Jurnal Matematika Vol. 20, No. 2, November 2021
ISSN: 1412-5056 / 2598-8980
http://ejournal.unisba.ac.id

Diterima: 30/09/2021
Disetujui: 04/12/2021
Publikasi Online: 07/12/2021

Application of Cluster Analysis Using Hierarchic Method for Classification of Municipalities/Cities in Kalimantan Based on Socio-Economic Variables

Ananto Wibowo[1], M. Rismawan Ridha[2]

[1]BPS Kabupaten Ciamis
[2]BPS Kabupaten Maluku Tengah

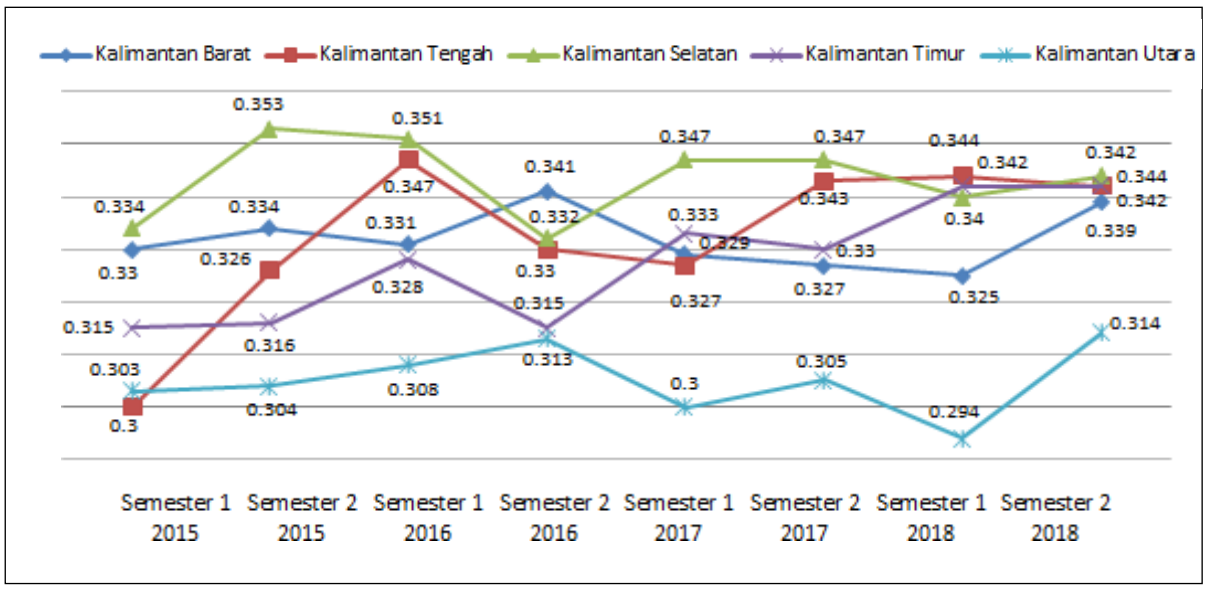
ananto.wibowo@bps.go.id[1], rismawan.ridha@bps.go.id[2]

Abstract. All provinces in Kalimantan show inequality level of income distribution that fluctuate, but some have increased from 2015 to 2018. This study aims to classify the regencies and municipalities based on the conditions of social and economic variables with hierarchical cluster analysis to minimize the disparities between regions. The data source comes from Statistics Indonesia (BPS) in each province of Kalimantan in 2018. The results show a correlation between variables, so that the Mahalanobis Distance application is more appropriate to use than Euclidean Square Distance. We formed three groups of regencies and municipalities based on the cluster analysis results. The first group (the good economic group) consisted of 48 observations, the second group (the good economic and social group) consisted of 7 observations, and there was only one observation in the third group (the good social group). Based on empirical research, we should optimize the focus of development in regencies and municipalities in the first and third groups.

Keywords: Cluster Analysis, Inequality, Economic, Social, Mahalanobis Distance

1. Introduction

Economic inequality refers to disparities among individuals' incomes and wealth, where there are significant differences between regions vertically and horizontally [6]. To help measure inequality, Statistics Indonesia (BPS) releases the Gini Index every semester at the provincial level and annually at the municipalities level. This indicator is used to answer one of the 17 major objectives discussed in the Sustainable Development Goals (SDGs). This includes reducing inequality within and between countries, with one target being to empower and encourage social, economic, and political equality for all.



(Source: Statistics Indonesia)

Figure 1. Trends of Gini coefficient in Kalimantan (semester 1 of 2015 - semester 2 of 2018)

Based on data from Statistics Indonesia, all provinces on Kalimantan Island show a level of inequality in expenditure distribution that fluctuates, but some have increased in the last four years. For instance, Kalimantan Tengah Province had a Gini coefficient of 0.3 in the first semester of 2015 and then increased significantly to 0.344 in the second semester of 2018. This increase was also experienced by Kalimantan Timur and Kalimantan Utara, as shown in Figure 1.

At the same time, the economic growth rate in several provinces has also increased, as experienced by Kalimantan Utara and Kalimantan Timur. Even Kalimantan Utara Province in 2015 grew by -1.2 percent then increased to 2.67 percent (Source: Statistics Indonesia). This situation slightly contrasts with the current state of inequality, where the increasing economic growth is offset by an increase in the value of the Gini Index. The development was unsuccessful if an increase in the standard of living and an uneven level of welfare does not follow it. This circumstance is quite reasonable because the problems that arise when inequality widens can trigger conflicts, economic inefficiency weakens social stability, and inequality is seen as unfair [8].

Based on the description above, it is important to conduct research that aims to narrow the gaps between regions. Therefore, this study aims to group municipalities/cities in Kalimantan using cluster analysis based on the conditions of social and economic variables so that development is more focused.

2. Research Method

Cluster analysis is a multivariate technique that is often used in various research and is applied in many domains, especially marketing applications, both for grouping people, products, and things that require advanced study [5]. This analysis aims to group n observation to form k homogenous cluster based on p variable [1]. Observations in one cluster are expected to have similar characteristics but differ from observations in other clusters.

In the clustering technique, it is necessary to have a distance for each pair of objects *i*-th and *j*-th. This step primarily focuses on quantifying their degree of dissimilarity. The most popular choices are Euclidean, Mahalanobis, Minkowski Metric, Canberra Metric, and Czekanowski Metric [2]. Suppose there are two observations with p dimensional, namely $x^T = [x_1, x_2, \dots, x_p]$ and $y^T = [y_1, y_2, \dots, y_p]$, hence, the Euclidean distance is as follows:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2} = \sqrt{((x-y)^T(x-y)} \quad (1)$$

Euclidean square distance is defined by:

$$d^2(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2 = (x-y)^T(x-y) \quad (2)$$

The Euclidean square distance requires no correlation between variables with the same unit size, where the results of standardized data have a mean of zero and a standard deviation of one [2]. This distance is rather difficult to apply because, in the case of many variables (multivariate), the existence of correlation is likely to occur. We can measure the correlation using the following formula [3]:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2)\,(\frac{1}{n}\sum_{i=1}^{n} {y_i}^2 - \bar{y}^2)}} \quad (3)$$

Where n is number of observations, while $\bar{x}$ and $\bar{y}$ are respectively the mean of variable. The further the value of r_{xy} from zero, the stronger the correlation.

As an alternative, Mahalanobis distance used in cluster analysis when there is a correlation between variables [7]. The statistical distance between the same two observations is of the form [9]:

$$d_{ij}{}^2 = \frac{1}{1-r^2}\left(\frac{(x_{i1}-x_{j1})^2}{S_1{}^2} + \frac{(x_{i2}-x_{j2})^2}{S_2{}^2} - \frac{2r(x_{i1}-x_{j1})(x_{i2}-x_{j2})}{S_1 S_2}\right) \quad (4)$$

Where r is coefficient of correlation, $S_1{}^2$ and $S_2{}^2$ are variances of variables 1 and 2, while x_{i1}, x_{j1}, x_{i2}, x_{j2} are each of observation ith of variable 1, observation jth of variabel 2 etc. In the case of multivariate, Mahalanobis distance formulated as:

$$d_{ij}{}^2 = [X_{ik} - X_{jk}]^T S^{-1}[X_{ik} - X_{jk}] \quad (5)$$

Where X_{ik} and X_{jk} are each ith-vector of kth-variable and jth-vector of kth-variable while $\boldsymbol{S}$ is variance-covariance matrix that can be described by:

$$S = \begin{bmatrix} S_1{}^2 & cov(x_1,x_2) & \dots & cov(x_1,x_p) \\ cov(x_1,x_2) & S_2{}^2 & \dots & cov(x_2,x_p) \\ \dots & \dots & \dots & \dots \\ cov(x_p,x_1) & cov(x_p,x_2) & \dots & S_p{}^2 \end{bmatrix} \quad (6)$$

Furthermore, the distance will be measured between groups and presented in a distance matrix (proximity) with:

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix} \quad (7)$$

There are two methods of grouping in cluster analysis, including the hierarchical method and the non-hierarchical method. In this study, a hierarchical method was used to classify observations in a structured manner based on the similarity of characteristics, and the number of groups was unknown. The number of groups will be formed from predetermined variables, and objects that have been grouped in one group cannot be transferred to another group.

In the hierarchical method, there are also two ways to form certain groups, namely merging (agglomerative) and separating groups (divisive). The agglomerative method is obtained by combining objects or groups gradually. First, suppose each object is a group, then the two closest groups are combined to form one cluster consists of adjacent objects. To combine two groups, a measure of dissimilarity (d_{uv}) is required, which is expressed in a function of distance [4]. A measure of dissimilarity (d_{uv}) in this study used the average linkage method with the following formula:

$$d_{(uv),w} = \frac{\sum_i \sum_k d_{ik}}{N_{(uv)} N_w} \quad (8)$$

Stages of grouping using hierarchical techniques will be presented in the form of tree diagrams or dendrograms that allow tracking of groupings of observed objects more easily and informatively. In addition, the grouping results obtained are not unique, because they will be determined by the set of objects, the type of variable, the scale of the variable, the size of the dissimilarity used, and the grouping technique.

3. Result and Discussion

Kalimantan Island consist of five provinces includes Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, and Kalimantan Utara, with 47 municipalities and 9 cities observed. These 56 observations were grouped using cluster analysis based on economic and social variables, namely economic growth variables (percent), poverty levels (percent), life expectancy (years), and the mean years of schooling (years). Meanwhile, the source of research data comes from Statistics Indonesia in each province in 2018. We then perform overall data processing by using the SPSS 20 software.

At the initial stage, we conduct Pearson correlation test between the variables, as shown in Table 1, to decide whether to use Mahalanobis distance or the Euclidean distance. This hypothesis test proposes $H_0: \rho = 0$ or there is no correlation between i-th variable and j-th variable. The results obtained that there are several H_0 which are rejected or there is correlation between variable mean years of schooling with life expectancy (0,406) and variable poverty rate with mean years of schooling (-0.375). The t-test on these two correlations is also significant with a limit of five percent. This indicates that the continuation of the cluster analysis will be better using the Mahalanobis distance.

Table 1. Correlation between research variables in Kalimantan

Variables	Economic Growth	Life Expectancy	Mean Years of Schooling	Poverty rate
Economic Growth	1	-0,219	-0,141	-0,029
Life Expectancy	-0,219	1	0,406*	0.202
Mean Years of Schooling	-0,141	0,406*	1	-0,375*
Poverty rate	-0,029	0,202	-0,375*	1

*) level of significance 5%

Furthermore, the result of cluster analysis in Kalimantan Island using *average linkage* method by using the Mahalanobis distance is shown in the dendrogram (Appendix 1). Researchers admitted dividing all observations into a group of three based on the ease of giving a special name for each group. Based on the results obtained, the number of municipalities/cities in group one consists of 48 observations, group two consists of seven observations, and group three consists of one observation (Table 3). By studying the characteristics that exist in each group, group one is called the "good economic group," while the second and third groups are called "good economic and social groups," respectively.

Table 2 shows the average mean of the centroid of each group based on the variables used in the study. This table discovers the characteristics of each group that distinguish it from other groups. Group one is an area with relatively good economic development results yet the worst social level compared to other groups. The municipalities/cities in this group have an average economic growth of 5.16 percent or 0.05 percent higher than the average throughout Kalimantan. As for the social sector, represented by the variables of the poverty level, life expectancy, and mean years of schooling, each obtained an average of 6.59 percent, 69.86 years, and 7.72 years. These three social variables have quite striking differences compared to the average variables in Kalimantan as a whole (Table 2).

Table 2. Average of variables within each groups

Variables	Group 1	Group 2	Group 3	Kalimantan Island
Poverty rate (%)	6,59	4,32	4,67	6,27
Life expectancy (Years)	69,86	72,85	73,94	70,31
Mean Years of Schooling (Years)	7,72	10,46	10,72	8,12
Economic Growth (%)	5,17	6,13	-4,18	5,12

The two social variables such as life expectancy and the mean years of schooling are components of the Human Development Index (HDI). This condition indicates that the two variables that make up the HDI at the municipalities/city level in group one have a low value below average quality. This situation is exacerbated by the highest average poverty rate compared to other groups. The first group

indicates the highest percentage and total observations in Kalimantan Barat and Kalimantan Tengah with 92.9 percent each, followed by Kalimantan Selatan at 84.6 percent (Table 3). In addition, there is an interesting finding that all observations in group one are municipalities categories (except Singkawang City). It can be concluded that all municipalities in the Kalimantan region and Singkawang City have social aspects that still need to be addressed. Indeed, this is very reasonable because, in terms of area, the government of municipalities has a relatively larger area with the condition that there are many underdeveloped villages. This circumstance means that adequate development will require more budget at the municipalities level.

Table 3. Distribution of Municipalities/Cities Based on Provinces and Groups

Province	Group 1	Group 2	Group 3	Total
Kalimantan Barat	13 (92,9)	1 (7,1)	0 (0)	14 (100)
Kalimantan Tengah	13 (92,9)	1 (7,1)	0 (0)	14 (100)
Kalimantan Selatan	11 (84,6)	2 (15,4)	0 (0)	13 (100)
Kalimantan Timur	7 (70)	2 (20)	1 (10)	10 (100)
Kalimantan Utara	4 (80)	1 (20)	0 (0)	5 (100)
Total	48 (85.7)	7 (12.5)	1 (1.8)	56 (100.0)

Group two has the best economic and social level compared to the other groups with a percentage of 12.5 percent of the total observations. On average, the variables of poverty rate, life expectancy, the mean years of schooling, and economic growth rate are 4.32 percent, 72.85 years, 10.46 years, and 6.13 percent (Table 2).

Based on spatial conditions, almost all areas categorize as cities included in group two except Singkawang and Bontang (Figure 2). Development in urban areas is very centralized and dispersed with a relatively small area so that various government programs can reach the entire area quickly. This condition is also in line with [10], where urban areas have large amounts of resources. In addition, economic growth is speeding up, jobs and the various services provided are very supportive of sustainable development.

In contrast to the first group, group three has good social conditions yet poor economic aspects. The results of the cluster analysis showed that only Bontang City that included in group three, where its poverty rates, life expectancy, mean years of schooling, and economic growth rates mounted respectively 6.27 percent, 70.31 years, 10.46 percent. and 6.13 percent (Table 2).

Specifically, Bontang City categorizes in the third group because the value of the economic growth variable is the only negative one. Although the three social variables owned are quite high, the economic growth of Bontang City is still very depressed. Furthermore, the state of economic growth in Bontang City has been negative for the last two years, namely -4.18 percent in 2018 and -2.18 in 2019 (Source: Statistics of Kalimantan Timur). These numbers mean that Bontang City has experienced a recession because negative growth occurred in two periods. The existence of natural resources such as crude oil, which continues to deplete, is closely related to the transformation of the economic structure that occurs in Bontang City. Therefore, the local government of Bontang needs to think about other sources of the economy to spur growth. Action is needed to slow down the growth to be negative, which could lead to a recession to a depression.

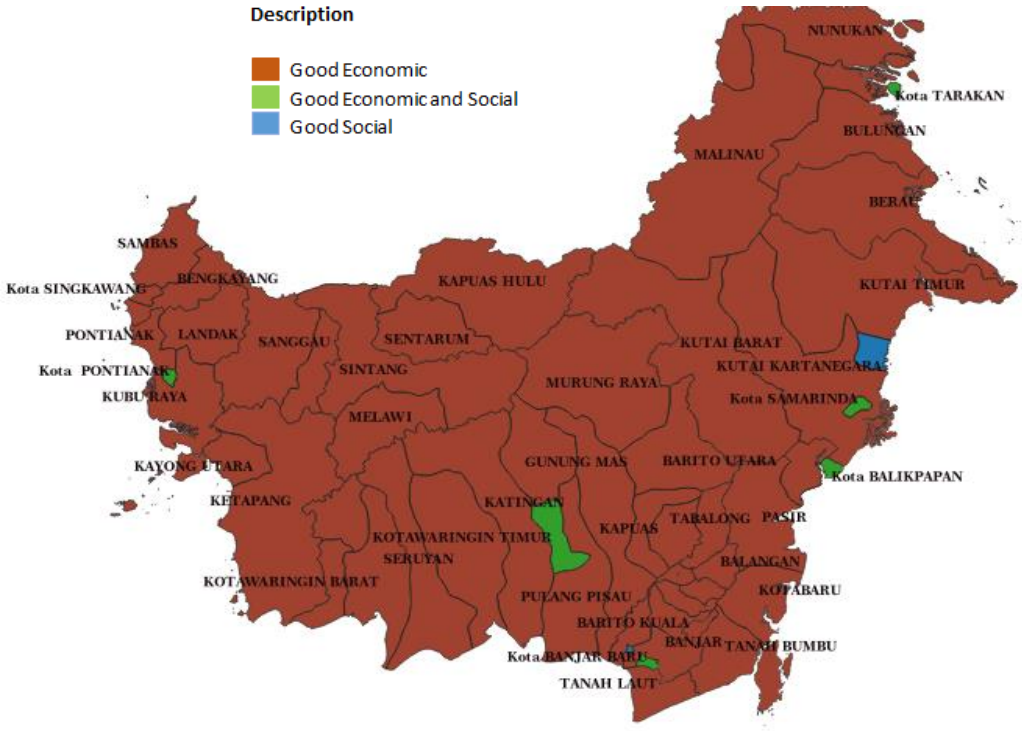


Figure 2. Map of municipalities/cities grouping based on social and economic variables in Kalimantan

4. Conclusion and Suggestion

Based on the results of the cluster analysis, three groups of municipalities/cities in Kalimantan formed including the first group (good economic group) consisting of 48 municipalities/cities, the second group (good economic and social group) consisting of 7 cities, and the third group (good social group) with only one city. All municipalities are in the first group, and almost all cities are in the second group.

The empirical results of the study pay special attention to the first and third groups. The utilization of the General Allocation Fund (DAU) by local governments at the municipalities and city levels of Singkawang should be optimized to boost development programs, especially those related to current social conditions. As for the local government of Bontang City, focusing on the economy will shift from relying on natural resources to tourism or trade-based sector to avoid a continuous recession. In addition, further research can add several other variables such as demographics, Gross Regional Domestic Product (GRDP), and unemployment rates so that grouping research can be more comprehensive.

5. Reference

[1] Johnson, Richard. And Wichern, Dean. 2007. *Applied Multivariate Statistical Analysis:Fifth Eddition.* New Jersey: Prentice Hall.
[2] Manly, B. F. J. (1986). *Multivariate Statistical Methods A Primer*. New York: Chapman and Hall.
[3] Obilor, Esezi Isaac & Amadi, Eric *Chikweru*. 2018. *Test for Significance of Pearson's Correlation Coefficient* (*r*). *International Journal of Innovative Mathematics, Statistics & Energy Policies* 6(1):11-23.
[4] Pulungan, Muhnifah Azmi. 2014. Pengelompokkan Kabupaten/Kota Di Pulau Sumatera Berdasarkan Hasil Pembangunan Ekonomi Dan Sosial Serta Infrastruktur Yang Membedakannya Tahun 2012. Jakarta: Sekolah Tinggi Ilmu Statistik.
[5] Punj, G., & Stewart, D. W. 1983. *Cluster analysis in marketing research: Review and suggestions for application*. *Journal of Marketing Research*, *20*(2), 134–148. https://doi.org/10.2307/3151680
[6] RS, Prawidya Hariani & Syahputri, Aulia Rizky. 2013. Analisis Ketimpangan Ekonomi Dan Pengaruhnya Terhadap Tingkat Kriminalitas Di Propinsi Sumatera Utara Tahun 2002-2013. Hal. 56-70

[7] Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley and Sons, Inc.
[8] Todaro, Michael P. 2000. Pembangunan Ekonomi di Dunia ketiga. Alih Bahasa Drs Han Munandar, M.A. Jakarta: Erlangga.
[9] Tohari, Amin. Analisis Cluster Psikografis Konsumen Kediri Town Square (*Cluster Analysis Psychographic Consumers Kediri Town Square*). Jurnal Matematika dan Pendidikan Matematika. Vol. 1. No. 2. (September 2016): 109-118.
[10] Widodo, W., dan Sunarti, S. (2019). Pola Perkembangan Perumahan di Kota Surakarta. Jur-nal Pembangunan Wilayah dan Kota (JPWK), 15 (4): 288-300.

Figure 1. The Dendrogram of Cluster Analysis of municipalities/cities in Kalimantan