pijmath
*Pattimura International Journal of Mathematics (PIJMath)*
*Volume 03, issue 01, pages 009-014, May 2024*
*E-ISSN: 2830-6791*
*RESEARCH ARTICLE* ◙ *OPEN ACCESS*
*DOI: https://doi.org/10.30598/pijmathvol3iss1pp09-14*

# Implementation of Centroid Clustering Method for Industrial Clusterization in Regencies and Cities in Maluku Province

**M. Y. Matdoan[1] , Rahmi Fadhilah[2*], N. S. Laamena[3], Dinda Ayu Safira[4], S. B. Loklomin[5]**

[1,3,5]*Statistics Study Program, FMIPA, Universitas Pattimura*
*Jl. Ir. M. Putuhena, Poka-Ambon, 97233, Maluku Province, Indonesia*

[2,4]*Department of Statistics, FSAD, Institut Teknologi Sepuluh Nopember*
*Jl. Arif Rachman Hakim, Sukolilo-Surabaya, 60111, East Java, Indonesia*

*Corresponding author's e-mail: [1]*6003231016@student.its@ac.id*

**Article History**

*Received: 07th February 2024*
*Revised: April 2nd, 2024*
*Accepted: April 27th, 2024*
*Published: May 1st, 2024*

**Keywords**
*Centroid;*
*Cluster;*
*Industry;*

**ABSTRACT**

*The industrial sector has a vital role in economic development. In addition to increasing state revenue, the industrial sector can also provide business opportunities that make a positive contribution to efforts to equalize community welfare. The limited employment opportunities available in Maluku Province need to be balanced with the increase in the labor force, which significantly impacts the high unemployment. Basically, the high unemployment rate will significantly impact economic development, which aims to improve the standard of living of the people in Maluku Province. Centroid Linkage is the average of all objects in the cluster, and the distance. The distance between the cluster centroids is what separates two clusters. Cluster centroid is the center value of observations on variables in a set of cluster variables. The purpose of this research is to cluster the distribution of industries in regencies and cities in Maluku Province using data from BPS Maluku Province. This study obtained the results that there are 3 clusters formed in the clusterization of industry in regencies and cities in Maluku Province, namely cluster 1 consisting of Tanimbar Islands Regency. Cluster 2 consists of Buru, South Buru, West Seram, East Seram, Central Maluku, Tual City, Southeast Maluku, and Aru Islands Regency. Furthermore, Cluster 3 consists of Ambon City and Southwest Maluku.*



**[1]*How to cite this article:***
M. Y. Matdoan , Rahmi Fadhilah*, N. S. Laamena, Dinda Ayu Safira, S. B. Loklomin., " IMPLEMENTATION OF CENTROID CLUSTERING METHOD FOR INDUSTRIAL CLUSTERIZATION IN REGENCIES AND CITIES IN MALUKU PROVINCE," *Pattimura Int. J. Math. (PIJMATH)*, vol. 03, iss. 01, pp. 009-014, May, 2024.

## 1. Introduction

Enhancing people's lives, creating jobs, balancing the income distribution in the community, boosting regional economic growth, and attempting to shift economic activity from the primary to the secondary and tertiary sectors are all goals of economic development [1]. Development is not only the responsibility of the government but also a shared responsibility between the government and the community. The industrial sector has a vital role in economic development. In addition to increasing state revenues, the industrial sector can also provide business opportunities that make a positive contribution to efforts to equalize community welfare [2]. Industrialization is a process of connection between international trade, innovation, specialization, and technical advancement, which ultimately promotes changes in the economic structure in line with rising public income [3].

The industrial sector has a role as a leading sector, meaning that the development of the industrial sector will spur and lift the development of other sectors, such as the agricultural sector and the service sector, which causes the expansion of employment opportunities [4]. The limited employment opportunities available in Maluku Province need to be balanced with the increase in the labor force, which significantly impacts the high unemployment. Basically, the high unemployment rate will significantly impact economic development, which has a policy that aims to improve the standard of living of the people in Maluku Province. Unemployment is a problem that is difficult to solve; this can be seen from the fluctuating number of unemployed people, which indicates that the availability of jobs is not optimal. Efforts to overcome the economic inequality that occurs are to treat particular policies in regencies or cities that have different economic levels. Grouping regencies or cities based on dispersed industries can be done with cluster analysis. One of the multivariate analytic techniques, cluster analysis, is mostly used to group objects according to their attributes. Individuals or study items are grouped according to their similarities and closeness to one another in a process known as cluster analysis. Clusters that form within one cluster tend to be homogeneous, while clusters that form outside of one another tend to be heterogeneous. The variables that were noticed are the basis for this categorization [5][6].

To get groups that are as homogeneous as possible, the basis for grouping is the similarity of the analyzed value scores. The smaller the distance of an individual to another individual, the greater the similarity of the individual. Data on the size of the similarity is then grouped so that it can be determined which individuals are in which group [7][8]. The linkage centroid is the average of all objects in the cluster. The distance between two clusters is the distance between the cluster centroids. Cluster centroid is the center value of observations on variables in a set of cluster variables. Using this technique, the centroid is instantly recalculated for every new cluster until a fixed cluster forms. Compared to other approaches, this one has the advantage that outliers do not significantly affect results [9][1[10].

## 2. Research Methods

### 2.1 Data Sources and Research Variables

The secondary data used in this study were taken from the Central Bureau of Statistics (BPS) of Maluku Province's publication of Maluku in Figures Year 2022. The variables used in this study consist of food industry variables, clothing industry, handicraft industry, electronic metal industry, and building material chemical industry variables.

### 2.2 Cluster Analysis

A multivariate technique called cluster analysis seeks to divide a sample of subjects into distinct groups according to a set of measured characteristics, placing comparable subjects together in a single group [11]. According to [12], A statistical analysis method called cluster analysis is used to divide a set of objects into two or more groups according to how similar they are to one another based on a variety of variables. According to [13], cluster analysis is a data mining technique for grouping a set of objects (dataset) into several clusters based only on the similarity of the characteristics of the attributes possessed by the object data so that object data in the same cluster are similar to each other but not similar to object data in different clusters. The results of cluster analysis are influenced by the objects being grouped, the variables observed, the similarity or dissimilarity measure used, the size scale used, and the clustering method used.

### 2.3 Centroid Clustering

The centroid method used the distance between two clusters. It is defined as the distance between the cluster centroids. The cluster centroid is the average variable value of all objects in a particular cluster. If

$$\bar{\boldsymbol{x}}_j = \sum_{i \in C_j} \frac{\boldsymbol{x_i}}{n_j} \tag{1}$$

is the centroid of $n_1$ member $C_1$ and $\bar{\boldsymbol{x}}_{\boldsymbol{2}}$ is the centroid of $n_2$ member $C_1$then

$$d_{(C1)(C2)} = P(\bar{\boldsymbol{x}}_{\boldsymbol{1}}, \bar{\boldsymbol{x}}_{\boldsymbol{2}}) \tag{2}$$

### 2.4 Research Stages

#### 2.4.1 Data Standardization

The standardization process was carried out if there was a significant difference in unit size between the variables studied. Striking differences in units can cause calculations in cluster analysis to be invalid [14]. Therefore, it is necessary to carry out a standardization process by transforming the original data before further analysis.

#### 2.4.2 Outlier Detection and Multicollinearity Assumption

Cluster analysis is essentially an algorithmic technique, not a statistical inference tool. Therefore, requirements such as normal distribution of data (in other statistical analyses) or linear relationships between variables are not required in cluster analysis. However, because the data processed in cluster analysis is usually only a small part of the population so the results can be generalized, the data processed should reflect the general picture or be representative [15]. Therefore, outliers must still be removed from the sample so that the results are not biased. Outlier detection is used to find data that is different from the majority of other data. Although they have different behavior from the majority of other data and are often considered noise, outliers often contain beneficial information. Not all data containing outliers can be transformed due to different data cases. However, by using the Centroid Linkage method, outliers are not significantly affected. In addition, the data used should not be correlated; in other words, there should be no multicollinearity. The reason is that in cluster analysis, each variable is given equal weight in the distance calculation. When some variables are correlated, the correlation will cause unbalanced weighting, which will affect the analysis results [16].

#### 2.4.3 Steps of Centroid Linkage Clustering

1. Creating $k$ Cluster. Each individual or observation unit becomes a group. Then, create a distance matrix (from $i$ to its group) with the formula:
$$D = \{dik\}$$
2. Find the smallest distance between pairs of clusters, $duv$ (distance between cluster $u$ and cluster $v$).
3. Merge cluster u and cluster v., and then update the distance matrix.
4. Repeat steps 2 and 3 for $N$-1 times. Record the distance value for each cluster merge.
5. Determine the *cut-off* value to determine the clusters formed. Do this by creating a dendrogram and determining the *cut-off* number of clusters.
6. Cluster naming is based on profiling, which looks at the characteristics of clusters formed on average.

## 3. Results And Discussion

### 3.1 Overview of Research Variables

Table 1 provides the following explanation of the variables used in this study.

**Table 1.** Descriptive Statistics

| | N | Minimum | Maximum | Average |
|---|---|---|---|---|
| Food Industry | 11 | 18.00 (Tual) | 202.00 (Tanimbar Islands) | 79.3636 |
| Clothing Industry | 11 | 3.00 (Central Maluku) | 218.00 (Tanimbar Islands) | 39.3636 |
| Craft Industry | 11 | 1.00 (Central Maluku) | 55.00 (Ambon) | 23.0000 |
| Electronic Metal Industry | 11 | 19.00 (SBB) | 101.00 (Ambon) | 39.2727 |
| Building Materials Chemical Industry | 11 | 21.00 (SBT) | 268.00 (Ambon) | 104.3636 |

### 3.2 Designing/Research Design

#### 3.2.1 Data Standardization

In the event that the variables under study exhibit large unit variances, data normalization is implemented. The outcomes of standardizing the study's variables are listed below.

**Table 2.** Standardization of Research Variables

| Ecotourism destination | X1 | X2 | ... | X6 |
|---|---|---|---|---|
| Food Industry | 1.93042 | 1.79425 | ... | -.35845 |
| Clothing Industry | .04150 | -.45931 | ... | -.38465 |
| Craft Industry | -.84000 | -.56880 | ... | -.68593 |
| Electronic Metal Industry | 1.55263 | -.44367 | ... | -.21435 |
| Building Materials Chemical Industry | -.49369 | -.31853 | ... | .34892 |
| Food Industry | -.88722 | -.38110 | ... | -.33225 |

### 3.2.2. *Outlier* Detection

Based on the research data that has been standardized, if there is data whose value is not between ± 2.5, it means that the data is an *outlier*. The results of observations of *outlier* data are shown in **Table 2** previously. Based on **Table 2**, it can be seen that there are no values that exceed ± 2.5. Consequently, it can be said that there are no outliers in the data used in this investigation.

### 3.3 Assumption Test

Some assumptions must be met in cluster analysis, namely:

**a. Sample Sufficiency Assumption**

To find out whether the sample used is sufficient for analysis, it can be seen from the Kaiser Meyer Olkin (KMO) value.

H0 : The sample is inadequate for further analysis
H1 : The sample is adequate for further analysis

Test Statistics

$$KMO = \frac{\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} r_{ij}^2}{\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} r_{ij}^2 + \sum_{i=1}^{\rho}\sum_{j=1}^{\rho} \rho_{ij}^2}$$

Test Criteria:
If the KMO value is > 0.5, it fails to accept H0, or the sample is suitable for further analysis.
The KMO test results can be seen in **Table 3** below:

**Table 3.** *KMO and Barlett's Test*

| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .598 |
|---|---|---|
| Bartlett's Test of Sphericity | Approx. Chi-Square | 33.510 |
| | Df | 10 |
| | Sig. | .000 |

According to **Table 3**, by examining the KMO MSA value, one may determine the viability of a variable and whether or not it can be processed further using this cluster analysis technique. The cluster analysis method can proceed if the KMO MSA value is higher than 0.50. Table 3 shows that the Barlett's Test of Sphericity (Sig.) value is 0.000 <0.05 and the KMO MSA value is 0.598> 0.50, indicating that the cluster analysis in this study can proceed as it satisfies the first condition.

**b. Multicollinearity Assumption**

The second assumption is multicollinearity. Correlation values in the correlation matrix can be used to determine if multicollinearity is present or not. If the correlation coefficient is more than 0.80, it is referred to as multicollinear. Additionally, Table 4 below displays the multicollinearity test results.

**Table 4.** Multicollinearity Test

| | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|
| X1 | 1 | .708 | .742 | .422 | .298 |
| X2 | .708 | 1 | .724 | .359 | .210 |
| X3 | .742 | .724 | 1 | .696 | .733 |
| X4 | .422 | .278 | .696 | 1 | .704 |
| X5 | .298 | .210 | .733 | .704 | 1 |

Based on **Table 4**, indicates that there are no instances of multicollinearity in the research variables, with a Pearson correlation value of less than 0.80 for each variable.

### 3.4 Industry Clusterization Analysis of Regencies and Cities in Maluku Province

The Centroid Clustering method is the average of all objects in the cluster. The distance between two clusters is the distance between the **cluster** centroids. Cluster centroid is the center value of observations on variables in a set of cluster variables. Here are the regency/city clusters that were formed.

**Table 5.** Regency/City Clusters Formed

| Case | 3 Clusters |
|---|---|
| 1 : Tanimbar Islands | 1 |
| 2 : Southeast Maluku | 2 |
| 3 : Central Maluku | 2 |
| 4 : Buru | 2 |
| 5 : Aru Islands | 2 |
| 6 : West Seram | 2 |
| 7 : East Seram | 2 |
| 8 : Southwest Maluku | 3 |
| 9 : South Buru | 2 |
| 10 : Ambon | 3 |
| 11 : Tual | 2 |

**Table 5** shows The three clusters with the following characteristics have emerged as a result of industrial clusterization in the regencies and cities of Maluku province.

Cluster 1 consists of the Tanimbar Islands Regency.

Cluster 2 consists of the regencies of Buru, South Buru, West Seram, East Seram, Central Maluku, Tual City, Southeast Maluku, and Aru Islands Regency.

Cluster 3 consists of Ambon City and Southwest Maluku Regency.

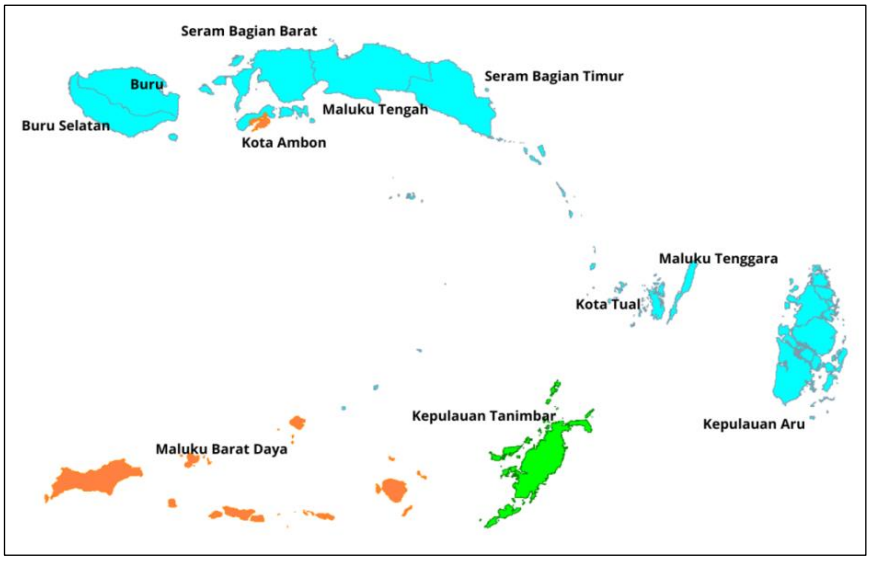


**Figure 2.** Visualization of the distribution of industrial clusterization in regencies and cities in Maluku Province

## 4. Conclusions

In light of the investigation and conversation that have been conducted, it is concluded that there are 3 clusters formed in industrial clusterization in regencies and cities in Maluku Province with details, namely cluster 1 consisting of Tanimbar Islands Regency. Cluster 2 consists of Buru, South Buru, West Seram, East Seram, Central Maluku, Tual City, Southeast Maluku, and Aru Islands Regency. Cluster 3 consists of Ambon City and Southwest Maluku Regency.

## References


[1] Saragih, A. H. (2018). Pengaruh penerimaan pajak terhadap pertumbuhan ekonomi di Indonesia [The effect of tax revenue on economic growth in Indonesia]. *Jurnal Sikap*, *3*(1), 277683.

[2] Nuraeni, (2018). Faktor-faktor yang mempengaruhi persistensi laba (studi kasus pada perusahaan property dan real estate yang terdaftar di bursa efek indonesia tahun 2013-2015) [Factors affecting earnings persistence (case study of property and real estate companies listed on the Indonesian stock exchange in 2013-2015). *Accruals (Accounting Research Journal of Sutaatmadja)*, *2*(1), 82-112.

[3] Hilman, A. M., & Ester, A. M. (2018). Peranan Sektor Industri Pengolahan Dalam Perekonomian Indonesia: Model Input-Output [The Role of the Manufacturing Industry Sector in the Indonesian Economy: An Input-Output Model]. *Media Ekonomi*, *26*(1), 63-76.

[4] Arsyad (2010). Pengaruh Perputaran Kas, Piutang dan Persediaan Terhadap Profitabilitas Perusahaan Sektor Aneka Industri yang Terdaftar di BEI [The Effect of Cash Turnover, Receivables and Inventories on the Profitability of Miscellaneous Industrial Sector Companies Listed on the IDX]. *Public Policy (Jurnal Aplikasi Kebijakan Publik & Bisnis)*, *2*(1), 57-74.

[5] Usman, H., & Sobari, N. (2013). Aplikasi Teknik Multivariate Untuk Riset Pemasaran [Application of Multivariate Techniques for Marketing Research], Jakarta: PT. *Prajagrafindo Persada*.

[6] Sitepu, R., Irmeilyana, I., & Gultom, B. (2011). Analisis cluster terhadap tingkat pencemaran udara pada sektor industri di Sumatera Selatan [Cluster analysis of air pollution levels in the industrial sector in South Sumatra]. *Jurnal Penelitian Sains*, *14*(3).

[7] Gudono (2014). Halo effect in analytical procedure: the impact of client profile and information scope. *Global Journal of Business Research*, *8*(1), 9-26.

[8] Metisen, B. M., & Sari, H. L. (2015). Analisis clustering menggunakan metode K-Means dalam pengelompokkan penjualan produk pada Swalayan Fadhila [Clustering analysis using the K-Means method in grouping product sales at Fadhila Supermarket]. *Jurnal media infotama*, *11*(2).

[9] Rahmawati, L. (2012). Analisis Kelompok Dengan Menggunakan Metode Hierarki Untuk Pengelompokan Kabupaten/Kota Di Jawa Timur Berdasar Indikator Kesehatan [Group Analysis Using Hierarchical Method for Grouping Regency / City in East Java Based on Health Indicators], Jurnal Matematika Vol.1 No.2 Universitas Negeri Malang,

[10] Ananda, R., & Yamani, A. Z. (2020). Determination of Initial K-means Centroid in the Process of Clustering Data Evaluation of Teaching Lecturers [Determination of Initial K-means Centroid in the Process of Clustering Data Evaluation of Teaching Lecturers]. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *4*(3), 544-550.

[11] Cornish (2007). A meta-analysis on the influence of inflammatory bowel disease on pregnancy. *Gut*, *56*(6), 830-837.

[12] Simamora, B. (2005). *Analisis multivariat pemasaran [Multivariate analysis of marketing*. Gramedia Pustaka Utama.

[13] Han, J., & Kamber, M. (2001). Data mining: concepts and techniques, second. *University of Illinois at Urbana Champaign: Morgan Kaufmann*.

[14] Budiman, I., Prahasto, T., & Christyono, Y. (2012). Data Clustering menggunakan metodologi Crisp-DM untuk pengenalan pola proporsi pelaksanaan tridharma [Data Clustering using Crisp-DM methodology for pattern recognition of tridharma implementation proportion]. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*.

[15] Sunaryo, S., Setiawan, S., & Siagian, T. H. (2011). Mengatasi masalah multikolinearitas dan outlier dengan pendekatan robpca (studi kasus analisis regresi angka kematian bayi di Jawa Timur) [Overcoming multicollinearity and outlier problems with robpca approach (case study of regression analysis of infant mortality rate in East Java)]. *Jurnal Matematika Sains dan Teknologi*, *12*(1), 1-10.

[16] Puspitasari, M. (2016). Pengelompokan Kabupaten / Kota Berdasarkan Faktor- Faktor Yang Mempengaruhi Kemiskinan Di Jawa Tengah Menggunakan Metode Ward Dan Average Linkage [Grouping Regencies / Cities Based on Factors Affecting Poverty in Central Java Using Ward and Average Linkage Methods], Jurnal Matematika Vol. 5 No. 6 Universitas Negeri Yogyakarta.