Forum Statistika dan Komputasi : *Indonesian Journal of Statistics*
ISSN : 0853-8115
Vol. 20 No.2, Oktober 2015, p: 53-59

available online at:
journal.ipb.ac.id/index.php/statistika

CLUSTERING PROVINCE IN INDONESIA BY COMMUNICATION TECHNOLOGY RELATED VARIABLES

Ahmad Nur Rohman[1], Erfiani[2], Muhammad Nur Aidi[2]

1. Graduate Student Applied Statistics, Bogor Agricultural University, Indonesia
2. Graduate Lecturer Bogor Agricultural Statistics, Indonesia
E-mail : an_rohman@yahoo.co.id[1]

ABSTRACT

Technological developments in Indonesia growth rapidly. Almost all systems used in daily life have been using the technology. One of its technology is communication technology. It because communication technology is a important tool for send information. All was done in order to communicate easier and faster. It is therefore important to research the condition of the existing communication technology in Indonesia. Communications technology also one of the focus of the government in national development. But not easy to know the state of communication technology in Indonesia because Indonesia has a large region and different geographically. The purpose of this research was to determine the grouping of provinces in Indonesia to increase the communication sector in order to support national development. The method used in this research is cluster hierarchical analysis method and criterion of determining the best method and many cluster optimal use Cubic Clustering Criterion (CCC). The data used is secondary data from the Statisctics Indonesia (BPS) and the Ministry of Communication and Information. The results showed that the number of cluster based on related communication technology variables are 3 cluster which 1st cluster members consist of 21 provinces, 2nd cluster members consist of 7 provinces and 3rd cluster members consist of 3 provinces.

Key words : Communications Technology, Cluster Analysis, Hierarchical Method, Cubic Clustering Criterion (CCC)

INTRODUCTION

Dengue Hemorrhagic Fever (DHF) is one Communication technology is a technology that makes it easy to everyone to communicate with others when performing daily activities to support life. Communication technology become one of the government's long-term plan, contained in Indonesian Undang-Undang No. 17 tahun 2007 on the long-term national development plan from 2005 to 2025, that Indonesian Information Society is projected to be realized in the third medium-term period, it is 2015 to 2019. Communication technology has also become one of the sectors that support the economy in Indonesia. Indonesian Information Society would be realized if the communication technology throughout Indonesia has been developed. But there are some things that become constraint to achieve, for example: a very wide area of Indonesia is 1910931.32 km^2 which consists of 17 504 islands (BPS, 2014) and has the different geographically. This cause government's difficulty to apply the policy in each region. Therefore it is necessary to know the conditions of each region in Indonesia in the field of communication technology.

One way to do is to group provinces in Indonesia according to the condition of each communication technology. Provincial grouping can be done by using a statistical tool that cluster analysis. Cluster analysis classifying objects based on similarities between the characteristics of these objects. There are a lot of research using cluster analysis that have been done, for example: grouping the respondents according to their characteristics seen from internet content is used and its central role in internet communications (Nechaev et al, 2010), Analytical cluster to classify districts / cities in Central Java is based on the production of crops (Safitri et al, 2012). While the development of methods in the cluster analysis has also conducted research that compares several approaches in the cluster analysis for nominal data, the data in large numbers as well as categorical data (Rezankova, 2010), as well as review and comparison of several methods in the cluster analysis and main characteristics of each method (Popat, 2014).

Based on the description above, the researchers want to do research to clustering the provinces in Indonesia based on variables related to communication technologies using the cluster analysis and is expected to provide input to the government in order to improve the communication sector to realize an information society appropriate national development goals.

.

Communication technology

Information and communication technology (ICT) is an umbrella term that covers all major technical equipment to process and communicate information (BPS, 2014). ICT has two aspects: information technology and communication technology. Information technology are all things that related to process, use, processing and management of information, communications technology is being everything to do with the use of tools to process and transfer data from one device to another.

Communication is also one of the subsectors in the national economy. Currently, communication is the one of national's development focus. The communications subsector has 22 economic activities. There are national postal, units of postal services, courier services, fixed network, mobile network terrestrial, mobile cellular networks, satellite mobile network, network call premium, radio service call to public, service radio trunking, kiosk, other telecommunications services, Internet service providers, communication system services, portal services, voice over internet protocol services, internet cafes, other multimedia services, specifically for its own telecommunication, telecommunications specifically for security, telecommunications specifically for broadcasting.

Cluster analysis

Cluster analysis is one of multivariate analysis to classify objects that have similar characteristics with one another into one group, and other objects into another group (Hair et al, 1995). In determining the group may use a measure that is a measure of distance, while the grouping there are two methods that can be used the hierarchical method and non-hierarchical methods.

Distance

Distance measurements were used to classify the objects that are similar in the same cluster. Distance measure commonly used include:

1. Euclidean Distance

   Euclidean distance is used for continu data. Euclidean distance between cluster i and j of p variables defined by

$$d_{euc}(x,y) = \left[\sum_{j=1}^{p}\left(x_j - y_j\right)^2\right]^{\frac{1}{2}}$$

2. Minkwoski Distance

   Minkwoski distance a common form of euclidean distance, manhattan distance and maximum distance. The distance function is as follows

$$d_{min}(x,y) = \left(\sum_{j=1}^{d}\left|x_j - y_j\right|^r\right)^{\frac{1}{r}} ,r \geq 1$$

   If the value of r of order 2 will be the euclidean distances, when r = 1 then a manhattan distance and when r = ∞ then it would be the maximum distance.

3. Mahalanobis Distance

   Mahalanobis distance useful to eliminate or reduce the difference in scale of each variable because this distance considering the value range and its covarian. Mahalanobis distance between two objects i and j is expressed in the form of vectors and matrices:

$$d_{mah}(x,y) = \left(\left(\mathrm{X}_{ik} - \mathrm{Y}_{jk}\right)\mathrm{S}^{-1}\left(\mathrm{X}_{ik} - \mathrm{Y}_{jk}\right)'\right)^{\frac{1}{2}}$$

With
S : covariance matrix of the data
X_i : vector of the object i
X_J : vector of the object j

Hierarchical Methods

Hierarchical method is a method used when the unknown number of cluster. There are two basic types used in hierarchical method that are agglomerative and divisive (Hair et al, 1995). Agglomerative method that considers all observation is cluster, so the two nearest objects that have similarities to one cluster and then proceed to another object. The procedure of divisive method is inverse of the agglomerative method that considers all observation is one cluster, then observations are not the same going out and form a new cluster. Hierarchical method used is agglomerative method. Agglomerative hierarchical method commonly used are single linkage, complete linkage, average linkage, Ward's method and centroid.

1. Single linkage

   Single linkage method uses clustering principle is based on the shortest distance between its members begins to look for 2 objects nearby and both create cluster first and then proceed to the next object served until each member has a cluster.
2. Complete linkage

   Complete linkage method is the opposite of single linkage method, using clustering principle is based on the longest distance between its members.
3. Average linkage

   Average linkage method based on the average distance obtained by the average of all the object distance in advance. Distance between groups (i, j) with k is

   $$d(ij,k) = \text{rata} - \text{rata}\ (d_{ik}, d_{jk})$$
4. Ward's method

   Ward's method is based on the cluster because miss information as a result of merging the object into a cluster, so as to find a cluster using this method measured total squared deviations from the mean cluster on each observation. Two objects will be one cluster if it has the smallest objective function among the possibilities. The objective function can be searched by the following formula.

   $$\text{ESS} = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j}\left(\sum_{i=1}^{n_j} x_{ij}^2\right)\right)$$

   With
   x_{ij} : the value of the object to-i in group j
   k : the number of groups each stage
   n_j : the number of the i-th group in the group j

5. Centroid method

   Centroid method is based by using the average distance to a group that is obtained by calculating the average for each variable for all objects. In this method each formed a new group then conducted a recount centroid to obtain a fixed group.

Cubic clustering criterion (ccc)

Sarle (1983) conducted a simulation to the development of the CCC, which is used in the determination of the amount of cluster. CCC is obtained by comparing the observed values of R^2 with the approximated of the expected value of R^2. CCC can be calculated using the formula

$$CCC = ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{(0.001+E(R^2))^{1.2}}$$

With
R^2 : proportion of varian accounted by cluster
$E(R^2)$: expected value of R^2
n : the numbers of observation
$p^* < p$, p is the number of variables

Positive CCC value indicates that the value of R^2 is greater than the expected value of R^2 means it can be used in the determination of many groups. CCC value of more than 2 or 3 indicates a good cluster, CCC value between 0 and 2 indicate a potential cluster, whereas the large negative CCC value indicates outliers.

RESEARCH METHODOLOGY

Data

This research used secondary data from the publication of *Badan Pusat Statistik* (BPS) in 2011 and the Data Communication and information related to the field of communications in 2011. The variables are used as part of the development goals of communication ministry of communications and informatics 2010-2014.

Method

This research use cluster analysis with hierarchical method. This is because the grouping condition is not known communication technology in Indonesia. Steps of cluster analysis are:

1. Data collection
2. Determine the distance to be used
3. Determine the method used
4. Determine the number of groups and the optimal method

5. Interpretation cluster

Table 1. Variables of Field Communication Technology

No	Variable	Symbol	Source of Data
1	The number of households own fixed wired telephone	X_1	BPS, Potensi Desa (2011
2	Number of villages own public telephone facility	X_2	
3	Number of villages own kiosks facility	X_3	
4	Number of villages own internet cafe facility	X_4	
5	Number of villages own Base Transceiver Station (BTS) tower	X_5	
6	Number of villages by strong mobile telephone signal reception	X_6	
7	Number of villages can be receive the television program / broadcast	X_7	
8	Percentage of population owns mobile telephone	X_8	BPS, Survei Sosial Ekonomi Nasional (2011)
9	Percentage of households owns computer	X_9	
10	Percentage of households ever accessing internet in the last 3 months	X_{10}	
11	Average Household consumption	X_{11}	
12	Average household consumption for telecommunications	X_{12}	
13	Number of the frequency band used	X_{13}	Kementerian Komunikasi dan Informatika (2011)
14	Number of Radio Frequency used	X_{14}	
15	Number of television frequencies used	X_{15}	
16	Number of of frequencies GSM/DCS used	X_{16}	
17	TV Channel Utilization	X_{17}	
18	Radio Channel Utilization	X_{18}	

RESULTS AND DISCUSSION

Data in this research has a description of the data can be seen in Table 2. Description of the data was conducted to determine the characteristics of the data so that it can be used a reference in using the method to be applied.

Table 2 shows that this research has a variable with a range of values that are very varied and have different measurement. Therefore, it is necessary to standardize the data to match all types of units to transform the data into a standard normal N ~ (0.1) so that the cluster becomes more valid analysis. Standardize data using equations.

$$Z = \frac{X_i - \bar{X}}{S}$$

With : X_i : Data -i
$\bar{X}$: Mean of Data
S : Standard Deviation

Cluster analysis which is used in this research are single linkage, complete linkage, average linkage, centroid linkage and ward linkage. One of these is selesction to be the best methods. In determining the many groups and the most appropriate method, according Sarle (1983) there is one of the criteria that can be used with a value of CCC.

Measure of distance used in analysis is euclidean distance for continuous scale data and the data has been standardized. Overall results of the analysis using hierarchical method can be seen in Table 3.

Table 3 shows that the method has a CCC that more than zero is Ward and complete linkage method, but the method ward is better because it has more than 1 value CCC more than complete linkage. Many clusters that optimal is 3 because the value of CCC started to be consistent at 3 cluster. To clarify the optimal number of cluster that can be seen in Figure 1. The members of each group are shown in Table 4.

Table2. Data Description

Variable	Minimum	Maximum	Mean	Std. Deviation
X_1	5488	1171832	173531.55	286654.37
X_2	9	802	151.29	189.09
X_3	22	1916	274.29	461.79
X_4	60	3322	540.48	804.90
X_5	112	3219	710.16	802.49
X_6	260	7356	1720.58	1907.99
X_7	65	22678	3488.55	5693.71
X_8	20.13	62.26	38.94	9.82
X_9	5.72	30.28	12.98	5.95
X_{10}	10.49	56.85	24.85	9.81
X_{11}	1913106.50	4813890.21	2775372.76	588980.83
X_{12}	61082.24	215484.93	96725.82	32267.72
X_{13}	476	47958	10702.42	11725.97
X_{14}	4	213	51.29	51.18
X_{15}	2	44	17.65	10.16
X_{16}	82	12272	2673.52	3068.89
X_{17}	8.25	100.00	30.99	25.41
X_{18}	3.57	100.00	22.41	23.65

Table 3. Value of CCC all method

Number of cluster	Single Linkage	Complete Linkage	Average Linkage	Centroid	Ward
2	-4.1	0.49	0.49	-4.1	0.49
3	0.10	0.10	0.10	0.10	1.03
4	-1.4	0.88	-0.56	-0.56	1.1
5	-1.9	1.01	-0.56	-1.4	1.24
6	-2.8	1.00	-1.0	-1.0	1.21

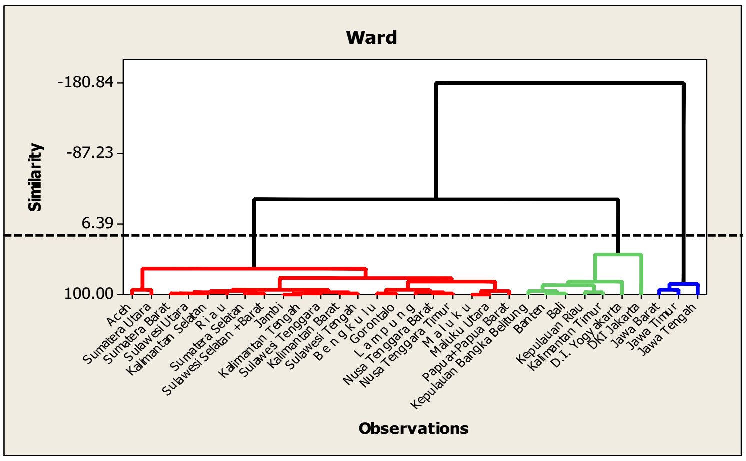

Figure 1 Dendogram with Ward method

Table 4. Members of Each Group

Cluster 1 (G1)		Cluster 2 (G2)	Cluster 3 (G3)
Aceh	Kalimantan Barat	DKI Jakarta	Jawa Barat
Sumatra Utara	Kalimantan Tengah	Kep Bangka Belitung	Jawa Tengah
Sumatra Barat	Kalimantan Selatan	Kepulauan Riau	Jawa Timur
Riau	Sulawesi Utara	DI Yogyakarta	
Jambi	Sulawesi Tengah	Banten	
Sumatra Selatan	Sulawesi Selatan + Barat	Bali	
Bengkulu	Sulawesi Tenggara	Kalimanta Timur	
Lampung	Gorontalo		
Nusa Tenggara Barat	Maluku		
Nusa Tenggara Timur	Maluku Utara		
	Papua+Papua Barat		

Characteristics of each cluster can be seen in Table 5.

Table 5. Characteristics value of each cluster

	G1	G2	G3
X_1	55506.95	252987.71	814306.00
X_2	139.43	93.14	370.00
X_3	127.19	141.14	1614.67
X_4	305.05	253.29	2858.67
X_5	512.62	363.57	2901.67
X_6	1411.38	574.29	6559.67
X_7	1883.05	1393.43	19615.67
X_8	34.80	51.66	38.25
X_9	10.34	21.97	10.51
X_{10}	20.45	38.07	24.80
X_{11}	2653228.34	3415576.85	2136574.22
X_{12}	87793.60	133556.81	73312.43
X_{13}	6228.52	11981.43	39035.33
X_{14}	36.71	36.14	188.67
X_{15}	16.29	14.29	35.00
X_{16}	1532.33	2744.86	10495.33
X_{17}	17.27	61.48	55.88
X_{18}	11.01	47.27	44.18

Based on Table 5, the characteristics of each cluster are :

Cluster 1 : Provinces are at cluster has a village with facilities that still are in terms of public phones, internet cafes, cell phone signals, television programs, use of radio frequencies and television, but has an average consumtion for telecommunications which is quite high as well

Cluster 2 : Province who are at cluster has a good economy looks of the percentage of the population and households have more mobile phones and computers, but it is also a lot of people who access the Internet and have the household consumption and telecommunication highest and has been optimized in the use of television and radio channels that have been available.

Cluster 3 : Province who are on this cluster of many villages that already have good communication technology facilities include telephone facilities, internet, Base transceiver Station, and a cellular phone signal reception program / broadcast television. In addition to owned facilities, the province also has a lot of use of radio frequencies, television or GSM / DCS

Based on the characteristics that have been held each cluster, then there are some things that should be improved in each cluster in order to advance the condition of communication technology :

Cluster 1 : Addition of facilities in the village, as well as increased utilization of frequencies available either frequency television, radio or GSM

Cluster 2 : Addition of facilities in the village such as public telephones, base stations, television programs

Cluster 3 : The increase in the use of television and radio channels as well as the addition of using computer and internet, as well as the budget for telecommunications

CONCLUSION

Based on the results of the grouping of provinces in Indonesia by using hierarchical cluster analysis method, it can be concluded that :

1. The number of clusters formed by variables related the field of communication technology are 3 cluster. 1st cluster consists of 21 provinces, 2nd cluster consists of 7 provinces, and 3rd cluster consists of 3 provinces.
2. 3rd cluster has the potential villages and the use of communication is good, 2nd cluster has social economy is good, 1st cluster still have to improve potential of village, the economy and the use of communication because there is less.

REFERENCES

[BPS] Badan Pusat Statistik. 2014. *Statistik Indonesia 2014*. Jakarta (ID): Badan Pusat Statistik

[BPS] Badan Pusat Statistik. 2014. *Statistik Perusahaan Informasi dan Komunikasi 2014*. Jakarta (ID): Badan Pusat Statistik

Hair JF, Anderson RE, Tatham RL, Black WC. 1995. *Multivariate Data Analysis with Readings*. Upper Saddle River, NJ (US): Prentice Hall International INC.

Nechaev VD, Brodovskaya EV, Kaira YV, Dombrovskaya AY. 2014. Classification of Russian Internet users: Preliminary result of Cluster analysis. *Life science Journal*. 11 (12): 330-335.

Popat, SK, Emmanuel M. 2014. Review and Comparative Study of Clustering Techniques. *International journal of computer science and information technologies*. 5(1):805-812.

Rezankova, H. 2014. Cluster Analysis of Economic Data. *Statistika*. 94 (1): 73-86.

Safitri D, Widiharih T, Wilandari Y, Saputra AH. 2012. Analisis Cluster pada Kabupaten/Kota Di Jawa Tengah Berdasarkan Produksi Palawija. *Media Statistika*. 5(1):11-16

Sarle WS.1983. *SAS Technical Report A-108* : *Cubic Clustering Criterion*. Cary, NC (US) :SAS Institute Inc.