JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 7. NO. 2 FEBRUARY 2022
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v7i2.2830

MULTIVARIATE ANALYSIS OF COMMODITY AVAILABILITY OF STAPLE FOODS USING COMPLETE LINKAGE HIERARCHICAL CLUSTERING METHOD

Arjon Sitio[1]; Anita Sindar Sinaga[2*)]; Akhyar Haikal[3]; Sumitra Dewi[4]

Informatic Engineering[1,2*),3,4]
STMIK Pelita Nusantara
http://penusa.ac.id/
arjon@yahoo.com[1], haito_ita@yahoo.com[2*)], haikal@gmail.com[3], sumitra@gmail.com[4]

(*) Corresponding Author

Abstract— The government directly supervises 11 basic food commodities. The system of interplay between the price of goods and the availability of staple food directly has an impact on the high price of food at certain times. It is necessary to classify the food that is most needed by the community on big holidays in Indonesia so that it can be a reference for the government in preparing market needs in the coming year. In this study, the grouping of staple food availability was based on hierarchical cluster analysis with complete linkage method. The availability of food commodities in the discussion of this research is sourced from production materials and daily prices for meat, eggs, cooking oil and rice commodities. Cluster interpretation results in cluster 1 indicating Fulfilled Availability of 88-89%, Cluster 2 showing Sufficient Commodity Availability of 90-93% and Cluster 3 showing Availability of Rare Commodities of 87%. The three clusters formed are depicted in the form of a dendogram as a visualization of the relationship between food availability groupings.

Keywords: Multivariate Analysis, Complete Linkage, Hierarchical Clustering, Food Commodities

Abstrak — Pemerintah secara langsung melakukan pengawasan terhadap 11 komoditas bahan pokok pangan. Sistem saling mempengaruhi antara harga barang dengan ketersediaan pokok pangan secara langsung berdampak bagi tingginya harga pangan pada waktu-waktu tertentu. Perlu pengelompokan pangan yang paling dibutuhkan masyarakat pada hari-hari besar di Indonesia sehingga dapat menjadi acuan pemerintah dalam mempersiapkan kebutuhan pasar pada tahun yang akan datang. Dalam penelitian ini pengelompokan ketersediaan pokok pangan berdasarkan analisa cluster hierarki metode complete linkage. Ketersediaan komoditi pangan dalam pembahasan penelitian ini bersumber dari bahan produksi dan harga harian komoditi daging, telur, minyak goreng dan beras. Interpretasi cluster menghasilkan cluster 1 menunjukkan Ketersediaan Terpenuhi sebesar 88-89%, Cluster 2 menunjukkan Ketersediaan Komoditi Cukup sebesar 90-93% dan Cluster 3 menunjukkan Ketersediaan Komoditi Langka sebesar 87%. Tiga cluster yang terbentuk digambarkan dalam bentuk dendogram sebagai visualisasi hubungan antara pengelompokan ketersediaan pangan.

Kata Kunci: Analisis Multivariat, Complete Linkage, Hirarki Clustering, Komoditi Pangan

INTRODUCTION

The current average price of local food is not competitive compared to other food sources such as rice, flour, and corn. The government directly supervises 11 basic food commodities at certain times [1]. The means for distributing food are limited, resulting in a lack of food production. In addition, with changing people's lifestyles, automatically the demand for food as consumers of food also changes [2]. This can result in changes in the prices of food products, especially before religious holidays. The government, through the Ministry of Trade, can control the prices of basic commodities and other important goods through the distribution of basic goods. The government is expected to improve the distribution of basic and other important goods. so that every time there is a price increase, the public can also monitor it [3]. This situation can be circumvented by a strategy of changing the habit of consuming foods from processed animal foods into foods that are low in fat, low in fiber but high in calories. In this study, the availability of 11 types of staple food was grouped based on hierarchical cluster analysis using the complete linkage method [4]. Logically, a good cluster is a cluster that has high homogeneity (similarity) between members in one cluster and high heterogeneity (difference) between one cluster and another [5]. The grouping of food types using the Hierarchy method works by determining two or more objects that have the closest similarities, forwarded to other objects, and so on until the cluster will form a tree, there is a clear level (hierarchy) between objects, from the most similar to the least similar [6]. In the complete linkage method, clustering is based on the furthest distance between one object and another [7] [8].

Grouping of food availability based on data on prices of basic commodities, data on food supply originating from production, trade (exports and imports), stock changes. Cluster analysis begins with the standardization process, if there are data (variables) that have a large difference in unit size, measuring the similarity between objects (similarity) with 3 ways of measuring correlation, distance, and size [9] [10].

After the cluster is formed, the next step is to interpret and validate the results of the cluster analysis. The purpose of the study was to map the relationship between selling prices and the availability of staple food commodities and to find clusters of food commodities that were sufficiently available according to consumer demand [11] [12].

Research on Cluster Analysis of People with Mental Disabilities in the Province of the Special Region of Yogyakarta describes in determining cluster categories, low, medium, and high categories seen from the average calculation value of the highest and lowest variables overall. The cluster results use the average linkage, complete linkage, single linkage, ward, and centroid methods [13]. Comparative Research of Single Linkage, Complete Linkage, and Average Linkage Methods in Grouping Districts Based on Variable Types of Livestock, Sidoarjo Regency explained that in the formation of clusters a matrix of distances between districts was formed against the data consisting of 18 districts with 11 types of livestock. The distance between districts is calculated by the Euclidean distance [14] [15].

Research on Poverty Analysis in the Agricultural Sector (Case Study of Rice Commodities in Malang Regency) describes clusters that will form a kind of tree, there is a clear level (hierarchy) between objects, from the most similar to the least similar. A tool that helps to clarify this hierarchical process is called a “dendrogram”. Research Using Matlab and Python in Data Clustering, Matlab and Python have enough libraries and toolboxes to help users cluster data, present graphs. The test results show that both programming languages can carry out the clustering process [16].

MATERIALS AND METHODS

The data inputted by the CSV file will enter the distance measurement process using the Euclidean Distance technique which will produce a distance matrix and then apply the Hierarchical Clustering Complete Linkage method.
Research implementation method:

a. Data collection

The availability of food commodities in the discussion of this study is animal food consisting of commodities with the availability of:

Food production materials source https://www.bps.go.id/ :

1) Beef Production per Province (Tons), 2018-2020.
2) Broiler Meat Production per Province (Tons), 2018-2020.
3) Laying hens Egg Production per Province (Tons), 2018-2020.
4) Cooking Oil Production per Province (Tons), 2018-2020.
5) Provincial Rice Production (Tons), 2018-2020.

Food price material source https://hargapangan.id/ :

1) Rice Food Prices for the period (daily) January 2020 – January 2021.
2) Chicken Meat Food Prices (daily) January 2020 – January 2021.
3) Beef Food Prices (daily) January 2020 – January 2021.
4) Egg Food Prices for the period (daily) January 2020 – January 2021.
5) Cooking Oil Food Prices (daily) January 2020 – January 2021.

b. Cleansing Data

The data set collected is not ready for use, so it is necessary to clean the data as data preparation. Data preparation consists of several processes such as data cleaning, data transformation, and data reduction. The data cleaning process includes identifying and removing outliers and correcting missing values.

c. Euclidean Distance Technique

Each data will undergo a distance measurement process to determine the cluster. The distance calculation is determined using the Euclidean distance technique which produces a distance matrix. After getting the distance matrix from the calculation of the distance between the data, the data will be processed using Hierarchical Clustering Complete Linkage.

d. Clustering

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

Hierarchical Clustering groups data that works by grouping two or more data that have similarities or similarities, then the process is continued to other objects that have proximity to two, this process continues until the cluster forms a tree, there is a clear hierarchy or level between objects from the most similar to least similar. Merging two clusters can still be continued if there are still other closest points that are possible to be combined. Next, as in the previous step, look for the two closest clusters to combine. The clusters that are combined are in the form of a single point in the first step or clusters which are a combination of two points/clusters. The process will end when all clusters have been merged into one large cluster. The Complete Linkage method groups the two objects that have the furthest distance first. Based on the distance matrix, then the data is grouped with complete linkage (farthest distance). By treating the data as a group, then determine the distance of the two smallest groups.

e. Multivariate Analysis
Analyzing the effect of several variables on other variables at the same time. A statistical method that allows researchers to research more than two variables simultaneously. Multivariate analysis techniques are classified into two, dependency analysis and interdependence analysis. Dependency analysis serves to explain or predict the dependent variable by using two or more independent variables. Multivariate analysis techniques are classified into two, namely dependency analysis and interdependence analysis. Dependency analysis serves to explain or predict the dependent variable by using two or more independent variables.

f. Interpretation of Cluster Analysis
Cluster interpretation is carried out to determine the profile of each group by using the average on each variable. Naming clusters or concluding is very subjective and depends on the research objectives.

g. Food Commodity Group
The results of grouping food commodities using the complete linkage method consisting of 3 clusters.

RESULTS AND DISCUSSION

Grouping using the complete linkage method is the process of merging two or more objects that have the furthest distance.
Step 1: Standardize the data on the complete linkage method. Standardization of data is needed if the data used in a study has various units.
Step 2: Determine the size of the similarity or dissimilarity between two objects in the complete linkage method. Pairs of objects that are closer together will be more 'similar' in characteristics compared to pairs of objects that are farther away. One method to measure the distance between objects is to use Euclidian Distance.
Step 3: Complete Linkage Cluster Analysis Process. Cluster analysis with the complete linkage method is the process of merging two or more objects that have a maximum distance or the farthest neighbor, the distance between one cluster and another cluster is measured based on the furthest distance of the members.
Step 4: Repair the Distance Matrix Using the Complete Linkage Method. Clusters are formed based on cluster pairs with the farthest distance.
Step 5: Determine the number of cluster members and their members in the complete linkage method. In determining cluster members, group objects into 3 clusters.
Step 6: Interpretation of the cluster on the complete linkage method. After the cluster is formed, the next step is to give specific characteristics to describe the contents of the cluster.

The types of food commodities discussed in this study consist of five commodities, namely beef, chicken, rice, eggs, and cooking oil. Dataset variables:
X1 = Beef Production by Province
X2 = Broiler Meat Production by Province
X3 = Egg Production of Layers by Province
X4 = Provincial Cooking Oil Production
X5 = Provincial Rice Production
X6 = Beef Prices for the Period Jan 2020 – Jan 2021.
X7 = Price of Broilers for Jan 2020 – Jan 2021.
X8 = Price of Laying Chicken Eggs Jan 2020 – Jan 2021.
X9 = Cooking Oil Price Jan 2020 – Jan 2021.
X10 = Rice Price Period Jan 2020 – Jan 2021.

The large percentage of imports on the one hand will benefit consumers with relatively affordable prices, but on the other hand it can reduce producer prices. Research involving multiple variables, data analysis techniques are often used multivariate analysis. All statistical methods that analyze several measurements (variables) that exist in each object in one or many samples simultaneously. Based on this definition, any analytical technique that involves more than two variables simultaneously can be considered as multivariate analysis.

A variation of a number of n weighted variables (X1 to Xn) can be expressed mathematically as follows variate value = w1X1+ w2X2+ w3X3+…+wnXn.
Daily commodity prices can be shown in the graph, shown in the Figure 1.

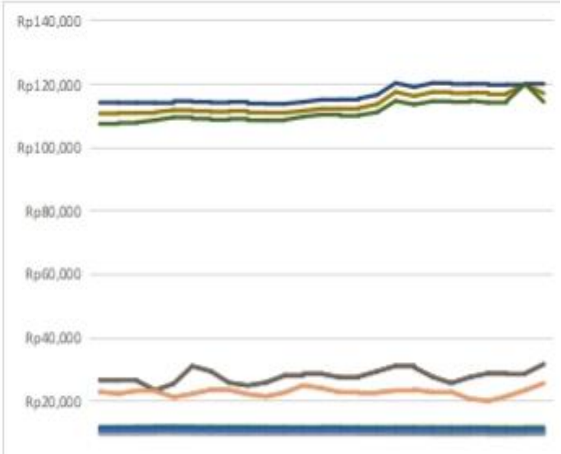

Figure 1. Graph of the average daily price

Computational linkage (y) representation of the vector from the distance matrix. Linkage checks if y is a Euclidean distance. Avoid this time-consuming check by entering X - y. The 'centroid' and 'median' methods can produce non-monotonic cluster trees. This result occurs when the distance from the combined two clusters, r, and s, to the third cluster is less than the distance between r and s. The path from the leaf to the root node takes several steps downward. In this case, cluster 1 and cluster 3 merge into a new cluster, and the distance between this new cluster and cluster 2 is smaller than the distance between cluster 1 and cluster 3.

The full dendrogram displays the progressive clustering of objects. If truncation has been requested, a broken line marks the level the truncation has been carried out. The truncated dendrogram shows the classes after truncation, shown in the Figure 2.

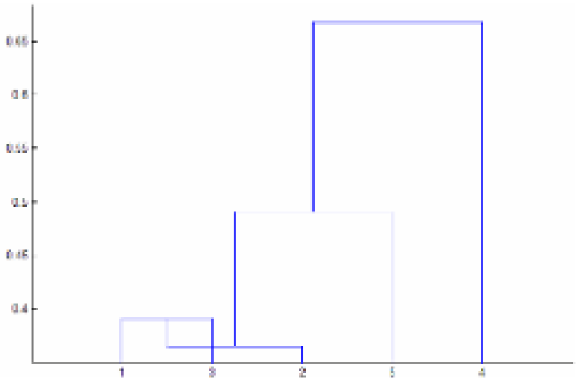

Figure 2. Non Monotonous Cluster Tree

Cluster interpretation is carried out to determine the profile of each group by using the average on each variable as follows:

1. The cluster with the lowest average is categorized as Rare Availability.
2. Clusters with an average higher than the lowest cluster mean are categorized as Sufficient Availability.
3. The cluster with the highest average is categorized as Fulfilled Availability.

The purpose of cluster interpretation for determine the number of clusters. Result of clustering of food commodities cluster interpretation consist of Rare Cluster, Fulfilled Cluster and Enough Cluster shown in the Table 1.

Table 1. Clustering of Food Commodities

Commodity	Total	%	Cluster
Beef Quality 1	308990	87	Rare
Beef Quality 2	220331	86	Fulfilled
Chicken meat	259300	86	Fulfilled
Chicken eggs	569392	89	Enough
Bulk Cooking Oil	359201	90	Fulfilled
Cooking Oil 1 kg	319313	92	Fulfilled
Cooking Oil 2 kg	422498	95	Enough
Super Quality Rice 1	108990	78	Fulfilled
Medium Rice 1	120331	86	Fulfilled
Bottom Rice 1	259300	92	Enough

Complete Linkage also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The steps of the complete method linkage is as follows:

1. Calculate the distance matrix between data with using Euclidean distance calculations. An example of calculating the distance matrix between cluster 1 and cluster 2.
   By using the calculation that the same, the distance matrix of cluster 1 and cluster 3, cluster 1 and cluster 4, and so on
2. Determine the smallest or closest distance from distance matrix. Calculate the combined cluster distance with other clusters.
3. Calculate the combined cluster distance with other clusters.
4. Create a new distance matrix based on previous calculations.
5. Repeat step (2) through step (4) to form four clusters.
6. Based on the distance matrix, four clusters have been obtained, so the grouping process stops. The clusters in this study are three clusters. The three clusters formed can be depicted in the form of a dendrogram, Figure 3.

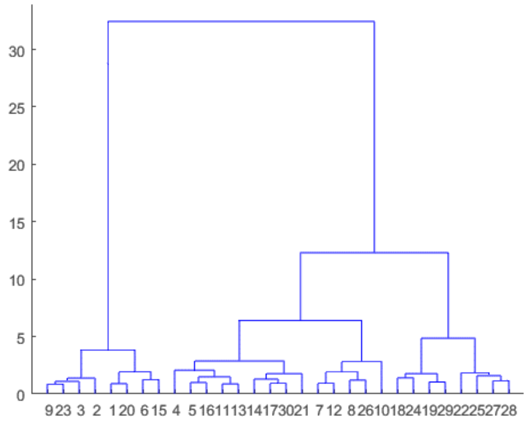

Figure 3. Complete Linkage

This allows us to find certain hierarchical clustering methods that can identify stronger clustering structures. These techniques do not let you explicitly set the number of clusters. Instead, you pick a distance value that will yield an appropriate number of clusters. This will be discussed further when we discuss the Dendrogram and the Linkage report.

Multivariate regression analysis is a statistical method that allows examining the relationship of more than two variables simultaneously. The continuous (real) value of the output t aims to predict the output accurately for new data. Regression analysis studies the form of the relationship between one or more independent variables (X) and one dependent variable (Y). In research, the independent variable (X) is usually the variable determined by the researcher independently. predictions that are often used on quantitative scale data (intervals or ratios) are Linear Regression. Linear regression technique can analyze the effect of several variables on other variables at the same time, Figure 4.

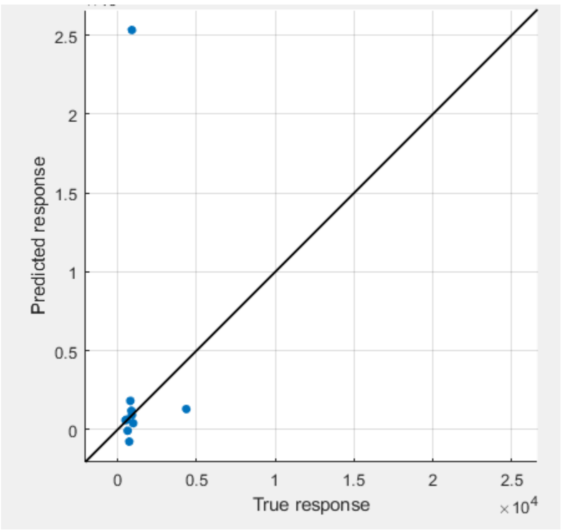

Figure 4. Prediction of Linear Regression

CONCLUSION

The availability of food commodities tends to fluctuate and demand continues to increase, especially on religious holidays in Indonesia, the application of complete linkage clustering can map the types of food needs from the distance between data set variants. The grouping results are displayed in the form of a dendrogram diagram. The dendrogram is used to clarify the grouping in the hierarchical method obtained 3 clusters of food availability. The application of linear regression in analyzing multivariate data results in precise data cluster modeling that can predict the output accurately. From the results of grouping food availability, it is found that there are types of food commodities that experience high prices due to high demand while the availability of food ingredients is not sufficient for consumer needs with the percentage of 85%-88%.

ACKNOWLEDGMENTS

The author would like be thank for STMIK Pelita Nusantara who has provided support for the implementation of this research.

REFERENCE

[1] A. S. Sinaga and A. S. Sitio, “Big Data Analysis of Covid-19 Spread Based on Distribution Map and Protocol Regulations with Business Intelligence,” vol. 15, no. 1, pp. 106–114, 2022.

[2] T. Märzinger, J. Kotík, and C. Pfeifer, “Application of hierarchical agglomerative clustering (Hac) for systemic classification of pop-up housing (puh) environments,” *Appl. Sci.*, vol. 11, no. 23, 2021, DOI: 10.3390/app112311122.

[3] M. Roux, “A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms,” *J. Classif.*, vol. 35, no. 2, pp. 345–366, 2018, DOI: 10.1007/s00357-018-9259-9.

[4] S. Patel, S. Sihmar, and A. Jatain, “A study of hierarchical clustering algorithms,” *2015 Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2015*, vol. 3, no. 10, pp. 537–541, 2015.

[5] M. N. Aziz and T. Ahmad, “Cluster analysis-based approach features selection on machine learning for detecting intrusion,” *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 233–243, 2019, doi: 10.22266/ijies2019.0831.22.

[6] H. Nguyen, X. N. Bui, Q. H. Tran, and N. L. Mai, “Corrigendum to ‘A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and cubist algorithms’ [Appl.

Soft Comput. 77 (2019) 376–386] (Applied Soft Computing Journal (2019) 77 (376,” *Appl. Soft Comput.*, vol. 100, p. 107123, 2021, doi: 10.1016/j.asoc.2021.107123.

[7] A. Achmad and R. Fernandes, “Comparison of Cluster and Linkage Validity Indices in Integrated Cluster Analysis with Structural Equation Modeling War-PLS Approach no. February, 2021.

[8] S. Aggarwal, P. Phoghat, and S. Maitrey, “Hierarchical Clustering- An Efficient Technique of Data mining for Handling Voluminous Data,” *Int. J. Comput. Appl.*, vol. 129, no. 13, pp. 31–36, 2015, doi: 10.5120/ijca2015907081.

[9] M. Nedyalkova and V. Simeonov, “Multivariate chemometrics as a strategy to predict the allergenic nature of food proteins,” *Symmetry (Basel).*, vol. 12, no. 10, pp. 1–19, 2020, doi: 10.3390/sym12101616.

[10] C. Etumnu and A. W. Gray, “A Clustering Approach to Understanding Farmers’ Success Strategies,” *J. Agric. Appl. Econ.*, vol. 52, no. 3, pp. 335–351, 2020, doi: 10.1017/aae.2020.4.

[11] P. Yildirim and D. Birant, “K-Linkage: A new agglomerative approach for hierarchical clustering,” *Adv. Electr. Comput. Eng.*, vol. 17, no. 4, pp. 77–88, 2017, doi: 10.4316/AECE.2017.04010.

[12] Vijaya, S. Aayushi, and R. Bateja, “A Review on Hierarchical Clustering Algorithms,” *Journal of Engineering and Applied Sciences*, vol. 12, no. 24. pp. 7501–7507, 2017.

[13] E. A. Leal Piedrahita, “Hierarchical Clustering for Anomalous Traffic Conditions Detection in Power Substations,” *Cienc. e Ing. Neogranadina*, vol. 30, no. 1, pp. 75–88, 2019, doi: 10.18359/rcin.4236.

[14] N. Apfel and X. Liang, “Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables,” 2021, [Online]. Available: http://arxiv.org/abs/2101.05774.

[15] G. Brandi and T. Di Matteo, “Higher-Order Hierarchical Spectral Clustering for Multidimensional Data,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12746 LNCS, pp. 387–400, 2021, doi: 10.1007/978-3-030-77977-1_31.

[16] A. M. Jarman, “Hierarchical Cluster Analysis : Comparison of Single linkage, Complete linkage, Average linkage and Centroid Linkage Method,” no. 2, 2020, doi: 10.13140/RG.2.2.11388.90240.