Profile of West Java Province Based on School Participation Rate Using Biplot's Principal Component Analysis

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

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

Go to school are those who are enrolled and actively participating in education both in the formal education level in the form of basic education namely Primary / equivalent and Junior High School / equivalent, secondary education, namely High School / equivalent and higher education, namely Higher Education / equivalent) and non-formal Package A equivalent to elementary school, package B is equivalent to junior high school and packet C is equivalent to senior high school) which is under the supervision of the Ministry of National Education, School (APS) is the proportion of school children at certain educational levels in the age group that matches the education level. Symptoms of a low level of education characterize that investment in human resources has not been carried out adequately. The purpose of this study was to determine the profile of the province of West Java by grouping districts and cities in West Java based on the School Participation Rate in order to provide recommendations to the relevant agencies. The usefulness of this research is to provide scientific references for the West Java government in making policies. The data used is secondary data collected by the Central Bureau of Statistics. The results of the analysis using Biplot's Principal Component Analysis, obtained a map of the similarity of districts and cities based on school participation rates, diversity of school participation rates, correlation of school participation rates.

Keywords:
School participation numbers, grouping maps, biplot's principal component analysis.

1. Introduction

Schooling is those who are registered and actively enrolled in education in the form of formal education in the form of basic education, namely Elementary School/equivalent and Junior High School/equivalent, secondary education namely High School/equivalent and higher education namely Higher Education or equivalent Package A equivalent to elementary school, package B is equivalent to junior high school and package C is equivalent to high school) which is under the supervision of the Ministry of National Education (Kemdiknas), School (APS) is the proportion of school children at certain levels of education in the age group (https: //www.bps.go.id).

Formal education is a structured and tiered educational path consisting of basic education, secondary education, and higher education (Depdiknas, 2003), covering elementary/MI/equivalent, junior/MTs/equivalent, high/MA/ equivalent and college (Machfoedz and Suryani, 2011). Non-formal education is a pathway to education outside formal education that can be carried out in a structured and tiered manner (Depdiknas, 2003). Includes life skills

education (courses), early childhood education (PAUD) or pre-school, youth education, women's empowerment education, literacy education, skills education and job training, equality education (package A, package B, and package C) and other education aimed at developing students' abilities. No/never attended school is not/has never been registered and is actively participating in education at a level of education, including those who have graduated/have not graduated from kindergarten who did not continue to elementary school. Graduating from school is completing a class/final level education in a public or private school with a diploma. A person who has not attended the highest class but if he takes the exam and passes it is considered graduated. Illiteracy Rate is the proportion of a population of a certain age who cannot read and or write Latin letters or other letters to a population of a certain age. Symptoms of low education levels characterize that investment in human resources has not been carried out adequately (Badan Pusat Statistik).

The purpose of this study was to find out the profile of West Java province by grouping districts and cities in West Java based on School Participation Figures in order to provide recommendations to the relevant agencies. The usefulness of this research is to provide scientific references for the West Java government in making policies.

The systematics of grouping districts and cities in West Java based on School Participation Figures were analyzed using the Biplot Principal Component Analysis (PCA) method which is part of multivariate 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; Jolliffe and Cadima, 2016). This analysis aims to present data in two-dimensional maps so that data behavior is easily seen and interpreted (Ginanjar et al., 2003; Rifkhatussa, 2014; Ukalski and Klisz, 2016).

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 variables of school enrollment rates aged 7-12 years, 13-15 years, 16-18 years, and 19-24 years. The data used is secondary data obtained from the Central Statistics Agency (BPS) in 2015 (Badan Pusat Statistik).

2.2 *Biplot Principal Component Analysis (PCA)*

Biplot Principal Component Analysis 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; Jolliffe and Cadima, 2016). This analysis aims to present data in two-dimensional maps so that data behavior is easily seen and interpreted (Ginanjar et al., 2003; Rifkhatussa, 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 steps in this analysis are as follows :

2.2.1 Data Matrix

Data in the form of objects as many as 27 districts and cities with 4 variables are presented in the initial **Y** matrix measuring $n \times p(27 \times 4)$.

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{14} \\ y_{21} & y_{22} & \cdots & y_{24} \\ \vdots & \vdots & \ddots & \vdots \\ y_{271} & y_{272} & \cdots & y_{274} \end{bmatrix}$$

The **Y** matrix is transformed against the average, becomes.

$$\mathbf{X} = \begin{bmatrix} y_{11} - \bar{y}_1 & y_{12} - \bar{y}_2 & \cdots & y_{14} - \bar{y}_4 \\ y_{21} - \bar{y}_1 & y_{22} - \bar{y}_2 & \cdots & y_{24} - \bar{y}_4 \\ \vdots & \vdots & \ddots & \vdots \\ y_{271} - \bar{y}_1 & y_{272} - \bar{y}_2 & \cdots & y_{274} - \bar{y}_4 \end{bmatrix}$$

$$= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{14} \\ x_{21} & x_{22} & \cdots & x_{24} \\ \vdots & \vdots & \ddots & \vdots \\ x_{271} & x_{272} & \cdots & x_{274} \end{bmatrix}$$

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 $\mathbf{X^T X}$ (Hair et al., 2010). Eigenvalue denoted by and eigenvector denoted by can be calculated as follows:

$$\left|\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda_i \mathbf{I}\right| = 0$$
$$\left(\mathbf{X^T X} - \lambda \mathbf{I}\right) = 0$$

2.2.3 Singular Value Decompotition (SVD)

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

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

with:

- $\mathrm{r} \le \{27,4\}$
- $\mathbf{U}$ 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 $\sqrt{\lambda_1} \ge \sqrt{\lambda_2} \ge \ldots \ge \sqrt{\lambda_r}$ which forms the matrix as follows:

  $$\mathbf{L} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_r} \end{bmatrix}$$

  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 :

  $$\mathbf{A} = \left[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_r\right]$$
- The columns of matrix $\mathbf{U}$ is obtained from the formula :

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

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 $\mathbf{X}^\mathbf{T}\mathbf{X}$
- $\mathbf{X}$: initial matrix that corrected against average

After the SVD results are obtained, the equation becomes:

$$\mathbf{X} = \mathbf{U}\mathbf{L}^{\alpha}\mathbf{L}^{1-\alpha}\mathbf{A}^{\mathbf{T}}$$

In determining $\mathbf{L}^{\alpha}$, for $0 \leq \alpha \leq 1$, then the diagonal matrix has diagonal elements $\sqrt{\lambda_1^{\alpha}} \geq \sqrt{\lambda_2^{\alpha}} \geq ... \geq \sqrt{\lambda_r^{\alpha}}$. Determinition of $\mathbf{L}^{1-\alpha}$ applies the same as diagonal elements $\sqrt{\lambda_1^{1-\alpha}} \geq \sqrt{\lambda_2^{1-\alpha}} \geq ... \geq \sqrt{\lambda_r^{1-\alpha}}$ [10].

Suppose $\mathbf{G} = \mathbf{U}\mathbf{L}^{\alpha}$ dan $\mathbf{H}^{\mathbf{T}} = \mathbf{L}^{1-\alpha}\mathbf{A}^{\mathbf{T}}$, then equation becomes:

$$\mathbf{G}\mathbf{H}^{\mathbf{T}} = \mathbf{U}\mathbf{L}^{\alpha}\mathbf{L}^{1-\alpha}\mathbf{A}^{\mathbf{T}} = \mathbf{U}\mathbf{L}\mathbf{A}^{\mathbf{T}} = \mathbf{X}$$

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

$$\mathbf{x}_{ij} = \mathbf{g}_i\mathbf{h}_j$$

with $\mathbf{g}_i$, i =1, 2,..., *27* and $\mathbf{h}_j$, j =1, 2, 3, 4 each is row of matrix $\mathbf{G}$ and coloumn 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 is taken are λ_1 and λ_2 so the amount of diversity explained is as follows:

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

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 (Grenaacre., 2010; 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

Table 1 shows the number of components formed, namely 2 components and Table 2 shows the diversity of data that can be explained by PCA Biplot maps using 2 components calculated from the cumulative eigenvalue is 0.83020, it can be concluded that the percentage of data that can be explained by PCA Biplot is 83,020%. This can be seen in the table as follows:

Table 1. Number of Components Formed

Component Matrix[a]

	Component	
	1	2
age 7-12	0.006	0.987
age 13-15	0.827	0.221
age 16-18	0.861	0.007
age 19-24	0.909	-0.214

Table 2 Data Diversity Explained by PCA Biplot

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.252	56.288	56.288
2	1.069	26.732	83.020

3	0.461	11.531	94.552
4	0.218	5.448	100.000

3.2 Identify information on Mapping Analysis Results

Figure 1 shows the results of analysis through maps, formed 2 groups of districts/cities with variable characteristics that have similarities. Group 1 consists of 18 districts and cities namely Cianjur district, Karawang district, Pangandaran, Majalengka, Indramayu, Purwakarta, Ciamis, Subang, Sukabumi, Sumedang, Bandung, Tasikmalaya, Bandung Barat, Kuningan, Bekasi, Tasikmalaya city, Banjar city, Bekasi city. Group 2 consists of 9 districts and cities namely Depok city, Sukabumi, Cirebon, Cimahi, Bandung, Bogor, Bogor district, Cirebon district, Garut district.

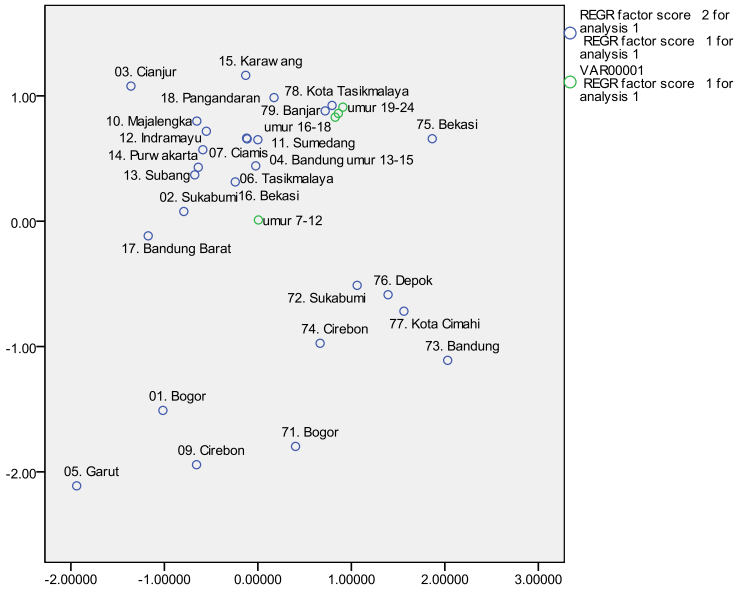

Figure 1. Mapping of Districts and Cities in West Java

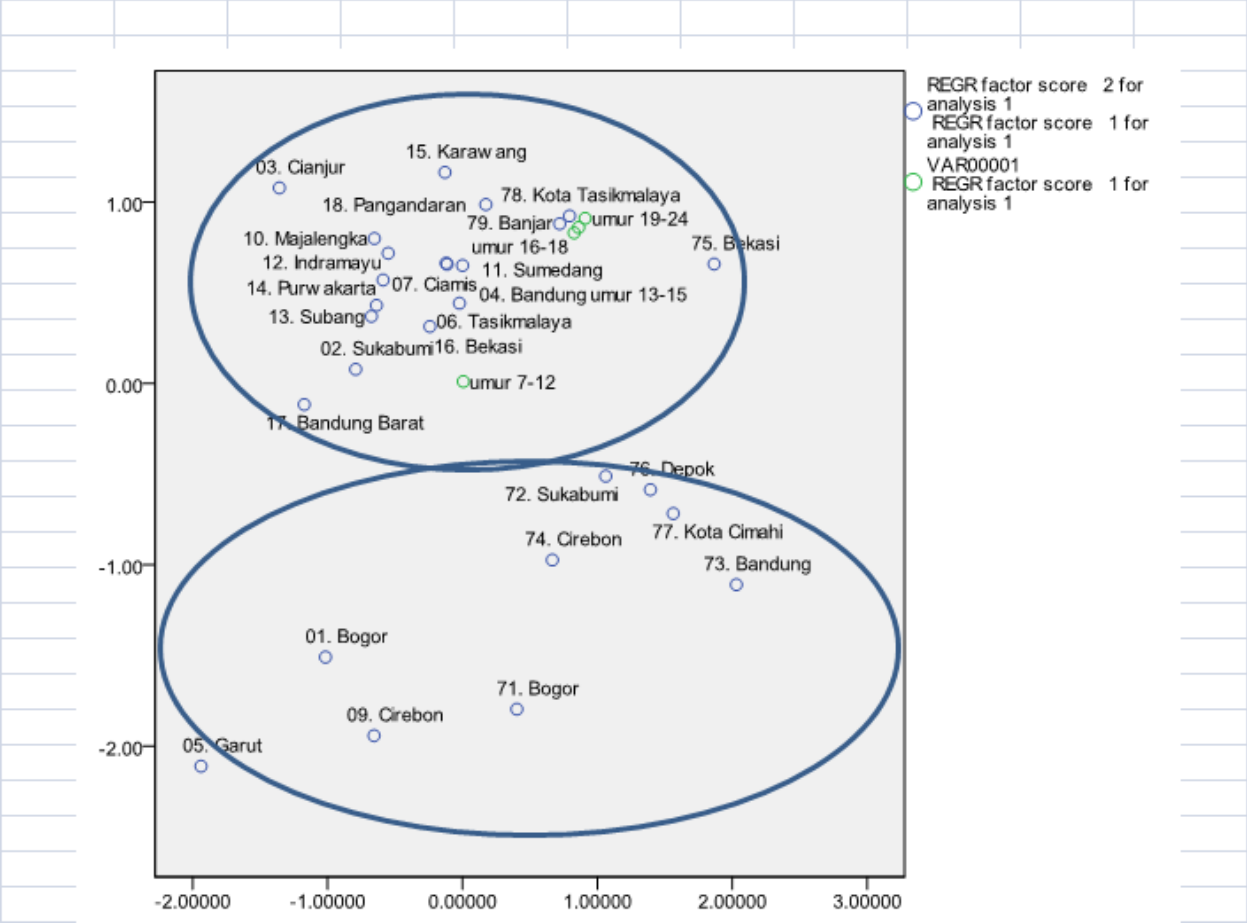

Figure 2. Map of City District Grouping

3.3 Relationship between Research Variables

Figure 3 shows the closeness of the relationship between variables observed based on the grouping of districts and cities in West Java. From the Figure, it can be seen that the school enrollment rates aged 13-15 years, aged 16-18 years, and aged 19-24 years have a very close relationship, this is indicated by the angle formed by the two lines that show school enrollment rates.

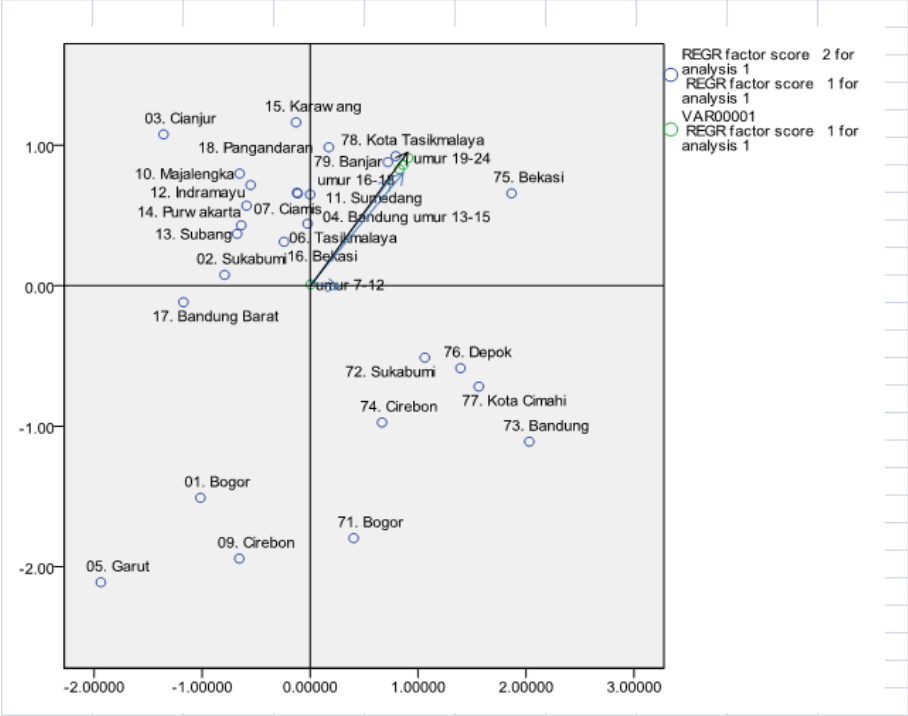

Figure 3. Inter-Variable Relationship Map

3.4 Variable Diversity

Variable diversities are described through long / short vector sizes. The variable that has the greatest diversity is the school enrollment rate aged 19-24 years, while the variable that has the smallest diversity is the school enrollment rate aged 7-12 years.

4. Conclusion

Based on the analysis using Biplot Principal Component Analysis, it was concluded that formed two groups of districts / cities with similar characteristics of variables, the percentage of diversity of data that can be explained by PCA Biplot was 83,020%, school enrollment rates aged 13-15 years, age 16- 18 years, and the age of 19-24 years has a very close relationship, this is indicated by the angle formed by the two lines which shows the school enrollment rate, the variable that has the greatest diversity is the school enrollment rate of 19-24 years, while the variable that has the smallest diversity is the school enrollment rate aged 7-12 years.

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

Depdiknas, Undang-undang RI No.20 tahun 2003 Tentang Sistem Pendidikan Nasional, 2003.

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.
Badan Pusat Statistik, https://www.bps.go.id.
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, 2016.
Machfoedz, I., and Suryani, E., *Pendidikan Kesehatan Bagian Dari Promosi Kesehatan*, Fitrayama, Yogyakarta, 2007.
Mattjik, A. A., and Sumertajaya, I. M., *Sidik Variabel Ganda*, IPB Press, Bogor, 2011.
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.
Ukalski, K., and Klisz, M., Application of GGE biplot graphs in multi-environment trials on selection of forest trees, *Folia Forestaia Polonica, Series A – Forestry,* Vol. 58 (4), pp. 228-239, 2016.

Biographies

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.
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.

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.