APPLICATION OF K-MEANS ALGORITHM FOR CLUSTER ANALYSIS ON POVERTY OF PROVINCES IN INDONESIA

Albert V. Dian Sano[1]; Hendro Nindito[2]

[1,2] Information Systems Department, School of Information Systems, Bina Nusantara University
Jln. K.H. Syahdan No 9, Palmerah, Jakarta Barat, 11480
[1]albert_vds@yahoo.com; [2]h_nindito@yahoo.com

ABSTRACT

The objective of this study was to apply cluster analysis or also known as clustering on poverty data of provinces all over Indonesia.The problem was that the decision makers such as central government, local government and non-government organizations, which involved in poverty problems, needed a tool to support decision-making process related to social welfare problems. The method used in the cluster analysis was k-means algorithm. The data used in this study were drawn from Badan Pusat Statistik (BPS) or Central Bureau of Statistics on 2014.Cluster analysis in this study took characteristics of data such as absolute poverty of each province, relative number or percentage of poverty of each province, and the level of depth index poverty of each province in Indonesia. Results of cluster analysis in this study are presented in the form of grouping of clusters' members visually. Cluster analysis in the study can be used to identify more quickly and efficiently on poverty chart of all provinces all over Indonesia. The results of such identification can be used by policy makers who have interests of eradicating the problems associated with poverty and welfare distribution in Indonesia, ranging from government organizations, non-governmental organizations, and also private organizations.

Keywords: *cluster analysis, k-means, poverty*

INTRODUCTION

Poverty is one of the social problems and also become a challenge for many communities around the world to always find a solution. At the global level, data on poverty regarding the number of poor people is dominated by developing countries. However, in developed countries like the United States as well, there are still poor people. So poverty is everywhere universally to be a problem with the community and the world. In the national context, Indonesia, in the New Order era that began in the mid or late 1960s to 1996, the poverty rate in Indonesia decreased drastically due to strong economic growth with poverty alleviation programs. Unfortunately, the economic crisis in 1997/1998 hit the Indonesian economy and raised the amount of poverty in Indonesia sharply.

The experience of poverty reduction in the past have shown many weaknesses, such as: (1) macro growth orientation without considering aspects of equity, (2) centralized policy, (3) more caricature than transformative, (4) positioning communities as objects rather than subjects, (5) orientation poverty alleviation tends to caricature and instantaneous than sustainable productivity, and (6) perspectives and solutions that are generic to the problems of poverty that exist regardless of diversity which exists. Because it is so diverse nature of the challenges that exist, the handling of the problem of poverty must touch the bottom of the source and root of the real problem, either directly or indirectly (Multifiah, 2011).

Inequality or inequality in poverty reduction is not an easy problem. Reducing inequality is more complex than just reducing poverty. In addition, it also has the potential imbalances bigger problem than the problem of poverty itself. Figures economic inequality is measured using the Gini index. Gini index figures are presented in the form of a number between 0 and 1. The higher the number of Gini index, the higher the degree of economic inequality. Gini index numbers are high is a serious social threat and could lead to social unrest. In some countries that have a Gini index above 0.5 tends to be social unrest and disintegration.

Table 1 Gini Index Indonesia (BPS, 2015)

INDONESIA	**1996**	**1999**	**2002**	**2005**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
	0.355	0.308	0.329	0.363	0.364	0.35	0.37	0.38	0.41	0.41	0.413

To reduce drawbacks of the problems related to inequality in poverty reduction, then this study aims to represent the poverty map based on provinces groupings all over Indonesia. This study will apply cluster analysis or commonly called clustering. Results of this study are expected to facilitate the stakeholders to see the poverty map visually so that the decision makers associated with poverty alleviation programs can more quickly and easily set policies for provinces deserving to get higher priorities and to receive major attention in poverty alleviation programs.

The problem is that the decision makers such as central government, local government and non-government organizations, which involve in poverty problems, need a tool to support decision-making process related to social welfare problems. Cluster analysis in this study will try to group provinces all over Indonesia into four clusters. This study will consist of two types of cluster analysis. The first is to perform cluster analysis based on a variable of the percentage of the number of poor people and a variable of depth's level of poverty. The second is to perform cluster analysis based on a variable of an absolute number of poor people and a variable of depth's level of poverty. Cluster analysis is an analysis aimed at grouping the data objects into clusters based on similar variables or characteristics. Data objects having high similarities would be put in the same cluster while those having low similarities or a big difference will be put into different clusters.

Poverty has a diverse concept. World Bank defines poverty using purchasing power, i.e., US $2 per capita per day. Meanwhile, Badan Pusat Statistik or BPS (BPS-Statistics Indonesia) defines poverty based on the poverty line. According to BPS (2015), the poor are people who have an average monthly expenditure below the poverty line per capita. The poverty line, according to BPS, is the sum of the food poverty line (FPL) and Non-Food Poverty Line (NFPL). FPL is the spending on minimum food needs equivalent to 2100 calories per day per capita. Basic needs package of food commodities is represented by 52 kinds of commodities, such as grains, tubers, fish, meat, eggs, milk, vegetables, legumes, fruits, oils, fats, and others. NFPL is the minimum requirement for housing, clothing, health and education. Another definition of poverty by Bappenas (Purwanto, 2007) is a condition in which a person or a group of men and women who are unable to meet their basic rights to maintain and develop a dignified life. The basic rights of human beings include the fulfillment for food, clothing, health, education, employment, housing, clean water, land, natural resources and the environment, safe treatment or safe from the threat of violence and the right to participate in political and social life. Because this study will use a database of BPS in 2014, then it will automatically refer to the concept of poverty by BPS.

Cluster analysis technique in this study will apply to some specific characteristics or attributes of data, that is, the absolute number of poverty, the relative number or percentage of poverty, and the

level of poverty's depth index. The absolute number of poverty is the number of people living below the poverty line. The relative number is the percentage of population below the poverty line. The poverty's depth index is the average gap of expenditure of the poor to the poverty line. The higher the index value, the higher the gap between the average expenditure from the poverty line.

"We are living in the information age" is a popular saying; however, we are living in an era of data. The data in terabytes or petabytes poured into our computer network, World Wide Web (www), and various data storage devices each day ranging from world business, community, science and engineering, medicine, and almost every other aspect of daily life. The explosive growth of the volume of existing data is the result of the process of computerization of our society and the rapid development of various devices the collection and storage of data which is terrific (Han & Kamber, 2012).

The explosive widely available growth of data really makes us aware that we are in the era of data. Various reliable and versatile tools are needed to automatically reveal valuable information from the large-volume data and transform it into the organized knowledge. This need has led to the birth of data mining. The field is still young, dynamic and promising. Data mining has been and will continue to make great strides in our journey from the era of data into the information age to come (Han & Kamber, 2012).

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions (Hossain *et al.*, 2013)

Data mining is a fun way to extract various kinds of patterns, which presents knowledge stored in large data sets implicitly and focuses on matters related to its feasibility, usefulness, effectiveness and scalability. Data mining can also be seen as a very important step in the process to find knowledge. Data is normally done through a pre-process data cleansing, data integration, selection and transformation of data and prepared for mining. Data mining can also be done on different types of databases and data storage, but the type of pattern is found determined by different types of functionality mining data such as descriptions, association, correlation analysis, classification, prediction, analysis of clusters, and so on (Tajunisha, 2010).

The concept of data mining involves three steps, i.e., capturing and storing the data, converting the raw data into information and converting the information into knowledge. Data in this context comprises all the raw material an institution collects via normal operation. Capturing and storing the data is the first phase that is the process of applying mathematical and statistical formulas to "mine" the data warehouse (Kumar, 2011).

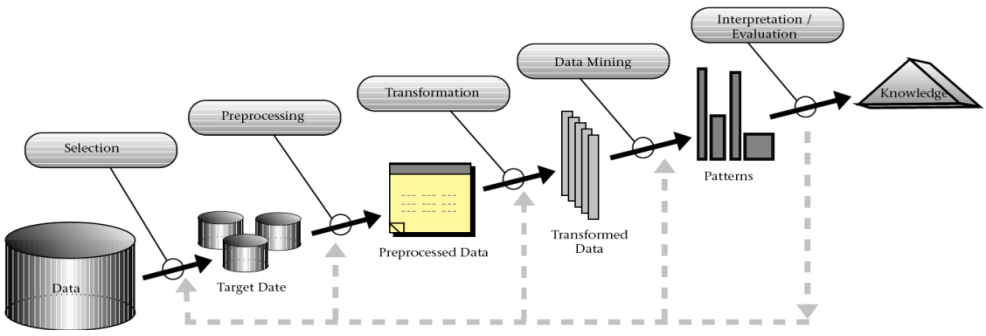

Figure 1 Data mining and Knowledge Discovery Process of Database
(Sources: Fayyad in Silwattananusarn & Tuamsuk, 2012)

Based on Figure 1 above, the knowledge discovery process consists of several sequential and iterative methods such as the following (Fayyad, Han, in Silwattananusarn & Tuamsuk, 2012): (1) Selection: Choosing yes ng Data relevant to the task of a database analyst. (2) Preprocessing: Delete the invalid data and inconsistent data; combine multiple sources of data. (3) Transformation: Transforming the data into a form suitable to perform data mining. (4) Data Mining: Choosing the data mining algorithm that matches the pattern in d nature of the data; extract various data patterns. (5) Interpretation / Evaluation: Interpret various patterns into knowledge by eliminating irrelevant various patterns and the same pattern and repetitive; translating a variety of patterns useful in terms that could be understood by ordinary people.

Clustering is an important method of data warehousing and data mining. It groups similar object together in a cluster (or clusters) and dissimilar object in other cluster (or clusters) or remove from the clustering process. But there are some special requirements for search results clustering algorithms, two of which most important is, clustering performance and meaningful cluster description (Gothai & Balasubramanie, 2012).

Cluster analysis also called clustering is the process of dividing a set of data objects (or object of observation) into several subsets. Each of these subsets is a cluster, such that the objects in a cluster are the objects that are similar to each other, but very different from the objects that are in another cluster. A set of clusters resulting from the cluster analysis such as clustering can be referred to as a clustering (Han & Kamber, 2012).

Cluster analysis offers a useful way to organize and present a complex dataset (Wang & Song, 2011). Analysis of the cluster can be regarded as the most popular techniques and foremost to solve problems that are unsupervised learning or the learning process undirected or unsupervised. So each technique used to solve problems with techniques like this, will certainly find a way of dealing with the structure of the data that has not been labeled (Tayal & Raghuwanshi, 2011).

One important component of the clustering algorithm is a measure of the distance between data points. If a component of the vector sample data is in the same physical unit, then it is likely that the simple Euclidean distance metric is sufficient to classify the data instants that are similar to each other. The distance between the two groups can be measured by (Tayal & Raghuwanshi, 2011) Euclidian and City Block or Manhattan.

In addition to the similarity and dissimilarity of the two types of measurement above, some of the other measurements are shown in Table 2 (Xu & Wunsch, 2005).

Table 2 Size of Similarity and Dissimilarity for Quantitative Variables

Measures	Forms
Minkowski distance	$d(x,y)=\left[\sum_{i=1}^{p}\lvert x_i-y_i\rvert^m\right]^{1/m}$
Euclidean distance	$D_{ij}=\sqrt{\sum_{k=1}^{p}(X_{ik}-X_{jk})^2}.$
City-block distance	$D_{ij}=\sum_{k=1}^{p}\lvert X_{ik}-X_{jk}\rvert$
Mahalanobis distance	$D_{ij}=(X_i-X_j)^t S^{-1}(X_i-X_j)$, where S is the within group covariance matrix

(Sources:Xu & Wunsch, 2005)

K-means is one of the learning algorithms undirected/unsupervised learning, the simplest used to solve various problems of the grouping. The procedure is by applying a simple and easy way to classify data that has been given to some clusters (such as clusters k) predefined (Tayal & Raghuwanshi, 2011).

K-means algorithm will define the midpoint of the cluster from the average value of the points in the cluster. Steps in k -means algorithm can be explained as follows. Per all, the algorithm will select k (central cluster) at random from various objects in D (dataset), which respectively represent the center of the cluster at the beginning or the first time. For any other object, each object is assigned or grouped into clusters that are most similar or the most closely based on the Euclidean distance between the object and the center cluster.

K-means clustering algorithm then iterates to improve or increase the separation distances or similarities in the cluster. For each cluster, this algorithm will calculate a new average using the objects are grouped into a cluster in the previous iteration. All objects will then be regrouped by using the average of the newly updated as the new cluster center. The iterations will continue until it reaches a stable grouping, which means that the clusters formed in the latest iteration are the same as the clusters formed in the previous iteration. K-means clustering procedure is summarized in Figure 2 below (Han & Kamber, 2012).

Algorithm: *k*-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial cluster centers;
2) **Repeat**
3) (re)assign each object to the cluster to which the object is the most similar,
4) based on the mean value of the objects in the cluster;
5) Update the cluster means, that is, calculate the mean value of the objects for
6) Each cluster;
7) **Until** no changes;

Figure 2 Summary of Procedure Algorithm K-Means
(Source: Han & Kamber, 2012)

As with any other algorithms, k-means also has some advantages and disadvantages. Here are the advantages and disadvantages of k-means algorithm according to Tayal and Raghuwanshi (2011). The advantages are (1) k-means is a simple algorithm that has been adapted to many domains. (2) More automated than making threshold imply a manual of an *image* or images. (3) This is an algorithm that can be good candidates for use as a continuation of the work relates to vectors that have the characteristics feature or vague (fuzzy).

On the other hand, the disadvantages are (1) though it can be demonstrated that the procedure will always end, k-means clustering algorithm does not always find the most optimal configuration, which is related to the global objective function. (2) This algorithm is also very sensitive to cluster centers randomly selected at the beginning. K-means algorithm can be run several times to reduce the impact on this problem.

1. MSE = largenumber;
2. Select initial cluster centroids $\{m_j\}_j$
   K = 1;
3. Do
4. OldMSE = MSE;
5. MSE1 = 0;
6. For j = 1 to k
7. mj = 0; nj = 0;
8. Endfor
9. For i = 1 to n
10. For j = 1 to$k$
11. Compute squared Euclidean distance $d^2(x_i, m_j)$;
12. Endfor
13. Find the closest centroid m_j to x_i
14. $m_{j=} m_j + x_i$, $n_j = n_j + 1$;
15. MSE1 = MSE1 + $d^2(x_i, m_j)$;
16. Endfor
17. For j = 1 to k
18. n_j = max (n_j , 1) ; $m_j = m_j / n_j$;
19. Endfor
20. MSE = MSE1; while (MSE<OldMSE)

Figure 3 Traditional k-means Algorithm
(Source: Oyelade, *et al.,* 2010)

METHODS

The method applied in this study generally includes three main stages: (1) data collection, (2) data pre-processing, and (3) data mining. In data collection, data collected in this study was taken from BPS-Statistics Indonesia's (Badan Pusat Statistik) website (www.bps.go.id). Next, data pre-processing is the most important task in data mining. This stage is often said to take almost 80% of the total time or task in data mining. Techniques and methods to be applied in this stage must be precise and correct. Data pre-processing used in this study is based on the theory by Jiawei Han and Michelin which includes: first, data cleaning: filling in the missing values, repairing data errors, identify or remove outliers, and fixing inconsistent data. Second, data integration: merging related data from tables, databases, cube, or files. Third, data selection: Selecting data only related to the process of analysis. The benefit of this step is to reduce less important or less relevant data in data mining processes.

Fourth, data transformation: Transforming data to support the process of analyzing the data that will be used. Fifth, data mining: This stage is the primary stage of the entire task in this study. As with the data collection as well as data pre-processing, this stage also applies the theory by Jiawei Han and Michelin which include: (1) Data Mining, this stage is the stage of the implementation of the modeling used in data mining. In this study, the model applied is k-means cluster analysis. (2) Pattern Evaluation, this is an evaluation of the pattern that has been processed. (3) Knowledge Presentation, this is a presentation of the results of the data mining process.

RESULTS AND DISCUSSIONS

The application of cluster analysis in this study applied four numbers of clusters. The analysis applies to two types of measurement of the number of poor people in all provinces all over Indonesia. The first one is a measurement based on the percentage of poor people or also known as the relative

number of poor people. The second is a measurement of the number of poor people as it is or also called as the absolute number of poor people.

The source of data for this study was taken from BPS-Statistics Indonesia's (Badan Pusat Statistik) website. Data applied in this study is the data that was last updated in September 2014. The data has 16 attributes. In this cluster analysis, the software used is RapidMiner Studio. Pre-processing will select and determine four relevant attributes to be analyzed, namely (1) Province attribute, (2) Number of Poor (Village + City) attribute, (3) Total Percentage of Poor (Village + City) attribute, and (4) P1 (Village + City) attribute. Province attribute will act as an identifier whereas P1 attribute describes the level of the depth of poverty. The study runs two cluster analysis processes applying k-means algorithm. The first process is to process Number of Percentage of Poor People vs. Depth Poverty Level while the second process is to process Absolute Number of Poor People vs. Poverty Depth's Level. The first process will apply attributes (1), (3) and (4) while the second process will apply attributes (1), (2), and (4).

Both cluster analysis in this study implements similarities and dissimilarities between data objects based on Euclidian distance measurement method. So suppose, i = (χi1, χi2, ..., χip) and j = (χj1, χj2, ..., χj1) are two objects described by a numerical attribute p, then to measure the Euclidian distance between these objects is (Han, 2012):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (1)$$

The similarity measurement technique using the Euclidean method as above also meets the mathematical properties such as the following (Han, 2012): (1) Non-negative: d (i, j) ≥ 0: The distance is not possible to be negative. (2) The identity of an indistinguishable: d (i, i) = 0: The distance of an object to itself is 0. (3) Symmetrical: d (i, j) = d (j, i): The distance is a function of symmetry. (4) Triangle inequality: d (i, j) ≤ d (i, k) + d (k, j): The distance of the object i to j cannot be greater than the distance going through k.

The plot of figure 4 below is the result of a cluster analysis based on the relative number of poor people and poverty depth's level in all provinces all over Indonesia.

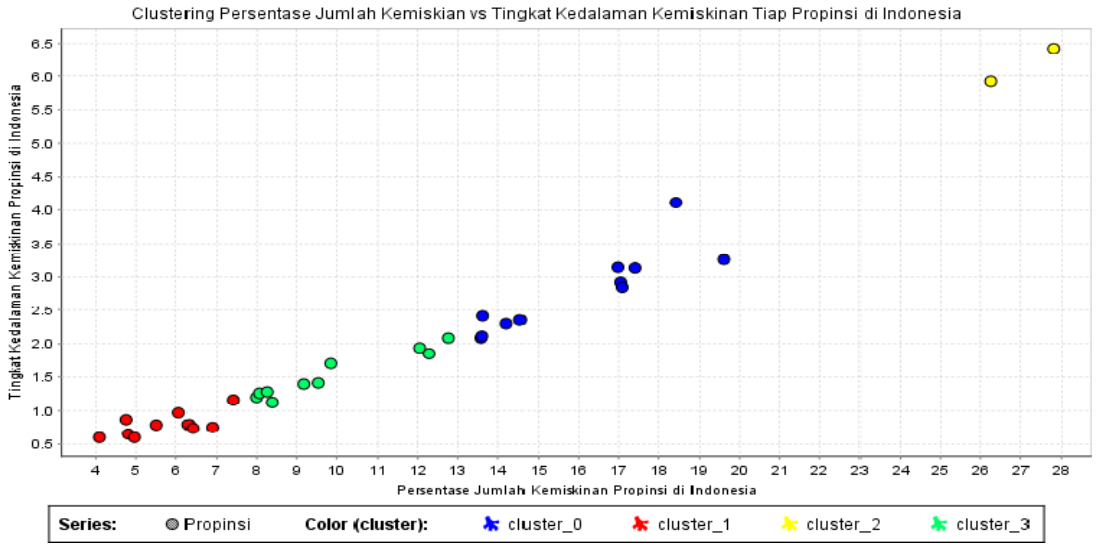

Figure 4 Plot of Cluster Analysis of the Percentage
of Poverty vs. Poverty's Depth for Each Province in Indonesia

Figure 4 describes that the graph between the percentages of poverty versus poverty depth's level tends to be linear. It means that the provinces having a large percentage of poverty also tend to have deep level of poverty. Thus, to set policies as to how to prioritize provinces coping up with poverty alleviation tends to be easier. The figure also describes that cluster_2 (yellow color) be a representation of two provinces having the greatest percentage of poor people and the deepest level of poverty. So both two provinces should receive top priority in poverty handling when viewed from the percentage and depth of poverty points of view. When viewed from the dataset, the two provinces are provinces of Papua and West Papua. Then the next provinces to highlight are provinces in cluster_0 (blue color). When we look at the dataset, provinces in the list of cluster_0 are the province of Aceh, South Sumatra Province, Bengkulu, Lampung, Central Java, Yogyakarta, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Gorontalo, and Maluku.

Figure 5 shows the result of a cluster analysis based on the absolute number of poor people (not the amount in percentage) and the depth level of poverty of all provinces in Indonesia.

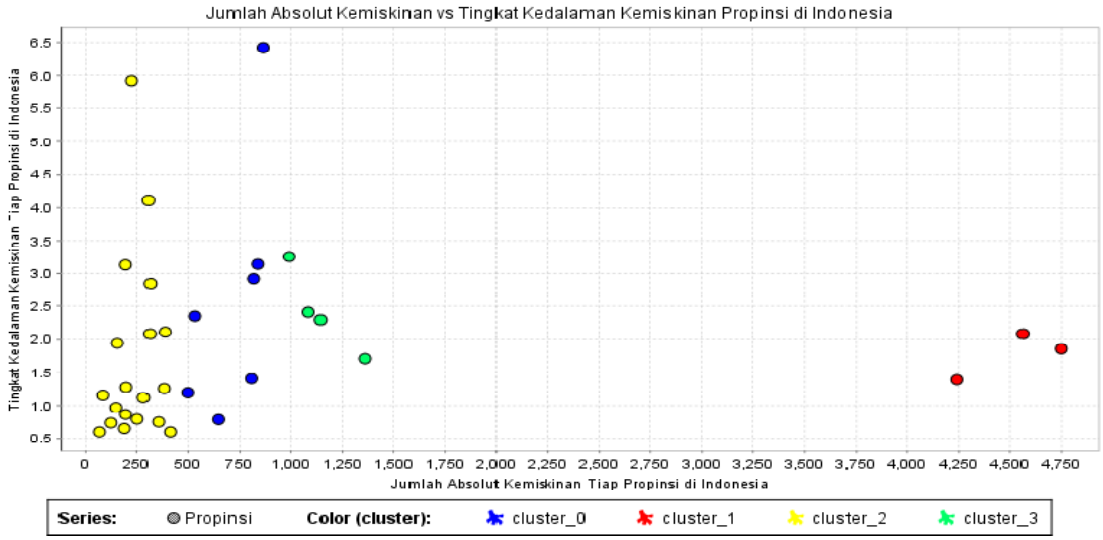

Figure 5 Plot of Cluster Analysis of the Percentage
of Poverty vs. Poverty's Depth of Each Province in Indonesia

Figure 5 describes that there is a quite outstanding cluster and in a remotely separated area, which is cluster_1 (red color). This cluster should also receive top priority in poverty reduction if poverty alleviation program refers to the absolute number of poor people. From the dataset, we know the provinces in the list of cluster_1 cluster are East Java, Central Java and West Java. The next priority to be considered is cluster_3 (green color). From the dataset, it can be identified easily that provinces in the list are the provinces of North Sumatra, Lampung, South Sumatra and East Nusa Tenggara. However the figure above also shows that the relationship between the numbers of absolute poverty versus the depth of poverty of provinces in Indonesia is not linear. This means that cluster_1 and some members of cluster_3 have a large number of absolute poverty but the depth of poverty is not high. If poverty reduction policies are more oriented towards poverty depth's level then some members of cluster_2 (yellow color) or cluster_0 (blue color) are the target of the policy. Members of the said clusters are the provinces of Papua (cluster_0) and West Papua (cluster_2).

A more comprehensive approach in setting policies for poverty reduction based on priorities can combine both types of cluster analysis above. If such a combination is applied then some provinces that should be prioritized in poverty reduction are Provinces of Papua, West Papua, East

Java, Central Java, West Java, Aceh, South Sumatra, Bengkulu, Lampung, Yogyakarta, West Nusa Tenggara, East Nusa East, Central Sulawesi, Gorontalo, and Maluku.

As with many other types of studies, this study certainly is not perfect. Some potential weakness in this study identified by the researchers is the need for comparing the accuracy of cluster analysis if the cluster analysis is run multiple times with a different number of clusters. Another potential weakness is the need for comparing accuracy of cluster analysis by using different models of cluster analysis such as k-medoids or other.

CONCLUSIONS

This cluster analysis study can provide information more quickly and efficiently on the distribution picture of the poor provinces all over Indonesia. Results of this cluster analysis of poverty in Indonesia provide visual information that is useful to see the map of the poorest provinces which should become the target of policies of stakeholders. Provinces that need attention and become the target of the policies ordered by priorities are provinces of Papua and West Papua, East Java, Central Java, West Java, Aceh, South Sumatra, Bengkulu, Lampung, Yogyakarta, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Gorontalo, and Maluku. Results of cluster analysis in this study could be used by stakeholders associated with poverty alleviation programs such as government agencies, non-governmental organizations and private institutions as part of the decision-making processes.

REFERENCES

BPS. (2015). *Gini Ratio Menurut Provinsi Tahun 1996, 1999, 2002, 2005, 2007-2013*. Retrieved August 3, 2015 from http://www.bps.go.id/linkTabelStatis/view/id/1493

BPS. (2015). *Konsep Penduduk Miskin*. Retrieved August 3, 2015 from http://www.bps.go.id/Subjek/view/id/23#subjekViewTab1|accordion-daftar-subjek1

Gothai, E., Balasubramanie, P. 2012. An efficient way for clustering using alternative decision tree. *American Journal of Applied Science, 9,* 531-534.

Han, J., Kamber, M. (2012). *Data Mining: Concepts and Techniques* (4th ed.). San Francisco: Morgan Kaufmann Publishers.

Hossain, J., Sani, N. F. M., Mustapha A., & Affendey L. S., 2013. Using feature selection as accuracy benchmarking in clinical data mining. *Journal of Computer Science, 9,* 883-888.

Kumar, S. P., Ramaswami, K. S. (2011). Fuzzy modeled k-cluster quality mining of hidden knowledge for decision support. *Journal of Computer Science, 7,* 1652-1658.

Multifiah. (2011). Telaah Kritis Kebijakan Penanggulan Kemiskinan Dalam Tinjauan Konstitusi. *Journal of Indonesian Applied Economics, 5*(1), 1-27. Retrieved August 4, 2015 from http://jiae.ub.ac.id/index.php/jiae/article/view/109

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. *International Journal of Computer Science and Information Security (IJCSIS), 7*(1). Retrieved on August 3rd, 2015 from http://arxiv.org/ftp/arxiv/papers/1002/1002.2425.pdf

Purwanto, E. A. (2007). Mengkaji Potensi Usaha Kecil dan Menengah (UKM) untuk Pembuatan Kebijakan Anti Kemiskinan di Indonesia. *Jurnal Ilmu Sosial dan Ilmu Politik, 10*(3), 295-324.

Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 2*(5). Retrieved on August 3, 2015 from http://arxiv.org/ftp/arxiv/papers/1210/1210.2872.pdf

Tajunisha, S. (2010). Performance analysis of k-means with different initialization methods for high dimensional data. *International Journal of Artificial Intelligence & Applications (IJAIA), 1(4),* 44-52. Retrieved on August 3, 2015 from https://www.academia.edu/12640770/Performance_analysis_of_k-means_with_different_initialization_methods_for_high_dimensional_data

Tayal, M. A., & Raghuwanshi, M. M. (2011). Review on Various Clustering Methods for the Image Data. *Journal of Emerging Trends in Computing and Information Sciences, 2,* 34-38, *Special Issue.*

Wang, H., & Song, M. (Desember, 2011). Ckmeans.1d.dp: Optimal k-means clustering in One Dimension by Dynamic Programming. *The R Journal, 3*(2), 29-32.

Xu, R., Wunsch, D. C. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16* (3), 645 - 678.