Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators

To cite this article: S D A Larasati *et al* 2021 *J. Phys.: Conf. Ser.* **1751** 012021

View the article online for updates and enhancements.

IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 110.137.37.28 on 28/01/2021 at 11:49

Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators

S D A Larasati[1,2], **K Nisa**[2*], **N Herawati**[2]

[1] Graduate School of Mathematics, University of Lampung, Jl. Sumantri Brojonegoro no 1, Bandar Lampung, Indonesia
[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung, Jl. Sumantri Brojonegoro no 1, Bandar Lampung, Indonesia

email: siskadiahayularasati@gmail.com[1,2], khoirin.nisa@fmipa.unila.ac.id[2*], netti.herawati@fmipa.unila.ac.id[2]

*corresponding author

Abstract. Cluster analysis is a multivariate technique for grouping observations into clusters based on the observed values of several variables for each individual. The existence of outliers in the data can heavily influence standard clustering methods, i.e. the outliers will cause the standard clustering results to be not optimal. Therefore, it is necessary to use a robust clustering method. Trimmed clustering is one of robust clustering methods which is non-hierarchical and known for its good performance in cluster analysis when data contain outlier. The purpose of this study is to classify 34 provinces in Indonesia based on the 2019 Human Development Index (HDI) indicators and see the achievements of human development in each province. The results of this study indicate that there are three optimal clusters. The first cluster consists of 17 provinces with good HDI criteria, the second cluster consists of 9 provinces with a fairly good HDI, and the third cluster consists of 7 provinces with the lowest HDI criteria.

Keyword: robust, trimmed cluster analysis, human development index

1. Introduction

The basic idea of human development is to position humans as the true assets of the nation and create growth in the economic, social, political, cultural and environmental fields that encourage the improvement of people's welfare. Based on this thinking, the main goal of human development is to be able to create an environment that allows people to have a long life, be healthy, and lead a productive life [1].

Achievement of human development is a summary measure of average achievement in key dimensions of human development namely: a long and healthy life, being knowledgeable and have a decent standard of living. The Human Development Index (HDI) is an index used to see human development in a long term. Human development in Indonesia continues to progress. In 2019, Indonesia's HDI reached 71.92. This figure increased by 0.53 points or grew by 0.74 percent compared to 2018 [2]. Based on this enhancement, it is necessary to group Indonesian provinces to find the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

provinces with the same characteristics and also differences between the groups of provinces. The result can be used as a basis for the government to determine specific policies or programs that suitable for each group.

One of the statistical methods that can be used for grouping objects is cluster analysis. According to Härdle & Simar [3], cluster analysis is a multivariate technique which has the main objective to group objects so that diversity within a cluster is minimum while between clusters is maximum. Cluster analysis is based on a distance matrix representing a similarity measure, and the most commonly used measure is Euclidean distances. There is assumption that must be fulfilled when using Euclidean distances for cluster analysis, i.e. all variables are uncorrelated, and this assumption is frequently ignored. An effective procedure that can be used in dealing with the correlation between the variables is by performing a principal component analysis (PCA) before calculating the Euclidean distances, one can see e.g. [4-7] for the use of PCA for clustering in various researches.

PCA is a multivariate technique which aims to reduce the dimensions of data (i.e. the number of the original variables) in order to obtain new variables (i.e. principal components) which are not correlated and contain most of the information of the original variables [8]. Principal components are new variables that are constructed as linear combinations of the original variables. These combinations are done in such a way that these new variables are uncorrelated and most of the information within the initial variables is stored into the first components. Even though the *k*-dimensional data give *k* principal components but PCA tries to put maximum possible information in the first ones.

Deviations from theoretical assumptions together with the presence of certain amount of outlying observations are common in many practical statistical applications. In this case, a robust procedure is needed. Robustness in statistics refers to stable behavior of methodology under small changes of data or models. Robustness is a desirable property for general statistical methodology and it has been studied by many authors. Some examples of robustness study can be seen in e.g. [9-12]. A small percentage of outliers can have a large impact on many statistical techniques. This is also the case when applying cluster analysis methods, where those troubles could lead to unsatisfactory clustering results. Therefore, in the case of the existence of outliers and correlation among variables, performing robust procedures of PCA and cluster analysis are highly recommended.

Let $\boldsymbol{X} = (X_1\ X_2\ \ldots X_k)'$ has a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ having eigen vectors $\boldsymbol{a}_j$, $j = 1, 2, \ldots, k$. The principal components are linear combinations of the *k* original variables which can be expressed as follows [8]:

$$PC_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{kj}X_k = \boldsymbol{a}_j'\boldsymbol{X}. \tag{1}$$

The covariance matrix $\boldsymbol{\Sigma}$ is known to be very sensitive to the presence of outliers. In many situations, outliers cannot be removed from data for some reasons such as the information they contain and the complicated procedure for outlier detection. To overcome this, a robust estimate of $\boldsymbol{\Sigma}$ is needed namely by replacing the classical sample covariance matrix $\mathbf{S}$ with a robust estimator. The use of robust covariance matrix estimate for the principal component constructing is the same as performing a robust PCA, see e.g. [13] for details.

One of the robust methods for covariance matrices is the Minimum Covariance Determinant (MCD). According to Rousseeuw & Van Driessen [14], the MCD estimator is a pair $(\bar{\boldsymbol{X}}_{MCD}, \boldsymbol{S}_{MCD})$, where $\bar{\boldsymbol{X}}_{MCD}$ is the mean vector and $\boldsymbol{S}_{MCD}$ is the covariance matrix that minimizes the determinant value of the sample covariance matrix $\boldsymbol{S}$ in a subsample containing exactly *h* members of *n* observations, i.e.

$$\boldsymbol{S}_{MCD} = \min\{\det(\boldsymbol{S}_j)\}, \qquad j = 1, \ldots, \binom{n}{h}. \tag{2}$$

The standard value of *h* is $[(n+k+1)/2]$ where *n* is the sample size and *k* is the number of variables in the data.

For robust cluster analysis, we used trimmed cluster technique called TCLUST [15-16]. TCLUST is a method in statistical clustering technique which is based on modification of trimmed *k*-means

clustering algorithm [17]. According to García-Escudero *et.al.* [18], the algorithm of TCLUST can be described as follow:

1. Randomly select starting values for the centers $\boldsymbol{m_j^0}$'s, the covariace matrices $\boldsymbol{S_j^0}$'s and the weights of the grup $\boldsymbol{p_j^0}$'s for $j = 1, \dots, k$.
2. From the $\boldsymbol{\theta^l} = (p_1^l, \dots, p_k^l, m_1^l, \dots, m_k^l, S_1^l, \dots, S_k^l)$ returned by the previous iterations:
   - Obtain $d_i = D(x_i, \theta^l)$ for the observation $\{x_1, \dots, x_n\}$ and keep the set H having the $[n(1-\alpha)]$ observations with largest d_i's where $D(x_i, \theta^l) = \max\{D_1(x_i, \theta^l), \dots, D_k(x_i, \theta^l)\}$ and $D_j(x_i, \theta^l) = \pi_j f(x; \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ with π_j - the group weight and $f(x; \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})$ is the probability density function (p.d.f.) of *k*-variate normal distribution.
   - Split H into $H = \{H_1, \dots, H_k\}$ with $H_j = \{x_i \in H : D_j(x_i, \theta^l) = D(x_i, \theta^l)\}$.
   - Obtain the number of data points n_j in H_j and their sample mean and sample covariance matrix, $\boldsymbol{m_j}$ and $\boldsymbol{S_j}$, $j = 1, \dots, k$.
   - Consider the singular-value decomposition of $\boldsymbol{S_j} = \boldsymbol{U_j'D_jU_j}$ where $\boldsymbol{U_j}$ is a orthogonal matrix and $\boldsymbol{D_j} = diag(\Lambda_j)$ is a diagonal matrix (with diagonal elements given by the vector $\boldsymbol{\Lambda}$). If the full vector of eigenvalues $\boldsymbol{\Lambda} = (\Lambda_1, \dots, \Lambda_k)$ does not satisfy the eigenvalues-ratio (ER) restriction, obtain (for instance) through Dyksra's algorithm [19-21] a new vector $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\Lambda}_1, \dots, \widetilde{\Lambda}_k)$ obeying the eigen restriction and with $\left\|\widetilde{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}^{-1}\right\|^2$ being as smaller as possible. ($\boldsymbol{\Lambda}^{-1}$ denotes the vector made up by the inverse of the elements of the vector $\boldsymbol{\Lambda}$). Notice that the eigen restriction for $\boldsymbol{\Lambda}$ corresponds exactly to the same eigen restriction applied to $\boldsymbol{\Lambda}^{-1}$.
   - Update θ^{l+1} by using:
     - $\boldsymbol{p_j^{l+1}} \hookleftarrow n_j/[n(1-\alpha)]$,
     - $\boldsymbol{m_j^{l+1}} \hookleftarrow \boldsymbol{m_j}$,
     - $\boldsymbol{S_j^{l+1}} \hookleftarrow \boldsymbol{U_j'\widetilde{D}_jU_j}$ and $\widetilde{\boldsymbol{D}}_\boldsymbol{j} = diag(\widetilde{\Lambda}_j)^{-1}$
3. Perform *F* iterations of the process described in step 2 (moderate values for *F* are usually enough) and compute the evaluation function $L(\theta^F; P_n)$ using $\theta \longmapsto L(\theta, P) \coloneqq E_p\left[\sum_{j=1}^k z_j(.;\theta) \log D_j\right]$,
4. Start from step 1 several times, keeping the solutions leading to minimal values of $L(\theta^F, P_n)$ and fully iterate them to choose the best one.

For more details and discussion on the trimmed cluster algorithm, one can see [15-18].

2. Methodology

The 2019 HDI data published by Central Bureau of Statistics of Indonesia consists of *n*=34 provinces. The variables used in the analysis are: life expectancy of birth in 2019 (LE), expected years of schooling of 7-year-old children (EYS), mean years of schooling of the population aged 25 years and over (AYS), and average spending per capita adjusted at the provincial level (ASC). The indicator variables used to compose the HDI index in Indonesia were adopted from the HDI indicators recognized by United Nations Development Program (UNDP).

The analysis was performed using R software, the analysis procedure can be described as follow:

- HDI data screening to detect the presence of outliers by using Mahalanobis distance $\boldsymbol{D(x_i)_{MCD}} = \sqrt{(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{MCD}})\boldsymbol{S_{MCD}^{-1}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{MCD}})}$ for *i*=1, 2, …, *34*;
- checking the correlation between variables LE, EYS, AYS and ASC;
- conducting robust principal component analysis using covariance matrix $\boldsymbol{S_{MCD}}$;
- performing robust clustering based on the robust principal component scores obtained using TCLUST;
- determining the optimal clusters of the provinces.

3. Result and Discussion

The first step in this research is assessing the presence of outliers in the data so that the analysis results obtained are correct. Outlier detection from the HDI data in 2019 is based on the robust squared Mahalanobis distance for 34 provinces in Indonesia as follows.

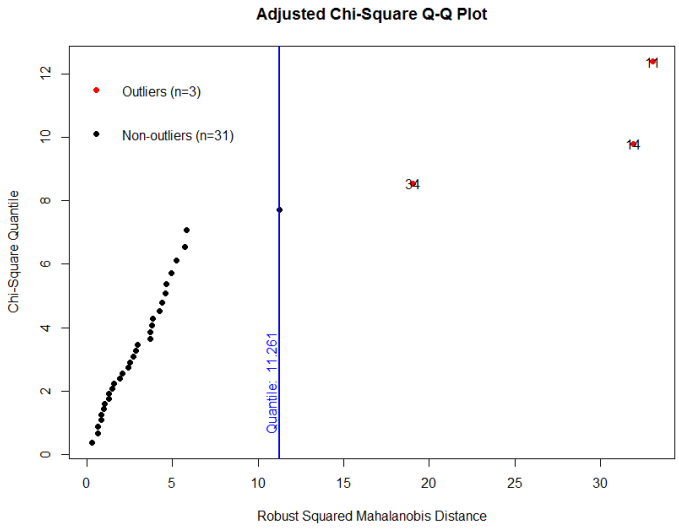

Figure 1. Outlier detection

In Figure 1, three provinces were identified as outliers. The suspected provinces are presented in the following table.

Table 1. List of Outliers

Province	LE	EYS	AYS	ASC
DKI Jakarta	72.79	12.97	11.06	18527
D.I Yogyakarta	74.92	15.58	9.38	14394
Papua	65.65	11.05	6.65	7336

In Table 1, DKI Jakarta and D.I Yogyakarta have very high values of the HDI indicators while Papua has the lowest HDI indicators in 2019. Furthermore, checking the correlation between variables is also considered important so that the characteristics of the clusters that are formed are optimal.

The following is the table of correlation matrix of data which shows correlation coefficients between variables. Based on Table 2, it can be seen that there are fairly small correlations between variables, which is between 0.2 and 0.5. To obtain independent variables for the cluster analysis, principal component analysis was performed.

Table 2. Correlation Matrix

	LE	EYS	AYS	ASC
LE	1.00000	0.25544	0.40547	0.58957
EYS	0.25544	1.00000	0.47532	0.17042
AYS	0.40547	0.47532	1.00000	0.58356
ASC	0.58957	0.17042	0.58356	1.00000

Because of the correlation and outlier problems, we used a robust method when performing PCA by using MCD mean vector $\bar{\boldsymbol{X}}_{MCD}$ and covariance matrix $\boldsymbol{S}_{MCD}$.

Determining the number of clusters to be formed can be done by looking at how the data spread and cluster. For that we looked at the scatter plots of the first and second principal components obtained from PCA. To ensure that the scatter plots of the components contain sufficient information to represent the original variables, we examined the proportions of their variances. T he variance of each principal component (PC) is equivalent to the eigenvalue of the covariance matrix $\boldsymbol{S}_{MCD}$. We obtain the eigen values of covariance matrix $\boldsymbol{S}_{MCD}$ were: 0.78220, 0.75539, 0.26012 and 0.18735, the corresponding proportion of variance of each PC is shown in Table 3.

Table 3. The proportion of variance of principal components

Eigen Value	Variance proportion	Cumulative variance proportion
0.78220	0.39404	0.39404
0.75539	0.38053	0.77457
0.26012	0.13104	0.90561
0.18735	0.09438	0.99999

The variance of first PC (PC1) is 0.78220 which explains 39.40% of total variance, the variance of the second PC (PC2) is 0.75539 and explains 38.05% of total variance. Thus, PC1 and PC2 are deemed sufficient to represent the data structure with a cumulative variance of 77.45%. However, this research requires 100% information to represent the data structure in the next analysis so that the results obtained are optimal. So this study uses all principal component scores, namely the PC1-PC4 score. The following is the plotting of the principal component scores (Figure 2) to see the spread of the data based on the first and the second principal component scores. Based on the spread of the data we suspected that they are divided into three clusters.

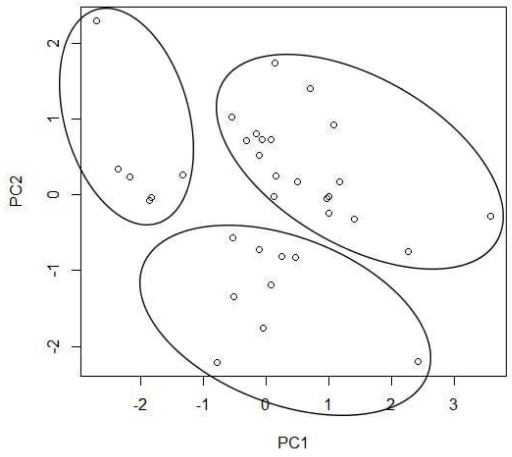

Figure 2. Principal Component Plot

Furthermore, for cluster analysis, this study also uses a robust method, namely trimmed clustering. The robust cluster analysis used is based on the principal component scores that have been obtained. According Md.Jedi & Adnan [17], most complex problem when applying non-hierarchical cluster analysis is to choose the number of clusters, *k*. It is certain that we must choose the initial number of cluster, but we did not really know what the best number of clusters that is supposed to be in the data. The same principal also applies to the trimming size, where we did not know exactly the true outlying level. Garcia-Escudero *et.al.* [18] introduced some classification trimmed likelihood curves as useful curve for choosing the number of clusters *k*. The *k-th* trimmed likelihood function is defined as:

$$\alpha \mapsto \ell_c^{\Pi}(\alpha, k) \text{ for } \alpha \in [0,1)$$

with $\ell_c^{\Pi}(\alpha, k) = \sum_{j=1}^{k} \sum_{i \in R_j} \log f(x_i; \mu_j, \Sigma_j)$. This curve function is allowed to measure $\Delta_c^{\Pi}(\alpha, k) = \ell_c^{\Pi}(\alpha, k+1) - \ell_c^{\Pi}(\alpha, k)$ where $\Delta_c^{\Pi}(\alpha, k)$ shouls be close to 0. Figure 3 shows the classification trimmed likelihood curve $\Delta_c^{\Pi}(\alpha, k)$ when k=1,2,3,4 and α range is [0, 0.2] and c=50. Because in Figure 3 it can be seen that no significant increase occurs when increasing k from 3 to 4 and it is supported by the results of the principal component plot in Figure 2, then for this data case, the optimal number of clusters will be selected as 3 clusters with α = 0.05, which means 5 % of trimmed data is data that is not part of the cluster formed.

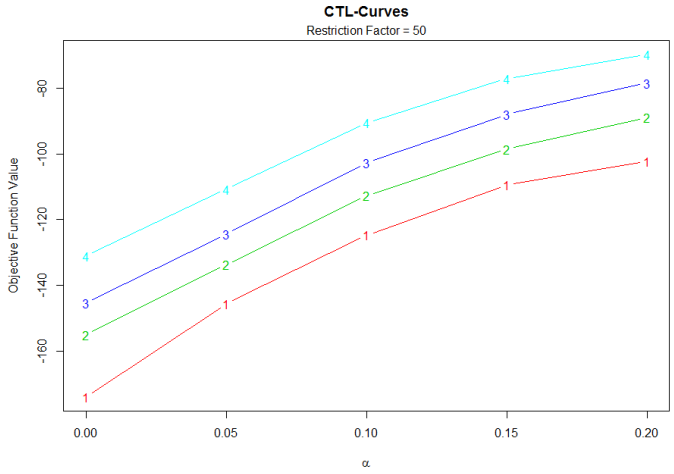

Figure 3. Classification trimmed likelihood curves $\ell_c^{\Pi}(\alpha, k)$ when k=1,2,3,4 and α range in [0, 0.2] and c=50

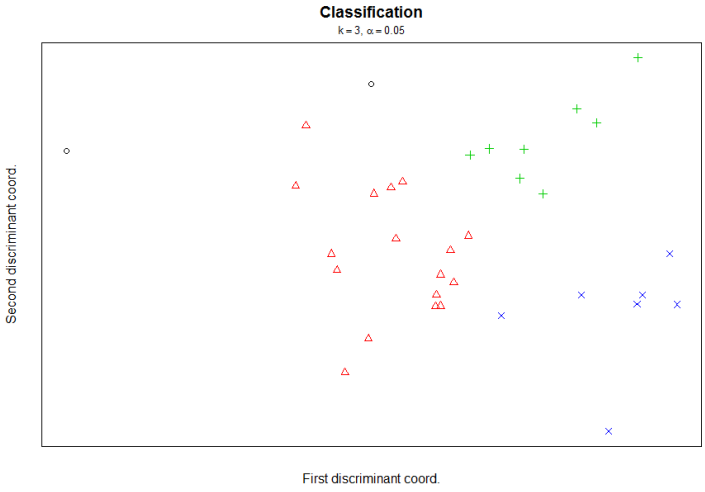

Figure 4. Classification Trimmed Plot

After trimmed cluster analysis on the principal component scores, two outliers namely DKI Jakarta and D.I Yogyakarta, were considered to have their own clusters. While the members of the three clusters are presented in Table 4.

Table 4. Cluster Members

Cluster	Provinces
1	Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Island, Riau Island, West Java, Central Java, East Java, Banten, Bali, Central Kalimantan, South

	Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, South Sulawesi
2	Aceh, North Sumatra, West Sumatra, Bengkulu, Central Sulawesi, Southeast Sulawesi, Maluku, North Maluku
3	NTB, NTT, West Kalimantan, Gorontalo, West Sulawesi, West Papua, Papua

Table 5. Cluster Centre ($\bar{X}_{(x)k}$)

Cluster	LE	EYS	AYS	ASC
1	71.178	12.825	8.568	11656.1
2	68.817	13.664	9.094	9727.6
3	66.856	12.665	7.377	8890.7

Cluster 1 is a group of provinces with life expectancy of birth in 2019 (LE) of approximately 71.2 years, the expected years of schooling of 7-year-old children (EYS) of approximately 12.8 years, mean years of schooling of the population aged 25 years and over (AYS) of 8.6 years, and average spending per capita adjusted at the provincial level (ASC) of IDR 11656.1. Cluster 2 is a group of provinces with LE approximately of 68.8 years, EYS of approximately 13.6 years, AYS of about 9 years and ASC of IDR 9727.6. Cluster 3 is a group of provinces with LE of about 66.8 years, EYS of about 12.6 years, AYS of about 7.4 years and ASC of IDR 8890.7.

To determine the variables that must be paid more attention to in each cluster so that the HDI achievement in the following year can be increased, it is necessary to carry out an analysis of the description in each cluster by calculating the average value of each indicator variable forming the 2019 HDI, namely LE ($\bar{X}_{LE} = 69.64$), EYS ($\bar{X}_{EYS} = 13.01$), AYS ($\bar{X}_{AYS} = 8.44$), ASC ($\bar{X}_{ASC} = 10569$). Then these values are compared with the cluster center ($\bar{X}_{(x)k}$) shown in Table 5.

If we get $\bar{X}_{(x)k} \leq \bar{X}_x$ where *x* is the observed variables (LE, EYS, AYS, ASC), it can be interpreted that the average value of the variables in the cluster is low or classified as "Bad" so that the observed variable must be further increased in each province. Conversely, if we get $\bar{X}_{(x)k} \geq \bar{X}_x$ then the average value of the variables in the cluster can be said as "Good" in the HDI forming indicators. Details are presented in the following table:

Table 6. Cluster Characteristics

Cluster	LE	EYS	AYS	ASC
1	Good	Bad	Good	Good
2	Bad	Good	Good	Bad
3	Bad	Bad	Bad	Bad

Table 6 shows that the members in Cluster 1 have fairly good HDI indicator characteristics on life expectancy of birth in 2019, mean years of schooling of the population aged 25 years and over, and average spending per capita adjusted at the provincial level, and obtained a low average score for expected years of schooling of 7-year- old children.

Cluster 2 members have fairly good HDI indicator characteristics on expected years of schooling of 7-year- old children, and mean years of schooling of the population aged 25 years and over, and obtained a low average score for life expectancy of birth in 2019 and average spending per capita adjusted at the provincial level.

Furthermore, members in cluster 3 are provinces that have very low HDI indicator characteristics. This is because the average of observed variables LE, EYS, AYS, ASC in the cluster center less than the actual average value. Thus, the provinces in Cluster 3 need more attention so that life expectancy of babies born in 2019, expected years of schooling of 7-year-old children, mean years of schooling of the population aged 25 years and over, and average spending per capita adjusted at the provincial level can be increased.

4. Conclusion

In this paper, we applied the robust principal component trimmed clustering for grouping Indonesian provinces based on HDI indicators. Based on the cluster analysis on the data using robust principal component trimmed clustering, the optimum number of clusters for the Indonesian provinces based on HDI indicators is three clusters.

5. References

[1] United Nations Development Programme (UNDP) 1990 *Human Development Report* (New York: UNDP)
[2] Badan Pusat Statistik (BPS) 2020 *Indeks Pembangunan Manusia 2019* (Jakarta: BPS)
[3] Härdle W K and Simar L 2019 *Applied Multivariate Statistical Analysis 5th Edition.* (New York: Springer)
[4] Rahman A S and Rahman A 2020 Application of principal component analysis and cluster analysis in regional flood frequency analysis: a case study in New South Wales, Australia. *Water* **12** 781
[5] Koij F S and Saba J 2015 Using cluster analysis and principal component analysis to group lines and determine important traits in white bean *Procedia Environ. Sci.* **29** 38 – 40
[6] Penkova T G 2017 Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety. *Procedia Comput. Sci.* **112** 99–108
[7] Suzana M, Zulkifli Y, Marhalil M, Rajanaidu N and Ong-Abdullah M 2020 Principal component and cluster analyses on Tanzania oil palm *Elaeis Guineensis* Jacq. germplasm. *J. Oil Palm Res.* **32** (1) 24-33
[8] Johnson R A and Wichern D W 2018 *Applied Multivariate Statistical Analysis 6th Edition* (New Jersey: Pearson Prentice Hall)
[9] Herawati N and Nisa 2017 A robust procedure for GEE model *Far East J. Math.Sci.***102** 645-654
[10] Nisa K and Herawati N 2017 Robust generalized estimating equation when data contain outliers *Int. S. Interdisc. Sci. Tech. INSIST* **02** (01) 1-5
[11] Olive J O 2017 *Robust Multivariate Analysis* 1st Edition (New York: Springer)
[12] Ortner I Filzmoser P and Croux C 2020 Robust and sparse multigroup classification by the optimal scoring approach *Data Min. Knowl. Dis.***34** 723-741
[13] Nisa K, Herawati N, Setiawan E and Nusyirwan 2006 Robust principal component analysis using the minimum covariance determinant estimator *Proc. ICMNS* 789-792
[14] Rousseeuw P J and Van Driessen K 1999 A fast algorithm for the minimum covariance determinant estimator *Technometrics* **41** (3) 212-223
[15] García-Escudero L A, Gordaliza A Matrán C and Mayo-Iskar A 2011 Exporing the number of groups in robust model-based clustering *Preprint* available at http://www.eio.uva.es/ infor/personas/langel.html
[16] Gallegos M T and Ritter G 2005 A robust method for cluster analysis *Ann. Stat.* **33** (1) 347–380
[17] Md.Jedi M A and Adnan R 2012 TCLUST: Trimming approach of robust clustering method *Mal. J. Fund. Appl. Sci* **8** (5) 253-258
[18] García-Escudero L A, Gordaliza A Matrán C and Mayo-Iskar A 2008 A general trimming approach to robust cluster analysis *Ann.Stat.* **36** (3) 1324-1345
[19] Akram M, Habib A and Alcantud J C R 2020 An optimization study based on Dijkstra algorithm for a network with trapezoidal picture fuzzy numbers. *Neural Comput. Applic.* https://doi.org/10.1007/s00521-020-05034-y
[20] Mukhlif F and Saif A 2020 Comparative study on Bellman-Ford and Dijkstra algorithms, *Int. Conf. Comm. Electric Comp. Net.* (ICCECN 2020) Kuala Lumpur Malaysia 1 – 2 February 2020
[21] Gbadamosi O A and Aremu D R 2020 Design of a modified Dijkstra's algorithm for finding alternate routes for shortest-path problems with huge costs *Int. Conf. Math. Comp. Eng. Comp. Sci.* (ICMCECS) Lagos Nigeria 1-6 doi: 10.1109/ICMCECS47690.2020. 240873