PROC. INTERNAT. CONF. SCI. ENGIN. ISSN 2597-5250
Volume 3, April 2020 | Pages: 291-296 E-ISSN 2598-232X

Multivariable Panel Data Cluster Analysis using Ward Method Gross Enrollment Ratio (GER) Data in West Java in the Year 2015-2018

Reka Ramadhan, Asep Solih Awalluddin, Rini Cahyandari
Mathematics Departement, Faculty of Science and Technology, Universitas Islam Negeri Sunan Gunung Djati Bandung,
Jl. A.H. Nasution No.105, Cipadung, Kec. Cibiru, Kota Bandung, Jawa Barat 40614. Telp. 022 – 7800525, Fax. 022-7803936

Abstract. The aim of this study is to determine cluster analysis for panel data with multivariable data structures. Choosing a Ward method Choosing a method in a cluster analysis hierarchical technique. Ward method is a method based on Sum Square Error (SSE) with a measure of homogeneity between two objects based on the minimum number of error squares. The measure of similarity used is the Euclidean distance squared. The Ward method is used to add variation between objects in one cluster and maximize variation with objects in another cluster. The steps of the analysis are described in the discussion of this study. The method of implementation uses education gross enrollment rate (GER) data in West Java Province in 2015-2018. The results of the study indicate that the grouping of education GER data in West Java in 2015-2018 using the Ward method produces four clusters. The first cluster consists of five regions, GER for Elementary school, junior and senior high school in the cluster are below the average APK in West Java. The second cluster consists of two regions, in contrast to the first cluster GER for elementary schools in this cluster according to the average GER in West Java but for junior and senior high school GER below the average GER in West Java. The third cluster consists of seven regions, the GER for elementary, junior and senior high schools in this cluster is above the average GER in West Java while the fourth cluster consists of five regions, the GER for elementary and junior high schools in this area is above the average GER in West Java.

Keywords : Cluster Analysis, Ward Method, Distance Matrix, Panel Data, Multivariable

INTRODUCTION

Multivariate analysis in statistics is an analysis used for data analysis with many variables. Some analyzes that can be used for multivariate data include cluster analysis, discriminant analysis, principal component analysis and factor analysis.

This study discusses the development of cluster analysis in multivariate analysis. Cluster analysis was first used by Tyron in 1939. Cluster analysis aims to allocate a group of objects to groups that are mutually free so that the objects in one group are similar to each other. In the grouping used a measure that can explain the similarity or closeness between data, that is a measure of distance or similarity. A measure of distance that is often used is a measure of distance called the Euclidean distance.

In cluster analysis the assumptions that must be considered are data that is free from outliers and there is no collinierity. In selecting objects into groups, cluster analysis is sensitive to outliers. The clusters obtained do not match the actual structure of the population if outliers are involved in data processing. Whereas if there is collinearity between variables before cluster analysis, the initial data is transformed through the main component technique into Z-Score.

The use of cluster analysis can be found in various fields including marketing, insurance, urban planning, education, psychology, language and others. In the field of marketing cluster analysis can be used to form marketing parties to determine specific clusters and create special programs for this cluster. In the area of urban planning, cluster analysis can be used to identify houses by type, price and location. In analytical education can be used to classify student data. For example data in the form of students, parents, gender, or GPA scores. So it can be grouped by students who have high, medium and low GPA.

Education is considered very important in supporting a country's economic growth. Several previous researchers have proven the importance of education for the community, including relating to poverty reduction, improving health status, increasing social and political participation, and so on. The participation rate in an important activity is known, by knowing the participation rate it can be assessed whether the activity is liked by the community or not. The greater the participation rate of an educational program means that the program, institution, region is of high quality, conversely it is less and many participants stop in the process of implementing the program meaning the program, institution and region are not qualified. School participation rates consist of 2 types of measurements, namely the pure enrollment rate (NER) and the gross enrollment rate (GER).

This research was conducted to group cities/ regencies in West Java into groups so that it is easy to observe to determine which cities / regencies have high and low GER. And also can see how far the difference in GER in these groups. This research was conducted by grouping objects against gross enrollment rate variables of elementary, middle, and high school.

METHOD

Multivaribel Panel Cluster Data Analysis With Ward Method

Cluster analysis is a multivariate statistical analysis that aims to find out the data structure by placing the same observational objects into one data group so that it can be distinguished between groups of one with another group or by separating cases / objects into several groups that have different characteristics between groups one with another. In this analysis each group is homogeneous between members in the group or it can be said that the variation of objects / individuals in one group is formed as small as possible.

The main purpose of cluster analysis is to place a group of objects into two or more groups based on the similarities of objects on the basis of various characteristics. The stages of analysis carried out in research in solving cluster analysis cases are as follows:

a. Data standardization
b. Calculate similarity
c. Determine the cluster method
d. Determine the number of clusters
e. Cluster interpretation

- ***Standarization of data***

Outlier is a value from another set of data or different than usual and does not describe the characteristics of the data. In statistics, outliers are observation points far from other observations. An outlier may be due to variability in measurements or may indicate that the latest experimental error is sometimes excluded from the data set. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a tail weight distribution. Standardization of data detection data with standardization in principle changes the value of data to form Z, by:

$$z = \frac{x_{i-\bar{x}}}{s} \tag{2.1}$$

Where x_i is the i data, $\bar{x}$ is the average and s is the standard deviation

- ***Calculating Euclidean Distance***

The distance between sample r and sample k in collectivity The distance between sample r and sample k in collectivity can be marked as d_{rk}, and d_{rk} must meet the following conditions: productivity can be marked as d_{rk}, and d_{rk} must meet the following conditions:

(1) $d_{rk} \geq 0$ jika $X_r = X_k$
(2) $d_{irk} = d_{kr}\ \forall\ X_r, X_k$
(3) $d_{rk} \leq d_{rj} + d_{kj},\ \forall\ X_r, X_k, X_j$

Then calculate the Euclidean distance

$$d_{rk} = \left\{\sum_{t=1}^{t}\sum_{j=1}^{p}\left[X_{rj}(t) - X_{kj}\right]^2\right\}^{\frac{1}{2}} \tag{2.2}$$

d_{rk} = distance between the r object and the k object
X_{kj} = data from the k subject on the j variable
p = number ofcluster variables
t = time in the cluster variable
X_{rj} = data from the r subject on the j variable

the results of the calculation of each distance that has been calculated and then made a distance matrix

$$\begin{matrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & \ddots & & \\ \cdots & \cdots & \cdots & \ddots & \\ [d_{N1} & d_{N2} & \cdots & d_{NN-1} & 0] \end{matrix}$$

d_{21} = distance between object 2 and object 1
d_{32} = distance between object 3 and object 2
d_{NN-1} = distance between object N and object N-1

- ***Determine the Cluster Method***

In this study the method used is the Ward method. The steps of the Ward method are:

Step 1. Begin by looking at N clusters that have one object per cluster (all objects are considered as clusters). SSE will be zero for the first stage because each object will form a cluster.

Step 2. The first cluster is formed by selecting two of the N clusters that have the smallest SSE value. This is in line with its objective function, which is to minimize heterogeneity. The SSE formula is:

$$S_g = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^g}\left[X_{ij}(t) - X^{-g}(t_j)\right]^2 \tag{2.3}$$

$X_{ij}(t)$ = value for object *i* on cluster to *j* at time *t*
$X^{-g}(t_j)$ = the average *j* index at *t* when all samples are of type *g*

Step 3. cluster clusters then consider again to determine two of these clusters that can minimize heterogeneity. Then calculate the distance between clusters by using

$$D_{rk} = \Delta S_{rk} = \frac{n_r+n_p}{n_r+n_k}S_{rp} + \frac{n_r+n_q}{n_r+n_k}S_{rq} - \frac{n_r}{n_r+n_k}S_{pq} \tag{2.4}$$

With

$$S_{rp} = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^r}\left[X_{ij}(t) - X^{-r}(t_j)\right]^2 \tag{2.5}$$

$$S_{rq} = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^r}\left[X_{ij}(t) - X^{-rq}(t_j)\right]^2 \tag{2.6}$$

$$S_{pq} = \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i\in i^r}\left[X_{ij}(t) - X^{-p}(t_j)\right]^2 \tag{2.7}$$

Where

S_r = SSE sample r and sample p
S_{rq} = SSE sample r and sample q
S_{pq} = SSE sample p and sample q
n_r, n_p, n_k = is the number of objects in the r, p and k clusters.

Step 4. Repeat steps 2 and 3, until one cluster or all objects are joined together in one cluster.

- ***Total number of clusters***

The main problem in cluster analysis is to determine how many clusters. Actually there are no standard rules to determine how many clusters, however there are number of clues that can be used, namely:

a. Theoritical, conceptual, practical considerations, might be suggested/suggested to determine how many clusters actually exist. For example, if the clustering objective is to identify/identify market segments, management might want.
b. The relative size of the cluster should be useful

- ***Interpretation***

The interpretation phase involves testing each cluster formed to provide a name or description exactly as a description of the nature of the cluster, explaining how they are relevant in each dimension. When starting the interpretation process the centroid of each cluster is used on each variable. To interpret clusters involves a study of centroids, namely the average value of objects contained in the cluster on each variable. The I centroid cluster is calculated using the following formula:

$$V_i = \frac{\sum_{i=1}^{n} y_i}{n}$$

Where V_i is centroid on i cluster. y_i is the i-object and n is the number

RESULTS AND DISCUSSION

The data used in this study is the APK data of West Java Province taken from the Indonesian Central Statistics Agency in several publications on GER in West Java by district / city. The data period used in this research is 2015-2018.

The results of grouping cities on GER data in West Java in 2015-2018 are illustrated in the form of dendograms. Dendograms are read from left to right where vertical lines show clusters joined together

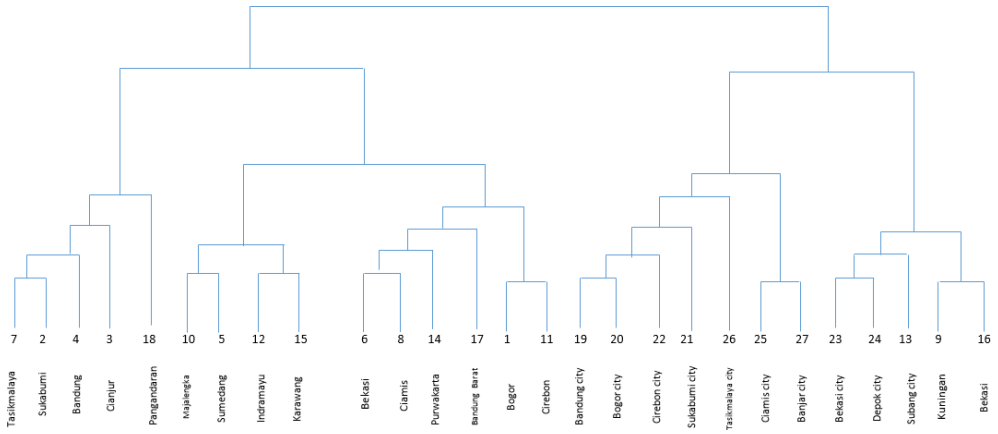

Figure 1. Dendogram with Ward method

From Figure 1 it can be concluded that the members of each cluster are

Table 1. Cluster Members.

No Cluster	Cluster Members
Cluster 1	Tasikmalaya district, Sukabumi district, Bandung district, Cianjur district and Pangandaran district
Cluster 2	Majalengka district, Sumedang district, indramayu district, karawang district, Bekasi district, Ciamis district, Purwakarta district, Bandung barat district, Cirebon district
Cluster 3	Bandung city, Bogor city, Cirebon city, Sukabumi city, Tasikmalaya city, Ciamis city, Banjar city
Cluster 4	Bekasi city, Depok city, Subang city, kuningan district, Bekasi district

After determining the number of clusters and their members, the next step is cluster interpretation. To interpret cluster profiles, it can be done by using the average of each cluster on each variable (centroid). In table 1 there is the first cluster consisting of Tasikmalaya district, Sukabumi district, Bandung district, Cianjur district and Pangandaran district. The centroid values for each variable in the first cluster can be seen in table 2.

Table 2. Centroid Values for the First Cluster.

Kabupaten / Kota	2015			2016			2017			2018		
	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA
Tasikmalaya	110,14	97,83	56,41	105,49	98,02	64,97	104,05	98,73	70,48	103,21	95,78	76,56
Sukabumi	109,92	96,07	54,59	106,27	96,10	64,24	105,82	99,66	69,74	101,83	97,33	76,34
Bandung	107,08	95,41	56,53	104,42	95,53	65,62	104,13	99,15	71,14	100,78	96,99	73,57
Cianjur	107,36	98,53	51,58	103,90	98,72	62,15	102,80	103,35	66,98	99,26	101,62	75,29
Pangandaran	101,10	99,37	51,71	97,94	99,37	62,61	97,50	98,76	67,97	96,81	98,76	68,60
Total	535,60	487,21	270,82	518,02	487,74	319,59	514,30	499,65	346,31	501,89	490,48	370,36
Centroid	107,12	97,44	54,16	103,60	97,55	63,92	102,86	99,93	69,26	100,38	98,10	74,07

Based on the results of the centroid analysis above. The first cluster consist of five regions GER for Elementary school, junior and senior high school in the cluster are below the average APK in West Java.

Next, recalculate the centroid value for the second cluster in table 3 consisting of Majalengka district, Sumedang district, indramayu district, karawang district, Bekasi district, Ciamis district, Purwakarta district, Bandung barat district, Cirebon district. The centroid values are as follows.

Table 3. Centroid Values for the Second Cluster.

Kabupaten / Kota	2015			2016			2017			2018		
	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA
Majalengka	109,01	100,36	63,44	105,37	100,40	74,24	104,70	102,30	80,74	101,47	99,90	84,79
Sumedang	110,52	101,40	68,25	107,29	101,43	72,90	107,37	101,77	79,79	104,10	99,48	83,10
Indramayu	107,23	96,09	68,87	104,13	96,26	77,63	103,61	96,51	78,50	101,48	94,33	78,25
Karawang	107,65	98,66	66,11	106,12	98,68	74,20	106,36	98,19	77,58	103,51	96,60	76,28
Garut	109,48	95,78	61,19	105,81	96,03	69,37	104,50	96,86	74,96	101,58	95,52	83,66
Ciamis	111,05	99,44	61,19	106,66	99,51	68,81	105,42	99,96	72,54	101,65	97,49	78,15
Purwakarta	107,44	101,39	59,96	104,65	101,50	70,60	104,94	102,64	72,13	102,12	102,07	77,71
Bandung Barat	101,98	98,81	57,72	99,49	98,89	69,88	99,71	101,11	76,83	96,59	99,58	81,82
Bogor	110,56	97,25	57,09	108,71	97,35	66,36	109,98	101,40	72,98	108,11	102,69	74,84
Cirebon	112,30	96,99	61,11	109,31	97,18	72,90	109,25	99,60	75,35	105,82	98,30	73,64
Total	1087,22	986,17	624,93	1057,54	987,23	716,89	1055,84	1000,34	761,40	1026,43	985,96	792,24
Centroid	108,72	98,62	62,49	105,75	98,72	71,69	105,58	100,03	76,14	102,64	98,60	79,22

The second cluster consists of two regions, in contrast to the first cluster GER for elementary schools in this cluster according to the average GER in West Java but for junior and senior high school GER below the average GER in West Java.

Next, recalculate the centroid value for the third cluster in table 4 consisting of Bandung city, Bogor city, Cirebon city, Sukabumi city, Tasikmalaya city, Ciamis city, Banjar city. The centroid values are as follows.

Table 4. Centroid Values for the Third Cluster.

Kabupaten / Kota	2015			2016			2017			2018		
	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA	SD	SMP	SMA
Bandung city	108,80	107,11	100,72	106,81	107,13	107,04	106,19	104,80	109,66	103,92	100,30	110,61
Bogor city	112,10	109,18	101,00	110,03	109,28	106,70	109,65	105,48	110,89	106,20	103,55	108,91
Sukabumi city	111,50	108,32	100,94	110,46	108,34	117,84	109,87	103,79	118,69	105,09	102,32	109,92
Cirebon city	110,39	108,22	99,53	107,54	108,41	105,07	106,82	106,56	106,84	102,80	105,25	106,06
Cimahi city	106,91	101,19	91,79	104,25	101,33	97,79	105,56	99,57	102,64	102,54	96,29	101,90
Tasikmalaya city	104,79	108,09	91,84	101,71	108,09	102,50	101,58	108,75	111,40	98,87	105,81	110,48
Banjar city	109,29	101,59	88,54	106,69	101,64	95,64	107,84	103,33	105,51	104,62	102,15	105,13
Total	763,78	743,70	674,36	747,49	744,22	732,58	747,51	732,28	765,63	724,04	715,67	753,01
Centroid	109,11	106,24	96,34	106,78	106,32	104,65	106,79	104,61	109,38	103,43	102,24	107,57

The third cluster consists of seven regions, the GER for elementary, junior and senior high schools in this cluster is above the average GER in West Java.

Next, recalculate the centroid value for the third cluster in table 5 consisting of Bekasi city, Depok city, Subang city, kuningan district, Bekasi district. The centroid values are as follows

Table 5. Centroid Values for the Fourth Cluster.

Kabupaten / Kota	2015			2016			2017			2018		
	SD	**SMP**	**SMA**	**SD**	**SMP**	**SMA**	**SD**	**SMP**	**SMA**	**SD**	**SMP**	**SMA**
Bekasi city	110,61	109,92	84,45	109,28	110,11	91,01	110,19	104,54	92,57	105,74	102,25	91,49
Depok city	104,60	109,00	81,75	104,66	109,12	93,01	106,46	104,98	94,90	103,44	100,89	96,35
Subang	109,32	109,87	74,47	106,05	110,19	86,66	105,23	105,01	92,17	101,63	103,78	97,43
Kuningan	109,43	99,50	79,31	106,41	99,55	85,96	105,98	101,57	92,53	102,78	98,86	92,25
Bekasi	108,44	96,60	73,13	107,06	96,76	82,40	109,05	97,48	86,72	106,79	96,93	87,26
Total	542,40	524,89	393,11	533,46	525,73	439,04	536,91	513,58	458,89	520,38	502,71	464,78
Centroid	108,48	104,98	78,62	106,69	105,15	87,81	107,38	102,72	91,78	104,08	100,54	92,96

The Fourth cluster consists of five regions, the GER for elementary and junior high schools in this area is above the average GER in West Java.

CONCLUSION

The case study results show that the grouping of education GER data in West Java in 2015-2018 using the Ward method produced four clusters. In cluster one consists of Tasikmalaya district, Sukabumi district, Bandung district, Cianjur district and Pangandaran district. On cluster two consisting Majalengka district, Sumedang district, indramayu district, karawang district, Bekasi district, Ciamis district, Purwakarta district, Bandung barat district, Cirebon district. Cluster three consists of Bandung city, Bogor city, Cirebon city, Sukabumi city, Tasikmalaya city, Ciamis city, Banjar city. Cluster four of Bekasi city, Depok city, Subang city, kuningan district, Bekasi district Of all the clusters that have good gross participation rates is the third cluster because this cluster consists has a gross participation rate above the average GER in West Java.

ACKNOWLEDGE

The authors are very grateful to the people that supplied suggestion during the research. This research has been partially supported by Mathematics Departement, Faculty of Science and Technology, Universitas Islam Negeri Sunan Gunung Djati Bandung.

REFERENCES

Agus Widarjono. 2007. *Ekonometrika Teori dan Aplikasi*. Yogyakarta: Ekonisia FE UII.

Badan Pusat Statistik Jawa Barat. *Jawa Barat Dalam Angka 2011*. Bandung.

Badan Pusat Statistik Jawa Barat. *Jawa Barat Dalam Angka 2013*. Bandung.

Badan Pusat Statistik Jawa Barat. *Jawa Barat Dalam Angka 2014*. Bandung.

Baltagi, B. H. *Econometrics Analysis of Panel Data (3rd ed)*. Chicester, England: John Wiley & Sons Ltd. 2005.

Barro R. J., dan Sala-i-Martin, Xavier. 2004. *Economic Growth, Second edition*, London: The MIT Press.

Billson Simamora. (2005). *Analisis Multivariat Pemasaran Edisi Pertama*. Jakarta: PT. Gramedia Pustaka Utama.

Bunkers, dkk. (1996). *Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique*. J.Climate: 130-146

Dillon, William R. dan metthew Goldstein.1984. *Multivariate Analysis Method and Application*.

United States of America John Wiley & Sons. Inc

Dwi Putra Abadi & Sutikno. (2013). *Pengclusteran Zona Musim (ZOM) dengan Agglomerative Hierarchical Clustering*. Diakses dari www.its.ac.id tanggal 5 Agustus 2016

Fadhli. (2011). *Analisis Cluster Untuk Pemetaan Mutu Pendidikan di Aceh*. Tesis. PPs-UGM.

Fress. Edward W, *Longidutinal and Panel Data*, New york: United States of America by Cambridge University Press, 2004

Gudono. (2011). *Analisis Data Multivariat Edisi Pertama*. Yogyakarta: BPFE.

Gujarati, Damodar, 2003, Ekonometri Dasar. Terjemahan: Sumarno Zain, Jakarta: Erlangga.

Hedeker, Donald. dan Gibbons, Robert D. *Longitudinal Data Analysis*. John Wiley & Sons, Inc. Canada. 2006.

Johnson, R.A & Wichern, DW. (1992). *Applied Multivariate Statistical Analysis Third Edition*. New Jersey: Prentice Hall International.

Jonathan Sarwono, 2007 *Analisis Jalur untuk Riset Bisnis dengan SPSS*, Yogyakarta: Andi Offset

Laeli Sofya, 2014 Analisis *Cluster* dengan *Average Linkage Method* dan *Ward's Method* untuk Data Responden Nasabah Asuransi Jiwa Unit Link

Luthfi Kurnia Hidayati & Lucia Aridinanti. (2013). *Pengclusteran Kabupaten/Kota di Jawa Timur Berdasarkan Faktor-Faktor Penyebab*

Perceraian tahun 2010. Diakses dari www.its.ac.id pada tanggal 5 Agustus,

Nurul & Muhammad Sjahid Akbar. (2013). *Pengclusteran Kabupaten/Kota di Provinsi Jawa Timur Berdasarkan Indikator Kemiskinan dengan Metode Cluster Analysis.* Diakses dari www.its.ac.id pada tanggal 5 Agustus 2016

Ross, Catherine E. dan Chia-ling Wu. 1995. "*The Links Between Education and Health." American Sociological Review*, 60 (5): 719-745.

Rencher, Alvin C. *Method of Multivariate Analysis. 2nd ed*. New York: John Wiley and Sons, inc; 2002.

Stromquist, Nelly P. 1989. "*Determinants of Educational Participation and Achievement of Women in the Third World: a review of the evidence and a theoretical critique."* Review of Educational Research, 59 (2): 143-183.

Supranto. (2004). *Analisis Multivariat Arti dan Interpretasi Edisi Pertama*. Jakarta: Rineka Cipta.

Santoso, Slamet, 2004, *Dinamika Kelompok*, Jakarta: Bumi Aksara

Wibisono, M. S. 2005. *Pengantar Ilmu Kelautan*. Jakarta: PT. Gramedia Widiasarana Indonesia World Bank. 1980. *World Development Report.* Washington, D.C.: World Bank