Agglomerative Hierarchical Clustering Analysis Based on Partially-Ordered Hasse Graph of Poverty Indicators in East Java

Ina Maya Sabara[1](✉), Fachrur Rozi[2], and Mohammad Nafie Jauhari[2]

[1] Magister Students of Mathematics Education Study Program, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia
220108210004@student.uin-malang.ac.id

[2] Mathematics Study Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Indonesia

Abstract. Poverty is a central issue in many countries, so one of the main goals of a country is to eradicate poverty. One of the efforts is to identify indicators that affect poverty using cluster analysis. In this research, we discuss cluster analysis using the agglomerative hierarchical clustering method based on the partially-ordered Hasse graph. Meanwhile, one form of facilitating cluster analysis is the Hasse graph. Therefore, this study was conducted to find out which areas have close or similar poverty indicators based on the partially-ordered Hasse graph and reduce the incidence of poverty in East Java. Before conducting cluster analysis, a multicollinearity test was carried out between poverty indicators, then the proximity between objects was determined using the Euclidean distance. Afterward, cluster analysis was performed using agglomerative methods (single linkage and complete linkage) to obtain the best cluster solution. The single linkage method provides the best solution consisting of five clusters. The results of the partially-ordered Hasse graph show that the fifth cluster becomes the top layer based on the Gini indicator. The fourth cluster becomes the top layer based on the depth index indicator. Last, the first cluster becomes the top layer based on the open unemployment rate indicator and life expectancy.

Keywords: Agglomerative Hierarchical Clustering · Hasse Graph · Poverty · Cluster Validity Test · Partially-ordered

1 Introduction

Statistics is a technique for collecting, processing, presenting, analyzing, and interpreting quantitative data. This technique not only provides a way of collecting, processing, and presenting the analysis of data but also provides a technique for concluding the data that has been analyzed [1]. If more than two objects are applied in statistics, it will be more difficult to apply statistical analysis. The solution to this problem is multivariate analysis.

E. Susanti et al. (Eds.): ICGT 2022, AER 221, pp. 460–469, 2023.
https://doi.org/10.2991/978-94-6463-148-7_46

One of the multivariate analyses that are usually applied is cluster analysis. There are two methods in cluster analysis, namely hierarchical and non-hierarchical. The hierarchical clustering method consists of the agglomerative (union) and divisive (spreading) methods. In the agglomerative method, there are several methods for forming clusters, including single linkage, complete linkage, average linkage, and ward methods. Best, which was then transformed into a partial ordering Hasse graph In this study, researchers will apply the agglomerative hierarchical clustering method to poverty indicators in East Java Province.

Cluster analysis is used to categorize items or data that have similar features. One use of cluster analysis is detecting poverty instances. By using cluster analysis, it will be possible to examine how the description of the situation of individuals in a group depends on their poverty level.

Several previous studies using the agglomerative hierarchical clustering method include the following: Fadliana [2] research using the single linkage method, the complete linkage method, the average linkage method, and the ward method in the case of regency/city classification in East Java Province based on the quality of family planning services (KB) was one of several previous studies that used the agglomerative hierarchical clustering method. Marcelino [3] applies the single linkage method, the complete linkage method, and the ward method in classifying the shares of the Syari'ah Jakarta Islamic Index (JII).

Afandi [4] applied the complete linkage method in regencies and cities in East Java Province on poverty indicators. Meanwhile, the study of Cong, et al. [5] investigated the hierarchical clustering method based on the partially-ordered Hasse graphs on online hotel data to measure the city's vitality. The Hasse graph is a type of partial ordering relational in the form of a directed graph with a more simplified form to produce clearer results.

In 2021, East Java Province will be the province with the number one poverty case in Indonesia compared to other provinces in Indonesia, with a total of 4,572,730 people [6]. Continuing previous research, the author is interested in conducting research by applying the agglomerative hierarchical clustering method to the case of poverty in East Java Province in 2021, based on partially-ordered Hasse graph. In this study, the authors compare the two agglomerative hierarchical clustering methods, namely the single linkage method and the complete linkage method, to know the best cluster, which is then formed in a partially-ordered Hasse graph.

The authors hope this research can provide the best solution for classifying districts and cities in East Java Province based on a partially-ordered Hasse graph of poverty indicators in East Java Province. An overview of the conditions of poverty in the region and city of East Java Province can be used as a reference to minimize poverty cases in each region in East Java Province. So that in the future, cases of poverty in East Java Province will be reduced, and in the end, the welfare of the community will increase.

This paper is organized in the following order: In Sect. 2, we present the research methods. Also, the briefly review of cluster analysis and the Hasse graph are presented in this section. A result and discussion are given in Sect. 3. Finally, last section provides a conclusion and recommendation for further research.

Table 1. The poverty indicators in Jawa Timur

No	Indicators (Variables)	Description
1.	Gini ratio (X_1)	measure the degree of inequality in population distribution (rasio)
2.	Poverty depth index (X_2)	average expenditure gap of each poor population towards the poverty line. (mean)
3.	The open unemployment rate (X_3)	percentage of unemployment to the number of labor force (percent)
4.	Life expectancy number (X_4)	the average estimate of the number of years that a person can live in life. (year)

2 Material and Methods

2.1 Data and Variables

The data used in this research is secondary data. The secondary data referred to in this study is poverty data in East Java Province in 2021, which was obtained from the official website of the Central Statistics Agency (BPS) East Java for observation units using 38 regions in East Java. The data contains several indicators (variables) of poverty, namely (Table 1):

2.2 Testing of Assumption

Before conducting cluster analysis, several assumptions must be met. Usually, researchers focus on two assumption tests, namely the multicollinearity test and the sample adequacy test [7].

a) Multicollinearity Test

The multicollinearity test is one of the assumption tests that shows a correlation or relationship between two or more independent variables in a multiple regression model [8]. The hypothesis testing as follows:
$H_0 : VIF \leq 10$ (There is nonmulticollinearity).
$H_1 : VIF > 10$ (Multicollinearity exists)with statistics test

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1}$$

where,
R_j^2 = coefficient of determination between X_j and independent variables X_i,

$$i \neq j, j = 1, 2, \ldots, p$$

b) Sample Sufficiency Test

The sample adequacy test is carried out to find out or ensure that the samples that have been collected are representative of the existing population. The sample adequacy test formula can be written as follows to calculate the KMO value using the Kaiser Meyer Olkin equation [9]:

$$KMO = \frac{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} r_{ij}^2}{\sum_{i=1}^{n}\sum_{j=1,j=i}^{n} r_{ij}^2 + \sum_{i=1}^{n}\sum_{j=1,j\neq 1}^{n} \alpha_{ij}^2} \tag{2}$$

where,

$i = 1; 2; 3; ...; n$ and $i = 1; 2; 3; ...; n$ and $i \neq j$.

r_{ij} = correlation value between variables i and j.

α_{ij}^2 = partial correlation value between variables i and j.

Because the data used by the researcher is population data, there is no need to test the adequacy of the sample.

2.3 Standardized Data

In cluster analysis, if the data used is in various units, data standardization is performed. The Z-score is commonly used for standardization. The formula for the Z-score is as follows [10]:

$$Z_{ij} = \frac{x_{ij} - \overline{x}_i}{s_i} \tag{3}$$

where,

x_{ij}= value of the object-i in the variable-j

$\overline{x}_i$ = average value of the object-i

s_i = standard deviation value of the object-i

2.4 Distance Calculation

Several dissimilarity measures or distance measurements are needed to measure the distance between two objects. In this study, we used the Euclidean distance with the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{4}$$

where,

d_{ij} = the distance between objects-i and j.

x_{ik} = value of object-i in the variable-k.

p = the number of variables.

2.5 Agglomerative Hierarchical Clustering

The agglomerative hierarchical clustering method is a hierarchical method that starts with the smallest cluster and then merges it with other clusters with comparable properties [7]. There are several forms of analysis in the agglomerative approach, including single linkage, average linkage, full linkage, and ward. This study employs two agglomerative methods:

1. **Single Linkage**

The single linkage method is a grouping procedure that is measured based on the object that has the smallest distance value with the following formula:

$$\boldsymbol{d_{(UV)W} = min(d_{UW}, d_{VW})} \tag{5}$$

where,

$d_{UW(\min)}$ = distance between the nearest neighbors of clusters U and W.
$d_{VW(\min)}$ = distance between the nearest neighbors of clusters V and W.

2. **Complete linkage**

The complete linkage method is a grouping procedure that is measured based on the object that has the largest distance value with the following formula:

$$\boldsymbol{d_{(UV)W} = max(d_{UW}, d_{VW})} \tag{6}$$

where,

$d_{UW(\max)}$ = distance between the largest neighbors of clusters U and W.
$d_{VW(\max)}$ = distance between the largest neighbors of clusters V and W.

2.6 Validity Test

Cluster validity testing is used to determine the quality of cluster analysis results. The benchmark that can be applied to check the validity of the results of hierarchical grouping is the cophenetic correlation coefficient. The cophenetic correlation coefficient is the correlation coefficient between each object of the Euclidean distance matrix and the object of each cophenetic matrix [11].

The cophenetic correlation coefficient is used for computing the validity test with the formula [12]:

$$r_{coph} = \frac{\sum_{i=1}^{n-1}\sum_{j>1}^{n}\left(c_{ij}-\overline{c}\right)\left(d_{ij}-\overline{d}\right)}{\left(\sum_{i=1}^{n-1}\sum_{j>1}^{n}\left(c_{ij}-\overline{c}\right)^{2}\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n-1}\sum_{j>1}^{n}\left(d_{ij}-\overline{d}\right)^{2}\right)^{\frac{1}{2}}} \tag{7}$$

where,

r_{coph} = cophenetic correlation coefficient.
c_{ij} = Euclidean distance of object i and j.
d_{ij} = Euclidean distance of object i and j.
$\overline{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n} c_{ij}$
$\overline{d} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n} d_{ij}$

2.7 Hasse Graph

If a binary relation R on a set A satisfies the properties of the relation (reflexive, antisymmetric, and transitive), then it is called a partial ordering relation. Partial sorting relations between sets can be represented in various ways, one of which is a graph. A graph is a set of vertices (minimum of one) and edges (minimum of zero).

The graph used in this study is the Hasse graph. A hasse graph is a mathematical diagram used to represent points on a directed graph. One of the functions of the Hasse graph is to facilitate groups of objects called hierarchies or tiers [13].

3 Results and Discussions

3.1 Descriptive Statistics

A descriptive analysis was carried out to describe the general characteristics of regencies in East Java Province in 2021 according to the percentage of poverty in each indicator.

Based on the measure of the amount of inequality in the distribution or income inequality of the poor (Gini ratio) has a mean of 0.33 with a standard deviation of 0.03 and a value range of 0.27 to 0.40. Based on the depth index indicator, or the total measure of the mean expenditure gap of each poor population to the poverty line, a mean value of 1.695 is obtained with a standard deviation of 0.94 and a value range of 0.39 to 4.33. The ratio of the number of unemployed to the number of people in the labor force (working-age population of 15 years or more who work) has a value range of 2.04 to 10.87, with a mean value of 5.52 and a standard deviation of 2.01.

Meanwhile, the mean number of years of life for those who still managed to reach a certain age, or the life expectancy indicator, has a mean of 71.72 with a standard deviation of 1.97 and a fairly large range of values is around 66.89 to 74.18.

3.2 Hierarchical Clustering

In this section, before performing cluster analysis, multicollinearity assumptions will be checked. The results showed that the overall value of VIF was less than 10, so there was no multicollinearity between variables. Afterward, the single linkage method produces five clusters from the results of classifying districts and cities in East Java Province on poverty indicators. In comparison, the complete linkage method obtains two clusters with different characteristics (Fig. 1) (Table 2).

Based on the results of the cophenetic correlation coefficient, it is found that the single linkage method (0,6842) provides a better cluster solution than the complete linkage method (0,6334).

While the form of partially-ordered the Hasse graph as follows.

Based on Fig. 2, we can see the results of the Hasse graph on the Gini ratio indicator. Cluster 5 creates a relationship that leads to cluster 3, cluster 4, and cluster 1. This means that areas in cluster 5 (Malang City) can help or solve problems based on the size of the distribution inequality or income inequality of the poor in cluster 3 (Sampang Regency), cluster 4 (Sumenep Regency), and cluster 1 (Lamongan Regency).

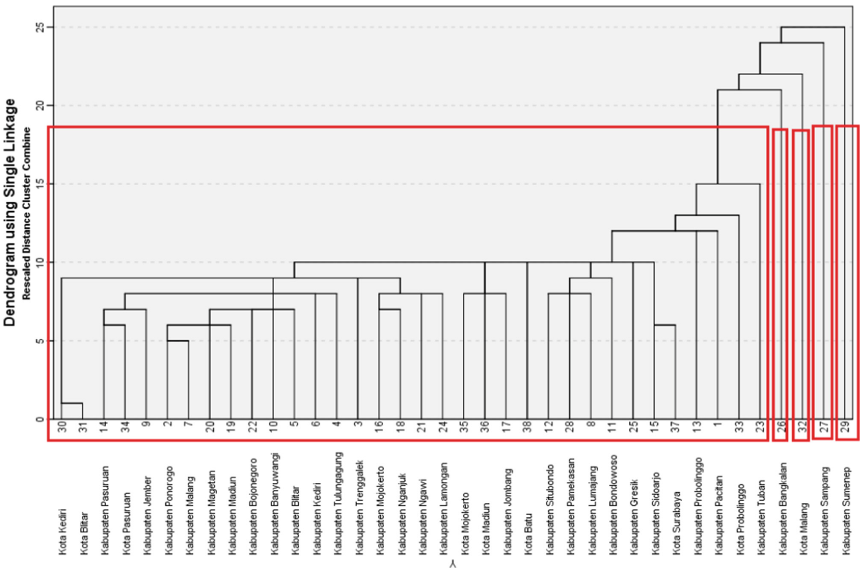

Fig. 1. Dendogram of cluster based on single lingkage method

Table 2. Members of Each Cluster Based on Single Lingkage Method

Cluster	#Members	Regency
1	34	Probolinggo, Bondowoso, Tuban, Pamekasan, Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Kediri, Malang, Lumajang, Jember, Banyuwangi, Situbondo, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Bojonegoro, Gresik, Lamongan, Kediri City, Blitar City, Pasuruan City, Probolinggo City, Mojokerto City, Batu City, Madiun City, Surabaya City
2	1	Bangkalan
3	1	Sampang
4	1	Sumenep
5	1	Malang City

Actually, Kediri Regency can also help Sampang Regency, Sumenep Regency, and Lamongan Regency in solving problems with the Gini ratio indicator. Meanwhile, based on the depth index indicator, areas in cluster 1 (Madiun City and Batu City) can help areas in cluster 2 (Bangkalan Regency), cluster 3 (Sampang Regency), and cluster 4 (Sumenep Regency).

Figure 3. Explains that for indicators of the open unemployment rate, cluster 1 (Sidoarjo Regency) and cluster 5 (Malang City) are related to cluster 4 (Sumenep Regency) and cluster 1 (Pacitan Regency and Pamekasan Regency). This means that Sidoarjo Regency and Malang City can help with poverty problems in Sumenep Regency,

Pacitan Regency, and Pamekasan Regency based on the ratio of the number of unemployed to the workforce. Meanwhile, due to the proximity of the region, Malang City is better positioned to assist Pacitan Regency.

For the last indicator, areas that have problems based on life expectancy indicators are in the same cluster. For example, the city of Kediri and the city of Surabaya can help with problems of public health status and the success rate of development in the health sector in Bondowoso and Probolinggo Regency, because these regencies still have a low life expectancy.

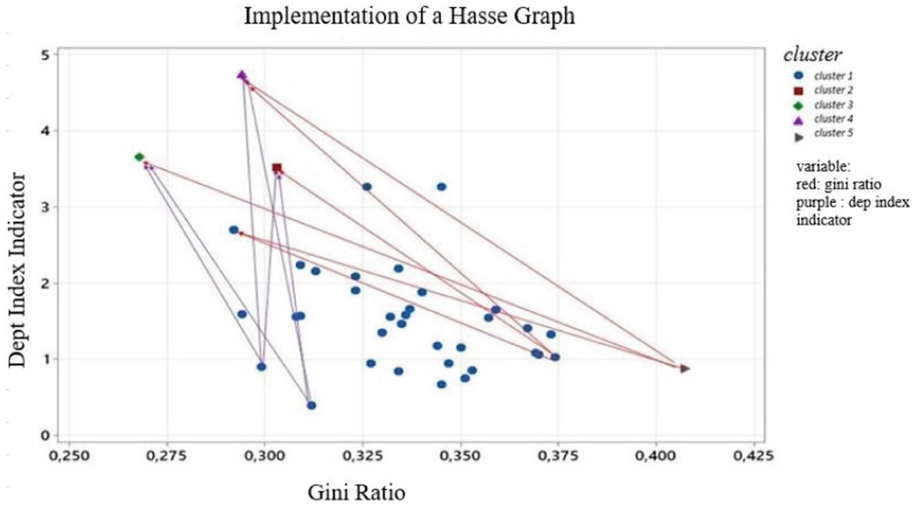

Fig. 2. Results of the Hasse Graph of Variables X_1 and X_2.

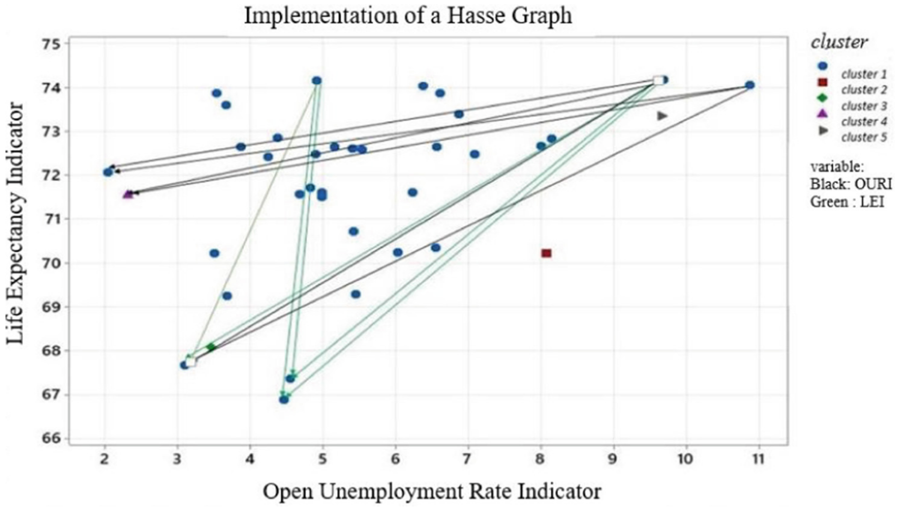

Fig. 3. Results of the Hasse Graph Between of Variables X_3 and X_4.

4 Conclusion

The conclusion of the application of agglomerative hierarchical clustering based on partially-ordered Hasse graph is to get five clusters using the single linkage method. The results of the partially-ordered Hasse graph show that the fifth cluster becomes the top layer based on the Gini indicator. The fourth cluster becomes the top layer based on the depth index indicator. Last, the first cluster becomes the top layer based on the open unemployment rate indicator and life expectancy.

References

1. Dajan, A. (1983). *Pengantar Metode Statistik.* Jakarta: LP3ES
2. Fadliana, A., & Rozi, F. (2015). Penerapan Metode Agglomerative Hie-rarchical Clustering untuk Klasifikasi Kabupaten/Kota Provinsi Jawa Timur berdasarkan Kualitas Pelayanan Keluarga Berencana. *Cauchy*, 4(1), 36-40.
3. Marcelino, R. (2018). *Perbandingan Kinerja Metode Single Linkage, Metode Complete Linkage, dan Metode Ward dalam Mengelompokkan Saham Syari'ah Jakarta Islamic Index(JII).* Yogyakarta: Universitas Islam Sunan Kalijaga.
4. Afandi, M. I. (2020). *Analisis Cluster Hirarki dengan Metode Complete Linkage pada Provinsi di Kabupaten/ Kota Jawa Timur berdasarkan Indikator Kemiskinan.* Malang: Universitas Islam Negeri Maulana Malik Ibrahim Malang.
5. Cong, W., Hongxin, L., & JiaJia, R. (2021). Research on Hierarchical *Cluster*ing Method Based on Partially-Odered. *Future Generation Computer System*, 125, 785-791.
6. BPS. (2020). *Kemiskinan dan Ketimpangan.* Surabaya: Badan Pusat Statistik Provinsi Jawa Timur.
7. Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In *Multivariate data analysis* (p. 785).
8. Ghozali, I. (2011). *Aplikasi Analisis Multivariate dengan Progam SPSS.* Semarang: Badan Penerbit Universitas Diponegoro.
9. Kaiser, H. (1974). An index of factor simplicity. *Psychometrika* , 31–36.
10. Moo-Young, M. (2011). *Comprehensiven Biotechnology (Second Edition).* Canada: Saint Louis.
11. Silva, A. R., & Dias, C. T. (2013). A cophenetic correlation coefficient for Tocher's method. *Pesquisa Agropecuaria Brasileira*, 48(6), 590-596.
12. Saraçli, S., Dogan, N., & Dogan, I. (2013). Comparison of Hierarchical *Cluster* Analysis Methods by Cophenetic Correlation. *Journal of Inequalities and Applications*, 203, 2-8.
13. Bruggemann, R., Halfon, E., Welzl, G., Voigt, K., & Steinberg . (2001). Applying The Concept of Partially Ordered Sets on Ranking of Near Shore Sediments by a Battery of Tests. *Chem. Inf. Comput,* 41(4), 918-925.