JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 1) JUNE 2024: 54 - 62
https://doi.org/10.21009/JSA.08105

MAPPING DOMESTIC AND FOREIGN TOURISTS IN EAST JAVA USING C-MEANS CLUSTERING

Marita Qori'atunnadyah [1*]

[1]*Informatics, Technology and Business Institute Widya Gama Lumajang*
4th Gatot Subroto St. No 4, Lumajang, East Java 67352, Indonesia

*Corresponding author's e-mail: * **maritaqori@gmail.com***

ABSTRACT

Article History:
Received: 14 April 2024
Revised: 19 May 2024
Accepted: 9 June 2024
Published: 30 June 2024

Keywords:
Tourism, Clustering, East Java, Covid 19

Tourism is a priority sector identified by the government for its potential to drive economic growth, job creation, community development, and regional progress. Although significant, it still requires a detailed mapping of tourist visit patterns to optimize regional tourism potential. This study uses the C-Means Clustering method to categorize districts and cities in East Java based on the number of domestic tourists and foreign tourists. Data from 2018 to 2022 is used to identify different patterns and groups. The methodology involves clustering the data based on similarities in the number of visitors, which provides insight into regional tourism dynamics. The results revealed three main groups of domestic tourists: high, medium, and low-visitation regions. For foreign tourists, five groups were identified, reflecting variations in the level of tourist visits. These groups help understand the distribution and concentration of tourists in different regions, which is important for targeted promotion strategies and efficient resource allocation. A limitation of this study is that it needs to delve more deeply into external factors affecting tourism, such as the COVID-19 pandemic. The originality of this research lies in its application of the C-Means Clustering method to map domestic and foreign tourists in East Java not simultaneously, thus providing valuable insights for policymakers and industry stakeholders to encourage collaboration and innovation in the tourism sector.

How to cite this article:
M. Qori'atunnadyah, "MAPPING DOMESTIC AND FOREIGN TOURISTS IN EAST JAVA USING C-MEANS CLUSTERING", Jurnal Statistika dan Aplikasinya, vol. 8, iss. 1, pp. 54 – 62, June 2024

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

The government has recognized tourism as one of the priority industries that is essential to fostering community development, job creation, economic growth, and regional development. The improvement of the tourism sector is a major focus as its progress is often used as an indicator of the economic stability and security of a region. Therefore, tourism sector development continues to be enhanced as part of the national development strategy [1].

In East Java Province, tourism plays an important role in the regional economy. With abundant cultural, historical, and natural resources, East Java has many tourist attractions spread across various districts and cities. To optimize this potential, mapping visitors to different tourist destinations is essential to understand visitation patterns and identify effective development strategies. In the period 2018-2022, there were significant dynamics in the number of visitors to various tourist destinations in East Java. The COVID-19 pandemic has had a significant impact on the tourism sector, as well as infrastructure and tourism promotion all influence this variation in the number of travelers. Therefore, mapping districts/cities based on the number of visitors to tourist attractions during this period using the C-Means Clustering method will provide in-depth insight into the pattern of tourist visits. This method can group areas with similar visitation characteristics to be used as a basis for more precise planning and decision-making in developing tourism potential in East Java.

Some previous research related to clustering includes grouping regions using the Hierarchical clustering Method based on road conditions [2], as well as clustering areas based on the teacher-student ratio using the K-Means Algorithm [3]. In addition, a study on regional clustering with the C-Means approach based on HDI indicators [4], and grouping manufacturing organizations using the C-Means approach based on firm value influencing elements [5]. The use of the C-Means approach unites these investigations.

Other clustering and tourism-related research include the use of K-Means on tourist visit data in Karawang Regency [6], clustering the foreign tourists with K-Means [7], and clustering tourist visits in Yogyakarta City using K-Means [8]. This research also includes mapping tourism destinations based on the carrying capacity of provincial tourism in Indonesia [9], clustering regional in Central Java Province based on tourism potential in 2020 [10], and applying K-Means Clustering to Bojonegoro Regency tourism [11].

By clustering using C-Means, it is expected to identify groups of regions that have high, medium, and low levels of tourist visits. This information is valuable for formulating more targeted promotional strategies, allocating resources more efficiently, and increasing the attractiveness of tourist destinations in East Java as a whole. This research will not only assist local governments in decision-making but also guide tourism industry players to collaborate and innovate in advancing the tourism sector in East Java.

2. METHODS

The methods consist of Subsection Materials and Data and Subsection Research Method.

Material and Data

This study employs acquired secondary data from the East Java Provincial Culture and Tourism Office Publication, namely East Java Culture and Tourism in Figures. The data used consists of 38 districts/cities which include the number of domestic tourists and foreign tourists in East Java Province Tourism Attractions in 2018-2022 [12].

Research Method

The C-Means procedure consists of starting with c groups each consisting of one random point, and then adding each new point to the group whose new point mean is closest. Once a point is added to a group, the group mean is adjusted to account for the new point. So, at each stage, C-Means is a means of the group it represents (hence the name C-Means) [13]. The C-Means algorithm is as follows [14].

(1) Choose a set of $c \geq 2$ unique places to act as the first cluster 'centroid' in the p-dimensional space under investigation. This might be a choice of one's observations or another well-separated site

recommended by experience. (Avoid beginning the cluster too close to prevent unstable cluster outcomes.)

(2) Locate the closest centroid, $\overline{z}_c$ for each observation z_i, $i = 1, \ldots, n$, and designate z_i to the cluster centered around $\overline{z}_c$ it.

(3) Calculate the new centroid $\overline{z}_c = \left[\overline{z}_{+1c} \quad \overline{z}_{+2c} \quad \cdots \quad \overline{z}_{+pc}\right]$ for the resulting cluster c.

(4) Repeat Steps (2) and (3) until the cluster assignment does not change.

(5) C-Means is an example of "greedy search" or "greedy descent" optimization; as it iterates over the expanding cluster, it looks for a better value of the cumulative sum of squares in the cluster. When it comes to computational efficiency, C-Means can significantly cut down on the amount of time needed to produce a c-cluster solution. It is also frequently the only workable choice for partitioned clustering, particularly for extremely large p (high-dimensional) or very large n ('big' data).

In the middle of the 20th century, C-Means, as a computer method, developed from a variety of sources and disciplines. Because of this, the algorithm has several variations and forms, each concentrating on a different way to manipulate the fundamental greedy descent method. Some of the earliest instances are from [15], and [16]. Hartigan and Wong (1979) provided a variation that is frequently preferred by data analysts, in which observations are blocked from transferring between two clusters if the final solution raises the cumulative sum of squares inside the cluster [17].

3. RESULTS

In this study, cluster analysis of Visitors to Tourist Attractions in East Java in 2018 - 2022 was carried out using the C-Means method. Tourist Attraction Visitors are divided into 2, namely Domestic Tourists and Foreign Tourists. East Java comprises 38 districts/cities, which will be arranged into many clusters. The most optimal cluster will be chosen when clustering is done utilizing as many as two to five clusters. Table 1 displays the outcomes of the grouping of domestic travelers.

Table 1. Clustering results for domestic tourists.

Cluster	Number of Cluster			
	2	**3**	**4**	**5**
1	6	9	3	2
2	32	26	25	26
3		3	8	1
4			2	1
5				8

Table 1 lists the number of districts/cities that make up each group in the clustering findings of 38 districts/cities in East Java Province using C-Means, with two to five clusters. The pseudo-f-statistic value with the highest value among the two to five clusters may be used to identify the optimal cluster. The pseudo-f-statistic value for each cluster is shown below.

Table 2. Pseudo F-Statistic value for domestic tourist clustering.

Number of Cluster	Pseudo F-Statistic
2	71.77864
3	**72.13407**
4	68.49494
5	55.43873

Table 2 displays how the C-Means method was used to get the pseudo-F-statistic value for two to five clusters. Based on the number of domestic visitors, three clusters are the ideal number to organize districts/cities in East Java. The greatest value of 72.13407 is the pseudo-statistic value in three clusters, indicating this.

Disparities in the features of each group are predicted when districts/cities in East Java Province are grouped using the C-Means approach based on the number of Indonesian visitors. The one-way ANOVA and one-way MANOVA techniques can be used to ascertain whether or not the features of the groups that were generated differ from one another. Utilizing Pillai's Trace test statistics and One-Way MANOVA, one may test for differences in attributes. The group created is the factor or treatment that is assumed to have an impact on the response variable in this study. The variable number of Indonesian visitors in the years 2018–2022 serves as the response variable for the One-Way MANOVA test.

Table 3. One-Way MANOVA Test Results for domestic tourist clustering.

Pillai's Trace Value	*F*	Hypothesis degrees of freedom	Error degrees of freedom	Sig.
1.140	8.486	10	64	0,000

The One-Way MANOVA test results in Table 3 have a test statistic value of $F = 8.486$. In contrast, $F_{50;320;0.05}$ has a value of 1.391. The decision to reject H_0 is made when the F value between the two values is more than $F_{30;320;0.05}$, indicating that the groups that were constructed vary.

To evaluate if group members' variables vary from one another, one-way ANOVA is utilized. The one-way ANOVA test yielded the following findings.

Table 4. One-Way ANOVA Test Results for domestic tourist clustering.

Variable	*F*	Sig.
2018 Domestic Tourist Number	88.670	0.000
2019 Domestic Tourist Number	62.304	0.000
2020 Domestic Tourist Number	57.842	0.000
2021 Domestic Tourist Number	44.810	0.000
2022 Domestic Tourist Number	95.517	0.000

Based on Table 4, each F value of each variable is known. The value will be compared with $F_{2;35;0.05}$ of 3.267. The five variables have a larger F value when compared to $F_{2;35;0.05}$, indicating that they impact the creation of groups and that there are differences in the five variables' features on the groups that are created. This leads one to reject the null hypothesis.

When districts and cities in East Java Province were grouped according to the number of domestic tourists using the C-Means method, three clusters were formed. One-way MANOVA testing revealed differences among the three groups, and the five variables had an impact on the differences between the groups based on the one-way ANOVA testing results.

The results of grouping domestic tourists using 3 clusters are presented in Table 5.

Table 5. District/City List for domestic tourist clustering.

Cluster	District/City			
1	Lamongan Kota Malang Blitar	Malang Banyuwangi	Kediri Gresik	Pasuruan Bangkalan
2	Blitar Pacitan Mojokerto Magetan Kota Mojokerto Tulungagung Lumajang	Jombang Sidoarjo Probolinggo Sumenep Kota Probolinggo Bojonegoro Situbondo	Ponorogo Trenggalek Ngawi Jember Madiun Bondowoso	Pasuruan Sampang Kota Madiun Pamekasan Kota Kediri Nganjuk
3	Kota Surabaya	Tuban	Batu	

After listing the districts in each group, a description of each created group may be found below.

Table 6. Characteristics for each domestic tourist clustering.

Cluster	**1**	**2**	**3**
N	9	26	3
2018 Domestic Tourist Number	2,966,663	927,264	6,708,863
2019 Domestic Tourist Number	4,074,400	926,765	7,265,356
2020 Domestic Tourist Number	1,285,409	402,382	2,793,489
2021 Domestic Tourist Number	1,244,995	402,318	3,135,962
2022 Domestic Tourist Number	3,094,431	905,182	6,107,802

Table 7 displays the outcomes of the clustering of foreign tourists.

Table 7. Clustering results for foreign tourists.

Cluster	**Number of Cluster**			
	2	**3**	**4**	**5**
1	1	1	5	1
2	37	32	31	31
3		5	1	4
4			1	1
5				1

Table 7 lists the number of districts/cities that make up each group in the clustering findings of 38 districts/cities in East Java Province using C-Means, with two to five clusters. The pseudo-F-statistic value with the highest value among the two to five clusters may be used to identify the optimal cluster. The pseudo-F-statistic value for each cluster is shown below.

Table 8. Pseudo F-Statistic value for foreign tourist clustering.

Number of Cluster	**Pseudo F-Statistic**
2	98.0934
3	169.9234
4	422.4898
5	**604.9411**

Table 8 displays how the C-Means technique is used to get the pseudo-F-statistic value for two to five clusters. Based on the number of national visitors, five clusters are the ideal number to organize districts/cities in East Java. The greatest value of 604.9411, the pseudo-f-statistic value at 5 clusters, illustrates this.

Disparities in the features of each group are predicted when districts/cities in East Java Province are grouped using the C-Means approach based on the number of foreign visitors. One-way ANOVA and one-way MANOVA can be used to see if the features of the groups that were generated differ from one another. To test for differences in attributes, use Pillai's Trace in One-Way MANOVA.

Table 9. One-Way MANOVA Test Results for foreign tourist clustering.

Pillai's Trace Value	***F***	**Hypothesis degrees of freedom**	**Error degrees of freedom**	**Sig.**
3.041	20.286	20	128	0,000

The One-Way MANOVA test results in Table 9 have a test statistic value of $F = 20.286$. In contrast, $F_{100;640;0.05}$ has a value of 1.268. The decision to reject H_0 is made when the F value between the two values is more than $F_{100;640;0.05}$, indicating that the groups that were created vary from one another.

To evaluate if group members' variables vary from one another, one-way ANOVA is utilized. The one-way ANOVA test yielded the following findings.

Table 10. One-Way ANOVA Test Results for foreign tourist clustering.

Variable	*F*	Sig.
2018 Foreign Tourists Number	407.844	0.000
2019 Foreign Tourists Number	881.175	0.000
2020 Foreign Tourists Number	60.173	0.000
2021 Foreign Tourists Number	39.848	0.000
2022 Foreign Tourists Number	62.505	0.000

Every variable's *F* value is known based on Table 10. The value and $F_{4;33;0.05}$ of 2.659 will be compared. The five variables have a bigger *F* value when compared to $F_{4;33;0.05}$, indicating that the five factors impact the creation of groups and that there are differences in the five variables' features on the groups that are created. This leads one to reject the null hypothesis.

When districts and cities in East Java Province were grouped according to the number of foreign visitors using the C-Means method, it was discovered that as many as five clusters of groups had formed. The results of one-way MANOVA testing indicated that there were differences among the five groups that had formed, and the five variables had an impact on the differences among the groups that had formed.

The results of grouping foreign tourists using 3 clusters are presented in Table 11.

Table 11. District/City List for foreign tourist clustering.

Cluster	District/City			
1	Kota Surabaya			
2	Tuban Kota Batu Lamongan Blitar Kediri Pasuruan Bangkalan Kota Blitar	Pacitan Mojokerto Magetan Kota Mojokerto Tulungagung Lumajang Jombang	Sidoarjo Sumenep Kota Probolinggo Bojonegoro Situbondo Ponorogo Trenggalek Ngawi	Jember Madiun Nganjuk Kota Pasuruan Sampang Kota Madiun Pamekasan Kota Kediri
3	Gresik	Kota Malang	Probolinggo	Bondowoso
4	Malang			
5	Banyuwangi			

After listing the districts in each group, below is a description of each group formed.

Table 12. Characteristic for each foreign tourist clustering.

Cluster	1	2	3	4	5
N	1	31	4	1	1
Number of Foreign Tourists 2018	80,475	1,907	45,834	7,897	126,251
Number of Foreign Tourists 2019	56,199	1,642	30,823	122,612	290,792
Number of Foreign Tourists 2020	15,447	301	5,365	3,412	15,517
Number of Foreign Tourists 2021	4,680	86	40	115	2,145
Number of Foreign Tourists 2022	10,869	622	14,199	2,495	30,232

4. DISCUSSIONS

Based on Table 6, the characteristics (average) of each group formed on the data of the number of domestic tourists in East Java Tourism Attractions in 2018-2022 are obtained. It is known that group 3 on average has the highest value among other groups each year. This indicates that group 3 is a group with a high number of domestic tourists, so it can be said that this group includes tourist destinations that are the main destinations for domestic tourists, with a very high number of visits. Destinations in this group may include famous and iconic places that are the main attraction. Group 1 is the group with the second highest average each year after Group 3, so it can be said that this group includes tourist destinations that are quite popular and have a stable number of visits. Destinations in this group may include places that are well-known to domestic tourists but not as big as group 3. Group 2 is the group with the lowest average in all years, so it can be said that this group includes tourist destinations that have a lower number of visits, but have the potential to develop further. Destinations in this group may need more promotion or facility development to attract more domestic tourists. Based on the average achievement in each year, the ranking status of each group of districts/cities formed can be given as follows Table 13 and Figure 1.

Table 13. The Status of Each Cluster for domestic tourists clustering.

Cluster	Status
1	Domestic Popular Tourist Destinations
2	Domestic Emerging Tourist Destinations
3	Domestic Main Tourist Destinations

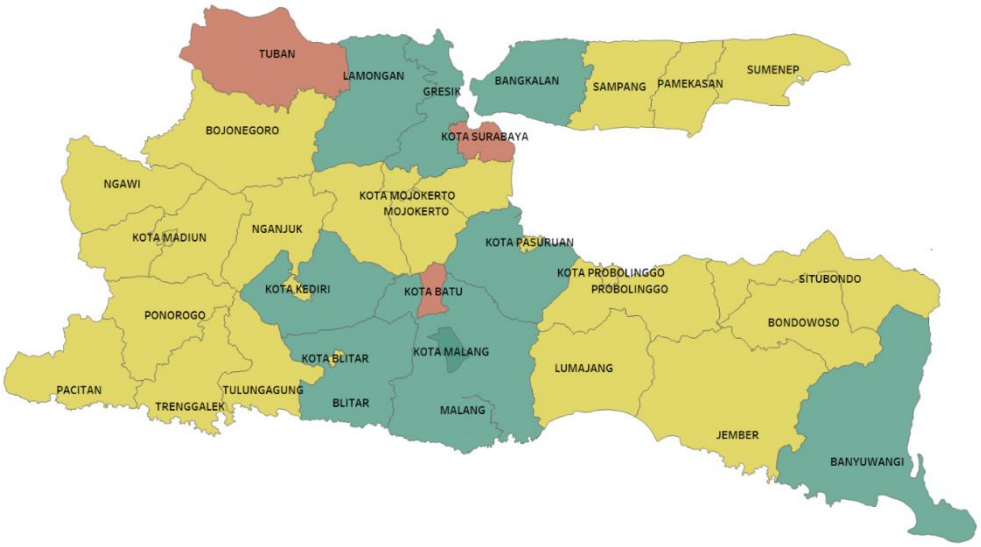

Figure 1. Domestic Tourists Clustering

Based on Table 12, the characteristics (average) of each group formed on the data of the number of foreign tourists in East Java Tourism Attractions in 2018-2022 are obtained. It is known that group 5 on average has the highest value among other groups in each year. This indicates that group 5 is a group with high foreign tourists, so it can be said that this group includes destinations that receive the highest number of foreign tourists in absolute terms in certain years, especially showing the highest peak visits. Group 1 is the group with the second highest average almost every year after Group 5, so it can be said that this group includes destinations that receive the highest number of foreign tourists consistently. It includes destinations that are most popular and attracts a large number of tourists. Group 4 is a group of destinations with high fluctuations in the number of tourists, experiencing large spikes in some years but still having a high average. Group 3 is a group of destinations that receive a medium number of foreign tourists, showing moderate popularity but not as high as the main group. Group 2 is the group with the lowest average across all years so it can be said that this group includes destinations that receive consistently low numbers of foreign tourists. These destinations may be less popular or less developed

than the other groups. Destinations in this group may need more promotion or facility development to attract more foreign tourists. Based on the average achievement in each year, the ranking status of each group of districts/cities formed can be given as follows Table 14 and Figure 2.

Table 14. The Status of Each Cluster for foreign tourists clustering.

Cluster	Status
1	Major International Tourism Destinations
2	Emerging International Tourism Destinations
3	Moderate International Tourism Destinations
4	Volatile International Tourism Destinations
5	Premium International Tourism Destinations

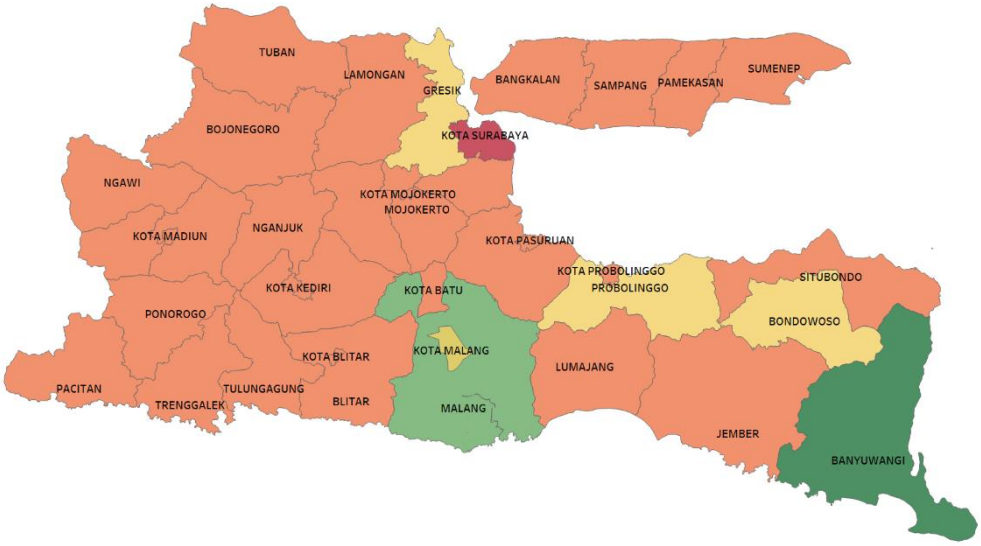

Figure 2. Foreign Tourists Clustering

5. CONCLUSION

Based on the clustering analysis of the number of Indonesian tourists in 2018-2022, three main groups were formed: Main Tourism Destinations of the Archipelago, Popular Tourism Destinations of the Archipelago, and Emerging Tourism Destinations of the Archipelago. The first group includes destinations with very high visitation and popular among domestic tourists. The second group includes destinations with moderate visitation, showing good stability and popularity. The third group includes destinations with low visitation that have the potential to be further developed. Meanwhile, the clustering analysis of the number of foreign tourists in 2018-2022 resulted in five main groups: Major International Tourism Destinations, Emerging International Tourism Destinations, Moderate International Tourism Destinations, Volatile International Tourism Destinations, and Premium International Tourism Destinations. The first and fifth groups show the highest-visited destinations, with the fifth group experiencing the highest peak visits in certain years. The second group includes low-visit destinations, while the third group shows medium-visit destinations. The fourth group includes destinations with significant fluctuations in tourist numbers. Future research can use other methods such as fuzzy c-means [18].

6. REFERENCES

[1] Badan Pusat Statistik, "Statistik Pariwisata Provinsi Jawa Timur 2022," 2023.
[2] M. Qori'atunnadyah and F. D. Rahmawati, "Pengelompokkan Kabupaten dan Kota Berdasarkan Kondisi Infrastruktur Jalan Menggunakan Hierarchical Clustering," *Journal of Informatics Development*, vol. 1, no. 1, pp. 1–5, 2022.

[3] M. Qori'atunnadyah, "Pengelompokkan Wilayah Berdasarkan Rasio Guru-Murid Pada Jenjang Pendidikan Menggunakan Algoritma K-Means," *Journal of Informatics Development*, vol. 1, no. 2, pp. 33–38, 2022.
[4] M. Qori'atunnadyah, "Metode C-Means untuk Pengelompokkan Kabupaten/Kota Provinsi Jawa Timur berdasarkan Indikator Indeks Pembangunan Manusia (IPM)," *Journal of Informatics Development*, vol. 1, no. 2, pp. 51–58, 2023, doi: 10.30741/jid. v2i2.1013.
[5] M. Qori'atunnadyah, F. S. Liyundira, and N. T. Indrianasari, "Grouping Manufacturing Companies Based on Factors Affecting Firm Value Using C-Means Clustering," 2023, pp. 28–31. doi: 10.2991/978-94-6463-346-7_6.
[6] K. Gustipartsani, N. Rahaningsih, R. Danar Dana, and I. Yulia Mustafa, "DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS PADA DATA KUNJUNGAN WISATAWAN DI KABUPATEN KARAWANG," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3595–3601, Feb. 2024, doi: 10.36040/jati. v7i6.8282.
[7] E. Satria, H. S. Tambunan, I. S. Saragih, I. S. Damanik, and F. T. E. Sitanggang, "Penerapan Clustering dalam Mengelompokkan Jumlah Kunjungan Wisatawan Mancanegara Dengan Metode K-Means," *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, vol. 1, p. 462, Sep. 2019, doi: 10.30645/senaris. v1i0.52.
[8] B. Setio Purnomo and P. Taqwa Prasetyaningrum, "PENERAPAN DATA MINING DALAM MENGELOMPOKKAN KUNJUNGAN WISATAWAN DI KOTA YOGYAKARTA MENGGUNAKAN METODE K-MEANS," *Journal of Computer Science and Technology (JCS-TECH)*, vol. 1, no. 1, pp. 27–32, Sep. 2022, doi: 10.54840/jcstech. v1i1.38.
[9] M. F. Masteriarsa and Riyanto, "PEMETAAN DESTINASI PARIWISATA BERDASARKAN DAYA DUKUNG KEPARIWISATAAN PROVINSI DI INDONESIA," *Jurnal TAMBORA*, vol. 7, no. 2, pp. 133–141, Jul. 2023, doi: 10.36761/jt. v7i2.3450.
[10] S. Maulidia and E. T. Astuti, "Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Potensi Pariwisata Tahun 2020," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 405–414, Nov. 2022, doi: 10.34123/semnasoffstat. v2022i1.1480.
[11] B. M. Al-Fahmi, E. Rahmawati, and T. Sagirani, "Penerapan K-Means Clustering Pada Pariwisata Kabupaten Bojonegoro Untuk Mendukung Keputusan Strategi Pemasaran," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 9, no. 2, pp. 141–149, Aug. 2023, doi: 10.25077/TEKNOSI.v9i2.2023.141-149.
[12] Dinas Kebudayaan dan Pariwisata Provinsi Jawa Timur, "Kebudayaan Dan Pariwisata Jawa Timur Tahun 2022 Dalam Angka," 2023.
[13] W. W. Piegorsch, *Statistical Data Analytics*, 1st ed. USA: John Wiley & Sons, Ltd, 2015.
[14] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised Learning," 2009, pp. 485–585. doi: 10.1007/978-0-387-84858-7_14.
[15] J. Macqueen, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.
[16] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans Inf Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.
[17] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl Stat*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.
[18] M. Qori'atunnadyah, "Fuzzy C-Means for Regional Clustering in East Java Province Based on Human Development Index Indicators," *J Statistika*, vol. 16, no. 2, p. 524, 2023.