Earthquake Point Clustering Using Self Organizing Maps (SOM) In Sulawesi and Maluku Regions

Irwan[1*], Ahmad Zaki[2], and Eka Janivia Widiyaningrum[3]
[1,2,3]Department of Mathematics, State University of Makassar, Makassar, Indonesia
*Corresponding author: irwanthaha@unm.ac.id

Received: 17 May 2025
Revised: 10 October 2025
Accepted: 30 October 2025

ABSTRACT —**Earthquakes pose a major threat in Indonesia, particularly in complex tectonic regions like Sulawesi and Maluku. To support disaster mitigation, this research employs the Self Organizing Maps (SOM) method—an unsupervised technique that reduces data dimensionality into an intuitive two-dimensional form—to cluster earthquake data using four key variables: longitude, latitude, magnitude, and depth. The dataset includes 5,275 earthquake records from 2022, sourced from the Meteorology, Climatology, and Geophysics Agency (BMKG). SOM training produced 25 neurons, which were then grouped into three optimal clusters using hierarchical clustering, validated by internal metrics: the lowest Connectivity Index (296.1512), highest Silhouette Index (0.3304), and a Dunn Index of 0.0058. Cluster 1, with 13 neurons, covers eastern Sulawesi and Maluku, featuring medium magnitude and depth. Cluster 2, with 11 neurons, represents central to southern Sulawesi, characterized by low magnitude and shallow depth. Cluster 3, comprising a single neuron, includes western regions with high-magnitude, very deep earthquakes.**
Keywords—**Clustering, Earthquake, Internal Validation, Self Organizing Maps (SOM).**

I. INTRODUCTION

Earthquakes are a natural phenomenon that often occurs in various parts of the world, including Indonesia [1]. According to a report by Indonesian Agency for Meteorological, Climatological and Geophysics in 2021, Indonesia is one of the most seismically active countries in the world due to the convergence of the Eurasian, Indo-Australian, and Pacific tectonic plates. The Sulawesi and Maluku regions are earthquake-prone areas because they are located in subduction zones and complex active faults [2]. Data from Center for Meteorology Climatology and Geophysics Region IV shows a significant increase in the number of earthquakes in this region, from 2,300 events per year to more than 3,000 since 2018. With the high frequency of earthquakes, further analysis is needed to understand their occurrence patterns to support disaster mitigation.

One effective approach to analyzing earthquake patterns is data clustering. This method aims to group objects based on similar characteristics so that hidden patterns in the data can be identified [3]. One of the artificial neural network-based clustering methods widely used in complex data analysis is the Self Organizing Map (SOM). SOM is able to reduce the dimension of data while maintaining its topological structure, making it suitable for earthquake distribution analysis [4].

Several previous studies have used SOM in earthquake clustering analysis. For example, one study clustered the level of damage and earthquake strength on the island of Java [5]. Another study clustered districts in West Java based on COVID-19 cases and identified two main clusters [6]. This study aims to cluster earthquake points in the Sulawesi and Maluku regions using the SOM method and analyze the characteristics of the resulting clusters.

II. LITERATURE REVIEW

A. Clustering

Clustering algorithms are used to separate data samples into groups so that samples belonging to a group have the most similarity with members of the same group and have the least similarity with members of other groups [7]. The basic concept of cluster analysis measurement is the concept of distance and similarity measurement. In this study, the distance measure used is the euclidean distance [8].

$$d_{(j)} = \sqrt{\sum_{I} \left(W_{ji} - x_i\right)^2}$$

$d_{(j)}$ = distance between the weight vector (W_{ji}) and the input vector (x_i)
W_{ji} = weight vector
x_i = input vector
I = many observed variables

B. Data Standardization

Data standardization is needed when there are differences in unit sizes between variables, which can affect the results of the analysis [9]. The goal is to reduce variation between variables so that there is no domination by variables with a larger scale [10]. One commonly used method is the z-score transformation, which is expressed in the following equation [9].

$$z = \frac{x_i - \bar{x}}{s}$$

z = data standardization value
x_i = data value i
$\bar{x}$ = mean value of variable X_i
s = standard deviation value

C. Cluster Analysis Assumption Test

In cluster analysis, one of the assumptions that need to be tested is multicollinearity, which is a strong linear relationship between two or more variables in the data [9]. High multicollinearity can reduce the accuracy of interpreting clustering results, so it is important to ensure variables are free from this problem before further analysis.

One method used to measure the relationship between variables is Pearson correlation, which assesses the strength and direction of the linear relationship between two variables [11]. If the correlation coefficient value is > 0.8, then the variables are considered to have a strong relationship. The value of r ranges from -1 to +1, where a positive value indicates a unidirectional relationship, while a negative value indicates an opposite relationship [10]. The Pearson correlation formula is as follows [11].

$$r_{xy} = \frac{N.\sum XY - (\sum X).(\sum Y)}{\sqrt{(N.\sum X^2 - (\sum X)^2).(N.\sum Y^2 - (\sum Y)^2)}}$$

r_{xy} = Correlation index number between variables X and Y
N = Number of objects
X = Value of variable X
Y = Value of variable Y

D. Cluster Validation

Cluster validation is the process of evaluating the results of cluster analysis or clustering to determine how good the resulting cluster is quantitatively and objectively. In this study, internal validation will be used, namely by using the connectivity index, silhouette index, and Dunn index.

1) Connectivity Index

The Connectivity Index is an index that shows the degree of cluster relationship based on the number of nearest neighbors and has values ranging from zero to infinity. The smaller the value, the better the resulting cluster [12]. The connectivity index is formulated as follows [12].

$$conn(c) = \sum_{i=1}^{N}\sum_{j=1}^{L} X_{i,nni(j)}$$

with,
$nn_{i(j)}$ = nearest neighbor observations from object j to object i
L = nearest neighbor count parameter
$X_{i,nni(j)}$ = the value of object i is 0 if object i and j are in one cluster and 1 for object j is not in one cluster.

2) Silhouette Index

The Silhouette Index is a good tool for evaluating the performance of clustering algorithms, especially on high-dimensional data sets where direct visualization of results is limited [13]. In certain observations, if the index value is close to 1, the cluster is considered well formed and if the index value is close to -1, the cluster result is considered unfavorable. The silhouette index is formulated as follows [12].

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max(a_{(i)}, b_{(i)})}$$

with,
$s_{(i)}$ = silhouette index of object i
$a_{(i)}$ = average similarity between object i and other objects in the cluster
$b_{(i)}$ = smallest value of average similarity between object i and other objects outside the cluster

3) Dunn Index

The Dunn Index aims to maximize the distance between clusters while minimizing the distance within clusters [13]. The Dunn index is formulated as follows [12].

$$Dunn = \frac{d_{min}}{d_{max}}$$

with,
d_{min} = smallest distance between observations in different clusters
d_{max} = largest distance in each data cluster

E. Self Organizing Maps (SOM)

SOM is a method introduced by Teuvo Kohonen in 1982, which combines artificial neural networks and prototype-based clustering [14]. SOM is used in cluster analysis and data visualization, working unsupervised by mapping the

weights according to the given input data [15].

In a SOM network, each neuron competes to be the winner, where the neuron that best matches the input pattern is selected as the winner neuron, and its weight is updated along with the surrounding neurons [16]. SOM has several advantages, including not requiring variable distribution assumptions, being able to handle complex nonlinear problems, and being effective in dealing with noise and missing data [17].

The stages in clustering using SOM are as follows [8].

1) Initialization in the form of weights (W_{ji}) obtained randomly for each node. After the weight (W_{ji}) is given, the network is given input (x_i).
2) After the input is received, the network will calculate the distance vector $d(j)$ obtained by summing the difference between the weight vector (W_{ji}) and the input vector (x_i) with the formula as follows [8].

$$d_{(j)} = \sqrt{\sum_I \left(W_{ji} - x_i\right)^2}$$

3) After the distance between the nodes is known, the minimum value of the distance vector $d(j)$ is determined, then the next step is to change the weight with the formula as follows [8].

$$W_{ji}(new) = W_{ji}(old) + \propto \left[x_i - W_{ji}(old)\right]$$

4) In the process of obtaining new weigts requires a learning rate (α) value of $0 \leq \alpha \leq 1$. The learning rate value at each epoch will decrease to $\alpha(i+1) = 0{,}5\alpha$. The test termination condition is carried out when the iteration has reached the maximum iteration that has been determined.

F. Earthquakes

An earthquake is a phenomenon of vibration or shaking that occurs on the earth's surface due to the sudden release of energy in the earth's crust. Tectonic earthquakes are the most common earthquakes that occur as vibrations resulting from the event of rock breaking due to the collision of two plates slowly that the accumulated energy of the collision exceeds the strength of the rock, then the rock below the surface [18].

Earthquakes are the biggest threat faced with the potential for tsunamis in them that can cause damage and even cause casualties [19]. Earthquake energy can cause great damage and disaster depending on the depth of the earthquake source, the strength of the earthquake, and the distance from the epicenter.

III. METHODOLOGY

This study conducted clustering using earthquake data in the Sulawesi and Maluku Island regions in 2022 totaling 5,275 data obtained from publications on the Meteorology, Climatology and Geophysics Agency (BMKG) website, https://www.bmkg.go.id. The data analysis method used is SOM analysis. SOM is an analysis method for high-dimensional data and no assumptions are needed and can produce visualization of the object. The variables used in this study are latitude, longitude, depth, and magnitude.

The research procedures carried out in this study are described as follows:

1) Collecting and inputting data on earthquake points in the Sulawesi and Maluku regions from 2022 obtained from the website of the Meteorology, Climatology and Geophysics Agency (BMKG).
2) Perform descriptive analysis to describe the variables used in the study, namely latitude, longitude, magnitude, and depth.
3) Performing a data standardization process for each research variable to have the same scale so that no variable dominates in the clustering process.
4) Performing assumption testing in the form of multicollinearity test to ensure that there is no strong linear relationship between independent variables.
5) Determining the best number of clusters through the validity test by comparing the values of the connectivity index, sillhouette index, and dunn index.
6) Perform clustering using the SOM method. First initialize the weights randomly, then calculate the distance of the data on the weights. From the results of the distance calculation, the minimum value will be determined, after which the weight changes and the weight is updated. Updating the weights requires a learning rate value, for the first iteration it has been determined while for each iteration the learning rate will be updated using the equation, repeat the steps until there is no change in weight in the previous iteration. Then visualization analysis is carried out with the SOM method.
7) Divide the SOM nodes into several clusters using the hierarchical clustering method according to the results of the cluster validity test. This process is carried out by calculating the distance between SOM nodes based on their weights, then the results of this process will determine the group of nodes that represent the earthquake distribution pattern in the SOM map.
8) The resulting output is the result of the clustering of earthquake points using SOM.

IV. RESULTS AND DISCUSSIONS

Before analyzing the clustering using SOM, the data preprocessing stage is first carried out by standardizing. This process is important because the earthquake data used has four variables with different measurement scales, namely

Longitude (118°-134° East), Latitude (8° LS - 4° LU), Magnitude (1-6 SR), and Depth (0-750 km). Using the Z-score method, the results of data standardization for the four variables are shown in Table 1.

Table 1 Data Standardization Results

Data	$z(X_1)$	$z(X_2)$	$z(X_3)$	$z(X_4)$
1	-0.4765	-1.4270	-0.6174	-0.5382
2	-0.4857	-1.4008	0.1024	-0.5382
3	-0.4748	-1.4034	-0.5108	-0.5382
4	0.7362	1.7548	1.8085	0.1986
5	-0.5665	-1.2905	-0.5241	0.3596
⋮	⋮	⋮	⋮	⋮
5275	-0.5739	0.7992	-0.9639	1.3758

Furthermore, multicollinearity assumptions were tested to identify whether there was a significant linear relationship between the independent variables used. By using the Pearson correlation formula, the overall correlation value between variables is presented in Figure 1.

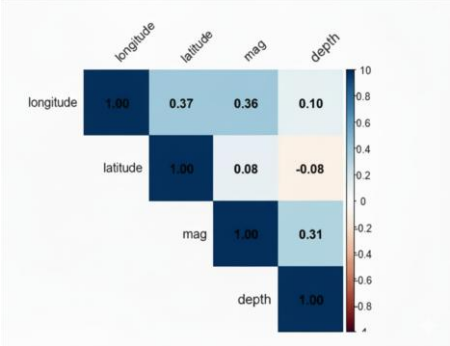

Figure 1 Correlation between variables

The results of multicollinearity testing using the correlation matrix in Figure 1. show that in general there is no serious multicollinearity problem in the variables to be used. The correlations between the independent variables tend to be low to medium, with the highest correlation found between Longitude and Latitude at 0.37, which is still within reasonable limits. A stronger correlation was only seen between Magnitude and depth at 0.31, which is also medium. As such, there is no strong linear relationship between the variables, so each variable can be considered to make a relatively independent contribution to the clustering model to be built.

Next, a cluster validity test was conducted to determine the optimal number of clusters. The cluster validity test used includes an internal validation test using the Connectivity, Silhouette, and Dunn indices. The number of clusters selected will be based on the smallest Connectivity index value, a large Silhouette index value, and a Dunn index value close to 1. The process of working on cluster validation and SOM cluster analysis is done using R software.

Table 2 Cluster Validation Output

Internal Validation Test	Cluster			
	3	4	5	6
Connectivity Index	**296.1512**	458.3206	523.8837	667.1020
Silhouette Index	**0.3304**	0.3094	0.3213	0.3043
Dunn Index	**0.0058**	0.0038	0.0037	0.0038

The number of clusters tested for validity is from 3 to 6 clusters. In the validity test conducted, it was found that the most optimal number of clusters was 3 clusters. This is proven through three comprehensive validation criteria: Connectivity Index with the lowest value of 296.1512 which indicates the closeness of objects in the cluster, Dunn Index of 0.0058 which illustrates the diversity of distances between clusters, and the highest Silhouette Index of 0.3304 which shows the quality of clustering with the level of similarity of objects in one cluster and differences with other clusters. The results of this analysis show that the division of data into 3 clusters for the SOM method provides the most coherent clustering structure, is able to capture the characteristics of the data in a representative manner, and provides a more in-depth perspective on the complex patterns and relationships in the dataset under study.

To obtain clusters of earthquake points in the Sulawesi and Maluku regions in 2022 and determine the characteristics of each cluster, SOM clustering analysis is used. The SOM network requires a training progress to minimize the average distance of an object to the nearest unit [20].

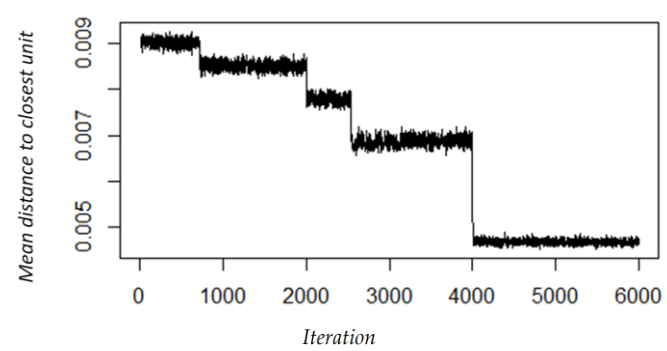

Figure 2 Training Progress Chart

Based on the graph in Figure 2, it is known that the training process lasts for 6000 iterations. The graph shows the value of the mean distance to closest unit (the average distance of each data to the nearest cluster unit) which decreases as iterations increase. At the 4000th iteration, the graph looks stable, which indicates that the training process using the SOM method has reached stability. This indicates that the resulting map is good enough to represent the trained data. The mean distance to closest unit value at the end of training is 0.0045, which means that on average each object in the data has a distance of 0.0045 to the nearest unit in the SOM map.

After the SOM model training process, 25 neurons were obtained that represented the final weights of the network. Each neuron describes the characteristics of earthquake data based on the four variables used, namely longitude, latitude, magnitude, and depth.

Furthermore, the clustering results using SOM were further divided into three clusters using the hierarchical clustering method. This cluster division is based on the results of the cluster validity test which shows that three clusters are the optimal number. The hierarchical clustering process was applied to the SOM neuron weights to group the 25 neurons into three main clusters.

In the process of the SOM algorithm, an SOM model is obtained in the form of a fan diagram as shown in Figure 3.

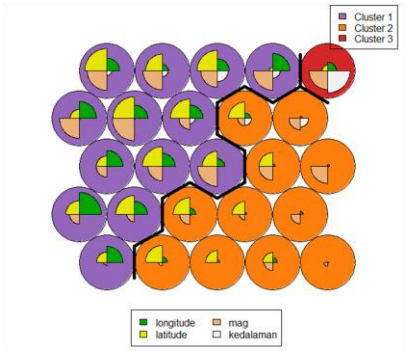

Figure 3 Fan Output Diagram

Based on Figure 3. obtained from the analysis using the SOM algorithm, we can see the clustering of neurons with a hexagonal topology that forms 3 clusters from 5275 earthquake data. In the diagram, there are 25 neurons divided into 3 clusters, where cluster 1 is marked in purple and has 13 neurons that are more dominant in the longitude, latitude and magnitude variables. Cluster 2 is marked in orange and has 11 neurons that are more dominant in the latitude and magnitude variables. Cluster 3, marked in red, has only 1 neuron with prominent characteristics in the depth and magnitude variables. This neuron clustering pattern illustrates the different characteristics between clusters based on their longitude, latitude, magnitude, and depth variables in representing the earthquake data analyzed.

The following are the clustering results of earthquake points in the Sulawesi and Maluku regions obtained with the help of the R program (Rstudio).

Table 3 Cluster Analysis Results

Cluster	Neuron List	Number of Members
1	Neuron 1, Neuron 6, Neuron 7, Neuron 11, Neuron 12, Neuron 13, Neuron 16, Neuron 17, Neuron 18, Neuron 21, Neuron 22, Neuron 23, Neuron 24	2457
2	Neuron 2, Neuron 3, Neuron 4, Neuron 5, Neuron 8, Neuron 9, Neuron 10, Neuron 14, Neuron 15, Neuron 19, Neuron 20	2732
3	Neuron 25	86

From the table, the number of each member of each cluster formed is known. The cluster division is obtained from the clustering results using the Kohonen SOM algorithm.

Table 4 Cluster Profilization

Variable	Mean		
	Cluster 1	**Cluster 2**	**Cluster 3**
Longitude	**127.78**	121.03	124.34
Latitude	**-0.88**	-4.25	-5.34
Magnitude	3.74	3.07	**4.33**
Depth	56.9	38.5	**469.3**

Based on the information contained in Table 4, each cluster can be interpreted as follows:

1. Cluster 1 represents earthquakes occurring mostly in the eastern part of Sulawesi and Maluku, characterized by moderate magnitude (≈3.74 SR) and moderate depth (≈56.9 km).
2. Cluster 2 corresponds to the central–southern region of Sulawesi, marked by lower magnitude (≈3.07 SR) and shallow depth (≈38.5 km).
3. Cluster 3 represents western and deeper regions, with high-magnitude (≈4.33 SR) and very deep earthquakes (≈469.3 km).

V. CONCLUSIONS AND SUGGESTIONS

Based on the results of the previous discussion, it can be concluded that the clustering results using SOM produce 25 neurons which are then divided into three optimal clusters based on the validity test with the smallest Connectivity Index value of 296.1512, the highest Silhouette Index of 0.3304, and the Dunn Index of 0.0058. Cluster 1 consists of 13 neurons with 2457 members, Cluster 2 consists of 11 neurons with 2732 members, and Cluster 3 consists of only 1 neuron with 86 members.

Future studies can expand this work by incorporating spatio-temporal earthquake patterns or integrating additional geophysical variables to improve clustering precision for disaster mitigation.

REFERENCES

[1] M. B. Siahaan and A. R. Rio, "Agglomerative Clustering of 2022 Earthquakes in North Sulawesi, Indonesia," *Buana Information Technology and Computer Sciences (BIT and CS*, vol. 4, no. 2, pp. 76–84, Jul. 2023, [Online]. Available: https://repogempa.bmkg.go.id/

[2] B. Rahman, T. W. Utami, and F. Fauzi, "Karakteristik Persebaran Kejadian Gempa Bumi di Pulau Sulawesi dan Maluku Berdasarkan Kovariat Gempa," *Prosiding Seminar Nasional Unimus*, vol. 4, no. 2, pp. 76–84, Jul. 2021.

[3] M. Faizan, M. F. Zuhairi, S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining: A Comparative Study," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 146–153, 2020.

[4] C. Gunawan and H. Sirait, "Analysis of Unemployment Clusters In Indonesia Using The Self Organizing MAP Method," *International Journal of Mathematics, Statistics, and Computing*, vol. 1, no. 3, pp. 8–15, 2023.

[5] Hartatik and A. Satria Dwi Cahya, "Clusterisasi Kerusakan Gempa Bumi di Pulau Jawa Menggunakan SOM," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 2, no. 02, pp. 25–34, 2020.

[6] A. R. Rayhani, W. Astuti, Z. Shufila, and E. Widodo, "Implementasi Self Organizing Map dalam Pengelompokkan Kabupaten di Jawa Barat berdasarkan Kasus Covid-19," *Jurnal Statistika dan Aplikasinya*, vol. 5, no. 2, 2021.

[7] S. Ilbeigipour, A. Albadvi, and E. Akhondzadeh Noughabi, "Cluster-Based Analysis of COVID-19 Cases Using Self-Organizing Map Neural Network and K-Means Methods to Improve Medical Decision-Making," *Inform Med Unlocked*, vol. 32, Jan. 2022, doi: 10.1016/j.imu.2022.101005.

[8] N. Nurul Halim and E. Widodo, "Clustering Dampak Gempa Bumi di Indonesia Menggunakan Kohonen Self Organizing Maps," *Seminar Nasional Integrasi Matematika dan Nilai Islami)*, vol. 1, no. 1, pp. 188–194, Jul. 2017.

[9] N. Ulinnuh and R. Veriani, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage dan Ward," vol. 5, 2020, doi: 10.30743/infotekjar.v5i1.2464.

[10] M. Khoncita Dasriana Bau, Y. Setyawan, and M. Titah Jatipaningrum, "Perbandingan Metode Algoritma K-Means Dan K-Medoids Pada Pengelompokan Kabupaten/Kota Di Provinsi Nusa Tenggara Timur Berdasarkan Dimensi Indeks Pembangunan Manusia Tahun 2020," *Jurnal Statistika Industri dan Komputasi*, vol. 08, no. 1, pp. 48–57, 2023, [Online]. Available: https://ntt.bps.go.id.

[11] F. Mayang Sari, R. Nur Hadiati, and W. Perinduri Sihotang, "Analisis Korelasi Pearson Jumlah Penduduk dengan Jumlah Kendaraan Bermotor di Provinsi Jambi," vol. 2, no. 1, p. 39, Jun. 2023, doi: 10.22437/multiproximity.v2i1.25568.

[12] M. Hernanda, A. Salma, D. Vionanda, and Z. Martha, "Penerapan Metode Self Organizing Maps (SOM) dalam Pengklasteran Berdasarkan Indikator Pemerlu Pelayanan Kesejahteraan Sosial (PPKS) Provinsi Jawa Barat," *UNP Journal of Statistics and Data Science*, vol. 1, no. 4, pp. 329–336, Aug. 2023, doi: 10.24036/ujsds/vol1-iss4/82.

[13] L. E. E. Awong and T. Zielinska, "Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification," *Sensors*, vol. 23, no. 18, Sep. 2023, doi: 10.3390/s23187925.

[14] V. Kotu and B. Deshpande, *Data Science: Concepts and Practice, Second Edition*. Morgan Kaufmann Publishers, 2019.

[15] S. M. Guthikonda, "Kohonen Self-Organizing Maps," 2005.

[16] S. I. Febianca and U. D. Wustqa, "Analisis Cluster Produksi Tanaman Perkebunan menggunakan Algoritma Self Organizing Map (SOM)," *Jurnal Kajian dan Terapan Matematika*, vol. 9, pp. 29–30, Mar. 2023.

[17] U. Asan and S. Ercan, *An Introduction to Self-Organizing Maps*, vol. 6. Istanbul: Atlantis Press Book, 2012.

[18] M. A. Nur, "Gempa Bumi, Tsunami Dan Migitasinya," *Jurnal Geografi*, vol. 7, no. 1, pp. 66–73, 2010.

[19] D. P. Utomo and B. Purba, "Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia," in *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, Sep. 2019, pp. 846–853.

[20] R. Wehrens and L. M. Buydens, "Self and Super-Organizing Maps in R: The Kohonen Package," *J Stat Softw*, vol. 21, pp. 1–19, 2007.