JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 6 No. 1 (2025)
https://doi.org/10.35877/454RI.jinav1571

Application of Cluster Analysis of Self Organizing Map (SOM) Method in the Community Literacy Development Index in Indonesia

Sanra Ariani*, Muhammad Nusrang, & Muhammad Kasim Aidid

Department of Statistics, Universitas Negeri Makassar, 90223, Makassar, Indonesia

Abstract

Self Organizing Map (SOM) is a method with a form of unsupervised learning, with Artificial Neural Network (ANN) training techniques that use a winner takes all basis, where only the neuron that is the winner will be updated. This study applies the cluster analysis of the SOM method in grouping provinces in Indonesia based on the characteristics of the Community Literacy Development Index (IPLM). The selection of the best cluster is based on internal validation i.e. connectivity, index Dunn and Silhouette. Based on the cluster validation results, 3 clusters were obtained that group provinces based on IPLM characteristics. of the 7 (seven) elements that make up the IPLM, 2 of them, namely energy and community visits, are shown in cluster 1. 5 other elements such as libraries, collections, SNP libraries, community involvement and library members are shown in cluster 3. Meanwhile, cluster 2 does not show significant IPLM-forming elements.

Keywords: cluster analysis, SOM, internal validation, IPLM, library.

Received: 17 January 2025 Revised: 30 May 2025 Accepted: 16 June 2025

1. Introduction

Cluster analysis is a method of grouping data (objects) based only on the information contained in the data by describing these objects and their relationships (Tan, 2006). Cluster analysis has a wide variety of methods, ranging from simple methods to complex methods. One of them is by using artificial intelligence such as Artificial Neural Network (ANN). Self Organizing Map (SOM) is one of the methods of ANN that can be used to cluster using unsupervised learning patterns (Setiani & Hakim, 2015).

As a method with a form of unsupervised learning, SOM does not require a special supervision so it is named self-organizing. The word map means that this method uses map in weighting input data. SOM is also commonly referred to as Self Organizing Feature Map, meaning that SOM uses the principle of "feature" or special characteristics in its basic principles that make it different from other methods (Guthikonda, 2005).

SOM is considered a spatial form of analysis of the K-Means group. The main advantage of SOM over K-Means is that SOM tends to get fewer branching results than using the K-Means algorithm. Another advantage is the acquisition of a topological sequence in which similar clusters are arranged together (Lobo, 2009).

Research related to the SOM method has been conducted by several previous researchers. As the research conducted by (Hafiludien & Istiawan, 2018) used the SOM method to obtain an overview of Persons with Social Welfare Problems (PMKS) which resulted in policy making in terms of setting targets and providing recommendations for interventions of persons with social welfare at the Central Java provincial level. In the study (Lumalessil et al., 2020) using the SOM method to group villages in Southwest Maluku Regency based on poverty characteristics and produced the best group with the smallest standard deviation ratio. Then the research conducted by (Apriliani et al., 2021) used the SOM method to group educational participation using two parameters, namely the rough participation rate (APK) and the pure participation rate (APM) which resulted in an accuracy of 81.25%.

* Corresponding author.
E-mail address: 1817141023@student.unm.ac.id

Based on the Programme for International Student Assessment (PISA) survey in 2015, it was noted that Indonesia was ranked 64th out of 72 countries with a literacy activity index of 37.32. During the period 2012-2015, the PISA score for reading in Indonesia only increased by 1 point from 396 to 397 (Tohir, 2016). The problem faced is in the form of improving and equalizing the quality of libraries. Therefore, it is necessary to develop and develop a library so that the impact felt is more pronounced and beneficial to the community. In implementing the library development program, it is necessary to identify based on the characteristics of the elements that make up the HDI of each province in order to make policies and strategies that are right on target and appropriate.

The SOM method was chosen because it is a visualization and analysis device for high-dimensional data into two-dimensional space. In addition, this method can also group category data as well as incomplete input data (Annas et al., 2007). Therefore, this study applies cluster analysis of the SOM method in grouping provinces in Indonesia based on IPLM characteristics.

2. Literature Review

2.1. Self Organizing Map (SOM)

Self Organizing Map (SOM) was first introduced by Teuvo Kohonen (1982), with an ANN training technique that uses a winner takes all basis, where only the neuron that is the winner will be updated. Although it uses an ANN base, SOM does not use the cluster target value, no cluster is assigned to each record. Characteristics like this then make SOM can be used for clustering purposes (ANN-based) (Prasetyo, 2012).

2.2. Distance Measures

Euclidean distance is one of the most frequently used measures of distance in cluster analysis to measure the distance between objects and cluster centers (Johnson & Wichern, 2002). Euclidean distance provides a straight distance between two pieces of data with N dimensions (Prasetyo, 2012). The formula for calculating the distance measure is as follows (Johnson & Wichern, 2002):

$$d_j = \sqrt{\sum_{i=1}^{n} \left(w_{ij} - x_i\right)^2} \tag{1}$$

i is the sum of the variables, d_j is the distance between the vector weight w_{ij} and the input vector x_i, w_{ij} is the weight of the j-th vector on the i-th variable, and x_i is the data of the input vector x on the i-th variable.

2.3. Cluster Validation

Cluster validation is the result of the evaluation procedure for cluster analysis quantitatively and objectively. There are three approaches to evaluate cluster validation (Hermayadi et al., 2013), namely external, internal and relative. The cluster validation used by researchers is internal validation, because this validation is based on the evaluation of clustering results where in quantitative concepts the IPLM data is in the form of quantitative data. There are several methods in internal validation, including the following (Susilowati et al., 2020):

a) Connectivity, forming the best number of clusters when the resulting value is smaller than the value of other clusters.

b) Dunn Index, produces the best cluster when the Dunn value obtained is greater and for Dunn Index values that are large or high, it explains that the cluster formed has been split between one cluster and another in an orderly and full or dense manner.

c) Silhouette, produces the best cluster when the value obtained is closer to the number 1. And for silhouette index values have values between -1 to 1 values.

3. Methods

The data used in this study is data from the publication of the 2021 IPLM Study which is sourced from the West Sumatera Provincial Bappeda website page.

The data were then analyzed using the SOM method with the following:

1) Initialization is in the form of a randomly obtained weight (w_{ij}). Once the weight (w_{ij}) is given, then the network is given input (x_i).
2) Specifies the parameters of the neighborly topology, learning rate and the number of epochs (iterations).
3) As long as the number of epochs (iterations) has not been reached, perform steps d and e.
4) For each input received, the network will perform the calculation of the distance of the d_j vector obtained by summing the difference between the weight vector (w_{ij}) and the input vector (x_i).
5) After the distance between the nodes is known, then the minimum value of the calculation of the vector distance d_j is determined, then the next stage performs a change in weight.

$$w_{ij}(new) = w_{ij}(old) + \alpha\left[x_i - w_{ij}(old)\right] \quad (2)$$

6) The process of obtaining a new weight requires a learning rate value () which is 0 1. The α≤α≤learning rate value at each epoch (iteration) will decrease to α (i+1) = 0.5 α.
7) The condition of stopping the test is carried out when it reaches a certain number of epochs (iterations). In addition, the termination of the test is carried out by calculating the difference between the weights of w_{ij} (new) and w_{ij} (old), if the test has reached convergence it can be stopped. Converging iterations when the mean distance to closest unit value is below 0.05.
8) Determine the best cluster using internal validation, namely connectivity, dunn index and silhouette. The best number of clusters is with the smallest connectivity value, the largest dunn index value and the silhouette value close to 1.
9) Perform an interpretation of the results of the formed clusters.

4. Result and Discussion

4.1. Result

The first step in the SOM method is to randomly initialize weights (w_{ij}) where $i = 7$ is the number of variables and $j = 3$ is the number of clusters. Random weights can be seen in Table 1 below.

Table 1. Random Weights

w_{ij}	j		
i	1	2	3
1	4	33	5
2	22	6	32
3	19	13	16
4	24	26	10
5	29	1	27
6	3	9	17
7	18	28	7

Second, data (provinces) are taken with input vectors (Table 2) which will be carried out the process of calculating the distance of the vector d_j.

Table 2. Input Vectors

Province	Library	Collection	Power	Visit	SNP	Community	Member
Central Java	2,964	1,178	1,885	1,010	0,257	4,242	2,277

Third, the calculation of the distance of the d_j vector obtained by summing the difference between the weight vector (w_{ij}) in Table 1 and the input vector (x_i) in the Table 2 so that random weights are obtained.

$$d_j = \sqrt{\sum_{i=1}^{n} \left(w_{ij} - x_i\right)^2}$$

$d_1 = \sqrt{(4-2{,}964)^2 + (22-1{,}178)^2 + (19-1{,}885)^2 + \cdots + (18-2{,}277)^2} = 48{,}281$ (random weight 1)

$d_2 = \sqrt{(33-2{,}964)^2 + (6-1{,}178)^2 + (13-1{,}885)^2 + \cdots + (28-2{,}277)^2} = 48{,}563$ (random weight 2)

$d_3 = \sqrt{(5-2{,}964)^2 + (32-1{,}178)^2 + (16-1{,}885)^2 + \cdots + (7-2{,}277)^2} = 46{,}201$ (random weight 3)

In this stage obtained distance d_3 as a random weight with a minimum value, which is 46.201.

Fourth, prosess to get a new weight requires a learning rate value (α) which is 0 1. $0 \leq \alpha \leq 1$. In this study, a learning rate of 0.05 was used. The learning rate value at each epoch (iteration) will decrease to α (i+1) = 0.5 α.

$w_{ij}(new) = w_{ij}(old) + \alpha\left[x_i - w_{ij}(old)\right]$

$w_{1,3} = 5 + 0{,}05(2{,}964 - 5) = 4{,}898$

$w_{2,3} = 32 + 0{,}05(1{,}178 - 32) = 30{,}459$

$w_{3,3} = 16 + 0{,}05(1{,}885 - 16) = 15{,}294$

$w_{4,3} = 10 + 0{,}05(1{,}010 - 10) = 9{,}550$

$w_{5,3} = 27 + 0{,}05(0{,}257 - 27) = 25{,}663$

$w_{6,3} = 17 + 0{,}05(4{,}242 - 17) = 16{,}362$

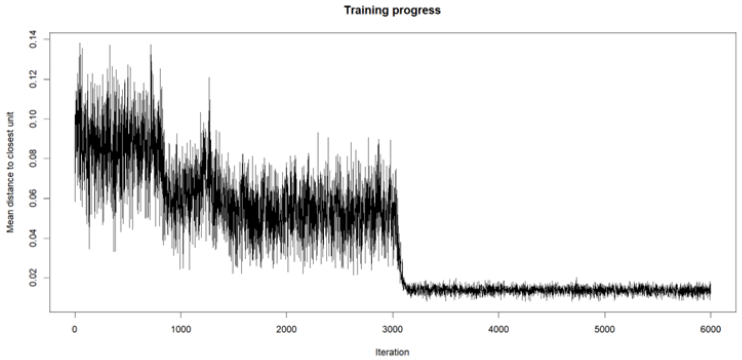
$w_{7,3} = 7 + 0{,}05(2{,}277 - 7) = 6{,}764$

So that a new weight is obtained as follows.

Table 3. New Weights

w_{ij}		j	
i	1	2	3
1	4	33	4,898
2	22	6	30,459
3	19	13	15,294
4	24	26	9,550
5	29	1	25,663
6	3	9	16,362
7	18	28	6,764

The training process is carried out repeatedly by calculating random weights to display a convergent iteration process. Convergent iteration is a number of processes carried out by software to get stable results. The mean distance to closest unit value starts to stabilize when it is below 0.05.



Figures 1. *Training Progress*

Next, to determine cluster best used validation internal that is connectivity, Dunn index and Silhouette.

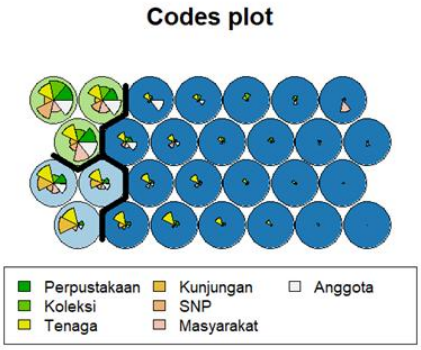
Table 4. Results of Grouping 2 Clusters

Number of Clusters	Connectivity	Dunn	Silhouette
2	7,9214	0,6135	**0,7101**
3	**7,6869**	**0,8341**	0,6982
4	29, 0119	0,0550	0,2142
5	25,9905	0,0670	0,2261

Based on the criteria of the best cluster which is with the smallest connectivity value, the Dunn index value is the largest and the Silhouette value is close to 1. So, this study used the number of clusters as many as 3 clusters to group provinces based on IPLM characteristics.

4.2. Discussion

Using the SOM method, a fan diagram is generated that shows the distribution of variables on the map. The fan diagram uses a hexagonal view with a 7x4 grid. The diagram is formed on the basis of the constituent elements of the IPLM.



Figures 2. *Fan* Diagram for 3 *Clusters*

In Figure 2 above, there are 3 different colors, which show the cluster results of each province. Each color has its own different characteristics. For more details can be seen in Table 5.

Table 5. Results of Grouping 2 Clusters

Cluster	Number of Members	Cluster Members
1	1	South Sulawesi
2	30	DKI Jakarta, South Kalimantan, Banten, Bali, Central Kalimantan, Bangka Belitung, Riau, South Sumatera, West Sumatera, West Nusa Tenggara, Southeast Sulawesi, North Sumatera, Aceh, Lampung, Riau Islands, West Kalimantan, DI Yogyakarta, East Kalimantan, Jambi, East Nusa Tenggara, Gorontalo, Maluku, Bengkulu, Papua, Central Sulawesi, North Maluku, West Papua, West Sulawesi, North Sulawesi and North Kalimantan
3	3	Central Java, West Java and East Java

From Table 5, it is known the number and each member of each formed cluster. The division of clusters is obtained from the results of clustering using the SOM algorithm. Furthermore, the average calculation of each cluster is carried out to recognize the characteristics of each cluster (profiling) based on the elements that make up the IPLM.

Table 6. Cluster Result Profiling

Cluster	Library	Collection	Power	Visit	SNP	Community	Member
1	8802	601374	1180	75560	3171	84879	305676
2	3520	494321	347	1400	694	56511	150687
3	24199	3083413	1068	26571	7700	679690	1111176

Cluster 1 consists of 1 member, namely South Sulawesi Province. Based on the average results, it shows that the area in cluster 1 has the most prominent IPLM forming elements with an average library power of 1,180 people and community visits with an average of 75,560 visits per day.

Cluster 2 consists of 30 members, namely DKI Jakarta, South Kalimantan, Banten, Bali, Central Kalimantan, Bangka Belitung, Riau, South Sumatera, West Sumatera, West Nusa Tenggara, Southeast Sulawesi, North Sumatera, Aceh, Lampung, Riau Islands, West Kalimantan, DI Yogyakarta, East Kalimantan, Jambi, East Nusa Tenggara, Gorontalo, Maluku, Bengkulu, Papua, Central Sulawesi, North Maluku, West Papua, West Sulawesi, North Sulawesi and North Kalimantan. Based on the average results, it shows that the area in cluster 2 has the IPLM forming element with the greatest impact compared to other clusters. This is because the areas in cluster 2 have the lowest average value compared to other clusters of all elements that make up the IPLM, such as the lack of libraries, collections, library personnel, community visits, SNP libraries, community involvement and library members.

Cluster 3 consists of 3 members, namely the provinces of Central Java, West Java and East Java. Based on the average results, it shows that the areas in cluster 3 have the most prominent IPLM forming elements with an average library of 24,199 libraries, 3,083,413 collections, 7,700 SNP libraries, 679,690 people involved in socialization and 1,111,176 library members.

5. Conclusion

Based on the results of data analysis and discussions that have been carried out, the result of cluster validation with connectivity methods, Dunn index and Silhouette states that the best grouping of IPLM forming elements is 3 clusters. In cluster 1, there is 1 member who shows the elements that make up the IPLM in the form of manpower and community visits. Cluster 2 has 30 members that do not show significant characteristics and have the lowest average value of all IPLM-forming elements compared to other clusters. While cluster 3 has 3 members who show the elements that make up the IPLM in the form of many libraries, collections, SNP libraries, community involvement and library members.

References

Annas, S., Kanai, T., & Koyama, S. (2007). PCA and SOM for Visualizing and Classifying Fire Risks in Forest Regions. *Agricultural Information Research, 16*(2), 44-51.

Apriliani, D. P., Khairat, U., & Syarli. (2021). Pengelompokan Partisipasi Pendidikan Menggunakan Metode Self Organizing Maps. *Jurnal Ilmiah Maju, 4*(2), 14-20. Retrieved from https://ojs.balitbang.sulbarprov.go.id/index.php/maju/article/view/102

Guthikonda, S. M. (2005). *Kohonen Self-Organizing Maps.* Wittenberg University.

Hafiludien, A., & Istiawan, D. (2018). Penerapan Algoritma Self Organizing Maps Untuk Pemetaan Penyandang Kesejahteraan Sosial (PMKS) di Provinsi Jawa Tengah Tahun 2016. *Proceeding of The 7th University Research Colloquium 2018*, (pp. 84-92).

Hermayadi, R., Purnomo, H. M., & Purnama, I. E. (2013, Februari 2). Pengelompokan Data Kordinat BTS Menggunakan K-Means dan Validasi Berbasis Google Map. (pp. 1-8). Prosiding Seminar Nasional Manajemen Teknologi XVII.

Johnson, & Wichern. (2002). *Applied Multivariate Analysis* (5th ed.). New Jersey: Prentice Hall.

Lobo, V. J. (2009). *Application of Self-Organizing Maps to the Maritime Environment.* Springer Berlin Heidelberg.

Lumalessil, F. L., Tomasouw, B. P., & Rijoly, M. E. (2020, November). Pengelompokan Desa di Kabupaten Maluku Barat Daya Berdasarkan Karakteristik Kemiskinan Menggunakan Metode Self Organizing Maps (SOM). *Zeta - Math Journal, 5*(1), 16-20.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Andi.

Setiani, D., & Hakim, R. F. (2015). Clustering Indikator Pembangunan Berkelanjutan di Indonesia Menggunakan Algoritma Self-Oragnizing Maps (SOMs) Kohonen). *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS*, 614-628.

Susilowati, T., Sugiarto, D., & Mardianto, I. (2020). Uji Validasi Algoritma Self-Organizing Map (SOM) dan K-Means untuk Pengelompokan Pegawai. *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi), 4*(6), 1171-1178. doi:https://doi.org/10.29207/resti.v4i6.2492

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson Addison-Wesley.

Tohir, M. (2016). Hasil PISA Indonesia Tahun 2015 Mengalami Peningkatan. *10*, 1-2. doi:10.17605/OSF.IO/KX4JV.