Comparison of hierarchical clustering methods (case study: data on poverty influence in North Sulawesi)

IOP Conf. Series: Materials Science and Engineering **567** (2019) 012048 doi:10.1088/1757-899X/567/1/012048

Comparison of hierarchical clustering methods (case study: data on poverty influence in North Sulawesi)

C E Mongi[1], Y A R Langi[1], C E J C Montolalu[1], N Nainggolan

[1]Department of Mathematics, Faculty of Mathematics and Natural Science,
University of Sam Ratulangi, Manado, Indonesia

Corresponding author: charlesmongi@unsrat.ac.id

Abstract. Grouping of Large data has been carried out in various fields. One method for grouping is cluster analysis where this method consists of hierarchy and non-hierarchy method. The aim of this study was to compare the use of cluster analysis on aspects of the causes of poverty data. The method used is agglomerative hierarchical clustering, that is, the average linkage, centroid methods and ward methods. The results obtained are compared with the RMSSTD value and the smallest value is the ward method with a value of 2.0937. So the ward method is good for this case.

Keywords : Hierarchical clustering, average linkage, centroid method, ward method.

1. Introduction

The problem of grouping large amounts of data has become important at this time. Grouping of Large data has been carried out in various fields, for example in the fields of soil and spatial [1][2], business [3], medicine [4], health [5] and other fields. One method for grouping is cluster analysis where this method consists of hierarchy and non-hierarchy method.

In cluster analysis we search for patterns in a data set by grouping the multivariate observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar but the clusters are dissimilar to each other[6]. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities)[7].

Hierarchical methods are among the traditional techniques of cluster analysis. They consist in successive aggregation or division of the observations and their subsets. Resulting from this kind of procedure there is a tree-like structure, which is referred to as dendrogram. The agglomerative techniques start from the set of observations, each of which is treated as a separate cluster. Clusters are aggregated in accordance with the decreasing degree of similarity (or the increasing degree of dissimilarity) until one, single cluster is established [8].

The application of the agglomerative hierarchy method will be used in data that affects poverty with aspects of population, economy, education and health. Population aspects are used data on population and unemployment, economic aspects are used gross regional domestic product data, educational aspects used net enrollment rate of elementary school, junior high school and senior high school, while the health aspect is used health insurance member data, health personal, and number of hospital.

In this article used 3 hierarchical method that is average linkage, centroid method, and ward method. In section 2 explained the formula of each method. In section 3 explained the use of the hierarchical method on the data that causes poverty in the North Sulawesi province. The application of the hierarchical method will be compared to each method using the RMSSTD value. The smallest RMSSTD value is a good method used for poverty influence data.

2. Hierarchical Methods

2.1. Average Linkage

The distance between two cluster is define by (1)

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{i \in C_L} d(x_i, x_j) \qquad (1)$$

D_{KL} is any distance or dissimilarity measure between cluster C_K and C_L. In average linkage the distance between two clusters is the average distance between pairs of obsevations, one in each cluster. Average linkage tends to join clusters with small variances, and it is slightly biased toward producing clusters with the same variance.

2.2. Centroid Method

The distance between two clusters is define by (2)

$$D_{KL} = \|\bar{x}_K - \bar{x}_L\|^2 \qquad (2)$$

In the centroid method, the distance between two clusters is defined as the Euclidean distance between their centroid or means.

2.3. Ward Method

The distance between two clusters is define by (3)

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{\frac{1}{N_K} - \frac{1}{N_L}} \qquad (3)$$

The distance between two cluster is the ANOVA sum of squares between the clusters added up over all the variables.

2.4. Evaluating the cluster

The root mean squared standard deviation of a cluster C_K is define on (4)

$$\text{RMSSTD} = \sqrt{\frac{\boldsymbol{W_K}}{\boldsymbol{v(N_K - 1)}}}, \qquad (4)$$

where $W_K = \sum_{\boldsymbol{i \in C_K}} \|\boldsymbol{x_i} - \boldsymbol{\bar{x}_K}\|^2$, N_K = number of observation in C_K [1]

$\boldsymbol{R^2 = 1 - \frac{P_G}{T}}$, P_G is $\sum \boldsymbol{W_J}$, where summation is over the *G* clusters at the *G*th level of the hierarchy

3. Aplication use influence poverty data in North Sulawesi Province

The data used consists of 15 objects which are 15 districts / cities in the province of North Sulawesi with variables namely the factors that cause poverty consist of population, unemployment, gross regional domestic product, elementary school net enrollment rate, junior high school net enrollment rate, senior high school net enrollment, health insurance member, personal health, number of hospital. From the data will be analyzed with three hierarchical methods. Data in the form of 9 variables and 15 objects, namely districts / cities in North Sulawesi province, were analyzed using the SAS University Edition program. Cluster formation results can be seen in section 3.1 for the average linkage, section 3.2 for the centroid method and section 3.3 for the ward method.

3.1. Result with Average linkage

In table 1. that can see the cluster formation process where column 1 is the number of clusters. The second column is the object that joins into a new cluster. The third column is the frequency of each cluster and the fourth column is the root mean square standard deviation value of each cluster. The fifth column is the norm value for root mean square distance. Figure 1 shows the dendogram using average linkage.

Table 1. Cluster history with average linkage

Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Norm RMS Distance
14	Talaud	Bolmut	2	0.3939	0.1313
13	Sitaro	Boltim	2	0.5114	0.1705
12	CL14	CL13	4	0.7350	0.2801
11	Minut	Bitung	2	1.0321	0.344
10	Sangihe	CL12	5	1.0511	0.4657
9	Minsel	CL11	3	1.3007	0.472
8	Bolmong	CL9	4	1.4852	0.5497
7	CL10	Mitra	6	1.4506	0.6752
6	CL7	Bolsel	7	1.6695	0.7067
5	Tomohon	Kotamo	2	2.2676	0.7559
4	CL8	CL6	11	1.9961	0.7647
3	CL4	CL5	13	2.1830	0.8629
2	CL3	Minahasa	14	2.3543	1.0649
1	CL2	Manado	15	3.0000	1.87

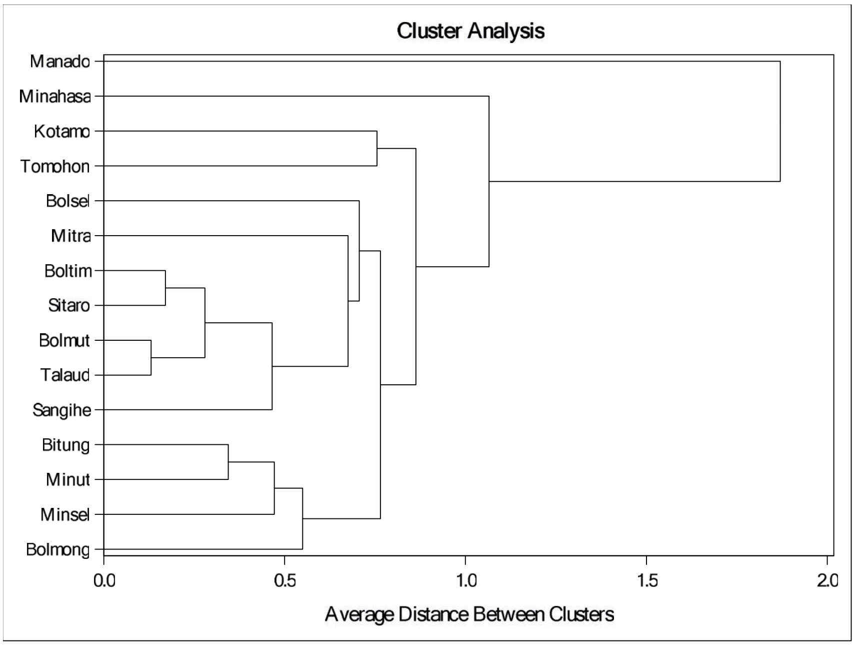

Figure 1. Dendogram used average linkage

3.2. Result with centroid method

Table 2. Cluster history using centroid method

Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Norm Centroid Distance
14	Talaud	Bolmut	2	0.3939	0.1313
13	Sitaro	Boltim	2	0.5114	0.1705
12	CL14	CL13	4	0.7350	0.2586
11	Minut	Bitung	2	1.0321	0.344
10	Minsel	CL11	3	1.3007	0.4396
9	Sangihe	CL12	5	1.0511	0.4409
8	Bolmong	CL10	4	1.4852	0.4894
7	CL8	CL9	9	1.7560	0.5837
6	CL7	Bolsel	10	1.8939	0.6856
5	CL6	Mitra	11	1.9961	0.6798
4	CL5	Tomohon	12	2.0753	0.6756
3	CL4	Kotamo	13	2.1830	0.7685
2	CL3	Minahasa	14	2.3543	0.9432
1	CL2	Manado	15	3.0000	1.7919

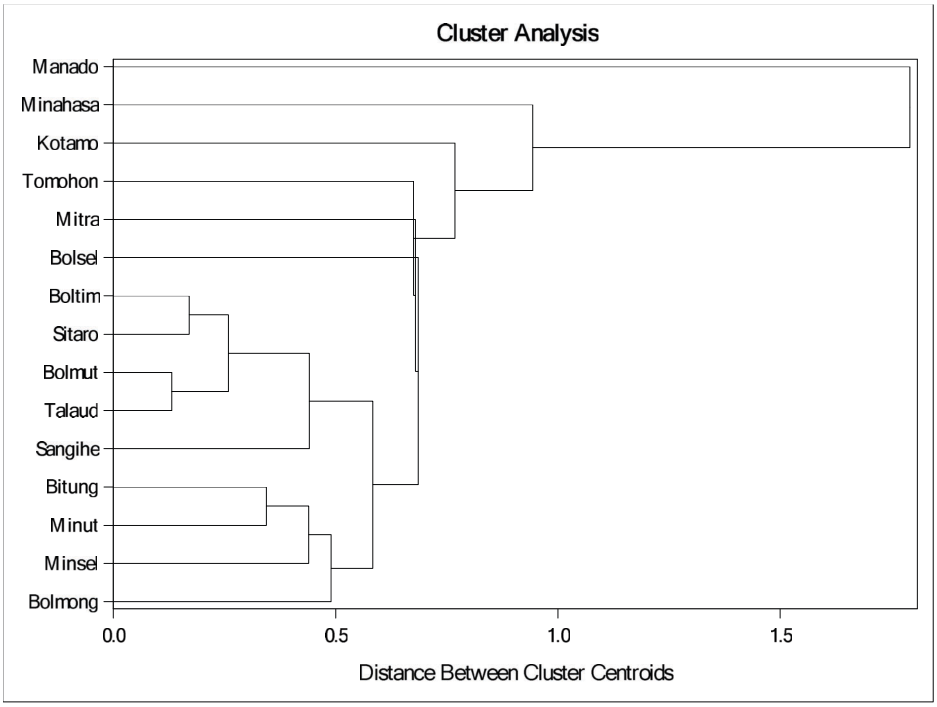

Figure 2. Dendogram used centroid method

3.3. Result with ward method

Table 3. Cluster history using ward method

Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square
14	Talaud	Bolmut	2	0.3939	0.0012	.999
13	Sitaro	Boltim	2	0.5114	0.0021	.997
12	Minut	Bitung	2	1.0321	0.0085	.988
11	CL14	CL13	4	0.7350	0.0096	.979
10	Sangihe	Tomohon	2	1.5075	0.0180	.961
9	Minsel	CL12	3	1.3007	0.0184	.942
8	Bolmong	CL9	4	1.4852	0.0257	.917
7	CL11	Bolsel	5	1.3194	0.0424	.874
6	CL10	Kotamo	3	1.9714	0.0437	.831
5	CL7	Mitra	6	1.6218	0.0491	.781
4	CL8	Minahasa	5	2.0185	0.0768	.705
3	CL6	CL5	9	2.0937	0.1123	.592
2	CL4	CL3	14	2.3543	0.1642	.428
1	CL2	Manado	15	3.0000	0.4281	.000

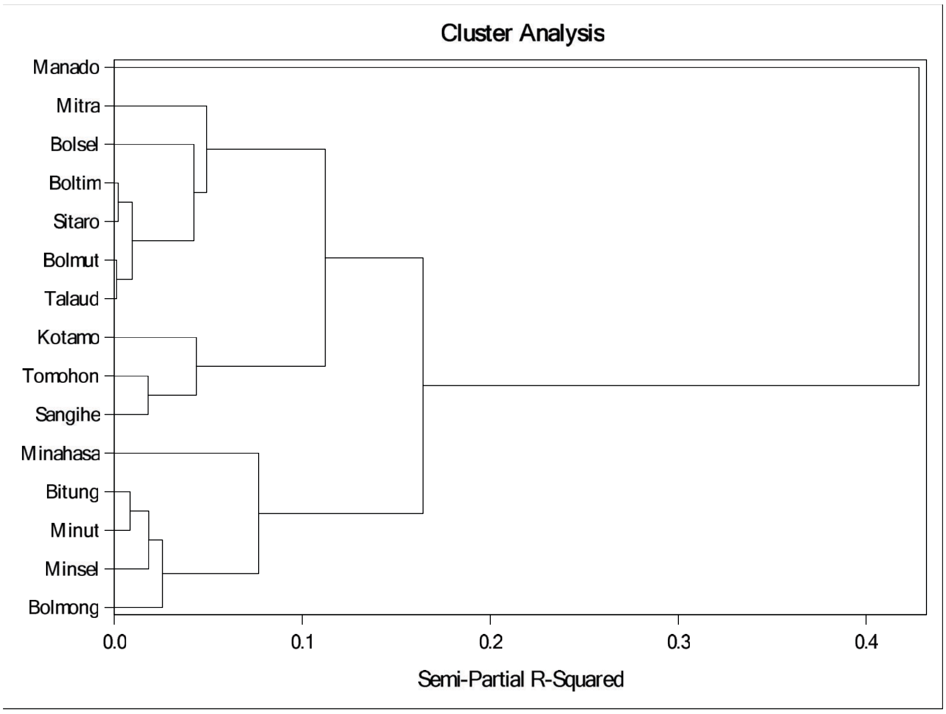

Figure 3. Dendogram used ward method

Based on the RMSSTD value obtained in Table 1,2,3 for the number of clusters 4 the smallest value at the average linkage is 1.9961. For the number of clusters 3 the smallest RMSSTD value is the ward method of 2.0937. Figure 2 and Figure 3 show the dendogram grouping using the centroid method and the ward method. In grouping using the ward method for 3 clusters, cluster 1 namely bolmong district, minsel, minut, bitung city, minahasa district. Cluster 2 is Sangihe District, Tomohon City, Kotamobagu City, Talaud District, Bolmut, Sitaro, Boltim, Bolsel, Mitra, and Cluster 3 only in Manado City.

4. Conclusion

A good method to use in the case of poverty data factor is the ward method based on the small RMSSTD value for the number of clusters selected by 3 clusters compared to other methods.

5. Acknowledgments

Thank you to the Sam Ratulangi University Manado ("Lembaga Penelitian dan Pengabdian Masyarakat"), for funding this research.

References

[1] A. M. Astel, L. Chepanova, and V. Simeonov. 2011. Soil contamination interpretation by the use of monitoring data analysis, *Water. Air. Soil Pollut.*, 216, 1–4, 375–390.

[2] L. Ke, L. Fan, and X. Kunqing. 2007. An efficient high dimensional cluster method and its application in global climate sets, *Data Sci. J.*, 6, SUPPL., S690–S697.

[3] H. Sun, M. Zhu, and F. He. 2013 . Cluster Analysis on PEAD Using SUE Model with Quarterly Data, *iBusiness*, 5, 31–34.

[4] M. Zhang, J. Zhang, Q. Gao, Y. Liu, and F. Lu. 2015. Evaluation Procedure for Quality Consistency of Generic Nifedipine Extended-Release Tablets Based on the Impurity Profile, *Am. J. Anal. Chem.*, 6, 776–785.

[5] C. Suhaeni, A. Kurnia, and Ristiyanti, 2018. Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki , K-Means Cluster , dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil), *J. media infotama*, 14, 1, 31–38.

[6] A. C. Rencher and W. F. Christensen, 2012. *Methods of Multivariate Analysis 3rd edition*. New Jersey: John Wiley & Sons, Inc.

[7] R. a. Johnson and D. W. Wichern, 2007. *Applied multivariate statistical analysis sixth edition*. New Jersey: Pearson Education, Inc.

[8] S. Wierzchon and M. Klopotek. 2018. *Modern Algorithms of Cluster Analysis*. Cham: Springer.

[9] SAS Institute Inc. 2017. *SAS/STAT® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.