JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e–ISSN: 2614–8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 19, No. 1, September 2022, pp. 79 - 89
DOI: 10.20956/j.v19i1.20882

Clustering Regencies/Cities in Kalimantan Island Based on Poverty Indicators using Agglomerative Hierarchical Clustering (AHC)

Ludia Ni'matuzzahroh[1], Andrea Tri Rian Dani[2], Narita Yuri Adrianingsih[3]

[1]Institut Teknologi Sepuluh Nopember, [2]Universitas Mulawarman, [3]Universitas Tribuana Kalabahi

***Email address**: [1]ludianimatuzzahroh@gmail.com, [2]andreatririandani98@gmail.com, [3]naritayuria98@gmail.com*

Abstract

Cluster analysis is a statistical analysis that can group objects of observation into several groups/clusters based on their similarity of characteristics. The grouping into several clusters is based on the information contained in the object under study. A cluster can be said to be good if it has high internal homogeneity and high external heterogeneity. The clustering method used in this study is the agglomerate hierarchical clustering (AHC) method, where the cluster formation algorithm used in this AHC method is average linkage, single linkage, complete linkage, and ward. Cluster analysis using the AHC method will be applied to poverty indicator data for Regencies/Cities in Kalimantan Island, which consists of several variables. This study aims to obtain the optimal results of grouping Regencies/Cities in Kalimantan Island, with the number of clusters that have been determined at the beginning, namely as many as 3 clusters. Based on the results of the analysis using the AHC method, the ward algorithm produces an agglomerate coefficient value of 0.89, where this value is close to 1, which means that the ward algorithm is the best in clustering Regencies/Cities in Kalimantan Island.

Keywords: *Agglomerative Hierarchical Clustering, Cluster, Poverty*

1. INTRODUCTION

In statistical analysis, there is one technique that aims to group the objects of observation into several groups based on their characteristics or similarities, which is called cluster analysis [9]. A cluster can be said to be good if the cluster has high homogeneity between members in one cluster and high heterogeneity between one cluster and another [16]. Cluster analysis is generally divided into two, namely hierarchical clustering methods and non hierarchical clustering methods. Then, in the hierarchical clustering method itself there are two parts, including the agglomerate (union) and divisive (spreading) methods. Agglomerative Hierarchical Clustering (AHC) is a better method in view of the connectivity relationship that fixes an object to one class, for similiar iterations [12]. Several methods of cluster formation in the agglomerate method,

including average linkage, single linkage, complete linkage, and ward. Several previous studies have developed cluster analysis, including [7], [8], [14], [15], [16], [19].

Cluster analysis, especially the Agglomerate Hierarchical Clustering (AHC) method, can be applied to various fields [15], one of which is related to social problems, especially poverty indicators. Poverty is one of the problems that the Indonesian government always pays attention to every year. Based on BPS data in 2021, the majority of Regencies/Cities in Kalimantan Island produced the lowest poverty rates compared Regencies/Cities on other islands, which is around 5% to 10% [1]. Even though it is classified as an island with the lowest poverty rate, there are also several Regencies/Cities that have high poverty rate, so that in this case the Regencies/Cities on the island of Kalimantan can be grouped based on poverty indicators. Based on this, what can be used as benchmarks for grouping Regencies/Cities on the island of Kalimantan are the average length of schooling, life expectancy, the open unemployment rate, the percentage of the poor, a number of households with access to proper sanitation, and a number of households with access to adequate sanitation.

Based on the description above, this study will discuss cluster analysis using the agglomerate hierarchical cluster method which is applied in clustering Regencies/Cities on the island of Kalimantan based on poverty indicators. The purpose of this study is to obtain optimal clustering results in the process of clustering Regencies/Cities on the island of Kalimantan based on poverty indicators.

2. PRELIMINARIES

2.1 Cluster Analysis

Cluster analysis is one of the tools used in the data mining process, which aims to identify homogeneous objects into groups called clusters. In this cluster analysis, identification is carried out on a group of objects that have similar characteristics that can be separated from other groups, and not to correlate one object with another object. The cluster formed has high internal homogeneity and external heterogeneity [9]. The number of clusters that can be identified depends on the number and variety of data objects [17]. A cluster can be said to have good results if it shows the following characteristics [18]:

1. High homogeneity between members in the same cluster (within-cluster)
2. High heterogeneity between one cluster and another (between-cluster).

2.2 Similarity Measure

The main purpose in identifying observational objects into a group/cluster is to find out how close an observation is and how far apart the observations are from one another. Two objects can be said to be close if the distance between the two objects are small or has a large value.

2.3 Agglomerative Hierarchical Clustering

One method that can be used in cluster analysis is the hierarchical clustering method, where the grouping method is done by building a group hierarchy. In general, the strategy in hierarchical grouping can be divided into two, agglomerative algorithm (bottom-up) and divisive algorithm (top-down). This paper focuses on the agglomerative algorithm (bottom-up). This agglomeration method is usually used in the social and economic fields of society [1]. The steps of the agglomerate hierarchical clustering method are [18]:

1. There are N clusters, where each cluster contains a single entity and a symmetric matrix $N \times N$ of distance by $\boldsymbol{D} = \{d_{ik}\}$
2. Find the distance matrix for the most similar (closest) cluster pairs, for example the distance between the most similar U and V clusters is written with the notation d_{UV}. The distance matrix can be using the Euclidean distance formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{2.1}$$

   where,
   d_{ij} : distance between object *i* and *j*
   x_{ij} : the value of object *i* in the *k* variable
   x_{jk} : the value of object *j* in the *k* variable
   p : number of observed variables
3. Merge clusters *U* and *V*. Gave the label from the newly formed cluster with (*UV*). And then updating the entries in the distance matrix by:
   a) Delete the rows and columns corresponding to clusters *U* and *V*
   b) Add a row and column giving the distances between the cluster (*UV*) and the remaining clusters.
4. Repeat for steps 2 and 3 until $N - 1$ times (all objects will be in a single cluster after the algorithm ends). Note the identity of the merged cluster and the levels (distance or similarity) at which the union is placed.

Grouping technique on agglomerate hierarchical clustering used in this study are as follows [9]:

1. Single linkage
   Single linkage is a grouping procedure based on the smallest/minimum distance between objects. The single linkage algorithm begins by selecting the smallest distance in the matrix $\boldsymbol{D} = \{d_{ij}\}$, then combining the corresponding objects, for example *U* and *V* to obtain a cluster (*UV*). The next step is to find the distance between (*UV*) and other clusters, for example *W*, it can be written as follows:

$$d_{(UV)W} = min(d_{UW}, d_{VW}) \tag{2.2}$$

   where d_{UW} is the closest distance from clusters *U* and *W*, and d_{VW} is the closest distance from clusters *V* and *W*.

2. Complete linkage
   Complete linkage is a clustering process based on the largest/maximum distance between objects. The complete linkage algorithm starts by selecting the largest distance in the matrix $\boldsymbol{D} = \{d_{ij}\}$, then merging the corresponding objects, for example *U* and *V* to obtain clusters

(*UV*). The next step is to find the distance between (*UV*) and other clusters, for example *W* so that it can be written as the following equation:

$$d_{(UV)W} = max(d_{UW}, d_{VW}) \tag{2.3}$$

where d_{UW} is the furthest distance from clusters *U* and *W*, and d_{UW} is the farthest distance from clusters *V* and *W*.

3. Average linkage
   Average linkage is a clustering process based on the average between objects. The average linkage algorithm starts by defining the matrix $\boldsymbol{D} = \{d_{ij}\}$ to obtain the closest object, for example *U* and *V*, then these objects are combined into clusters (*UV*) then the distance between (*UV*) and other clusters are *W*, so it can be written as equation following:

$$d_{(UV)W} = \frac{d_{(UW)} + d_{(VW)}}{n_{(UV)}n_W} \tag{2.4}$$

   where $n_{(UV)}$ is the number of members in the cluster (*UV*), and n_W is the number of members in the cluster *W*.
4. *Ward method*
   Ward method is a clustering process that seeks to minimize variations between objects in one cluster, or it can also be called the minimum variance method [10]. The ward method algorithm begins by defining the matrix $\boldsymbol{D} = \{d_{ij}\}$ to get the most similar objects, for example *U* and *V*, then these objects are combined into clusters (*UV*) then the distance between (*UV*) and other clusters are *W*, so it can be written as the following equation:

$$d_{(UV)W} = \frac{\left[(n_W + n_U)d_{(UW)} + (n_W + n_V)d_{(VW)}\right] - n_W d_{(UV)}}{n_W + n_{(UV)}} \tag{2.5}$$

   The cluster formation procedure will continue to iterate, until all objects are joined in the specified number of clusters.

2.3 Cluster Validity Test

After obtaining the cluster results from the clustering process, then perform cluster validity tests. This clusters validity test needs to be done with the aim of seeing the goodness or quality of the results of the cluster analysis. In this study, a measure that can be used to test the validity of the results of clustering with the hierarchical method is to use the agglomerate coefficient. Agglomerative coefficients measures the dissimilarity of an object from clustering in the first cluster, divided by the dissimilarity in the final grouping in the cluster analysis, which is averaged by the total sample [13].

3. MAIN RESULTS

3.1 Descriptive Statistics

This study uses variables regarding poverty indicators on the island of Kalimantan, which consist of the average length of schooling, life expectancy, the open unemployment rate, the percentage of the poor, a number of households with access to proper sanitation, and a number of households with access to proper drinking water. The data on the research variables were obtained from the recapitulation of the Badan Pusat Statistik (BPS) [2]-[6]. Descriptive statistical analysis can be used to see the initial pattern of the data for each variable used. More details on the results of descriptive statistical analysis are shown in Table 3.1.

Table 3.1. Descriptive statistics of research variables

Variable	Minimum	Maximum	Mean	Variance
Average Length of School	5.15	11.53	8.34	1.70
Life expectancy	64.10	74.76	70.75	5.96
Open Unemployment Rate	2.30	12.38	4.99	4.10
Percentage of Poor Population	2.89	12.01	6.29	5.28
Percentage of Households with Access to Adequate Sanitation	49.23	97.17	79.80	127.34
Percentage of Households with Access to Adequate Drinking Water	48.85	99.85	77.13	210.96

3.2 Spatial Distribution Mapping

Spatial distribution mapping is used as a visualization of the spatial distribution of each variable used. Spatial distribution mapping of research variables is shown in Figure 3.1 to Figure 3.6.

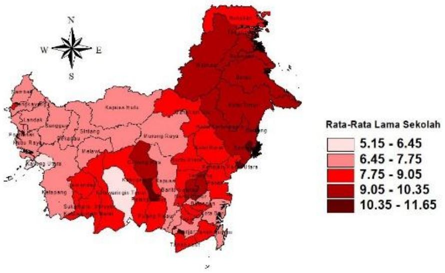


Figure 3.1. Spatial distribution mapping variable Average Length of School

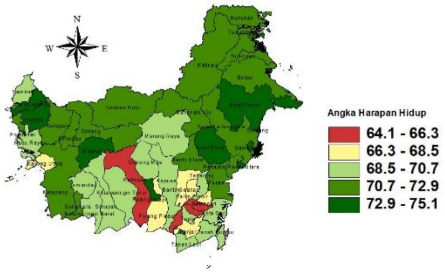


Figure 3.2. Spatial distribution mapping variable Life Expectancy

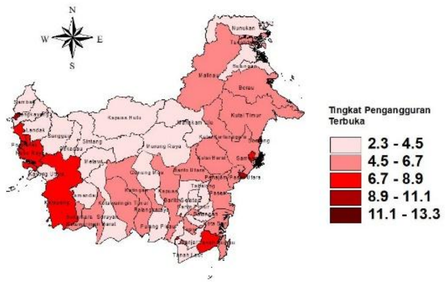


Figure 3.3. Spatial distribution mapping variable Open Unemployment Rate

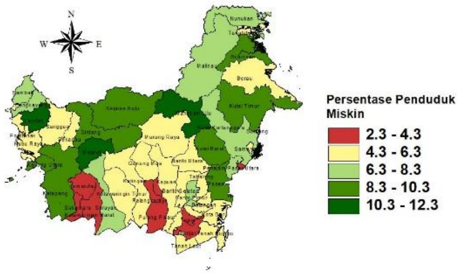


Figure 3.4. Spatial distribution mapping variable Percentage of Poor Population

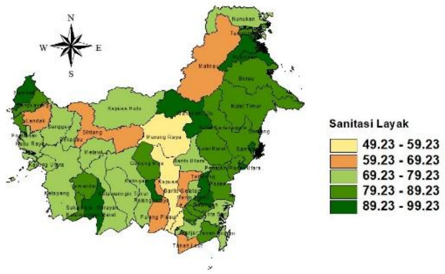


Figure 3.5. Spatial distribution mapping variable Percentage of Households with Access to Proper Sanitation

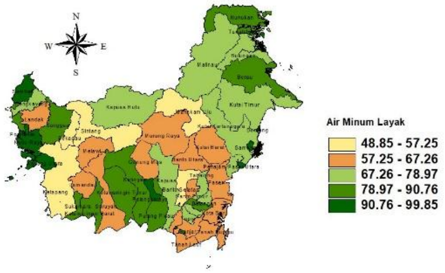


Figure 3.6. Spatial distribution mapping variable Percentage of Households with Access to Proper Drinking Water

For example, for the average length of schooling variable in Figure 3.1, based on the mapping results, it is known that there is still a serious gap between the average length of schooling at the district and city levels. It is of particular concern that there is a need for equitable access and improvement in education sector in some areas that have low average years of schooling. Furthermore, in Figure 3.2, the variable life expectancy. Based on the results of the mapping, it is known that there are still some areas on the island of Kalimantan where life expectancy is low. This can be seen from the red areas.

In Figure 3.3, the variable unemployment rate is open. Based on the mapping results, it is known that the percentage of the open unemployment rate in many districts/cities on the island of Kalimantan is low, this is indicated by many areas that are white and pink. However, it remains a concern that there are still some districts/cities that are red and need special attention. In Figure 3.4, the variable percentage of the poor population. Based on the results of the mapping, it is known that there are districts/cities on the island of Kalimantan that have a high percentage of poor people. This is indicated by the dark green area. We can see that there is still a serious gap between the percentage of poor people at the district and city levels. Of course, it is a special concern that there is a need for equity in various sectors to reduce the percentage of the poor.

In Figure 3.5, the variable a number of households with access to proper sanitation and Figure 3.6, the variable a number of households with access to proper drinking water. Based on the mapping results, it can be seen that the percentage of access to proper sanitation and proper drinking water in many districts/cities on the island of Kalimantan is already high, this is indicated by many areas that are lighted green and dark green. But keep in mind, that there are still some areas that are colored orange.

3.3 Clustering using the Agglomerative Hierarchical Clustering (AHC)

The process of clustering Regencies/Cities on the island of Kalimantan uses Agglomerative Hierarchical Clustering (AHC) with 4 tried algorithms, namely: single linkage, complete linkage, average linkage, and ward. The clustering process is carried out using R software with the packages "tidyverse", "cluster", and "factoextra" and then the Agglomerative Coefficient value is calculated which is shown in Table 3.2. Agglomerative Coefficient value close to 1 indicates that the clustering algorithm is the best in clustering an object.

Table 3.2. The results of the agglomerative coefficient calculation

Algorithms	Agglomerative Coefficient
Average Linkage	0.68
Single Linkage	0.41
Complete Linkage	0.80
Ward	**0.89**

Based on Table 3.2, it can be seen that the ward algorithm is a clustering algorithm that has an Agglomerative Coefficient value close to 1, which is 0.89. Also shown is the dendogram of the grouping results for each algorithm used in Figures 3.7 to 3.10.

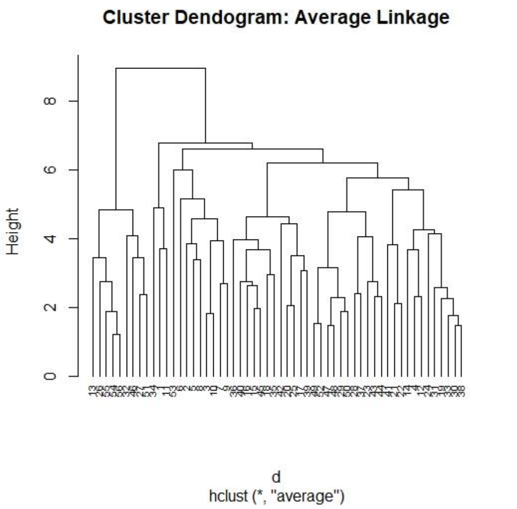


Figure 3.7. Average linkage algorithm

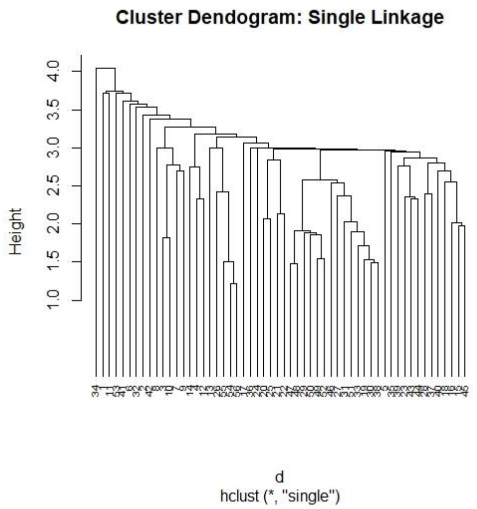
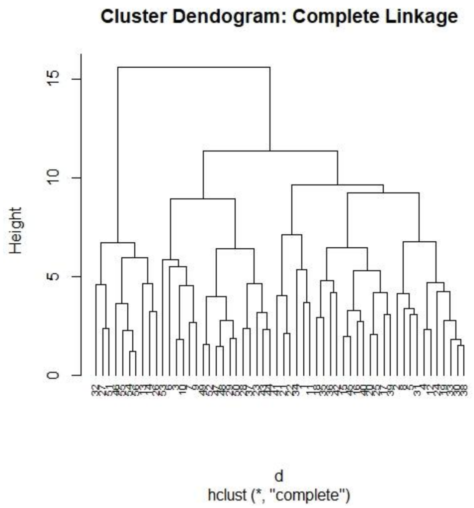


Figure 3.8. Single linkage algorithm



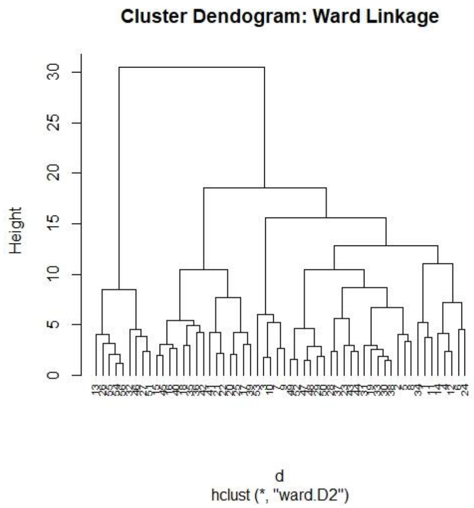
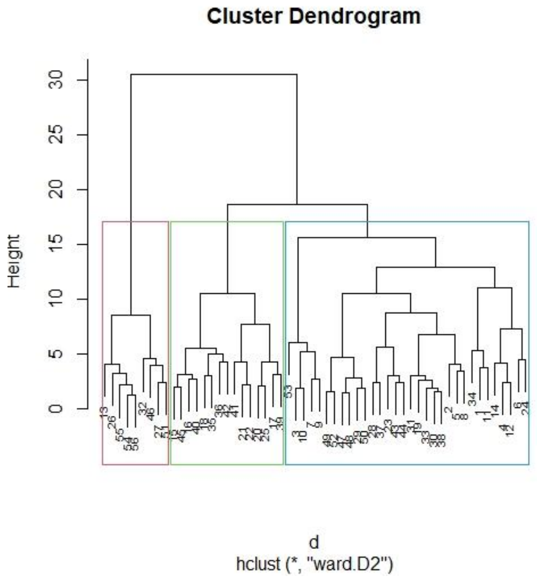


Figure 3.9. Complete linkage algorithm

Figure 3.10. Ward algorithm

Based on the Agglomerative Coefficient in Table 3.2, the best clustering algorithm is the ward algorithm. In this study, the number of clusters has been determined at the beginning, namely 3, so that 3 clusters will be formed based on the results of clustering using the best algorithm, in this case the ward algorithm. The dendogram of the grouping results and the cluster plot based on the ward algorithm are shown in Figure 3.11 and Figure 3.12



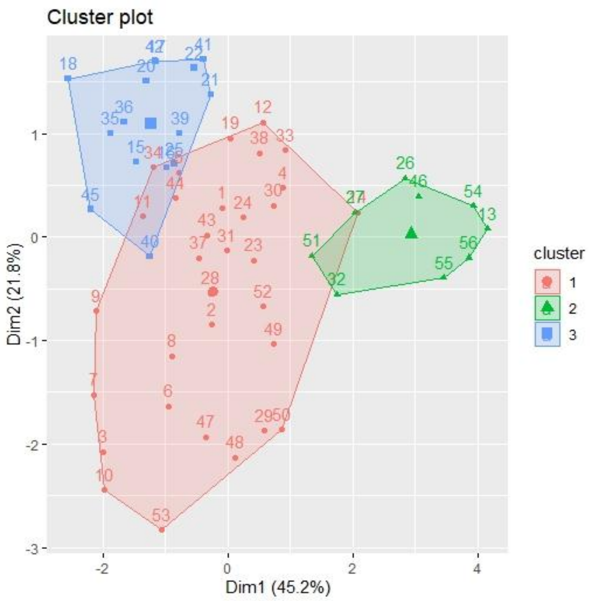


Figure 3.11. Dendrogram of grouping results based on the Ward algorithm

Figure 3.12. Ward algorithm cluster plot

The profiling of Regencies/Cities results in Kalimantan Island based on Figure 3.12 is presented in Table 3.3.

Table 3.3. Profiling of Regencies/Cities cluster results on the island of Kalimantan

No.	Cluster	Regency/City
1.	*Cluster* 1	Sambas, Bengkayang, Landak, Mempawah, Sanggau, Ketapang, Sintang, Kapuas Hulu, Sekadau, Melawi, Kayong Utara, Kubu Raya, Singkawang, Tapin, Tabalong, Tanah Bumbu, Malinau, Bulungan, Tana Tidung, Nunukan, Kotawaringin Barat, Kotawaringin Timur, Barito Utara, Sukamara, Gunung Mas, Barito Timur, Paser, Kutai Barat, Kutai Kartanegara, Kutai Timur, Penajam Paser Utara, Mahakam Ulu
2.	*Cluster* 2	Pontianak, Banjarmasin, Banjar Baru, Tarakan, Palangka Raya, Berau, Balikpapan, Samarinda, Bontang
3.	*Cluster* 3	Tanah Laut, Kota Baru, Banjar, Barito Kuala, Hulu Sungai Selatan, Hulu Sungai Tengah, Hulu Sungai Utara, Balangan, Kapuas, Barito Selatan, Lamandau, Seruyan, Katingan, Pulang Pisau, Murung Raya,

Based on Table 3.3, it can be seen that there are 32 regencies/cities included in cluster 1. In cluster 2 there are 9 regencies/cities, while in cluster 3 there are 15 regencies/cities.

Table 3.4. Characteristics for each cluster

Variable	Average		
	Cluster 1	Cluster 2	Cluster 3
Average Length of School	7.90	10.50	7.94
Life expectancy	71.47	73.33	67.66
Open Unemployment Rate	4.79	7.78	3.76
Percentage of Poor Population	7.25	4.74	5.19
Percentage of Households with Access to Adequate Sanitation	80.13	91.08	72.33
Percentage of Households with Access to Adequate Drinking Water	74.57	95.77	71.42

Based on Table 3.4, it is known that the regencies/cities in Cluster 1 are regencies/cities that need special attention in several sectors or indicators. Regencies/cities located in Cluster 1 tend to have a low average length of schooling and a high percentage of the poor. Cluster 2 is dominated by cities, some of which later become provincial capitals. It can be seen that in cluster 2, the thing to note is the open unemployment rate which is very high compared to other clusters. Furthermore, in Cluster 3, there are several indicators that also need attention. Regencies/cities that are members of Cluster 3 tend to have a low life expectancy. A number of households with access to proper sanitation and access to proper drinking water also needs to be a warning because it has low average compared to cluster 1 and cluster 2. Based on the characteristics for each cluster, of course this can be input for the Government in overcoming and prioritizing equitable development based on priority scale. The visualization of the results of the grouping of regencies/cities is shown in the form of spatial mapping in Figure 3.13.

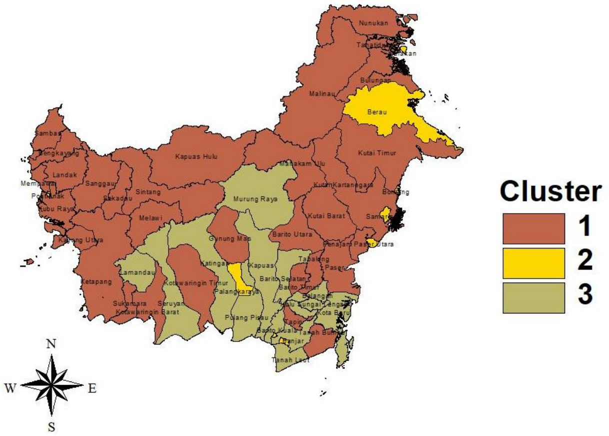


Figure 3.13. Visualization of regencies/cities clustering results

4. CONCLUSION

The process of clustering regencies/cities on the island of Kalimantan based on poverty indicators using the Agglomerative Hierarchical Clustering (AHC) has been successfully carried out. Based on the analysis, it is found that the ward algorithm is optimal with the agglomerative coefficient value of 0.89. From the three clusters that were formed, which was then profiled the cluster results, we can find out which regencies/cities need special attention in improving poverty indicators.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest

REFERENCES

[1] BPS., 2021. *DATA DAN INFORMASI KEMISKINAN KABUPATEN/KOTA*. Badan Pusat Statistik, Jakarta.

[2] BPS Provinsi Kalimantan Barat., 2021. Kalimantan Barat dalam Angka 2021. Badan Pusat Statistik, Kalimantan Barat.

[3] BPS Provinsi Kalimantan Selatan., 2021. Kalimantan Selatan dalam Angka 2021. Badan Pusat Statistik, Kalimantan Selatan.

[4] BPS Provinsi Kalimantan Tengah., 2021. Kalimantan Tengah dalam Angka 2021. Badan Pusat Statistik, Kalimantan Tengah.

[5] BPS Provinsi Kalimantan Timur., 2021. Kalimantan Timur dalam Angka 2021. Badan Pusat Statistik, Kalimantan Timur.

[6] BPS Provinsi Kalimantan Utara., 2021. Kalimantan Utara dalam Angka 2021. Badan Pusat Statistik, Kalimantan Utara.

[7] Dani, A. T. R., Wahyuningsih, S., and Rizki, N. A., 2019. Penerapan Hierarchical Clustering Metode Agglomerative Pada Data Runtun Waktu. *Jambura Journal of Mathematics*, Vol. 1, No. 2, 64-78.

[8] Dani, A. T. R., Wahyuningsih, S., and Rizki, N. A., 2020. Pengelompokan Data Runtun Waktu Menggunakan Analisis Cluster (Studi Kasus: Nilai Ekspor Komoditi Migas dan Nonmigas Provinsi Kalimantan Timur Periode Januari 2000-Desember 2016). *Jurnal EKSPONENSIAL*, Vol. 11, No. 1, 29-37.

[9] Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E., 2010. *Multivariate Data Analysis*, 7th Edition. Pearson Prentice Hall, New Jersey.

[10] Jain, A.K. & Dubes, R.C., 1988. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, New Jersey.

[11] Johnson, R.A. & Wichern, D.W., 2007. *Applied Multivariate Statistical Analysis*, 6th Edition. Prentice Education, Inc., New Jersey.

[12] Kamalha, E., Kiberu, J., Nibikora., I., Mwasiagi, J. I., and Omollo, E., 2017. Clustering and Classification of Cotton Lint Using Principle Component Analysis, Agglomerative Hierarchical Clustering, and K-Means Clustering. *Journal of Natural Fibers*, 1-11. DOI: 10.1080/15440478.2017.1340220

[13] Kaufman, L. & Rousseeuw, P.J., 1990. *Finding Groups in Data an Introduction to Cluster Analysis*. John Wiley & Sons Inc Publication, New Jersey.

[14] Liu, N., Xu, Z., Zeng, X. J., and Ren, P., 2021. An Agglomerative Hierarchical Clustering Algorithm for Linear Ordinal Rankings. *Information Sciences*, 170-193.

[15] Novidianto, R. and Dani, A. T. R., 2020. Analisis Klaster Kasus Aktif Covid-19 Menurut Provinsi di Indonesia Berdasarkan Data Deret Waktu. *Journal of Statistical Application and Computational Statistics*, Vol. 12, 15-24.

[16] Nurseptiani, A., Satria, Y., and Burhan, H., 2020. Application of Agglomerative Hierarchical Clustering to Optimize Matching Problems in Ridesharing for Maximize Total Distance Savings. *Journal of Physics: Conference Series*, 1-7. DOI: 10.1088/1742-6596/1821/1/012016

[17] Prasetyo, E., 2012. *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. Penerbit Andi, Yogyakarta.

[18] Santosa, B., 2007. Data Mining: Teknik Pemanfaatan Data untuk Memprediksi Kriteria Nasabah Kredit. *Jurnal Komputer dan Informatika*, Vol. 1, No. 1, 53-57.

[19] Zahrotun, L., 2015. Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan Metode Clustering K-Means dan Agglomerative Hierarchical Clustering (AHC). *JURNAL INFORMATIKA*, Vol. 9, No. 1, 1039-1047.