Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 2, No. 1 (2020), pp. 51–58
P-ISSN 2686-5335, E-ISSN 2716-2478
doi: 10.15408/inprime.v2i1.14167

Regency grouping in East Java based on Variable Type of Agriculture uses Hybrid Hierarchical Clustering Via Mutual Cluster Method

Sulthan Fikri Mu'afa* and Nurissaidah Ulinnuha
Department of Mathematics, Faculty of Science and Technology
Universitas Islam Negeri Sunan Ampel Surabaya
Email: fikrimuafa@gmail.com

Abstract

East Java Province is one of the provinces that has the largest agricultural resources in Indonesia. The Government of East Java needs to produce superior commodities in each region. This study aims to group districts in East Java Province based on variable types of agriculture with the hybrid hierarchical clustering via mutual cluster method that combines the merging of bottom-up clustering advantages and top-down clustering advantages. Mutual cluster is a grouping with the largest distance between small groups of the shortest distance for each point outside the group. In this research, the calculation uses Euclidean distance. The data used in this study are from the East Java Central Statistics Agency (BPS) in 2017. The division calculation is obtained by finding the minimum s_w (standard deviation of intra cluster) value and the maximum s_b (standard deviation of inter clusters) value and using the analysis of variance calculation. The grouping results obtained were nine groups with s_w value of 725.934, s_b value of 1.475.978 and F_{count} value of 7,908.
Keywords: agriculture; Hybrid Hierarchical Clustering; mutual cluster; Euclidean distance; analysis of variance.

Abstrak

Provinsi Jawa Timur merupakan salah satu provinsi yang memiliki sumber daya pertanian terbesar di Indonesia. Pemerintah Jawa Timur perlu mengembangkan komoditi unggulan di tiap daerah di Jawa Timur. Penelitian ini bertujuan untuk mengelompokkan kabupaten di Provinsi Jawa Timur berdasarkan variabel jenis pertanian dengan metode hybrid hierarchical clustering via mutual cluster yaitu menggabungkan kelebihan bottom-up clustering dan kelebihan top-down clustering. Mutual cluster yakni pengelompokkan dengan jarak terbesar antara bagian dalam kelompok yang kecil dari jarak yang terpendek kepada tiap titik di luar kelompok. Dalam penelitian ini, perhitungan jarak menggunakan jarak Euclidean. Data yang digunakan dalam penelitian ini dari Badan Pusat Statistik Jawa Timur tahun 2017. Perhitungan pembagian didapat dengan mencari nilai s_w (simpangan baku dalam klaster) yang minimal dan nilai s_b (simpangan baku antar klaster) yang maksimal, serta digunakan perhitungan analyze of varians. Hasil pengelompokkan yang diperoleh didapatkan sebanyak sembilan kelompok dengan nilai s_w sebesar 725.934, nilai s_b sebesar 1.475.978 dan nilai F_{hitung} sebesar 7,908.
Kata Kunci: pertanian; Hybrid Hierarchical Clustering; mutual cluster; jarak Euclid; analisis variansi.

1. INTRODUCTION

Agriculture has a vital role in the national economy i.e. an absorbent of labor, a source of economic growth, a contributor to foreign exchange, a determinant of price stability, and a producer of community food [1]. One of the provinces in Indonesia that has the largest agricultural resources is East Java. This province has the largest household of food farmers i.e. 5,163,979 households. This province is also famous for its food crop farming center that has a role in contributing to the largest national food stock [2]. In addition, East Java's rice production was 13.13 million tons or 16.1% of the total rice production, which reached 81.38 million tons [3].

East Java is currently developing an agribusiness center by forming centers for agricultural products and processing industry. Expanding agribusiness product marketing is expected not only as a national granary but also to meet international needs. Therefore, the agricultural sector must receive more special attention than others [2].

The agricultural sector, with all its potential commodities, requires integrated management. As the first step in management is identifying the potential of commodities in each region using cluster analysis to group the agricultural commodities in each region in East Java.

The cluster analysis is a multivariate technique in the science of statistical analysis that is useful in grouping objects that have similar uniqueness to smaller groups [4]. One method of clustering is Hybrid Mutual Clustering, where the method combines the advantages of top-down clustering and bottom-up clustering introduced by Chipman and Tibshirani [5]. Mutual cluster is a grouping that uses the maximum distance between partners in smaller groups over the minimum distance to the point in the outer group.

Research on grouping has been done by Alfira, who analyzed hybrid hierarchical clustering using Euclidean distances in 2016 [6]. In addition, Raharja grouped subdistricts in Lamongan sourced agricultural variables using hybrid hierarchical clustering via mutual cluster in 2011 [7]. The other research conducted by Mariyani [8]. She grouped regencies in East Java-based on agricultural variables using hybrid hierarchical clustering via mutual cluster with Pearson distance calculation.

In this study, we will group the districts on East Java based on the agricultural variables using hybrid hierarchical clustering through mutual cluster methods to identify the best agricultural commodities. We use the Euclidean to calculate the distance and complete linkage in the bottom-up clustering approach.

2. METHOD

In this study, we use 29 regencies data in East Java Province from the Central Statistics Agency (BPS) [16]. The variable of agriculture divided into four subsectors: food, agriculture, livestock, and fisheries (Table 1).

The step to cluster the agriculture commodities are follows:

1. Standardize data using the standardized z-scores [10]:

$$Z_i = \frac{x_i - \bar{x}}{s} \tag{1}$$

Variables in data must be standardized to avoid problems caused by non-uniform scale values between object grouping variables [9].

2. Test multicollinearity to know the condition of agricultural data. Multicollinearity is a condition where there is a correlation between the predictor variables when the type of regression used exceeds one predictor [11]. Multicollinearity arises when the predictor variable values have a correlation value (r value) more than 0.95 [12].
3. Conduct the bottom-up clustering (complete linkage), and determine the mutual cluster.
4. Conduct top-down clustering (k-means) by maintaining mutual clusters from bottom-up clustering grouping.
5. Conduct hybrid hierarchical clustering by grouping mutual cluster results based on top-down clustering.

Table 1. Research Variables.

Sub-sector	Variable	Information	Sub-sector	Variable	Information
Food crop subsector variables	x_1	paddy rice harvest area	Livestock subsector	x_{21}	cow population
	x_2	paddy rice production		x_{22}	goat population
	x_3	harvested rice field area		x_{23}	sheep population
	x_4	harvested rice production		x_{24}	native chicken population
	x_5	corn harvest area		x_{25}	layer chicken population
	x_6	corn production		x_{26}	broiler population
	x_7	soybean harvest area		x_{27}	duck population
	x_8	soybean production		x_{28}	beef production
	x_9	peanut harvest area		x_{29}	goat meat production
	x_{10}	peanut production		x_{30}	lamb meat production
	x_{11}	green bean harvest area		x_{31}	kampong chicken production
	x_{12}	green bean production		x_{32}	laying chicken production
	x_{13}	cassava harvest area		x_{33}	broiler meat production
	x_{14}	cassava production		x_{34}	duck meat production
	x_{15}	sweet potato harvest area		x_{35}	chicken eggs production
	x_{16}	sweet potato production		x_{36}	duck eggs production
Plantation subsector	x_{17}	sugarcane harvesting area	The fisheries subsector	x_{37}	fish consumption ponds area
	x_{18}	sugarcane production		x_{38}	fish consumption pond production
	x_{19}	coffee harvesting area		x_{39}	fish consumption pool area
	x_{20}	coffee production		x_{40}	fish consumption pool production

The hybrid hierarchical clustering is a cluster grouping method where the merging of particular objects originates from the collaboration between top-down clustering and bottom-up clustering methods. Bottom-up algorithms are good at grouping small sample sizes, while top-down algorithms are good at grouping large sample sizes [6]. Bottom-up clustering is the process of grouping from small groups into large groups [13], while top-down clustering is a grouping process that breaks large groups into smaller groups [6].

Mutual cluster is a grouping where the largest distance between parts in the smallest group is used from the shortest distance at each point outside the group. Because of that, the maximum

distance between objects in the mutual cluster is lower than the minimum distance from the collection of objects outside the mutual cluster. Data that exists in a mutual cluster cannot be separated [5].

Mutual cluster basically uses the largest distance between data elements in S, which is smaller than the smallest distance in the S element for data that is not included in S. It is formulated as follows:

$$x \in S, y \in S, d(x,y) > diameter(S) = \max_{w \in S, z \in S} d(w,z), \quad (2)$$

where d is a distance function between two data objects, S is a subset of data, x is an element of S, y is another element that is not included in S. w and z is a data object that has the closest distance [5]. In this paper, we use Euclidean distance that calculated using [4]:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_{ik} - y_{jk})^2} \quad (3)$$

6. Determine the number of the best grouping results using the values of minimum standard deviation in the cluster (S_w), the maximum standard deviation between clusters (S_b) and F_{count} from analysis of variance (ANOVA) using [14]:

$$S_w = \frac{1}{n}\sum_{k=1}^{n} S_k, \quad (4)$$

$$S_b = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(\bar{x}_k - \bar{X})^2}, \quad (5)$$

where S_k is the k-cluster standard deviation, $\bar{x}_k$ is the k-cluster cluster average, $\bar{X}$ is the overall cluster average, and k is the total members in each cluster. The determination of the number of groups can also be calculated using analysis of variance with the null hypothesis is there is no difference between groups. Null hypothesis will be rejected if $F_{count} > F_{(\alpha,a-1,N-1)}$ in analysis of varince table [14].
7. Describe cluster results using district map.

3. RESULTS AND DISCUSSIONS

The first step in our research is standardize data using the z-score. We use this data to test the multicollinearity. Table 2 shows the results of multicollinearity test. This table shows that there are ten variables that have r value more than 0.95 i.e. x_1, x_3, x_7, x_9, x_{11}, x_{13}, x_{17}, x_{19}, x_{25} and x_{27}. This indicate that those variables have multicollinearity with another variable. Therefore, these ten variables must be excluded.

After excluding the multicollinearity variables, there are 30 variables for the next step. We group these 30 variables with bottom-up clustering using the complete linkage rule, which is treated as a maximum distance between all two groups. The mutual cluster is determined using the bottom-up clustering. The result is shown in Figure 1. The x axis represent the province code. This figure shows that the rhombic structure signifies a mutual cluster. There are four mutual clusters i.e. (26, 27), (16, 20), (28, 11, 12, 13) and (18, 19, 21). The district groups are (Bangkalan, Sampang), (Mojokerto, Magetan), (Pamekasan, Bondowoso, Situbondo, Probolinggo), and (Nganjuk, Madiun, Ngawi).

Table 2. Ten variables that have *r* value more than 0.95 in the multicollinearity test.

1st variable	2nd variable	*r* value	1st variable	2nd variable	*r* value
x_1	x_2	0,989	x_{13}	x_{14}	0,958
x_3	x_4	0,988	x_{17}	x_{18}	0,999
x_7	x_8	0,985	x_{19}	x_{20}	0,988
x_9	x_{10}	0,953	x_{25}	x_{35}	0,994
x_{11}	x_{12}	1,000	x_{27}	x_{36}	0,994

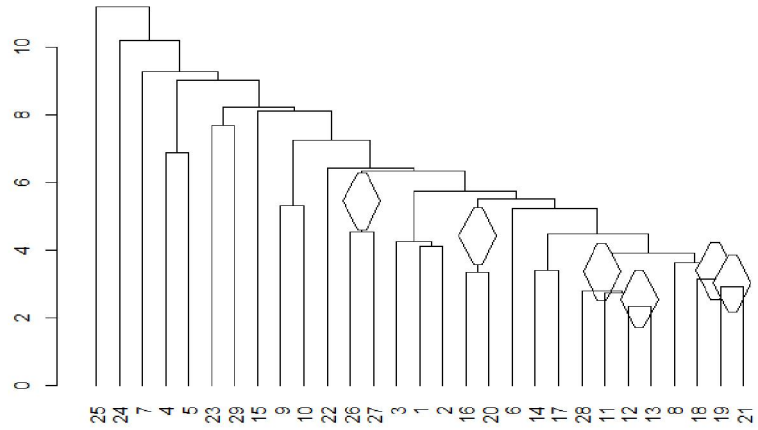

Figure 1. Bottom-up clustering with mutual clusters.

After bottom-up clustering is identified with mutual clusters, then top-down clustering is grouped by maintaining mutual clusters to obtain the hybrid hierarchical clustering results via mutual clusters (Figure 2). The best group in hybrid hierarchical clustering via mutual cluster determine using the smallest S_w, the highest S_b and the analyze of variance (see table 3). Table 3 shows the best group where the $F_{count} < F_{table}$. The map of these nine cluster is shown in Figure 3. The name of district based on table 3 are shown in table 4. This table shows the potential for leading commodities in the agricultural sector for each group.

Table 3. The results of S_w, S_b and F_{count} for the best group.

Group	S_w	S_b	S_w/S_b	F_{count}	F_{tabel}
2	1.121.466	1.721.318	0,652	36,049	4,21
3	924.683	1.297.065	0,713	18,272	3,37
4	823.775	1.194.451	0,690	11,769	2,99
5	844.357	1.430.553	0,590	16,312	2,78
6	731.907	1.386.258	0,528	13,067	2,64
7	719.086	1.277.569	0,563	11,203	2,55
8	683.420	1.231.766	0,555	9,253	2,49
9	725.934	1.475.978	0,492	7,908	2,45
10	596.079	1.479.760	0,403	8,559	2,55

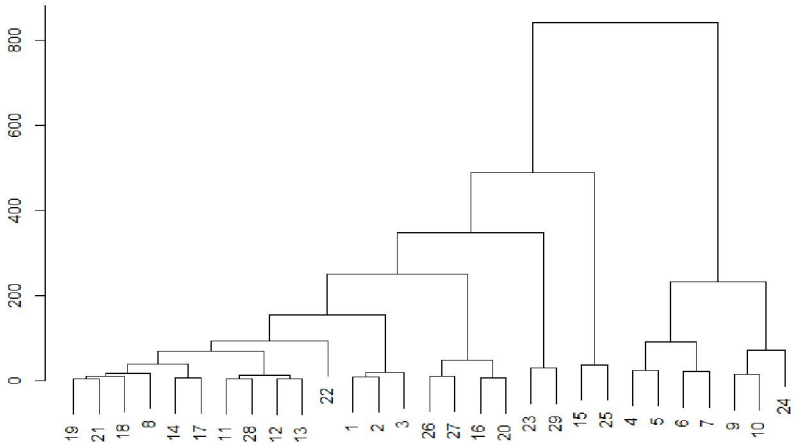

Figure 2. Grouping hybrid hierarchical clustering via mutual cluster.

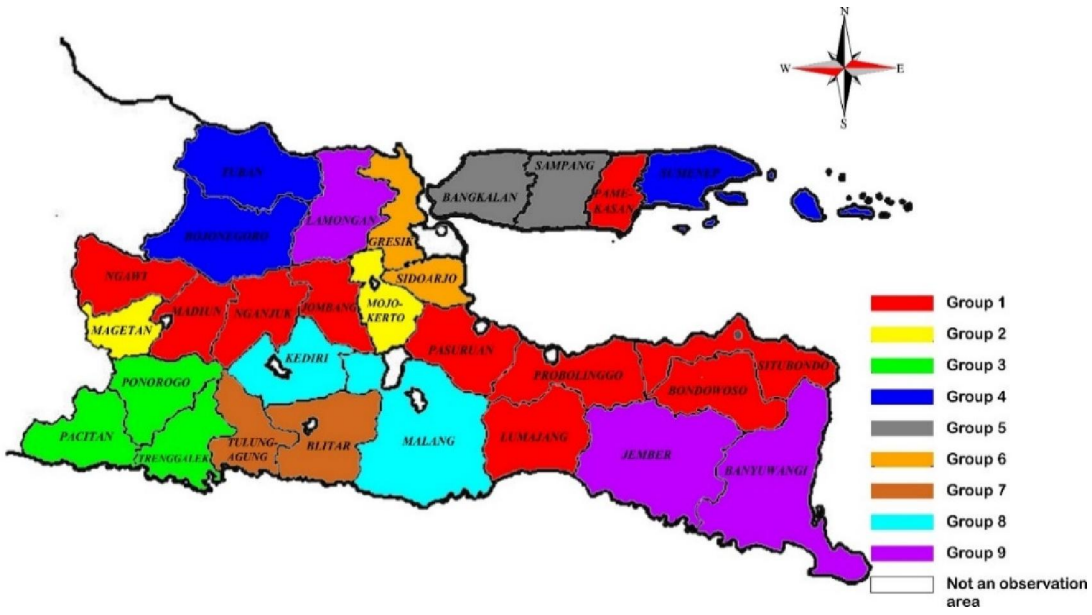

Figure 3. Map of Regency Grouping in East Java Province.

Table 4. Main Commodities of Each District Group in East Java Province.

Group	District	Main commodity
1	Madiun, Ngawi, Nganjuk, Lumajang, Pasuruan, Jombang, Bondowoso, Pamekasan, Situbondo, Probolinggo	-
2	Mojokerto, Magetan	sweet potato harvest area, sweet potato production
3	Pacitan, Ponorogo, Trenggalek	cassava production, goat population
4	Bojonegoro, Tuban, Sumenep	corn harvest area, corn production, green bean production, cow population, sheep population
5	Bangkalan, Sampang	peanut production
6	Sidoarjo, Gresik	beef production, fish consumption ponds area, fish consumption pond production, fish consumption pool area, fish consumption pool production
7	Tulungagung, Blitar	harvested rice production, native chicken population, goat meat production, kampong chicken production, duck meat production, duck eggs production
8	Kediri, Malang	sugarcane production, broiler meat production
9	Jember, Banyuwangi, Lamongan	paddy rice production, soybean production, coffee production, broiler population, lamb meat production, laying chicken production

4. CONCLUSION

Based on the results and discussion, the following conclusions are on multicollinearity test, there are several variables that experience multicollinearity, so it must exclude 10 variables, that is x_1, x_3, x_7, x_9, x_{11}, x_{13}, x_{17}, x_{19}, x_{25} and x_{27}. And the results of grouping hybrid hierarchical clustering method via mutual cluster that formed have the best performance with nine groups with s_w value of 725.934, s_b value of 1.475.978 and F_{count} value of 7,908. The grouping results of this research will help the government make decisions regarding the development of potential in each region and emerging regions whose potential is not yet fully developed.

REFERENCES

[1] BPS, *Hasil Survei Pertanian Antar Sensus (SUTAS) 2018*. Jakarta: Badan Pusat Statistik Indonesia, 2018.
[2] BPS, *Potensi Pertanian Provinsi Jawa Timur*. Surabaya: BPS Jawa Timur, 2013.
[3] N. Pigai, "Pertanian di Pulau Jawa Kritis, Lahan Habis untuk Infrastruktur," *Law Justice*, 2019. [Online]. Available: https://www.law-justice.co/artikel/60530/pertanian-di-pulau-jawa-kritis-lahan-habis-untuk-infrastruktur/. [Accessed: 15-Oct-2019].
[4] S. Nugroho, *Statistika Multivariat Terapan*. Bengkulu: UNIB Press Bengkulu, 2008.
[5] H. Chipman and R. Tibshirani, "Hybrid hierarchical clustering with applications to microarray data," *Oxford Univ.*, vol. 7, no. 2, pp. 286–301, 2006.
[6] A. Alfira, F. Hermin, and E. D. Wiraningsih, "Analisis Hybrid Mutual Clustering menggunakan Jarak Square Euclidean," pp. 9–15, 2016.
[7] A. K. Raharja and D. Agus, "Pengelompokan Kecamatan di Lamongan Berdasarkan Variabel

Sektor Pertanian dengan Metode Hybrid Hierarchical Clustering Via Mutual Cluster," pp. 1–6.
[8] D. Mariyani, S. W. Purnami, and W. S. Winahju, "Penerapan Hybrid Hierarchical Clustering melalui Mutual Cluster dalam Pengelompokan Kabupaten di Jawa Timur Berdasarkan Variabel Sektor Pertanian," pp. 1–10, 2011.
[9] J. Moeller, "A word on standardization in longitudinal studies: don't," *Front. Psychol.*, 2015, doi: 10.3389/fpsyg.2015.01389.
[10] Walpole, E. Ronald, and R. H. Myers, *Ilmu Peluang Dan Statistika untuk Insinyur dan Ilmuawan*, 4th ed. Bandung: Institut Teknologi Bandung, 1995.
[11] G. Rosenthal and J. A. Rosenthal, *Statistics and Data Interpretation for Social Work*. New York: Springer, 2012.
[12] R. Hocking, *Methods and Application of Linier Models*. New York: John Wiley & Sons, 1996.
[13] A. Rencher, *Methods Of Multivariate Analysis Second Edition*. Wiley Series In Probability and Mathematical Statistics, 2002.
[14] A. R. Barakbah and K. Arai, "Determining Constraints of Moving Variance to Find Global Optimum and Make Automatic Clustering Determining Constraints of Moving Variance to Find Global Optimum and Make Automatic Clustering," no. October, 2004.
[15] D. C. Montgonery, *Design and Analysis of Experiment, fifth edition*. New York: John Wiley & Sons, 2001.
[16] BPS Provinsi Jawa Timur, "Provinsi Jawa Timur Dalam Angka 2018," *BPS Provinsi Jawa Timur*, p. 390, 2018.