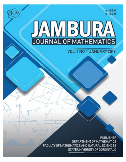
JAMBURA JOURNAL OF MATHEMATICS

Jambura J. Math. Vol. 4, No. 1, pp. 167–176, January 2022

Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom

DOI: https://doi.org/10.34312/jjom.v4i1.11883



Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia

Suwardi Annas[1*], Uca[2], Irwan[3], Rahmat Hesha Safei[4], Zulkifli Rais[5]

[1,4,5]*Department of Statistics, Faculty of Mathematic and Natural Science, Universitas Negeri Makassar, Makassar 90224, Indonesia*

[2]*Department of Geography, Faculty of Mathematic and Natural Science, Universitas Negeri Makassar, Makassar 90224, Indonesia*

[3]*Department of Mathematics, Faculty of Mathematic and Natural Science, Universitas Negeri Makassar, Makassar 90224, Indonesia*

[*]*Corresponding author. Email: suwardi_anas@unm.ac.id*

ABSTRACT

Air pollution is an important environmental problem for specific areas, including Makassar City, Indonesia. The increase should be monitored and evaluated, especially in urban areas that are dense with vehicles and factories. This is a challenge for local governments in urban planning and policy-making to fulfill the information about the impact of air pollution. The clustering of starting points for the distribution areas can ease the government to determine policies and prevent the impact. The k-Means initial clustering method was used while the Self-Organizing Maps (SOM) visualized the clustering results. Furthermore, the Geographic Information System (GIS) visualized the results of regional clustering on a map of Makassar City. The air quality parameters used are Suspended Particles (TSP), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), Surface Ozone (O_3), and Lead (Pb) which are measured during the day and at night. The results showed that the air contains more CO, and at night, the levels are reduced in some areas. Therefore, the density of traffic, industry and construction work contributes significantly to the spread of CO. Air conditions vary, such as high CO levels during the day and TSP at night. Also, there is a phenomenon at night that a group does not have SO_2 and O_3 simultaneously. The results also show that the integration of k-Means and SOM for regional clustering can be appropriately mapped through GIS visualization.

Keywords:

k-Means; Geographic Information Systems; Air Pollution; Self-Organizing Maps

How to Cite:

A. Annas, U. Uca, I. Irwan, R.H. Safei, and Z. Rais, ”Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia”, *Jambura J. Math.*, vol. 4, No. 1, pp. 167–176, 2022, doi: https://doi.org/10.34312/jjom.v4i1.11883

1. Introduction

Air pollution is a serious environmental problem for specific regions, including developing countries, since it causes chronic impacts on human health. Moreover, the decline in air quality can raise concerns about the impact. This is because every

Received: 30 October 2021 | Accepted: 4 January 2022 | Online: 12 January 2022

substance in the atmosphere under certain conditions can harm humans in terms of health and the environment [1].

WHO in 2019 reported that at least 7 million deaths occur annually due to air pollution. It can be caused by climate change, the greenhouse effect, and significant air pollutants such as particulate matter, sulfur dioxide, nitrogen dioxide, carbon monoxide, and ozone [2]. The increasing population and high energy consumption in urban areas cause air pollution to worsen. More than half in urban areas is caused by transportation. The high level of air pollution causes the depletion of the ozone layer and increased global warming [3–6].

The increase in this pollution should be monitored and evaluated, especially in urban areas densely packed with vehicles and industrial factories. This is a big challenge for the government in urban planning and policy-making to provide information about the dangers of air pollution. Therefore, preliminary knowledge about the conditions in urban areas is needed.

In this case, studies are conducted on air pollution in the Makassar City area, Indonesia. Makassar suffers severe air pollution as one of the urban areas that cause population growth to be very rapid. In addition, changes in air quality in this city are caused by the construction of industrial centers and the volume of vehicles increasing every day. Therefore, efforts are needed to facilitate the government in setting environmental policies by describing air pollution conditions by clustering points (regions).

Clustering is the process of collecting objects with similarities into a cluster. The results show that objects in one cluster are more homogeneous, while those between clusters are more heterogeneous [7]. There are two classical cluster analysis methods: hierarchical and non-hierarchical clustering methods. Determination of the total clusters formed for these two methods is conducted subjectively. First, the dendrogram's cut-off or cut point was determined using cluster visualization. In the non-hierarchical clustering method, the determination of the total clusters was conducted by the knowledge and experience of researchers [8, 9]. These methods are based on interval or ratio scale data [10], and one of the non-hierarchical clustering methods is k-Means.

A related research that uses k-Means as a clustering method includes [11] applying the k-Means clustering algorithm in human infectious diseases. The k-Means analyzed the spread of infectious diseases in humans based on several variables formed per sub-district in 32 health centers in the Majalengka Regency. Furthermore, a research by Ardillah, *et.al* [12] used permutation entropy, k-Means clustering, and multilayer perceptron in detecting epilepsy in humans. This creates a system that can predict whether a person has seizure-free or seizure epilepsy.

Other swarming methods are also developed using artificial intelligence. Artificial Neural Network (ANN) is an information processing paradigm inspired by biological systems, namely neurons, such as the brain, processing information [13]. The key to ANN is the structure of the information processing system, which consists of several elements (neurons) integrated to solve specific problems. There are two learning processes for weight changes in artificial neural networks: supervise and unsupervised learning [14, 15].

Self-Organizing Maps (SOM) is a topology form of Unsupervised ANN. Clustering using the SOM algorithm obtains the characteristics of the observed objects. Furthermore, this algorithm can be used for large and small data and visualize the clustering results in a lower dimension. SOM can group data for all data types (categorical and numeric) [13], and the visualization capability can overcome the shortcomings of other clustering methods that are difficult when the data size is large, such as the use of dendrograms.

However, the SOM clustering method has limitations because it needs to define the command line's code point (region). Research conducted by Annas and Rais [16] used k-Means and Geographic Information System (GIS) in visualizing the results of clustering natural disaster areas in Indonesia. The k-Means were used as the initial clustering and then developed the SOM method to visualize the results. Furthermore, a Geographic Information System (GIS) is used to visualize air pollution on a map of Makassar City. It can visually map by area of environmental quality [17, 18]. Therefore, ease the government in formulating policies to deal with air pollution problems in Makassar City.

2. Methods

The data used are the results of air condition measurements obtained from the Environmental Service, which are randomly distributed and have gone through laboratory analysis of the Makassar City Health Department. The data obtained are the results of air quality parameter measurement values. Explanation of air quality distribution in a map uses an interpolation approach; therefore, the concentration and level of air quality at other locations that are not measured can be known.

The Makassar City pollution measurement was carried out in two stages, namely Ambien I and Ambien II, respectively, in 2018. Each ambient was measured during the day and night, and the variables used were Total Suspended Particles (TSP), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Carbon Dioxide (CO), and Ozone Layer (O_3). The locations for measuring air quality are presented in Table 1.

Table 1. List of air sampling locations

No	Location	Longitude	Latitude	Code
1	Pettarani and ST. Alauddin T-junction	119.431	-5.174	A
2	Front of the PLN Region VII Hertasning Office	119.448	-5.167	B
3	Front of Kejasdem Wirabuana Military Region Command XIV	119.420	-5.148	C
4	Front of PT Eastern Pearl Flour Mills Office	119.411	-5.117	D
5	Front of Panampu Market	119.426	-5.118	E
6	Makassar Mall	119.413	-5.130	F
7	Karebosi Field	119.414	-5.135	G
8	Front of Prima Hotel	119.416	-5.152	H
9	Front of A. Mattalatta Stadium	119.414	-5.158	I
10	Front of the Mayor's Rujab	119.409	-5.147	J
11	Urip Sumiharjo – Pettarani intersection	119.440	-5.137	K
12	Front of the Al-Markaz mosque	119.426	-5.132	L
13	Front of the Governor's Office	119.452	-5.141	M
14	Front of Wirabuana Hall, Urip Sumiharjo Street	119.462	-5.144	N

The data were analyzed by combining k-Means and SOM to classify areas based on air pollution in Makassar City, Indonesia. The analysis stage carried out was theinitial cluster center determination

1. Calculation of the distance to the center of the cluster using the Euclidean distance

$$D_{(i,j)} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + ... + (x_{ki} - x_{kj})^2}$$

   where:
   D(i,j) = Distance of data to I to cluster center j-th
   x_{ki} = Data to i-th on attribute data to k-th
   x_{kj} = The j-th center point on the k-th attribute
2. Clustering data, the smaller the distance from the center of the cluster, the higher the similarity of the data.
3. Defining a new cluster center, with the following calculations:

$$Z_c = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad i = 1, 2, 3, ..., n$$

4. Carrying out until the results converge.

Based on the results of the initial clustering of k-Means to the distribution points of air volume, the SOM algorithm [2] was used to visualize the clustering results with the following steps.

1. Creating a training data matrix
2. Creating a grid and SOM model
3. Creating a SOM plot based on:
   (a) U-matrix, the distance between objects
   (b) Component planes, the characteristics of each variable
4. Determining the number of clusters using a decrease in variance in clusters (within-sum squares, WSS) that is no longer significant or the formation of an elbow (elbow) and the highest silhouette coefficient as a comparison for WSS.

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

   where:
   s(i) = i-th object silhouette coefficient
   a(i) = the distance between object *i* and other objects in the same cluster
   b(i) = the minimum value of the average distance of the i-th object with other objects from different clusters
5. SOM plots are based on codes, the characters of each cluster.
6. Adding a cluster delimiter in step 6 using the k-Means algorithm as follows:
   (a) Determining randomly the coordinates of the location of k centroids $c_1, c_2, \ldots, c_k$ as the center of the cluster, where the number of c is equal to the variable p.
   (b) Calculating the distance (d_{ij}) of each object (x_i) to each centroid (c_j)

$$d_{ij} = \sqrt{(x_i - c_j)^2}.$$

   (c) Putting object (x_i) into a j-th cluster if d_{ij} is less than dij', j $\neq$ j', j,j'∈k.
   (d) Updating the coordinates $c_1, c_2, \ldots, c_k$ where c is the vector mean of p variables from x_i that are members of the jth cluster.
   (e) Repeating steps b, c, and d until there is no change or the change in c_j is no longer significant.

A visualization of the air quality distribution map by region was carried out using GIS software based on using k-Means and SOM for regional clustering.

3. Results and Discussion

3.1. Air Quality During the Day

The parameters used to measure air quality in Makassar City are TSP, SO_2, NO_2, CO, and O_3. The first measurement was carried out during the day, and the statistical results of the descriptive analysis are presented in Table 2. It showed that the air content during the day with the highest average was CO and far above the average content of other parameters due to the density of vehicles during the day. Subsequently, industrial and construction workers operating in urban areas also contribute to high levels of air pollution [1, 2].

Table 2. Summary of measurement statistics during the day

Variable	Average	Standard Deviation	Minimum	Maximum
TSP	9.25	6.53	2.14	26.67
SO_2	29.42	3.71	26.01	36.57
NO_2	1.81	0.83	0.46	3.88
CO	775.40	39.83	718.00	835.10
O_3	11.19	11.18	0.95	41.93

Based on air quality parameters using k-Means, there are seven regional clusters. The members and their characteristics are described in Table 3, and the results are then visualized using SOM, which consists of a U-matrix and visualization of clusters formed through Component Plants. The results of the SOM visualization can be seen in Figures 1 and 3.

Table 3. Results of clustering k-Means and the characteristics of each

Cluster	Member	Characteristics
1	M, N	Filled by areas with high CO levels
2	E	Filled with areas with high levels of TSP, SO_2, NO_2, CO, and O_3
3	A, B, C, D	Filled with areas with high levels of TSP and NO_2 and high SO_2
4	I, J, K	Filled by areas with high enough CO levels
5	H	Filled by areas with high levels of NO_2
6	F	Filled by areas with high CO levels
7	G, L	Filled by areas with high CO and O_3 levels

Figure 1 shows the CO parameter content in group 2, marked in red (Figure 3). This condition shows the high CO content in the area, while the content of other air quality parameters is in the medium and low categories.

The optimum number of clusters is determined using WSS with an elbow compared with the silhouette coefficient. Based on the WSS graph shown in Figure 2 on the left, elbows are formed at points 3, 4, 6, and 7. On the right, the silhouette coefficient is at points 7 and 0.285 or a WSS value of 0.132. Based on this, there are seven optimal clusters.

The results of clustering and visualization with maps are shown in Figure 3. Each color represents a cluster; therefore, each area with the same cluster has a similar color. For example, cluster 1 is blue with only one region, while 2 is red with only one region. Other areas with similar characteristics are the result of interpolation. Cluster 3 is marked in green with four regions, while 4 is marked with a yellow color consisting of three regions.

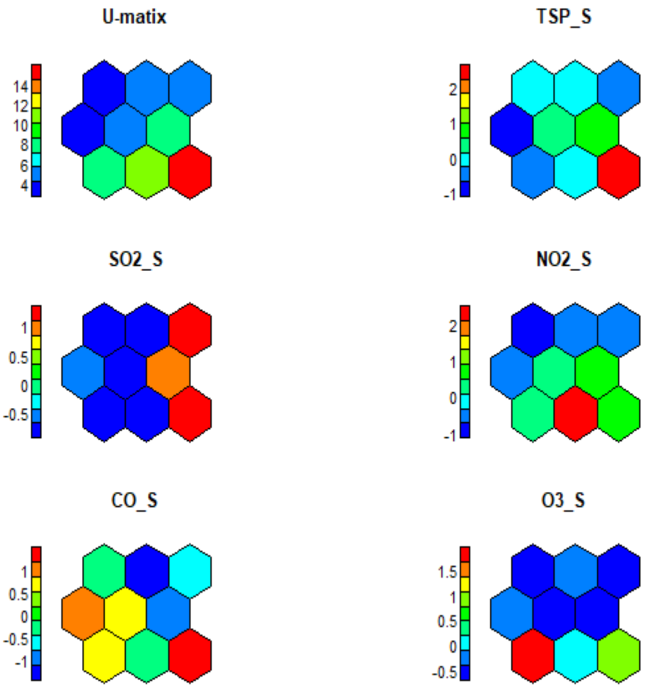


Figure 1. U-matrix and component planes

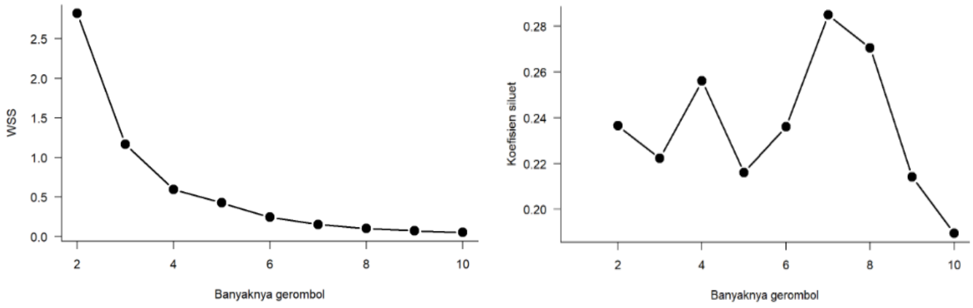


Figure 2. Determination of the number of clusters

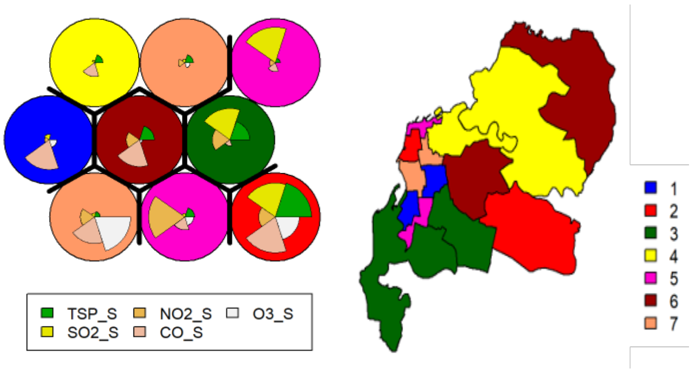


Figure 3. SOM results and air quality distribution map for Makassar City

Clusters 5 and 6 are highlighted with purple and brown colors, consisting of one region, while 7 consisting of two regions, is indicated in purple.

Figure 3 shows a map of the distribution of air quality resulting from SOM clustering using GIS, and the description of each cluster/region is presented in Table 2. Cluster 1 consists of two areas, namely M and N, with a high vehicle density level due to a relatively wide main road. Cluster 2 only consists of one area, E. This is the Pannampu Market area, and its surroundings have high levels of TSP, SO_2, NO_2, CO, and O_3. Cluster 3 consists of regions A, B, C, and D, which are characterized by high levels of CO. Cluster 5 consists of an H region with high levels of NO_2, while Cluster 6 consists of an F region with high CO levels. The last cluster consists of regions G and L with high CO and O_3 levels.

3.2. *Air Quality at Night*

The air quality measurements in Makassar City at night have characteristics as presented in Table 4. It can be seen that the average TSP, SO_2, NO_2, CO, O_3, and Pb tend to be smaller than the measurements during the day. The clusters formed can be descriptively different as the clustering results in measurements.

Table 4. Summary of measurement statistics at night

Variable	Average	Standard Deviation	Minimum	Maximum
TSP	7.93	3.45	2.14	14.76
SO_2	28.50	3.10	24.74	33.18
NO_2	1.15	0.67	0.23	2.41
CO	772.40	55.62	714.80	867.60
O_3	3.29	3.64	0.01	11.89

Table 4 shows descriptive measurements of air conditions. The air content at night with the highest average is CO due to the density of vehicles in Makassar City. However, the CO content in the air at night remains well above the average for other variables.

Table 5. k-Means clustering results and characteristics of each cluster

Cluster	Member	Characteristics
1	A, B, C, D, E, K	Filled by areas with high levels of NO_2, SO_2, and CO
2	L, M, N	Filled by areas with high levels of CO and O_3
3	G, H, J	Filled by areas with high levels of O_3 and SO_2
4	F, I	Filled by areas with high levels of TSP and SO_2

The clustering k-Means obtained 4 groups, and their characteristics are described in Table 5. The results are then visualized using SOM, consisting of a U matrix and visualization of the groups formed, as seen in Figures 4 and 6.

In Figure 4, the CO content is located in group 1, marked red. This condition also shows the high content of CO levels compared to other air quality parameters. The optimum number of clusters is determined using WSS with an elbow compared with the silhouette coefficient. Based on the WSS chart shown in Figure 5 on the left, elbows are formed at points 3, 5, 6, and 7. Therefore, the highest silhouette coefficient of 0.284 is at point 4 or the WSS value of 0.32. Based on this, the optimal number of clusters is 4.

The results of clustering and cluster visualization with maps are shown in Figure 6. Cluster 1 is blue with six regions, and other areas with similar characteristics result from interpolation. Clusters 2 and 3 are marked in red and green with three regions, while Cluster 4, consisting of two regions, is highlighted in yellow.

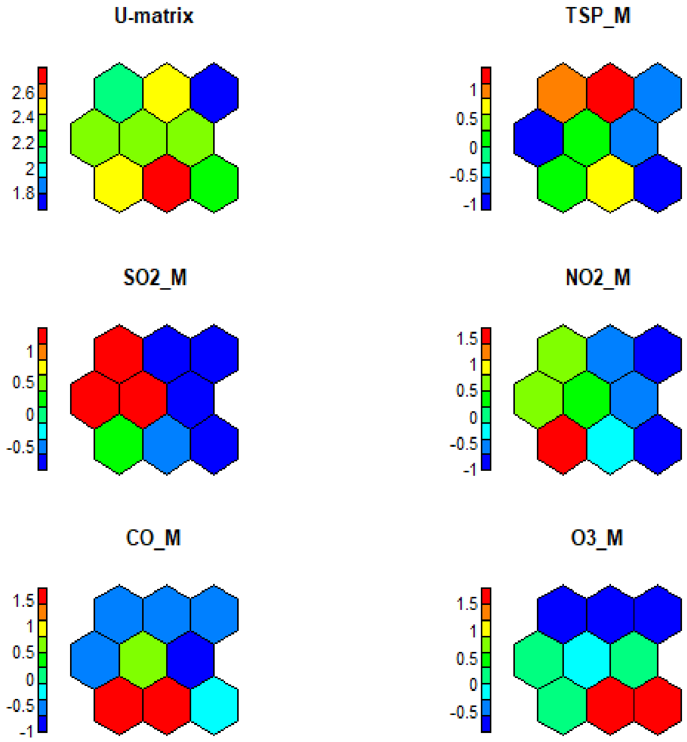


Figure 4. U-matrix and component planes

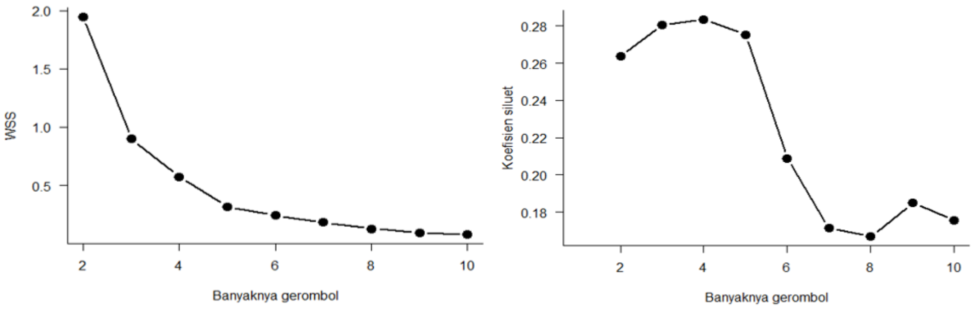


Figure 5. Determine the number of clusters

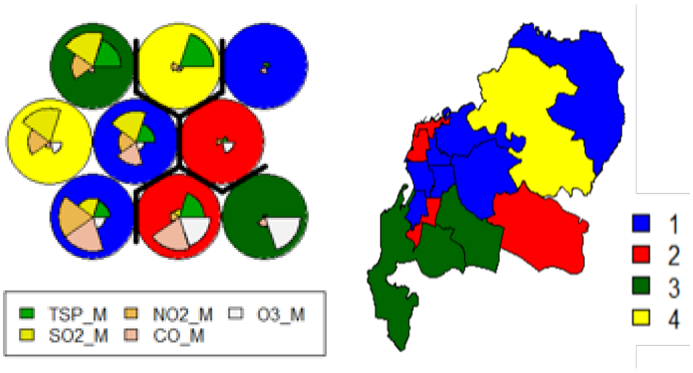


Figure 6. SOM results and air quality distribution map for Makassar City

Figure 6 shows a map of the distribution of air quality resulting from SOM clustering using GIS, and the description of each cluster/region is presented in Table 5. Cluster 1 consists of regions A, B, C, D, E, and K with high NO and SO. Cluster 2 consists of regions L, M, and N with high CO and O_3. Cluster 3 consists of regions G, H, and J, characterized by high O_3 and SO_2. Cluster 4 consists of regions F and I with high levels of TSP and SO.

4. Conclusion

The use of k-Means and SOM has been applied in regional clustering based on air quality parameters in Makassar City. The clustering results provide information about the air pollution in each region during the day and at night. The lower air quality in a region group indicates higher pollution. Determining the number of clusters based on the WSS graph at night, the optimal number of clusters is 7 regional clusters, while at night, 4 regional clusters are formed. This indicates differences in the distribution of pollution characteristics in each regional group formed between day and night. The unequal distribution of air pollution in each regional group is caused by differences in traffic density, industry, and development work. Although the integration of k-Means and SOM has grouped regional characteristics in Makassar City, mapping air pollution distribution per region can be adequately visualized using GIS.

Acknowledgements

We would like to express our deepest gratitude to the PNBP UNM which has assisted in funding, so that this research can be carried out properly.

References

[1] S. K. Dash and A. K. Dash, "Determination of Air Quality Index Status near Bileipada, Joda Area of Keonjhar, Odisha, India," *Indian Journal of Science and Technology*, vol. 8, no. 35, pp. 1–7, dec 2015, doi: http://dx.doi.org/10.17485/ijst/2015/v8i35/81468.

[2] A. Kurniawan, "Pengukuran Parameter Kualitas Udara (CO, NO2, SO2, O3 dan PM10) di Bukit Kototabang Berbasis ISPU," *Jurnal Teknosains*, vol. 7, no. 1, pp. 1–13, jul 2018, doi: http://dx.doi.org/10.22146/teknosains.34658.

[3] O. H. Adedeji, O. Oluwafunmilayo, and T.-A. O. Oluwaseun, "Mapping of Traffic-Related Air Pollution Using GIS Techniques in Ijebu-Ode, Nigeria," *Indonesian Journal of Geography*, vol. 48, no. 1, pp. 73–83, aug 2016, doi: http://dx.doi.org/10.22146/ijg.12488.

[4] S. N. Behera, M. Sharma, P. Mishra, P. Nayak, B. Damez-Fontaine, and R. Tahon, "Passive measurement of NO2 and application of GIS to generate spatially-distributed air monitoring network in urban environment," *Urban Climate*, vol. 14, no. 3, pp. 396–413, dec 2015, doi: http://dx.doi.org/10.1016/j.uclim.2014.12.003.

[5] R. Darmawan, "Analisis Resiko Kesehatan Lingkungan Kadar NO2 serta Keluhan Kesehatan Petugas Pemungut Karcis Tol," Skripsi, Universitas Airlangga, 2018.

[6] F. A. Farisi, B. Budiyono, and O. Setiani, "Pengaruh Sulfur Dioksida (SO2) pada Udara Ambien Terhadap Resiko Kejadian Pneumonia Pada Balita," *Jurnal Kesehatan Masyarakat*, vol. 6, no. 4, pp. 439–446, 2018.

[7] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. New Jersey: Pearson Education, Inc., 2007.

[8] M. Fujino and M. Yoshida, "Development and Validation of a Method of Forestry Region Classification Using PCA and Cluster Analysis together with the SOM Algorithm." *Journal of the Japanese Forest Society*, vol. 88, no. 4, pp. 221–230, 2006, doi: http://dx.doi.org/10.4005/jjfs.88.221.

[9] J. Hair, R. Anderson, R. Tatham, and W. Black, *Applied Multivariate Statistical Analysis*, 5th ed. New Jersey: Prentice-Hall, 1998.

[10] B. Sartono, D. Bodro, and G. Dito, *Teknik Eksplorasi Data yang Harus Dikuasi Data Scientist*. Bogor: IPB Press, 2020.

[11] A. Bastian, “Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka),” *J. Sist. Inf.*, vol. 14, no. 1, pp. 28–34, 2018.

[12] Y. Ardillah, H. Tjandrasa, and I. Arieshanti, “Deteksi Penyakit Epilepsi dengan Menggunakan Entropi Permutasi, K-Means Clustering, dan Multilayer Perceptron,” *J. Tek. ITS*, vol. 3, no. 1, pp. A70–A74, 2014, doi: http://dx.doi.org/10.12962/j23373539.v3i1.5486.

[13] J. Siang, *Jaringan Saraf Tiruan dan Pemrogramannya Menggunakan MATLAB*. Yogyakarta: Andi Offset, 2005.

[14] S. Annas, T. Kanai, and S. Koyama, “Principal Component Analysis and Self-Organizing Map for Visualizing and Classifying Fire Risks in Forest Regions,” *Agricultural Information Research*, vol. 16, no. 2, pp. 44–51, 2007, doi: http://dx.doi.org/10.3173/air.16.44.

[15] D. Klobucar and M. Subasic, “Using self-organizing maps in the visualization and analysis of forest inventory,” *iForest - Biogeosciences and Forestry*, vol. 5, no. 1, pp. 216–223, oct 2012, doi: http://dx.doi.org/10.3832/ifor0629-005.

[16] S. Annas and Z. Rais, “k-Means and GIS for Mapping Natural Disaster Prone Areas in Indonesia,” in *Proceedings of the Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, 12 October 2019, Bandung, West Java, Indonesia*. EAI, 2020, pp. 1–7, doi: http://dx.doi.org/10.4108/eai.12-10-2019.2296336.

[17] M. N. DeMers, *GIS For Dummies*, 1st ed. Amazon.com Services LLC, 2009.

[18] A. Kumar, I. Gupta, J. Brandt, R. Kumar, A. K. Dikshit, and R. S. Patil, “Air quality mapping using GIS and economic evaluation of health impact for Mumbai City, India,” *Journal of the Air and Waste Management Association*, vol. 66, no. 5, pp. 470–481, may 2016, doi: http://dx.doi.org/10.1080/10962247.2016.1143887.