ROBUST BIPLOT ANALYSIS OF NATURAL DISASTERS IN INDONESIA FROM 2019 TO 2021

Hilda Venelia[1], Khoirin Nisa[1*], Rizki Agung Wibowo[1], Mona Arif Muda[2]

[1] Department of Mathematics, Faculty of Mathematics and Natural Sciences,
University of Lampung, Lampung, Indonesia
[2]Department of Informatics Engineering, Faculty of Engineering, University of Lampung, Indonesia
e-mail: khoirin.nisa@fmipa.unila.ac.id

Abstract

Indonesia is one of the most natural disaster-prone countries in the world, frequently exposed to a range of hazards. Currently, Indonesia has 34 provinces and natural disasters that occur in each province are different, therefore it is necessary to analyze the mapping of natural disasters that often occur in each province to provide scientific analysis for risk management of the natural disasters. One of the quick steps in describing data that can be used is biplot analysis, as biplot analysis can describe a lot of data then summarized it into the form of a two-dimensional graph. The aim of this research is to map 34 provinces in Indonesia based on the incidence of natural disasters from 2019 to 2021 using robust biplot analysis to see which provinces that have a high risk of natural disaster. Based on the result, robust biplot analysis can explain 87,9% of the information on natural disasters in every province in Indonesia. Lampung, Bengkulu, Bangka Belitung, Special Region of Yogyakarta, North Sulawesi, West Sulawesi, Southeast Sulawesi, Gorontalo, East Nusa Tenggara, Bali, Maluku, West Maluku, Papua, and West Papua are provinces that have similar natural disaster characteristics. Flood, tornado and forest and land fires are natural disasters that often occur in Indonesia. The provinces that have the highest risk of flood, landslide, and tornado were West Java, Central Java, and East Java. Then, the provinces with the highest risk of forest and land fires were Aceh and South Kalimantan.

Kata kunci: Natural disasters, risk management, biplot.

INTRODUCTION

Based on Law of Republic of Indonesia number 24 of 2007 concerning Disaster Management, disaster is an event or series of events that threaten and disrupt people's lives and livelihoods, caused by natural and/or non-natural and human factors that result in human casualties, environmental damage, property losses, and psychological impacts. Natural disaster is a catastrophic event with atmospheric, geological, and hydrological origin (e.g., droughts, earthquakes, floods, hurricanes, landslides) that can cause fatalities, property damage and social environmental disruption (Xu, *et al*., 2016). Every country in the world has been hit by natural disaster, either high, medium, or low risk. Indonesia is one of the countries with a high disaster-prone level in most of its territory. Geographically, Indonesia is an archipelagic country located at the confluence of four tectonic plates, namely Asian Continent plate, Australian Continent plate, Indian Ocean plate and Pacific Ocean. These conditions are potential disaster-prone such as volcanic eruptions, earthquakes, tsunamis, floods and landslides (Arnold, 1986).

Indonesian National *Agency* for *Disaster Management* (*Indonesian*: Badan Nasional Penanggulangan Bencana (BNPB)) noted that the annual trend of natural disasters in Indonesia tends to increase. In period of 2010 until 2020, the highest number of annual disasters occurred in 2019, which was 3.814 incidents. Disasters that hit Indonesia are generally caused by hydrometeorology. Flood, landslide, and tornado have dominated the natural disasters that have occurred over the past decade. The number of missing and dead victims annually reaches hundreds to thousands of people. The peak occurred in 2018 with 3.397 disasters which caused the number of victims to reach 6.240 people. That year, Indonesia experienced a series of major disasters, such as Lombok earthquake (West Nusa Tenggara), Palu earthquake and tsunami (Central Sulawesi), and Sunda Strait tsunami. Meanwhile, from 31 December 2019 to 1 January 2020 there was rain in the Greater Jakarta (i.e. Jakarta Bogor Tangerang Bekasi, abbreviated to Jabodetabek) which resulted in flooding on 1-3 January 2020. The incident killed at least 48 people, and more than 31.000 people were evacuated. Then, floods and landslides also occurred in East Nusa Tenggara in early April 2021 due to Cyclone Seroja which killed 183 people.

Therefore, many provinces in Indonesia are affected by natural disasters in terms of social, economic, environmental, and other aspects. This is due to the lack of readiness of government and society in disaster management before the disaster occurs or better known as Disaster Risk Reduction (DRR). The main objective of DRR is to reduce losses due to disaster impacts by increasing society capacity and reducing exposure and increasing community resilience through preparedness, emergency response, and recovery. The first thing that must be considered in DRR is to map the province in Indonesia based on the level of risk from the natural disaster and find out what natural disasters often occur in each province in Indonesia. This is because, by mapping the provinces in Indonesia based on their level of natural disaster risk, it can be seen which provinces that have high, medium, and low risk of natural disaster. Thus, natural disaster management can be prioritized to provinces that have a high risk of being affected by natural disasters.

There is one of descriptive statistical methods that can be used for mapping purposes, namely biplot analysis. Biplot analysis is one of descriptive statistical methods that can simultaneously present plots of n observations and p variables in two-dimensional graph (Jolliffe, 2002). Biplot analysis has been used in various research fields, one can see e.g. (Makhya, *et al*., 2020; Fitry, *et al*., 2021; Suryowati, JP and Nasution, 2021; Ghazvini, 2021). However, the classical biplot is sensitive to the presence of outliers in the data, therefore when data contain outliers it is recommended to use robust biplot.

METHODOLOGY

The data used in this study is Natural Disasters in Indonesia data from January 1st 2019 until August 14th 2021 obtained from Indonesian National Disaster Mitigation Agency (BNPB). The data consists of $n = 34$ provinces and $p = 6$ natural disasters in Indonesia. The variables used in this study are the types of natural disasters, namely Flood (F), Landslide (LS), Earthquake (EQ), Tidal Wave/Abrasion (TW), Forest and Land Fires (FF), and Tornado (T).

The computation of biplot analysis based on Singular Value Decomposition (SVD) (Gabriel, 1971) that can be written as follows:

$$\mathbf{X} = \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{V}^{\mathrm{T}}, \qquad (1)$$

where $\mathbf{X}$ is data matrix with rank r, $\mathbf{U}_{(n\times r)}$ and $\mathbf{V}_{(p\times r)}$ are columnar orthonormal matrices $(\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I_r})$ and $\boldsymbol{\Lambda}_{(r\times r)} = \mathrm{diag}\left(\sqrt{\lambda_1, \lambda_2, \dots, \lambda_r}\right)$ with $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_r}$ and $r \leq p \leq n$. The element $\lambda_i,\ i = 1, 2, \dots, r$ are eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (Johnson and Wichern, 2014). Matrix $\mathbf{V}$ is a matrix whose columns consist of eigenvectors $\boldsymbol{v_i}$ corresponding to λ_i of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. The columns of $\mathbf{U}$ can be calculated as:

$$\mathbf{U} = \frac{1}{\sqrt{\lambda_i}} \times \boldsymbol{v_i}.$$

Define $\boldsymbol{\Lambda}^{\alpha} = \mathrm{diag}\left(\sqrt{\lambda_1^{\alpha}}, \sqrt{\lambda_2^{\alpha}}, \dots, \sqrt{\lambda_r^{\alpha}}\right)$ with $\alpha \in [0,1]$ and suppose $\mathbf{G} = \mathbf{U}\,\boldsymbol{\Lambda}^{\alpha}$, $\mathbf{H} = \mathbf{V}\,\boldsymbol{\Lambda}^{1-\alpha}$ than (1) can be written as (Gabriel, 1971):

$$\begin{aligned}\mathbf{X} &= \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{V}^{\mathrm{T}}\\ &= (\mathbf{U}\,\boldsymbol{\Lambda}^{\alpha})(\boldsymbol{\Lambda}^{1-\alpha}\,\mathbf{V}^{\mathrm{T}})\\ &= \mathbf{G}\,\mathbf{H}^{\mathrm{T}}.\end{aligned}$$

Therefore, element (i,j) of data matrix $\mathbf{X}_{(n\times p)}$ can be expressed as:

$$\boldsymbol{x}_{ij} = \boldsymbol{g}_i^T\,\boldsymbol{h_j}$$

where $\boldsymbol{g}_i^T, i = 1, 2, \dots, n$ and $\boldsymbol{h_j}, j = 1, 2, \dots, p$ are row vectors of $\mathbf{G}$ and $\mathbf{H}$, respectively, with r elements. Here, n rows of $\mathbf{G}$ correspond to rows of $\mathbf{X}$ and p rows of $\mathbf{H}$ correspond to columns of $\mathbf{X}$.

Outliers are data points that deviate far from the majority of the data (Filzmoser, 2004). Multivariate outliers can be identified using Mahalanobis distance (Majewska, 2015) as follows:

$$D_i(\boldsymbol{x}_i, \overline{\mathbf{X}}) = \sqrt{(\boldsymbol{x_i} - \overline{\mathbf{X}})^{\mathrm{T}}\mathbf{S}^{-1}(\boldsymbol{x_i} - \overline{\mathbf{X}})},$$

where $\boldsymbol{x_i}$ represents the i-th object, $\overline{\mathbf{X}}$ is the mean vector and $\mathbf{S}$ is the sample covariance matrix (Ghorbani, 2019). If $D_i^2(\boldsymbol{x}_i, \overline{\mathbf{X}}) > \chi^2_{p;1-\alpha}$ with p the number of variable in the data and α the significant level, then the object can be identified as outlier.

However, not all outliers can be removed from data because it can cause the loss of information contained in data. Furthermore, the covariance matrix also very sensitive to outliers (Islami and Sihombing, 2021; Larasati, *et al.*, 2021). To overcome this problem, biplot analysis can be generated using a robust covariance matrix by estimating eigenvalues and eigenvectors of $\mathbf{U}$ and $\mathbf{V}$ such that the predicted results are resistant to outliers (Hawkins, *et al.*, 2001). In this case the calculation of SVD uses robust estimator of mean vector $\overline{\mathbf{X}}$ and covariance matrix $\mathbf{S}$. Here we used the Minimum Covariance Determinant (MCD) estimator which is knows as a very robust method for mean vector and covariance matrix estimation (Hubert and Debruyne, 2010). The MCD estimator with the fast-MCD algorithm proposed by Rousseeuw and Van Dreissen (1999) can be done with the following steps:

1. Take a random subset of matrix $\mathbf{X}$, suppose the subset as $\mathbf{K_1}$ with the number of elements as much as h,
$$h = \frac{(n+p+1)}{2}$$
where n and p are the sample size and the number of variables in the data respectively.
2. Calculate the mean vector $\overline{\mathbf{X}}_\mathbf{1}$ and the covariance matrix $\mathbf{S_1}$ of $\mathbf{K_1}$ using:
$$\overline{\mathbf{X}}_1 = \frac{1}{h}\sum_{i\in H}\boldsymbol{x}_i$$
$$\mathbf{S}_1 = \frac{1}{h-1}\sum_{i=1}^{h}(\boldsymbol{x}_i - \overline{\mathbf{X}}_1)^T(\boldsymbol{x}_i - \overline{\mathbf{X}}_1)$$

3. Calculate the determinant of $\mathbf{S}_1$.
4. Calculate the relative distance of each observation to $\bar{\mathbf{X}}_1$ and covariance $\mathbf{S}_1$ using Mahalanobis distance.
5. Sort the observation by the distance of Mahalanobis, from the smallest to largest.
6. Take the elements of h observations with the smallest distance based on step 5 to become subset of $\mathbf{K}_2$, repeat steps 2 until 5 so that it is found the subsets are converge and have the smallest determinant of the covariance matrix, namely:

$$|\mathbf{S}_{n+1}| < |\mathbf{S}_n|$$

7. Based on the elements of h, the next data is weighted as:

$$w_i = \begin{cases} 1, & (\boldsymbol{x}_{ij} - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\boldsymbol{x}_{ij} - \bar{\mathbf{X}}) < \chi^2_{p;1-\alpha} \\ 0, & \text{others} \end{cases}$$

8. Based on the weighted above, the fast-MCD estimators are:

$$\bar{\mathbf{X}}_{MCD} = \frac{\sum_{i=1}^{n} w_i \, \boldsymbol{x}_{ij}}{\sum_{i=1}^{n} w_i}$$

$$\mathbf{S}_{MCD} = \frac{\sum_{i=1}^{n} w_i (x_{ij} - \bar{\mathbf{X}}_{MCD})^T (x_{ij} - \bar{\mathbf{X}}_{MCD})}{(\sum_{i=1}^{n} w_i) - 1}$$

Robust biplot analysis in this research was carried out using R software, the analysis procedure can be described as follows:

1. data screening to detect outliers by using Mahalanobis distance;
2. performing robust biplot analysis based on the covariance matrix $\mathbf{S}_{MCD}$ using *robustbase* package;
3. plotting data using *ggbiplot* package;
4. interpretation of the result.

RESULT & DISCUSSION

In this research, the first step was to detect outliers in data based on the robust Mahalanobis distance. The plot of the robust squared Mahalanobis distances and the chi-square quantiles is shown in Figure 1.

The result showed that there were ten provinces that were identified as outliers as can be seen in Figure 1. The list of data that identified as outliers is presented in the following table 1.

In Table 1, Aceh, North Sumatra, West Sumatra, South Kalimantan, and South Sulawesi are provinces that experience flood and tornado with high risk, especially in West Java, Central Java, and East Java that experience flood, landslide, and tornado with very high risk than other provinces. While East Kalimantan and Bali are provinces that experience flood and tornado with low risk than others.

Based on the report above it can be concluded that the data of natural disasters in Indonesia from 2019 until 2021 contains outliers. Therefore, it is appropriate to use robust biplot analysis in this research. The resulted robust biplot describes the characteristics of object (Provinces) against the variables (Natural Disasters) as can be seen in the following figure.

The important thing that needs to be considered in biplot analysis is how good the biplot can explain the information

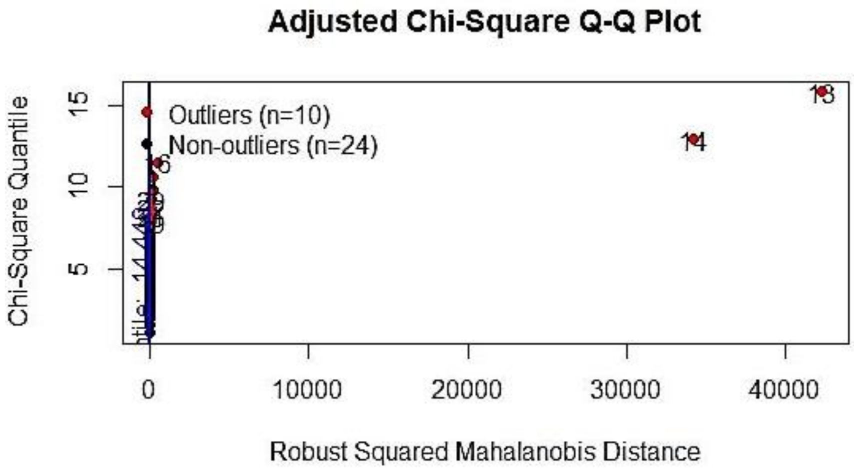

Figure 1. Outlier detection based on robust Mahalanobis distance

Table 1. Provinces that are considered as outliers and their number of natural disasters

Province	Flood	Landslide	Earthquake	Tidal Wave	Forest and Land Fires	Tornado
Aceh	217	33	4	13	245	193
North Sumatra	108	22	1	1	27	96
West Sumatra	133	68	1	7	24	164
West Java	452	883	26	3	83	895
Central Java	446	788	3	11	95	791
East Java	344	99	30	2	195	394
South Kalimantan	110	17	0	3	168	112
East Kalimantan	41	37	0	1	105	13
Bali	14	48	5	4	14	79
South Sulawesi	119	54	1	7	28	159

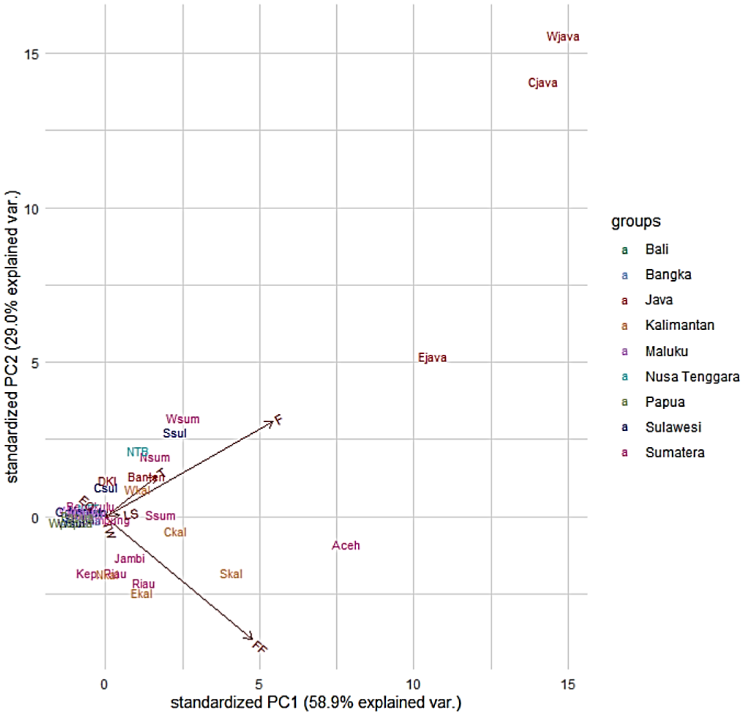

Figure 2. Robust biplot charts of natural disasters.

contained in data. Based on Figure 2, it can be explained that the information in component 1 was 58,9% and component 2 was 29%. Therefore, the cumulative information that can be explained by biplot against data based on two components was 87,9%. It means that the resulted biplot was very capable in describing the characteristics of natural disaster data in Indonesia. Furthermore, there are four information obtained from biplot in Figure 2 as described as follows.

Proximity Between Objects

This information can be used as a guide to identifying objects that have similar characteristics to other objects. Two objects that have similar characteristics will be described as two points with adjacent positions. Based on Figure 2, West Sumatra and South Sulawesi are two provinces with similar characteristics. West Nusa Tenggara and North Sumatra are two provinces with similar characteristics. DKI Jakarta and Central Sulawesi are two provinces with similar characteristics. Then, Banten and West Kalimantan are two provinces with similar characteristics. Also with Riau and East Kalimantan have similar characteristics. Jambi, Riau Islands, and North Kalimantan are three provinces with similar characteristics. Meanwhile, Lampung, Bengkulu, Bangka Belitung, Special Region of Yogyakarta, North Sulawesi, West Sulawesi, Southeast Sulawesi, Gorontalo, East Nusa Tenggara, Bali, Maluku, West Maluku, Papua, and West Papua form their own group and have similar characteristics.

Variability of Variables

This information is used to see if there are variables that have almost the same diversity. Variables with small variance are represented by short vectors, whereas if the variance is high, they are represented by long vectors. Based on Figure 2, flood and forest fire are variables with the highest variance compared to other variables because they have the longest vector. It means that the level of risk of flood and forest fires varies widely in each province in Indonesia. Then, tidal wave is variable with the smallest variance represented by the shortest vector. This means, tidal wave was a natural disaster with the lowest risk in every province in Indonesia from 2019 until August 2021.

Correlation Between Variables

Correlation between variables shows how one variable is related to another. In biplot, variables are presented as vector lines. If two vectors form an acute angle ($< 90^\circ$) or have the same direction, then it can be stated that the two variables are positively correlated. Whereas if two vectors form obtuse angles ($> 90^\circ$) or have opposite directions, it can be stated that the two variables are negatively correlated. Then, if the angle formed between two vectors is a right angle, then it can be stated that the two variables are not correlated.

Based on Figure 2, flood has positive correlation with forest and land fires, tornado, and landslide. This means that from 2019 until August 2021, the level of flood risk in each province in Indonesia was directly proportional to the level of risk of forest and land fires, tornado, and landslide. Earthquake has negative correlation with flood, landslide, and forest and land fires which means from 2019 until August 2021, the level of earthquake risk was inversely proportional to the level of risk of flood, landslide, and forest and land fires. But earthquake, tornado, and forest and land fires have no correlation.

Relative Position of Object to Variable

This information is used to see the advantages of each object. Objects that are in the same direction as the variable vector, indicate that the value of the object is above the average, if it is in the opposite direction, it means that the value is below the average, if it is almost in the middle, it means that the value is close to the average. Based on Figure 2, information about natural disasters in Indonesia from 2019 until August 2021 shows that West Java, Central Java, and East Java are provinces with the highest risk of flood, tornado, and landslide than other provinces. Whereas, West Sumatra, North Sumatra, South Sumatra, Banten, DKI Jakarta, West Kalimantan, West Nusa Tenggara, South Sulawesi, and Central Sulawesi are provinces with a medium risk of flood and tornado.

Aceh and South Kalimantan were at the highest risk of forest and land fires. Meanwhile, South Sumatra, Jambi, Riau, Riau Islands, Central Kalimantan, North Kalimantan, and East Kalimantan have a medium risk of forest and land fires and a lower risk of earthquake. Meanwhile, Lampung, Bengkulu, Bangka Belitung, Special Region of Yogyakarta, North Sulawesi, West Sulawesi, Southeast Sulawesi, Gorontalo, East Nusa Tenggara, Bali, Maluku, West Maluku, Papua, and West Papua are provinces with a medium risk of earthquake and a lower risk of forest and land fires.

CONCLUSION

Based on results and discussion, there are several points that can be concluded as follow:

1. The information obtained from robust biplot analysis is 87,9% which means very good to describe the characteristics of natural disasters in Indonesia.
2. Flood, tornado, and forest and land fires are natural disasters that often occur in Indonesia. Meanwhile, tidal wave is a

natural disaster that is very rare occurrence in Indonesia.

3. Lampung, Bengkulu, Bangka Belitung, Special Region of Yogyakarta, North Sulawesi, West Sulawesi, Southeast Sulawesi, Gorontalo, East Nusa Tenggara, Bali, Maluku, West Maluku, Papua, and West Papua are provinces that have similar natural disaster characteristics and have medium risk of earthquake.
4. West Java, Central Java, and East Java have the highest risk of being affected by flood, tornado, and landslide. Meanwhile, Aceh and South Kalimantan are provinces at high risk of being affected by forest and land fires.

Based on these information, it is recommended for further research to examine the factors and impacts of natural disasters that occur in Indonesia.

REFERENCES

Arnold, E.P. 1986. Series on Seismology Volume V: Indonesia. South-East Asia Association of Seismology and Earthquake Engineering.

Filzmoser, P. 2004. *Multivariate Outlier Detection Method*. Austria: Department of Statistics and Probability Theory.

Filzmoser, P. and Gregorich, M. 2020. Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional, and Cellwise Outliers. *Mathematical Geosciences*, Vol. 52 (8), 1049-1066.

Fitry, R., Rostini, N., Hersanti and Zubair, A. 2021. Biplot and Correlation Analysis of 189 Superior F2 Genotypes Chili-Unpad in Indonesia. *Advances in Biological Sciences Research*, Vol. 9, 10-14.

Gabriel. K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, Vol. 58 (3), 453–467.

Ghazvini, H., Bagherikia, S., Pour-Aboughadareh, A., Sharifalhossaini, M., Razavi, S.A., Mohammadi, S., Ghasemi, K. M., Fathihafshejani, A. and Khakizade, G. 2021. GGE Biplot Analysis of Promising Barley Lines in the Cold Regions of Iran. *Journal of Crop Improvement*, Vol. 35, 1–12.

Ghorbani, H. 2019. Mahalanobis Distance and Its Application for Detecting Multivariate Outliers. *Facta Universitatis: Mathematics and Informatics*, Vol. 34, 583.

Hawkins, D. M., Liu, L. and Young, S. S. 2001. Robust Singular Value Decomposition. *Linear Algebra and its Applications*, Vol. 417 (122), 370-380.

Hubert, M. and Debruyne, M. 2010. Minimum covariance determinant. *Wiley Reviews: Computational Statistics*, Vol. 2 (1), 36-43.

Islami, R. L. and Sihombing, P. R. 2021. Application Biplot and K-Medians Clustering to Group Export Destination Countries of Indonesia's Product. *Advance Sustainable Science, Engineering and Technology*, Vol. 3 (1), 0210105.

Johnson, R. and Wichern D. 2014. *Applied Multivariate Statistical Analysis 6th ed*. England: Pearson Education Limited.

Jolliffe, I. T. 2002. *Principal Component Analysis*, *Second Edition*. New York: Springer-Verlag.

Larasati, S.D.A., Nisa, K. and Herawati, N. 2021. Robust Principal Component Trimmed Clustering of Indonesia Provinces Based on Human Development Index Indicators. *Journal of Physics: Conference Series*, Vol. 1751 (1).

Mahalanobis, P. C. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, Vol. 2 (1), 49-55.

Majewska, J. 2015. Identification of Multivariate Outliers Problems and Challenges of Visualization Methods. *Informatyka i Ekonometria* 247, 69-83.

Mattjik, A. A. and Sumertajaya, I. M. 2011. *Sidik Peubah Ganda dengan Menggunakan SAS*. Bogor: IPB Press.

Rousseeuw, P. J. and Driessen, K. V. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, Vol. 41 (3), 212-223.

Suryowati, K, Titah, J.P., and Nasution, N. 2021. Application Biplot Analysis on Mapping of Non-Convertive Diseases in Indonesia. *Parameter: Journal of Statistics*, Vol. 1 (2), 11-20.

Syarief, M., Nisa, K., Larasati, S., Suripto and Suprihatin, A. 2020. Government policy in the case of provision of commercial bank loans based on business fields per Sumatra region. *TEST Engineering and Management*, Vol. 82, 6843-6851.

Xu, J., Wang, Z., Shen, F., Ouyang, C. and Tu, Y. 2016. Natural disasters and social conflict: A systematic literature review. *Int. J. Disaster Risk Reduct*, Vol. 17, 38–48.