups and mentoring roles for high-productivity groups.

***Keywords**: k-means, lecturer publication productivity, outliers, robust clustering, SINTA.*

Feby Wulandari Sembiring
Master of Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: febywulandari421@gmail.com

Eko Hariyanto,
Eko.hariyanto@dosen.pancabudi.ac.id .
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

The development of the Three Dharma of higher education is a mandatory role that must be fulfilled by educators or lecturers as a means of contributing to both the advancement of knowledge and society. [1]. Not only implementing the tri dharma (research, service and teaching) but lecturers play a fundamental role in spreading the tri dharma as a form of dissemination of research results and service to the community. In Indonesia, the main medium for disseminating lecturers' work is through publication in scientific journals [2].

The dissemination of information through such publications is governed by regulations. Kementrian Pendidikan, Kebudayaan, Riset dan Teknologi (Kemendikbudristek) as stated in Kepmendikbudristek No. 500/2024 and Permendikbudristek No 44/2024 The regulations state that publication serves as a performance indicator for faculty members, a matter of academic integrity, and a requirement under the faculty workload (BKD) framework for producing scholarly outputs; consequently, publication is inseparable from the tri dharma) of higher education teaching, research and community service that faculty members are expected to uphold. [3].

To monitor the performance of lecturers and to map and document their publications in Indonesia, a platform known as the SINTA portal serves as a primary assessment metric. The SINTA (Science and Technology Index) portal is an information system managed by the Ministry of Higher Education; it tracks the official digital records used to measure the performance, productivity, and academic reputation of lecturers, drawing from various sources including research, community service, books, intellectual property rights (IPR), Scopus, Web of Science (WoS), Garuda, and Google Scholar. [4]. Through the SINTA portal, the rankings and scores of both lecturers and higher education institutions can be viewed with transparency.

Publication data recorded in SINTA is generally used for evaluating faculty performance, ranking institutions, and formulating policies for the development of research human resources. One technique frequently employed to map such performance is cluster analysis grouping entities based on shared characteristics, such as the number of indexed articles, citation counts, h-index, and SINTA scores. The results can serve as a foundation for more targeted training initiatives, such as providing support to groups with lower productivity or fostering collaboration among highly productive groups.

However, when measuring productivity through publications, faculty publication data typically exhibits a right-skewed distribution. In terms of citations, most documents receive few or no citations, while a small number of documents garner a very high number of citations. Because nearly all bibliometric distributions are asymmetric, the use of an arithmetic mean can distort the representation of typical achievements. [5]. Therefore, it is suggested that the performance assessment of research units employ metrics more robust against extreme values (outliers) such as an adjusted mean to account for exceptionally high or low figures.

In the SINTA dataset, extreme high values may arise due to collaborations on multi-authored articles, specialization in fields with a culture of high citation rates, incomplete profiles, unsynchronized publication links, or recording errors. These extreme values are known as outliers. An outlier is a data point that differs significantly from the mean of the existing dataset. Outliers are generally classified as anomalies because their values and distribution deviate from the rest of the data. In research [6]. The outlier is a data point that exhibits characteristics significantly different from the rest of the data, arising from various factors such as measurement errors, rare events, or extreme (maximum or minimum) values, which can affect a model's performance.

Extreme values (outliers) are significant because they do not necessarily indicate errors. While some outliers represent noise that should be removed, others reflect exceptional performance precisely the kind of data that warrants the primary attention of policymakers. Eliminating all outliers risks the loss of critical information and can compromise both data

accuracy and observed patterns; conversely, leaving them unaddressed may distort the results of an analysis. [7].

In clustering, the influence of outliers is very significant. Centroid-based algorithms, such as K-Means, calculate cluster centers by taking the mean; consequently, the centroids can easily shift toward areas of data imbalance. [7], the boundaries between clusters become biased, and the majority cluster becomes less representative [8]. To address this, robust clustering has been developed a method that ensures clustering results remain stable even in the presence of outliers. One such approach is "robust clustering," which utilizes actual data objects as cluster centers (medoids). By minimizing the sum of pairwise dissimilarities rather than the sum of squared Euclidean distances, robust clustering methods such as K-Medoids are more resilient to noise and outliers compared to K-Means. [9].

Previous research by Dahnial et al., titled "Implementation of K-Means Clustering Method to Lecturers based on Publication of National Journals and Accredited SINTA," applied the K-Means algorithm to cluster lecturers based on their publications in national journals and SINTA-accredited journals. The study categorized lecturers according to the weightings assigned to these national and SINTA-accredited publications. The findings indicate that clustering via K-Means is effective, as it yields more optimal groupings. Although extreme values were identified, the approach taken was to remove them. [10]. Outliers have become a concern in clustering techniques, such as in research that detects outliers in data clusters, where the research states that outliers need to be detected in data sets because outliers greatly influence the sensitivity of the model. [11]. Furthermore, in a study by Lestari et al. titled “Assessment Clusterization of Teacher Performance with K-Means Algorithm and Agglomerative Hierarchical Clustering,” the researchers successfully categorized teacher performance and evaluated teacher effectiveness using clustering algorithms. However, the study has a limitation: no data cleaning was performed on the 1,944 teacher records. Consequently, the presence of outliers forced data points into the established clusters, resulting in groupings that do not accurately represent the actual data [12].

The aforementioned research indicates that clustering algorithms even those with high accuracy may fail to accurately represent the true structure of the data if outliers are not properly addressed. The K-means algorithm, in particular, is highly sensitive to extreme values; because it relies on calculating means, the centroid can be pulled toward extreme data points, thereby distorting the overall representation of the dataset. In the context of evaluating lecturer performance via the SINTA platform, data anomalies are common, necessitating the use of a robust clustering algorithm. A review of existing literature reveals several research gaps:

First, clustering studies involving SINTA data typically focus on applying K-means without treating outliers as objects of analysis. In cases where outliers are considered, they are often simply removed a process that inadvertently excludes high-performing individuals from the mapping.

Second, comparisons between K-means and K-medoids generally utilize non-bibliometric data (such as transaction records, student data, budgetary figures, or regional indicators), and the results have been inconsistent; consequently, there is no specific evidence regarding their application to lecturer publication data.

Third, there is a lack of clear distinction between "informative" outliers and those resulting from data errors. In the SINTA dataset, an extremely high value might signify genuine, exceptional achievement, whereas an extremely low value could indicate a profile that has not yet been synchronized.

The novelty of this research lies in several areas. It introduces robustness to the clustering of SINTA-based lecturer publication data by comparing K-means (as a baseline) with K-medoids (as a more robust method). Furthermore, this study treats outliers as an integral part of the analysis distinguishing between exceptional performance and potential data errors rather than simply discarding them. Additionally, because the cluster centers in K-

medoids are actual lecturers (data points) rather than calculated averages, the resulting cluster profiles are easier to interpret and can serve as more practical benchmarks for professional development. Based on the aforementioned background, this study aims to analyze the handling of outliers in faculty publication data sourced from SINTA using a robust clustering approach. By comparing the k-means and k-medoids algorithms, the research seeks to determine which model yields more robust results for SINTA data. This study is expected to produce more accurate and representative groupings of faculty members, provide insight into the characteristics of outliers within publication data, and serve as a reference for institutions in designing more targeted strategies for enhancing research performance.

Literature Review

To elaborate on the core subject matter of this research, the following literature review provides definitions for various points that will be discussed throughout the study:

1. **Data Extrem (*Data Outlier*)**
   An outlier is a data point with characteristics that differ significantly from the rest of the dataset. There are two types of outliers: univariate and multivariate. A univariate outlier occurs within a single variable and is typically assessed using Z-scores or the interquartile range. A multivariate outlier arises from a combination of values across two or more variables and is generally detected using Mahalanobis distance. Since the classical Mahalanobis distance is sensitive to outliers which can impair its performance a robust version is used instead.
2. **SINTA (Science and Technology Index)**
   Namely a portal integrated with Scopus and Google Scholar, designed to provide insight into the productivity profiles of lecturers. The primary functions of SINTA include indexing and citation tracking, determining the accreditation status of scientific journals, mapping research strengths, and facilitating collaboration through a directory of experts comprising authors affiliated with higher education institutions and research organizations.
3. **Segmentasi**
   Based on PUEBI (Pedoman Umum Ejaan Bahasa Indonesia), The term "segmentation" refers to the division or grouping of a structure or unit of information into homogeneous parts. According from Tjiptono & Chandra in suriansah (2025) Segmentation is the process of grouping heterogeneous data into groups/segments that have similarities in terms of objectives, characteristics and responses [13]. Thus, segmentation in machine learning can be defined as a technique for grouping data with heterogeneous characteristics into sets of homogeneous data to achieve a specific objective. The benefits of segmentation in machine learning include::
   a. Efficiency time and costs, helping to focus efforts on the right group without wasting resources on the wrong targets.
   b. Personalization, provide recommended actions that are relevant to the characteristics of each group
   c. Discovering hidden patterns: in machine learning, segmentation can uncover hidden patterns that are difficult to detect manually through conventional approaches.
4. **Publication**
   Publication is the process of disseminating information, knowledge, ideas, or findings to the general public through publishing media. According to Becher & Trow, as cited in research by Abdul Aziz (2025) states that publication is a means to document, communicate, and uphold scholarship within a community [14]. In the academic sphere, publications serve as a symbol of a faculty member’s performance and scholarly competence in fulfilling the *Tri Dharma* (Three Pillars) of Higher Education. Within this context, publications take the form of scholarly works that have

undergone a rigorous review process by a board of reviewers and an editorial team. Furthermore, the concept of publication in higher education constitutes an academic obligation, as stipulated in the Regulation of the Minister of Research, Technology, and Higher Education (Permenristekdikti) No. 20 of 2017 concerning the enhancement of scholarly publications.

5. **Publication Productivity Lecture**

This refers to the level of success or the volume of scholarly work produced by a faculty member such as articles in national or international journals, conference proceedings, books, or intellectual property rights (IPR) within a specific period. Publications serve as one of the tangible outputs of a faculty member's academic expertise within the framework of the Tri Dharma (the three pillars of higher education). A faculty member's productivity is influenced by several factors, namely: [15]:

a. Academic skills and literacy
b. Time management and motivation
c. Institutional support
d. collaboration
e. Used of technology

The urgency for faculty members to publish thereby fostering productivity stems from several factors, including institutional requirements, advancement in academic rank, faculty workload expectations, and the enhancement of the university's reputation, which is bolstered when the volume and quantity of its faculty's publications increase.[16].

Research Methodology

In conducting an analysis to determine the segmentation of faculty publication productivity through clustering, this study involves several stages, including:

1. **Literature Study:** Identifying reference sources and examining gaps in the area of faculty productivity a step that is essential for highlighting the research objectives..
2. **Data Collection:** This stage is essential for identifying data patterns and the nature of issues within real-world data, ensuring that the analysis is not based on assumptions [17]. The data was obtained from the Royal University SINTA Portal to facilitate accurate analysis. The data collected for this study consists of faculty publications from the last five years; a summary of this data is presented below.

Table 1. Dataset Segmentation Lecturer Productivity

Lecture Code	Study Program	Publication 2020	Publication 2021	Publication 2022	Publication 2023	Publication 2024	Publication 2025
DR001	Information system	10	10	18	12	11	28
DR002	Information system	3	5	5	4	4	1
…	…	…	…	…	…		
DR098	Biology	0	0	0	1	2	1
DR100	Law	0	0	0	1	1	8

3. **Data Prepocessing:** Data cleaning involves removing duplicate values, handling missing values, and addressing inconsistent formatting; this stage also identifies outliers, ensuring the data is more stable for segmentation by the model. Following the data cleaning process, the data is normalized to standardize its scale in this case, the Z-Score technique is employed. The Z-Score technique is used to transform various disparate scales into a single, standardized scale. The Z-Score normalization formula is as follows:

$$z = \frac{x - \mu}{\sigma}$$

Keterangan:
z = Z-score values (result of normalization)
x = Original data value to be calculated
μ = Mean value of the entire dataset
σ = Standart deviation of the entire dataset.

4. **Feature Selection:** Feature selection is used to identify variables that influence the model analysis for productivity value segmentation, based on SINTA data. Furthermore, it aims to determine the correlations between these variables and the research objectives.
5. **Exploratory Analysis:** Analysis to determine the distribution and spread of data can be performed using visualizations such as box plots or histograms; these visualizations make it easy to clearly detect data patterns, including extreme values or outliers, within a dataset.
6. **Segmentation / *Clustering***: Determining the segmentation of faculty productivity groups requires an accurate clustering algorithm. This study analyzes the k-means algorithm identifying the limitations of standard clustering as well as the robust clustering algorithm, K-Medoids, for data grouping. Robust clustering is essential for handling outliers, ensuring that the resulting clusters are significantly more accurate and precise than those produced by standard clustering algorithms..
   a. K-Means
      Namely A clustering algorithm that partitions data into (k) groups based on the similarity of characteristics between data points or their proximity to one another [18]. K-Means uses averages to determine the center of each cluster (centroid); typically, these centroid values are new decimal numbers that may not necessarily exist within the original dataset . The steps for the K-Means algorithm are as follows:
      1. Inisialization: Determine the number of k required based on needs.
      2. Initial Centroid Determination: Randomly selecting k points to serve as the initial centroids.
      3. Cluster Assignment: For each data object, the distance to the centroid is calculated, and the object is assigned to the cluster with the nearest centroid. This step employs Euclidean distance, as shown in the formula below [19]:

$$d(x_i, m_j = \sqrt{\sum_{l=1}^{p} (x_{il}-m_{jl})^2}$$

         Description:
         x_{il} = the lth attribute of the i-th object
         m_{jl} = the l-th attribute of the medoid of cluster j.
      4. Centroid Update: Recalculate the new centroid position for each cluster by finding the average of all data points contained within that cluster..

$$C_j = \frac{1}{|C_j|} \sum_{x_i \in C_j}$$

      5. Iteration : The steps are repeated until a convergent condition is reached, indicating that the data membership in each cluster has become stable.K-Medoids
   b. Robust Clustering

Robust clustering is a clustering method that offers advantages in handling outliers and noise within a dataset. In general, the robustness of this method can be achieved through two primary approaches:

1. Robust scaler pre-processing: used prior to the clustering process, this technique leverages the median and interquartile range (IQR) to mitigate the impact of outliers.
2. Mahalanobis Distance Metric: The use of this distance matrix enables calculations between objects that account for the variance and correlation between variables, thereby making cluster analysis more accurate.

7. **Value Mapping**
   Following a series of grouping stages, the resulting data distribution and value mappings will be visualized using a scatter plot to reveal data patterns and clusters.
8. **Evaluation of Clustering Algorithms:** Following the segmentation process, the model's accuracy is assessed using the Davies-Bouldin Index (DBI). The segmentation results obtained before and after applying robust clustering are then compared. This process identifies the algorithm that performs best for segmenting the publication productivity of faculty members. In terms of interpreting the Davies-Bouldin Index, a lower value (closer to 0) indicates higher quality in the clusters generated by the model. The Davies-Bouldin Index is calculated as follows [20]**:**

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i + s_j}{d(c_i, c_j)}\right)$$

Description:
k = Number of cluster
s_i dan s_j = The average distance distribution of all data points within cluster i and I relative to their respective cluster centers
$d(c_i, c_j$ = The distance between cluster center (centroid/medoid) i and cluster center j
max = DBI takes the worst-case ratio for each cluster i when compared to other clusters.
Next, the model's performance in segmenting data into distinct groups is evaluated using the silhouette score. To determine the best clustering model, one examines the silhouette score; a value closer to 1 indicates an excellent clustering structure.

$$S(i) = \frac{b(i)\text{-}a(i)}{\max (a(i), b(i))}$$

Results

The results of this study are presented according to the stages outlined in the preceding methodology section, ensuring a clear understanding of the outcomes from each phase of the segmentation of faculty publication productivity. The findings of this study include:

4.1 Data Normalization

When performing data segmentation using clustering algorithms, the initial step is to standardize the data scale to ensure stability for algorithmic processing. In this study, the data preprocessing stage involves normalization using the Z-score method. The features to be normalized include publication data from 2020 to 2025. These values are normalized using the Z-score standard scaler.

Table 2. Normalization Feature by Z-Score

Publication 2020	Publication 2021	Publication 2022	Publication 2023	Publication 2024	Publication 2025
4.0831	2.2002	2.5361	2.1110	4.0114	7.0845
0.5149	0.4984	-0.0918	-0.0946	0.7260	0.1015

…	…	…	…		
0.5149	1.8599	3.5458	1.2838	3.0750	-0.4570
0.005	0.8388	0.7167	1.2838	1.1958	0.3808

4.2 Exploratory Analysis Using Box Plot

To assess the state of the data specifically regarding the presence of outliers a box plot visualization analysis is performed. In this study, box plot visualizations are categorized based on the use of a standard clustering algorithm (K-Means) before handling outlier and k-means after handling outlier

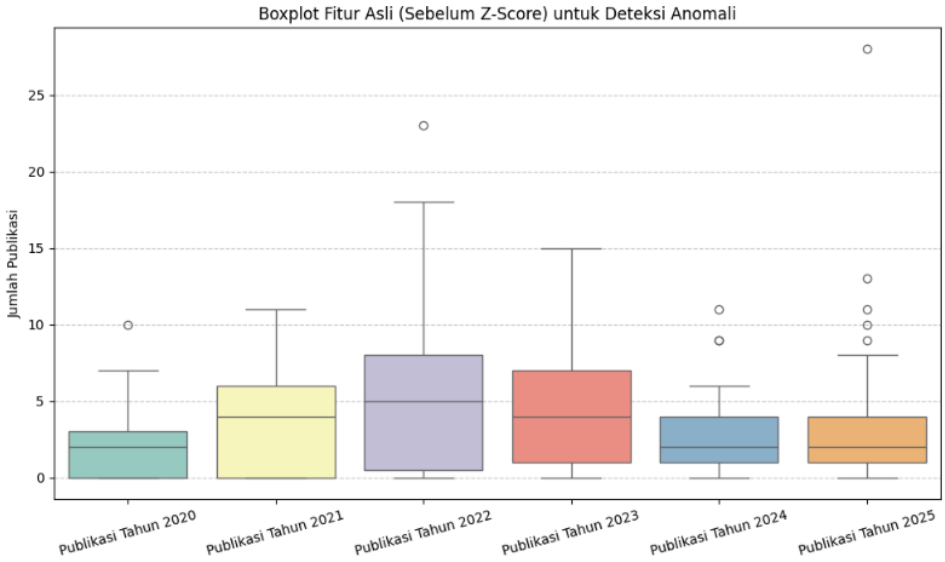


Figure 1. Box Plot Data before handling outlier

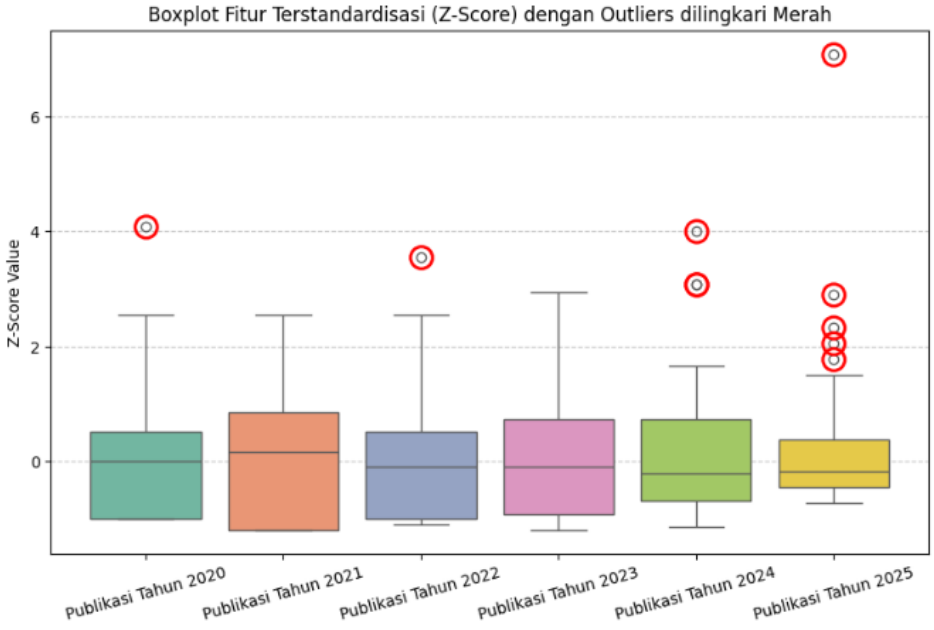


Figure 2. Box Plot after handling outlier

In Figure 1, the boxplot of the original data reveals that the number of publications for each year (2020–2025) exhibits a non-uniform distribution, with a right-skewed tendency. Most values are concentrated within a lower range; however, there are several extreme values such as approximately 23 publications in 2022 and 28 in 2025 that lie well above the majority of the data. Furthermore, the significant variation in ranges across different years makes visual comparison disproportionate, as years with a wider spread compress the representation

of other years. This situation indicates the presence of potential outliers that could influence analysis results if left unaddressed.

Following standardization via Z-score (Figure 2), all features are brought to a common scale, with medians clustered around zero, allowing for a more equitable comparison across years. Identified outliers are marked with red circles; the highest concentration appears in the 2025 data, where Z-scores approach 7 consistent with the extreme values observed in the original dataset. Nevertheless, standardization merely rescales the data without removing extreme values; consequently, these outliers remain present and require further handling for instance, through techniques such as capping or winsorizing, or by retaining them if they can be justified as valid data points.

4.3 Determining k Using the Elbow Method and Silhouette Score

To determine the optimal number of clusters (k) based on the data, the elbow method and silhouette score were employed to ensure that the value of k is more representative of the dataset. Through these techniques, it was determined that the optimal number of clusters for categorizing faculty publication productivity based on SINTA data is 3 (k = 3).

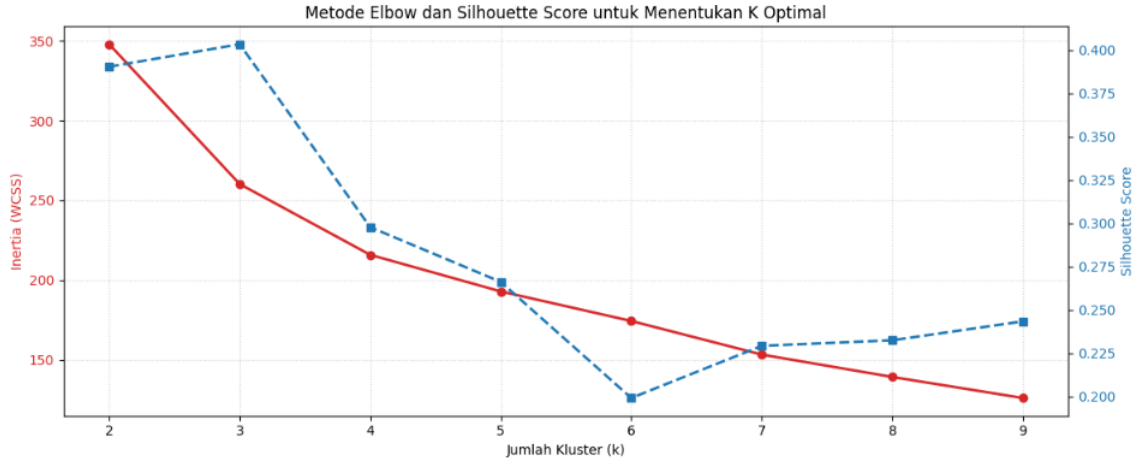


Figure 3. Elbow Method and Silhoutte score for decision k

4.4 Data Segmentation

a. K-Means Algorithm

The following are the results of the publication data segmentation using the K-Means clustering algorithm; these results are presented through a scatter plot visualization, as shown in Figure 4, and a spatial map illustrating the distribution of faculty productivity by study program, as shown in Figure 5.

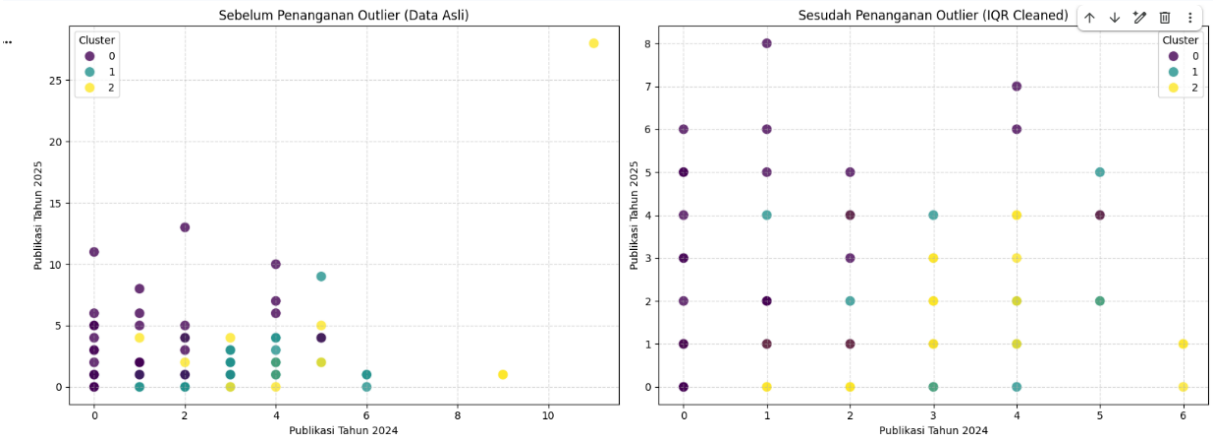


Figure 4. K-Means Visualization Using Scatter Plot

In the panel representing the data before outlier handling (the original dataset), the distribution of publications for 2024 and 2025 shows that most data points are concentrated within a low range approximately 0–6 publications for both years. However, several points lie far from the main cluster, with extreme values appearing at coordinates such as (11, 28), as well as other points like (2, 13), (0, 11), (4, 10), (5, 9), and (9, 1). The presence of these points expands the range of the axes, making it difficult to observe the cluster structure in the denser areas. Furthermore, these extreme values can skew the position of centroids in distance-based algorithms like K-Means, potentially degrading the quality of the resulting clusters.

After handling outliers using the IQR method, the data range narrows to approximately 0–6 for 2024 publications and 0–8 for 2025 publications, making the patterns between clusters more distinct. Cluster 0 tends to be associated with lower publication counts in 2024 and moderate-to-high counts in 2025; Cluster 2 tends to correspond to higher publication counts in 2024 and low-to-moderate counts in 2025; while Cluster 1 occupies the space between them. Nevertheless, some overlap between clusters remains which is expected, given that the visualization displays only two of the several features used in the clustering process. Consequently, the evaluation of cluster quality should be supplemented with quantitative metrics, such as the silhouette score or the Davies-Bouldin index.

4.5 K-Means Sectoral Spatial Analysis (by Study Program)

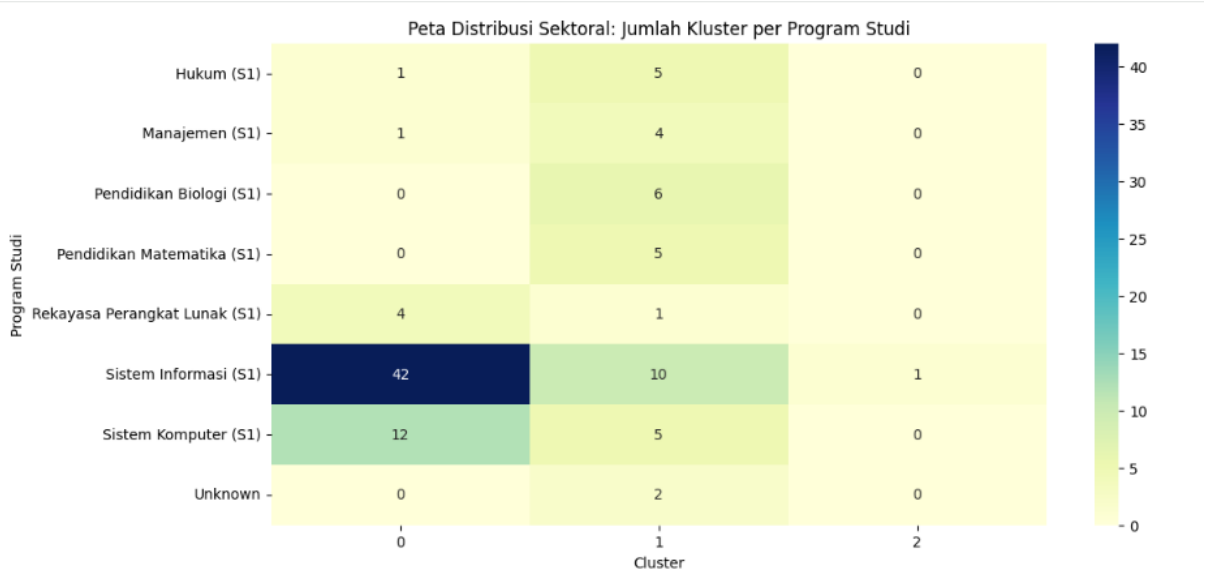


Figure 5. K-Means Sectoral Spatial Analysis for Productivity Segmentation

The distribution of clusters across study programs on the heatmap shows a fairly clear pattern. Out of 99 records in total, Cluster 0 contains 60, Cluster 1 contains 38, and Cluster 2 contains only 1. Information Systems is the largest study program (53 records), and most of its members (42 records, about 79%) fall into Cluster 0. Computer Systems shows a similar tendency, with 12 of 17 records (about 71%) in Cluster 0 and the remainder in Cluster 1. Software Engineering also leans toward Cluster 0 (4 of 5 records), although its sample is small. Overall, these three computing-related programs account for roughly 90% of the members of Cluster 0, indicating that this cluster is dominated by information technology-based study programs.

In contrast, study programs outside the computing field tend to be concentrated in Cluster 1. Biology Education (6 records) and Mathematics Education (5 records) are entirely in Cluster 1, with no members in any other cluster. Law (5 of 6 records, about 83%) and

Management (4 of 5 records, 80%) show the same pattern, with only one member each falling into Cluster 0. The "Unknown" category, which contains 2 records, is also entirely in Cluster 1. This pattern suggests that non-computing programs have publication productivity characteristics that differ from those of the computing group, so the clustering algorithm tends to separate the two.

Nevertheless, several points deserve attention. Information Systems is not fully homogeneous, since 10 of its records still fall into Cluster 1 and 1 into Cluster 2, which indicates variation within the program. Cluster 2 contains only a single member from Information Systems, so it is better regarded as an outlier cluster than as a substantively meaningful group. It should also be noted that the number of records in most non-computing programs is very small (5-6 records), so percentages that appear large are not necessarily statistically robust. In addition, Cluster 2 in this figure contains only 1 record, whereas in the previous scatter plot Cluster 2 appeared to have quite a few points. If the two figures come from different clustering runs (for example, before and after outlier removal), this should be explained in the manuscript to avoid confusing readers. Therefore, the final interpretation should be linked to the average publication profile of each cluster and supported by further testing, such as a Chi-Square test, to determine whether the relationship between study program and cluster is statistically significant.

4.6 Evaluation of Results Using the Davies-Bouldin Index and Silhoutte Score

Based on the modeling execution results and the Davies-Bouldin Index (DBI) evaluation metric (where a lower DBI value indicates a better cluster structure with clearer separation) and Silhoutte Score:

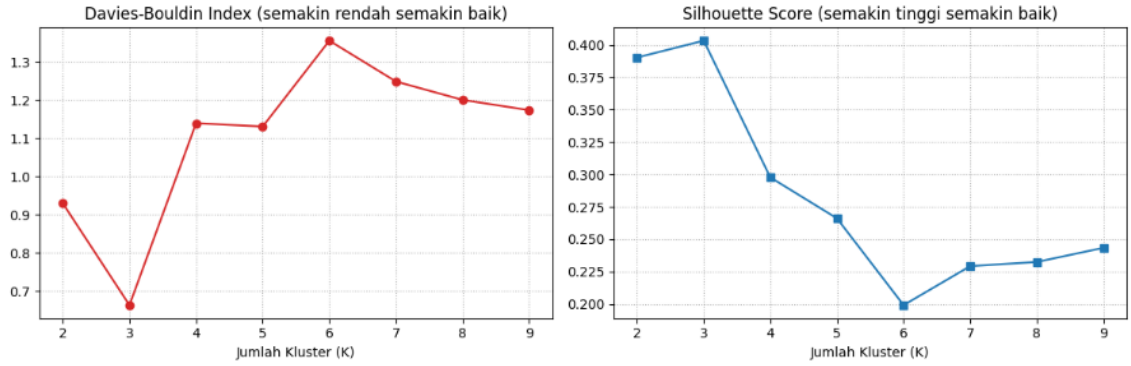


Figure 6. K-Means Sectoral Spatial Analysis for Productivity Evaluation

K	DBI Value	Silhoutte Score
2	0.929	0.390
3	0.663	0.403
4	1.139	0.297
5	1.130	0.266
6	1.355	0.199
7	1.248	0.229
8	1.200	0.232
9	1.173	0.243

Based on the analysis of the evaluation results, an optimal outcome is achieved when the segmentation is set to k=3; the Davies-Bouldin Index (DBI) value at k=3 is the best, at 0.663, and the silhouette score is approximately 0.4 a value approaching 1 thereby confirming that the decision to use k=3 clusters in this study is appropriate.

4.7 Comparison Evaluation Value of Model by Handling Outlier

Evaluation	Before Handling Outlier	After Handling Outlier
Davies Bouldin Index	1.0837	0.6636
Silhoutte Score	0.3877	0.4032
Calinski Harabaz Index	66.7273	61.5589

The evaluation results show that outlier handling improves clustering quality. The Davies-Bouldin Index decreased from 1.0837 to 0.6636 and the Silhouette Score increased from 0.3877 to 0.4032, indicating more compact and better separated clusters. The Calinski-Harabasz Index decreased from 66.7273 to 61.5589, which can be explained by the reduction in variance between clusters due to the removal of the influence of extreme values. Overall, the model after outlier handling is considered better and supports the determination of K = 3 as the appropriate number of clusters.

Conclusion

This study segmented the publication productivity of Royal University lecturers based on SINTA data into three clusters (k = 3), determined using the elbow method and silhouette score. Exploratory analysis showed right-skewed publication data with extreme values, especially in 2025, which can bias K-Means centroids. Handling outliers with the IQR method produced better clusters: DBI fell from 1.0837 to 0.6636 and the silhouette score rose from 0.3877 to 0.4032, indicating more compact and better-separated clusters. The decrease in the Calinski-Harabasz Index (66.7273 to 61.5589) reflects the reduced variance after extreme values were removed. Spatial analysis showed that computing-related study programs (Information Systems, Computer Systems, Software Engineering) dominate one cluster, while non-computing programs tend to fall in another. One cluster contained only a single lecturer and is better treated as an outlier group than as a meaningful segment.

References

[1] N. Amalia, "Akademik Dan Masyarakat Berpradaban," *Karimah Tauhid*, vol. 3, no. 4, pp. 4654–4663, 2024.

[2] T. N. Fitria, M. Muqorobin, S. L. Pardanawati, I. E. Prastiwi, M. H. Ma'ruf, and L. Kristiyanti, "Pelatihan update profil Sinta dan pembuatan ID Garuda untuk dosen Institut Teknologi Bisnis AAS Indonesia," *J. BUDIMAS*, vol. 5, no. 1, pp. 1–7, 2023, [Online]. Available: http://dx.doi.org/10.29040/budimas.v5i1.7678

[3] "Kepmendikbudristek 500/2024: Standar Indikator Kinerja Dosen - Ebizmark Blog." Accessed: Oct. 15, 2026. [Online]. Available: https://ebizmark.id/artikel/kepmendikbudristek-500-2024-standar-indikator-kinerja-dosen/?srsltid=AU7gw4UPxQ0VCb2x8l1BJ3Yxw5hThE2MO5_CKIvVEKq02TTfCJ6g6hhj

[4] G.A.Pamelia, "Sekilas tentang Science and Technology Index (SINTA) - IAES." Accessed: Oct. 15, 2026. [Online]. Available: https://iaes.or.id/sekilas-tentang-science-and-technology-index-sinta/

[5] R. Maya and M. Sarbini, "Analisis Publikasi Ilmiah Dosen STAI Al-Hidayah dalam Profil Google Scholar," *Khidmatul Ummah J. Pengabdi. Kpd. Masy.*, vol. 4, no. 01, pp. 1–16, 2023, doi: 10.30868/khidmatul.v4i01.4810.

[6] E. S. A. Simorangkir, A. P. U. Siahaan, L. Marlina, D. Nasution, and Z. Sitorus, "Deteksi Outlier Hasil Clustering Algoritma K-Medoids Menggunakan Metode Boxplot Pada Data KIP Kuliah," *J. Comput. Syst. Informatics*, vol. 5, no. 4, pp. 893–902, 2024, doi: 10.47065/josyc.v5i4.5479.

[7] N Tahta Phudjashakty, Hasbi Firmansyah, Wahyu Asriyani, and Ali Sofyan, "Segmentasi Pelanggan Grosir Menggunakan K-Means: Analisis Outlier dan Ketidakseimbangan Data," *J. Pengabdi. Masy. dan Ris. Pendidik.*, vol. 4, no. 3, pp. 15993–16002, 2026, doi: 10.31004/jerkin.v4i3.4771.

[8] K. abdi Sinuraya, Andrinata, and N. Azzuma, "Penanganan Data Outlier Menggunaan Robust Cluster Dalam Pemetaan Dataset Stunting," vol. 4307, no. August, pp. 137–140, 2025.

[9] T. Akbar, G. M. Tinungki, and Siswanto, "Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application With Noise Using Silhouette Coefficient Test," *Barekeng*, vol. 17, no. 3, pp. 1605–1616, 2023, doi: 10.30598/barekengvol17iss3pp1605-1616.

[10] D. Dahnial, "Implementation of K-Means Clustering Method to Lecturers Based on Publications of National Journals and Accredited Sinta," *JEECS (Journal Electr. Eng. Comput. Sci.*, vol. 8, no. 1, pp. 27–40, 2023, doi: 10.54732/jeecs.v8i1.4.

[11] J. Teknik, I. C. I. T. Medicom, E. Bu, R. Syahputra, E. Sabrina, and A. Simorangkir, "Outlier detection in the clustired data," vol. 16, no. 6, pp. 342–352, 2025.

[12] S. Lestari, "Assessment Clusterization Teacher Performance with K-Means Algorithm Clustering and Agglomerative Hierarchical Clustering (AHC)," vol. 9, no. 1, pp. 357–365, 2025.

[13] M. I. Maulana, Z. Ruma, and M. I. Wardhana, "Analisis Segmentation , Targeting , Dan Positioning (STP) Terhadap Pembelian Pada Restoran Mbak Daeng Abdesir," 2025.

[14] A. Aziz, "Konsep dan urgensi publikasi ilmiah serta jenis dan ruang lingkup artikel ilmiah," vol. 4, no. 5, pp. 2022–2029, 2025.

[15] D. Grahita, P. Prijana, and K. Kusnandar, "PRODUKTIVITAS PUBLIKASI ILMIAH DOSEN PERPUSTAKAAN DAN SAINS INFORMASI UNIVERSITAS PADJADJARAN : ANALISIS BIBLIOMETRIK Pendahuluan," vol. 29, no. 2, 2025, doi: 10.37108/tabuah.v29i2.2403.

[16] E. al. Jennie Fransiska, "Strategi Peningkatan Produktivitas Publikasi Ilmiah Dosen Di Iain Curup Dalam Mendukung Reputasi Akademik Perguruan Tinggi," vol. 15, no. 2, pp. 269–278, 2026.

[17] F. K. Zulham Sitorus, Eko Hariyanto, "IMPLEMENTASI MACHINE LEARNING SEBAGAI SISTEM PEMETAAN DAERAH RAWAN BANJIR DESA PAHLAWAN," pp. 1–50, 2023.

[18] D. Sartika, M. Iqbal, and Z. Sitorus, "Analisis Algoritma Exponensial Smoothing dan K-Means untuk Optimal-isasi Penerimaan Mahasiswa Baru di Universitas Haji Sumatera Utara," *J. Multimed. dan Teknol. Inf.*, vol. 7, no. 02, pp. 1–10, 2025, [Online]. Available: https://journal.cattleyadf.org/index.php/jatilima/article/view/1269

[19] S. Sutiono, R. F. Wijaya, and M. S. Novelan, "Analisis Metode K-Means Clustering Untuk Pengelompokan Jadwal Mengajar Guru Di SMKN 1 Stabat Dengan Visualisasi Data," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 8, no. 2, pp. 596–602, 2025, doi: 10.31539/intecoms.v8i2.15154.

[20] S. M. Rambe and Z. Sitorus, "Segmentation and Prediction of the Development of Small and Medium Industries (SMEs) in North Sumatra Using K-Means and Random Forest Algorithms Based on OSS RBA Data," pp. 1409–1424.