Grouping of Regencies and Cities in West Java Using the Biplot's Principal Component Analysis

Yuyun Hidayat and Titi Purwandari
Department of Statistics, Faculty of Mathematics and Natural Sciences,
Universitas Padjadjaran, Indonesia
yuyunhidayat@unpad.ac.id; titipurwandari@yahoo.com

Sukono and Herlina Napitupulu
Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Universitas Padjadjaran, Indonesia
sukono@unpad.ac.id; napitupuluherlina@gmail.com

Abdul Talib Bon
Department of Production and Operations,
University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

The main problem in regional development lies in the resources and potential they have in order to create an increase in the number and type of employment opportunities for regional communities. West Java is a potential province, the center of the development of science, technology and education. Through this potential, the achievement of development in West Java is still sub-optimal; this is shown by the West Java Human Development Index (HDI) which is still below the national average. Symptoms of low levels of education and health characterize that investment in human resources in West Java has not been done adequately. When high economic growth is not accompanied by a proportional increase in welfare, this indicates that economic growth, processes and benefits have not spread relatively evenly due to social problems. Based on the foregoing, a regional development program is needed that is focused on accelerating development in areas where social, cultural, economic, regional finance, accessibility, and infrastructure availability conditions are still lagging behind other regions. The purpose of this study is to classify districts and cities in West Java based on a number of criteria in order to provide recommendations to the West Java government regarding policy policies that need to be carried out. The usefulness of this research is to provide scientific references for the West Java government in making policies. The data used are secondary data collected by the Central Bureau of Statistics and the Ministry of Finance of the Republic of Indonesia. Using the main biplot's component analysis, a map of regencies and cities is grouped based on a number of criteria as a basis for decision making.

Keywords:
Principal component, biplot's analysis, grouping map.

1. Introduction

The progress of regional development and people's welfare in Indonesia is not always the same and evenly distributed, these results make a gap between regions. The condition is caused by differences in geographical conditions, natural resources, infrastructure, socio-cultural, and capacities of human resources (Naibaho, 2016). The main problem in regional development lies in the resources and potential they have in order to create an increase in the number and types of employment opportunities for the local community. West Java is a potential province, the center of science, technology and education development. Through this potential, the achievement of development in West Java is still sub-optimal; this is indicated by the West Java Human Development Index (HDI) which is still below the national average. Symptoms of low education and health levels characterize that investment in human

resources in West Java has not been carried out adequately. When high economic growth is not accompanied by a proportional increase in welfare, this indicates that economic growth, processes and benefits have not spread relatively evenly due to social problems. Based on the foregoing, regional development programs are needed that are focused on accelerating development in areas where social, cultural, economic, regional finance, accessibility, and infrastructure availability are lagging behind other regions. Solving inter-regional gaps requires a policy, program, and activity that are consistent, integrated and cross-sectorial. A policy considers the suitability of regional spatial planning, legal systems and reliable institutions as well as coordinating and collaborating between ministries / agencies and regional unit work units in planning, budgeting, implementing, monitoring, and evaluating (*Konferensi Pembangunan Jawa Barat*, 2015).

The purpose of this study is to group districts and cities in West Java based on a number of criteria in order to give recommendations to the West Java government regarding policies that need to be done and the usefulness of this research is to provide scientific references for the West Java government in making policies.

2. Method

2.1 Object and Variable Research

The object of observation in this study is districts and cities in West Java Province, which are as many as 26 districts and cities consisting of seventeen districts and nine cities. The variables in this study are the variables used by the Ministry of Development of Disadvantaged Regions and Transmigration (KNPDT). The data used in this study are secondary data obtained from the Central Statistics Agency (BPS) and the Ministry of Finance of the Republic of Indonesia, in the form of Village Potential Data Collection (PODES) in 2014, National Socio-Economic Survey (SUSENAS) in 2014 and Regional Financial Capability data 2014 for districts and cities in West Java (Naibaho, 2016).

The variables used in this study are as follows (*Kementerian Negara Pembangunan Daerah Tertinggal*, 2014):

1) Percentage of poor people
2) Per capita expenditure / consumption
3) Life expectancy
4) Average school length
5) Literacy rates
6) Number of villages with the largest type of asphalt / concrete road settlement
7) Number of villages with the widest type of road settlement is hardened
8) Number of villages with the widest type of residential land
9) Number of villages with the other widest type of road settlement
10) Percentage of household users of electricity
11) Percentage of household telephone users
12) Percentage of household users of clean water
13) Number of villages that have a market without permanent buildings
14) Amount of health infrastructure per 1000 inhabitants
15) Number of doctors per 1000 residents
16) Number of elementary and middle school per 1000 population
17) Regional financial capability
18) Average distance from the village office to the district office in charge
19) Number of villages with access to health services> 5 km
20) Distance of villages to basic education services
21) Percentage of earthquake villages
22) Percentage of village landslides
23) Percentage of villages flooded
24) Percentage of other disaster villages
25) Percentage of villages to protected forest areas
26) Percentage of villages with critical land
27) Percentage of conflict villages in the past year.

2.2 Biplot PCA Analysis

Biplot Analysis PCA is a mapping method in multivariate analysis that contains information in a data table, which shows the main structure of the data (Grenaacre, 2010; Hair et al., 2010; Jollife and Cadima, 2010). This analysis aims to present data in two-dimensional maps so that data behavior is easily seen and interpreted (Ginanjar et al., 2017; Rifkhatusa, 2014). Biplot analysis requires data from a number of objects with interval or ratio scale variables. This method is based on the Singular Value Decomposition (SVD) of a data matrix that has been corrected by the average (Jolliffe, 2010). The biplot main component analysis steps are:

2.2.1 Matrix of Data

Data in the form of objects as many as 26 districts and cities with 27 variables are presented in the initial **Y** matrix measuring *n* x *p* (26 x 27*).*

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{133} \\ y_{21} & y_{22} & \cdots & y_{233} \\ \vdots & \vdots & \ddots & \vdots \\ y_{261} & y_{262} & \cdots & y_{2633} \end{bmatrix} \tag{1}$$

The Y matrix in equation (1) is transformed against the average, becomes

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} y_{11} - \bar{y}_1 & y_{12} - \bar{y}_2 & \cdots & y_{127} - \bar{y}_{27} \\ y_{21} - \bar{y}_1 & y_{22} - \bar{y}_2 & \cdots & y_{227} - \bar{y}_{27} \\ \vdots & \vdots & \ddots & \vdots \\ y_{261} - \bar{y}_1 & y_{262} - \bar{y}_2 & \cdots & y_{2627} - \bar{y}_{27} \end{bmatrix} \\ &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{127} \\ x_{21} & x_{22} & \cdots & x_{227} \\ \vdots & \vdots & \ddots & \vdots \\ x_{261} & x_{262} & \cdots & x_{2627} \end{bmatrix} \end{aligned} \tag{2}$$

2.2.2 Eigenvalue and Eigenvector

Before searching for the singular decomposition value (SVD) it is necessary to calculate the eigenvalue and eigenvector of the data matrix $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ (Hair et al., 2010; Johnson, 2007). Eigenvalue denoted by λ and eigenvector denoted by $\boldsymbol{a}$ can be calculated as follows:

$$\left|\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda_i \mathbf{I}\right| = 0 \tag{3}$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda_i\right)\boldsymbol{a} = 0 \tag{4}$$

2.2.3 Singular Value Decompotition

The direct approach to obtain singular decomposition values (SVD) is as follows (Johnson, 2007):

$$\mathbf{X}_{(\mathrm{n}\times\mathrm{p})} = \mathbf{U}_{(\mathrm{n}\times\mathrm{r})}\mathbf{L}_{(\mathrm{r}\times\mathrm{r})}\mathbf{A}^{\mathbf{T}}{}_{(\mathrm{r}\times\mathrm{p})} \tag{5}$$

with:

- $r \leq \{26,27\}$
- $\mathbf{U}$ and $\mathbf{A^T}$ is a matrix with otonormal column so that $\mathbf{U^T U = A^T A = I_r}$ ($\mathbf{I_r}$ is matrix identity dimension *r*)
- $\mathbf{L}$ is matrix (*r* x *r*) with its diagonal elements is the square root of eigenvalue $\mathbf{X^T X}$, with $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq ... \geq \sqrt{\lambda_r}$ which forms the matrix as follows :

$$\mathbf{L} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_r} \end{bmatrix} \quad (6)$$

  Diagonal elements of matrix are called singular matrix value $\mathbf{X}$.
- The columns of matrix $\mathbf{A}$ is eigenvector of matrix $\mathbf{X^T X}$ which corresponds to eigenvalue λ_i i.e.:

$$\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_r] \quad (7)$$

- The columns of matrix $\mathbf{U}$ is obtained from the formula:

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} \mathbf{X} \boldsymbol{a}_i, \; i = 1, 2, ..., r \quad (8)$$

with:

- $\boldsymbol{u}_i$: The elements of matrix $\mathbf{U}$
- $\boldsymbol{a}_i$: The elements of matrix $\mathbf{A}$
- λ_i : *i*-th eigenvalue of matrix $\mathbf{X^T X}$
- $\mathbf{X}$: initial matrix that corrected against average

After the SVD results are obtained, equation (5) becomes:

$$\mathbf{X = U L^{\alpha} L^{1-\alpha} A^T} \quad (9)$$

In determining $\mathbf{L}^{\alpha}$, for $0 \leq \alpha \leq 1$, then the diagonal matrix has diagonal elements $\sqrt{\lambda_1^{\alpha}} \geq \sqrt{\lambda_2^{\alpha}} \geq ... \geq \sqrt{\lambda_r^{\alpha}}$. Determination of $\mathbf{L}^{1-\alpha}$ applies the same as diagonal elements $\sqrt{\lambda_1^{1-\alpha}} \geq \sqrt{\lambda_2^{1-\alpha}} \geq ... \geq \sqrt{\lambda_r^{1-\alpha}}$ (Mattjik and Sumertajaya, 2011).

Suppose $\mathbf{G = UL^{\alpha}}$ and $\mathbf{H^T = L^{1-\alpha} A^T}$, equation (9) becomes:

$$\mathbf{GH^T = UL^{\alpha} L^{1-\alpha} A^T = ULA^T = X} \quad (10)$$

The *(i, j)*-th element in matrix $\mathbf{X}$ can be written as follows:

$$\mathbf{x}_{ij} = \mathbf{g}_i \mathbf{h}_j \quad (11)$$

with $\mathbf{g}_i$, i =1, 2, ..., 26 and $\mathbf{h}_j$, j =1, 2,..., *27* each is row of matrix $\mathbf{G}$ and column of matrix $\mathbf{H}^{\mathbf{T}}$. In $\mathbf{g}_i$ and $\mathbf{h}_j$ have r dimensions. The first two columns of matrix $\mathbf{G}$ can be used for object mapping, while the first two columns of matrix $\mathbf{H}^{\mathbf{T}}$ can be used for variable mapping.

2.2.4 Identify Data Diversity Percentage

If matrix **X** has more than two ranks then eigenvalue that are taken are λ_1 and λ_2 so the amount of diversity explained is as follows:

$$\tau = \frac{(\lambda_1 + \lambda_2)}{\sum_{i=1}^{p} \lambda_i} \tag{12}$$

with:

- λ_1 : The first biggest eigenvalue
- λ_2 : The second biggest eigenvalue
- λ_i : The i-th Eigenvalue of $\mathbf{X}^{\mathbf{T}}\mathbf{X}$; i = 1, 2, …, *27*.

If the value of τ is getting closer to the value of 1, means the biplot obtained from approaches matrix with rank = 2 will provide a better presentation of the information contained in the actual data. So based on τ value, the resulting grouping map can be used in decision making. The resulting grouping map can give an idea of the position of the proximity of one object to another object and the proximity of the variable to the object.

2.2.5 Identify Principal Component Analysis Biplots Mapping Results (Biplot PCA)

The mapping results from PCA Biplot are as follows (Grennace, 2010; Khusnah, 2012; Mattjik and Sumertajaya, 2011):

1) Proximity (similarity) between research objects.
   The closer the position of the two object points then becomes more similar, the further the position of the two points the object then increasingly different..
2) Variable Diversity.
   Variables are described as trending lines (vectors). Variables with small diversity are described as short-sized vectors while variables with large diversity are described as long-sized vectors.
3) Relations or correlation between variable.
   The relationship between variables can be identified based on the angle formed by two vector variables on the map axis. If two variable vectors coincide with the axis of the map in the same direction (close to 0^o or 360^o), then it has a very close positive correlation. If two variable vectors coincide with the axis of the map in the opposite direction (close to 180^o), then it has a very close negative correlation, if two variable vectors are perpendicular to the axis of the map (close to 90^o or 270^o), then the two variables are not correlated
4) Value of variables on an object .
   Objects that are in the same direction of variable's direction, say that the object is above the average value. Conversely, if another object is located opposite the direction of the variable, then the object has a value below the average. While objects that are almost in the middle, have a value close to the average.

3. Result and Discussion

3.1 Identify Data Diversity Percentage

The diversity of data that can be explained by Biplot PCA maps using 4 components calculated from the cumulative eigenvalue is 0,71131 , it can be concluded that the percentage of data that can be explained by PCA Biplot is equal to 71,131%.

3.2 Identify information on Mapping Analysis Results

From the results of analysis through maps, 4 groups of districts / cities were formed with variable characteristics having similarities, namely Group 1 is the city of Bandung, Bekasi, Depok, Cimahi, Cirebon, Bogor, Sukabumi, Bogor district. Group 2 is the city of Bekasi, West Bandung, Sukabumi, Cianjur. Group 3 is the city of Tasikmalaya, Garut, Purwakarta, Karawang. Group 4 is city/district Banjar, Ciamis, Majalengka, Sumedang, Subang, Indramayu, Kuningan, Cirebon district. It can be seen on figure 1.

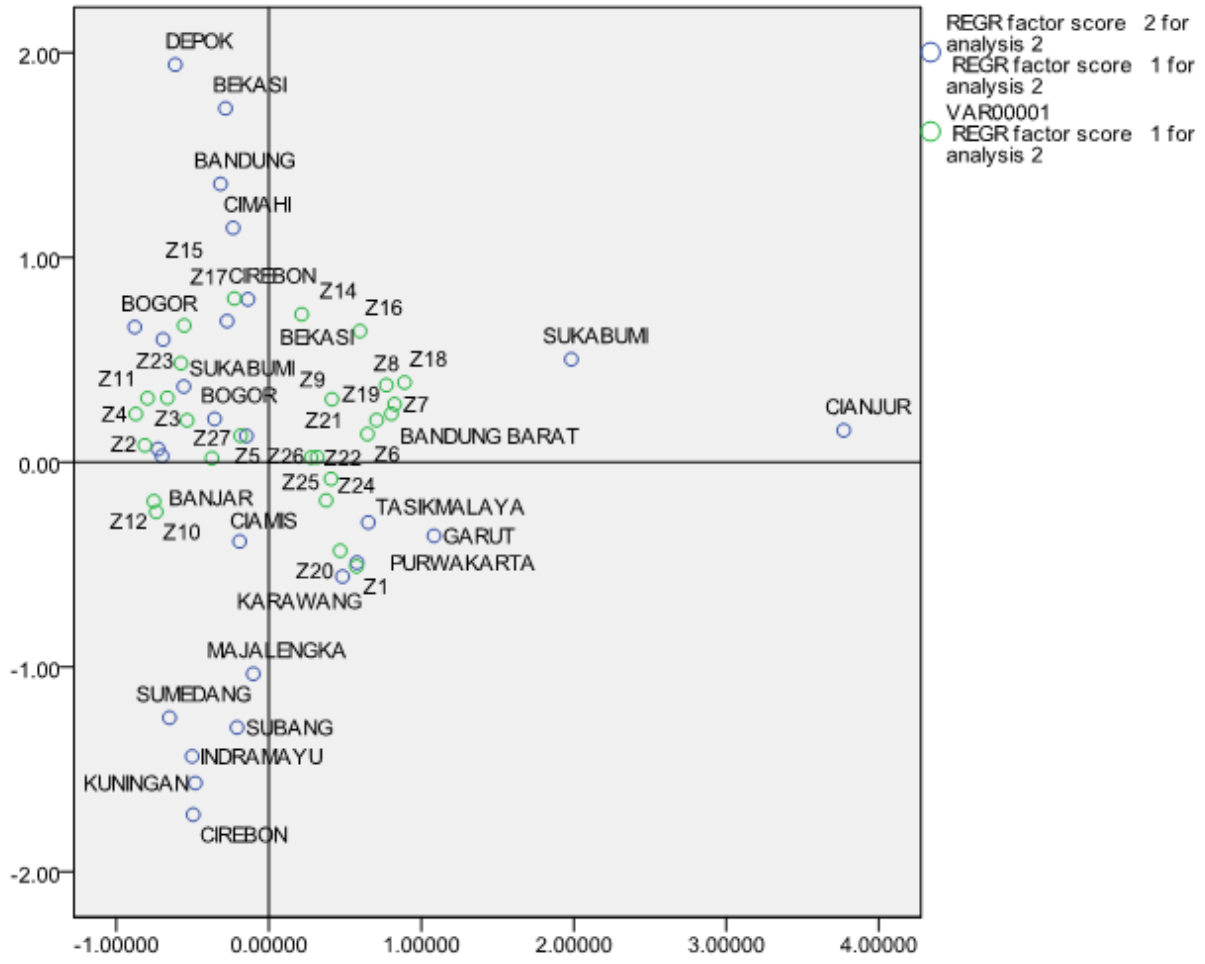

Figure 1. City / District Grouping Map Based on Variables

3.3 Relationship between Research Variables

Group 1, variables that have a positive correlation are per capita expenditure/consumption, life expectancy, average school length, percentage of telephone user households, percentage of conflict villages in the past year. In group 2 are variables Percentage of earthquake villages, percentage of village land landslide, average distance from village office to district office, number of villages with access to health services > 5 km, number of villages with the type of asphalt / concrete widest settlement, number of villages with the widest type of residential settlement, number of villages with type of land widest settlement. In group 3 is variable percentage of other disaster villages, percentage

of villages to protected forest areas, percentage of poor people, and distance of villages to basic education services. In group 4 is variable percentage of household users of electricity and percentage of household use clean water. It can be seen on figure 2.

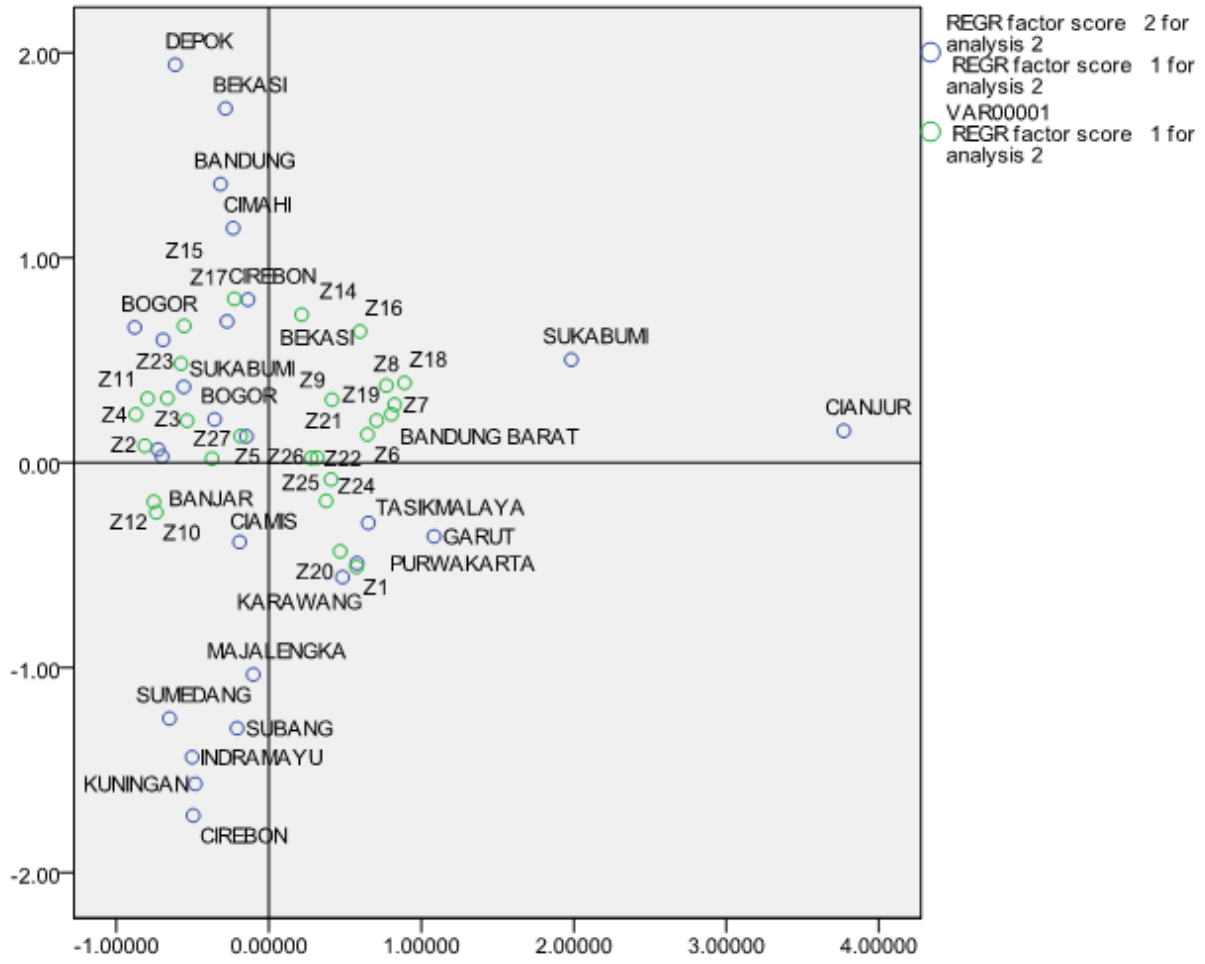

Figure 2. District/city grouping maps and research variables

3.4 Variable Diversity

Variable diversities are described through long/short vector sizes. The variable that has the greatest diversity in group 1 is the variable number of doctors per 1000 population, in group 2 is the variable number of elementary and middle school per 1000 population, in group 3 is the variable distance of the village to basic education services, percentage of poor people, in group 4 is variable percentage of household users of clean water. This can be seen in Figure 2.

3.5 District/city groups and research variables

District/city groups that contribute to the research variables can be seen in Table 1, Table 2, Table 3, and Table.4 as follows:

Table 1. District/city groups and research variables

Group	City/District	Variable/Characteristic
Large contributions of	The city of Bandung, Bekasi, Depok,	Per capita expenditure/consumption,

variable values above average	Cimahi, Cirebon, Bogor, Sukabumi, Bogor district.	Life expectancy, Average school length, Percentage of telephone user households, Percentage of conflict villages in the past year

Table 2. District/city groups and research variables

Group	City/District	Variable/Characteristic
Large contributions of variable values above average	Bekasi, Bandung Barat, Sukabumi, Cianjur	Percentage of earthquake villages, Percentage of village landslides, Average distance from village offices to supervising district offices, Number of villages with access to health services more 5 km, Number of villages with the type of asphalt / concrete widest settlement, Number of villages with type the widest road settlement is hardened, the number of villages with the widest type of residential land.

Table 3. City/District grouping maps and research variables

Group	City/District	Variable/Characteristic
Large contributions of variable values above average	Tasikmalaya, Garut, Purwakarta, Karawang.	Percentage of other disaster villages, Percentage of villages to protected forest areas, Percentage of poor people, Distance of villages to basic education services.

Table 4. City/District grouping maps and research variables

Group	City/District	Variable/Characteristic
Large contributions of variable values above average	Banjar, Ciamis, Majalengka, Sumedang, Subang, Indramayu, Kuningan, Cirebon.	Percentage of electricity user households, Percentage of users of clean water users.

4. Conclusion

Based on the results of data analysis, it can be concluded as follows:

1) The percentage of diversity of data produced by PCA Biplot is equal to 71.131% means that the resulting grouping map can represent information contained in the actual data.
2) Formed 4 groups of districts and cities in West Java based on a number of criteria.
3) The results of grouping districts and cities formed based on a number of criteria, giving recommendations to the government of West Java and related institutions in formulating policies and strategies.

Acknowledgements

Acknowledgments are conveyed to the Rector, Director of Directorate of Research, Community Involvement and Innovation, and the Dean of Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, with whomthe Internal Grant Program of Universitas Padjadjaran was made possible to fund this research.The grant is a means of enhancing research and publication activities for researchers at Universitas Padjadjaran.

References

Ginanjar, I., Pasaribu, U. S., and Indratno, S. W., A measure for objects clustering in principal component analysis biplot: A case study in inter-city buses maintenance cost data, *Statistics and its Applications AIP Conf. Proc. 1827*, 020016-1–020016-7, 2017.

Grenaacre, M. J., *Biplots in Practice.* Barcelona: BBVA Foundation, The Pompeu Fabra University Barcelona, 2010.

Hair, J. F., Anderson, R. E., Tatham, R. L., .and Black, Inc. W. C., *Multivariate Data Analysis*, Seven Edition, . Prentice Hall, New Jersey, 2010.

Johnson, W., *Applied Multivariate Statistical Analysis,* Sixth Edition, Prentice Hall, New Jersey, 2007.

Jolliffe, I.T, *Principal Component Analysis*, Second Edition, Springer, New York, 2010.

Jolliffe, I.T., and Cadima, J.,Principal component analysis: a review and recent developments, *Phil. Trans. R. Soc. A*.374: 20150202, 2010.

Kementerian Negara Pembangunan Daerah Tertinggal, http://kemendesa.go.id/hal/300027/183-kab-daerah-tertinggal, 2014 . [Accessed on: 20/08/2015]

Khusnah, R L., *Penentuan Subsektor Lapangan Usaha Potensial Di Wilayah Jawa Barat Menggunakan Metode Pricipal Component Analysis Biplots (Biplot PCA)*, Skripsi,Universitas Padjadjaran,Bandung, 2012.

Konferensi Pembangunan Jawa Barat, *Pokok-pokok Persoalan Pembangunan di Jawa Barat*" www.unpad.ac.id/.../konperensi-pembangunan-jawa-b, 2015.

Mattjik, A. A., and Sumertajaya, I. M., *Sidik Variabel Ganda*, IPB Press, Bogor, 2011.

Naibaho,E., *Perbandingan Backpropagation Neural Network dan Learning Vector Quantization*, Tesis,Universitas Padjadjaran,Bandung, 2016.

Rifkhatussa, E. F., Yasin, H., and Rusgiyono, A., Analisis Biplot Komponen Utama pada Bank Umum (*Commercial Bank*) yang Beroperasi di Jawa Tengah, *Jurnal Gaussian*, 61-70, 2014.

Biographies

Yuyun Hidayat is a lecturer at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. He currently serves as Deputy of the Head of Quality Assurance Unit, a field of Statistics, with a field of Quality Control Statistics.

Titi Purwandari is a lecturer at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, the field of Statistics, with a field of Quality Control Statistics.

Sukono is a lecturer in the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran. Currently serves as Head of Master's Program in Mathematics, the field of applied mathematics, with a field of concentration of financial mathematics and actuarial sciences.

Herlina Napitupulu is a lecturer at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, the field of algebra and optimization analysis, with a field of concentration of computational optimization.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.