Eduvest – Journal of Universal Studies
Volume 3 Number 2, February 2023
p- ISSN 2775-3735- e-ISSN 2775-3727

APPLICATION OF BLOCK-BASED K-MEDOIDS AND WARD'S METHOD TO CLASSIFY PROVINCES IN INDONESIA BASED ON ENVIRONMENTAL QUALITY INDEX

Kariyam[1,2], Abdurakhman[2], Adhitya Ronnie Effendie[2]
[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
Email: kariyam@uii.ac.id, rachmanstat@ugm.ac.id, adhityaronnie@ugm.ac.id

ABSTRACT

Environmental quality is an essential aspect of the world's sustainability. This paper presents a combination of several methods to obtain a profile of provinces in Indonesia according to the environmental quality indices. We apply the block-based k-medoids algorithm, Ward's method, and the Hartigan index to determine the number of clusters. Based on Hartigan index suggested classifying 34 provinces in Indonesia into four groups. A comparison of the mean vector of four groups using multivariate analysis of variance concluded that three areas have the best environment quality index and eight regions require serious attention to improve their environmental quality methods.

KEYWORDS *block-based k-medoids, Ward, environmental, clustering*

INTRODUCTION

The Environmental Performance Index (EPI) is a numerical and quantitative-based method for measuring the environmental performance of a country's policies. The 2022 EPI use forty performance indicators and ranks 180 countries on their national efforts to protect environmental health, enhance ecosystem vitality, and mitigate climate change. These indicators measure how closely governments meet internationally established sustainability targets for specific environmental issues

How to cite: Kariyam, Abdurakhman, Adhitya Ronnie Effendie (2023). Application of Block-Based K-Medoids and Ward's Method To Classify Provinces In Indonesia Based On Environmental Quality Index. Journal Eduvest. *Vol 3* (2): Page 374-384
E-ISSN: 2775-3727
Published by: https://greenpublisher.id/

(Wolf et al., 2022). Based on the 2022 EPI ranking, Indonesia is 164th out of 180 countries. Meanwhile, Indonesia ranks 22 out of 25 countries in the Asia Pacific region. This rating is certainly not encouraging. Indonesia's ecosystem vitality indicator strongly correlates with the EPI score.

Environmental quality management is crucial for the sustainable development and preservation of natural resources in a region, such as a province in Indonesia. Effective management of the environment ensures clean air and water, wildlife and habitat protection, and pollution and waste reduction. Additionally, it supports public health, economic growth, and tourism and helps mitigate the effects of climate change. By implementing sound environmental practices, a province in Indonesia can secure a better future for its citizens and the planet. The government is responsible for implementing policies and regulations that protect the environment and ensure its preservation for future generations. Effective environmental quality management is essential for Indonesia's livable and thriving province.

The Ministry of Environment and Forestry in Indonesia uses the Environmental Quality Index (EQI) as an indicator of environmental management performance. Three aspects used to measure EQI are Air Quality Index (AQI), Water Quality Index (WQI), and Land Cover Quality Index (LCQI) (Zhang, 2018; Inhaber, 1976; Wear et al., 1998; Berlemann, 2013; and Zhang, 2015). EQI is used as information material to support policy-making processes related to environmental protection and management. Through EQI, we can find out how far the condition and status of the environmental quality of a region are in terms of water quality, air quality and land cover quality. Indonesia is an archipelagic country consisting of 34 provinces. The Ministry of Environment and Forestry in Indonesia conducted a separate analysis of these three indicators (MEF RI, 2019). The grouping of the province based on the environmental quality index can be carried out simultaneously to facilitate the monitoring of environmental quality.

Clustering provinces in Indonesia based on EQI can bring several benefits, such as the government can better monitor and manage the different regions, ensuring the implementation of appropriate policies and programs for each group. Based on the EQI, authorities can prioritize and allocate resources to areas that require the most attention, leading to effective and efficient use of resources. The environmental quality index provides a transparent and objective measure of environmental performance, which can hold governments accountable for their actions and decisions. The EQI also provide a comprehensive picture of the state of the environment, which can inform decision-making and policy development at the local, regional, and national levels.

A block-based k-medoids (Block-KM) algorithm is one of the center-based clustering methods (Kariyam et al., 2022). As distance variance block KM (Var-KM) (Kariyam et al., 2022), a Block-KM is a variant of flexible k-medoids (FKM) (Kariyam et al., 2022), which consist of two phases. The first stage of Block-KM is similar to FKM. This process guarantees no empty cluster and some identical objects in the same group. While the second phase of Block-KM is similar to Var-KM. The second stage of Block-KM uses initialization as a basis to update final medoids only once from the initial groups, so this method is more efficient than

FKM. In comparison, Ward's method is one of the popular methods in hierarchical clustering. Ward's method minimizes the total variance in the distance between newly formed clusters. It works by iteratively combining the two closest clusters into a larger cluster and updating the distances between all other clusters to reflect this change.

This study aims to classify provinces in Indonesia according to the environmental quality index simultaneously. We use Block-KM and Ward's method to cluster it. Then we identify groups of regions with the best environmental quality index and regional classes that require serious handling to improve environmental quality. To achieve these objectives, we combined a clustering method and multivariate analysis of variance. By comparing and benchmarking against other group provinces, regions with lower environmental quality scores can be motivated to improve their performance and adopt best practices.

RESEARCH METHOD

Materials

In this study, we use secondary data on the 2019 EQI from the Center for Data and Information published by the Ministry of Environment and Forestry of the Republic of Indonesia (MEF RI, 2019). The Environmental Quality Index is a generalization of the EQI of all districts/cities and provinces in Indonesia. The EQI consists of three indicators: the water, air, and land cover quality indices. The weights of the three indicators of EQI are determined by:

$$EQI = (30\% \,.\, WQI\,) + (30\% \,.\, AQI) + (40\% \,.\, LCQI), \qquad (1)$$

where WQI is the water quality index, AQI is the air quality index, and LCQI is the land cover quality index.

Water quality indicators are calculated based on seven parameters, namely Total Suspended Solid (TSS), Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Fosfat, Fecal Coli, and Total Coliform (MEF RI, 2019; Inhaber, 1976; Berlemann, 2013). The water quality index (WQI) calculation uses the pollutant index method with the concept that the higher the pollutant index value, the worse the water quality. This method can determine the status of the WQI monitored against water quality standards with one data series, so it only requires a little cost and time. The quality standard used in the analysis of the pollutant index is the classification of class II water quality standards based on Government Regulation 82 of 2001 (IGR RI, 2001). Furthermore, the WQI value of each province was calculated from the WQI average of all samples in that province.

Air quality index (AQI) has two parameters, namely Sulfur Dioxide (SO2) and Nitrogen Dioxide (NO2) (MEF RI, 2019; Zang et al., 2018). The AQI value is calculated based on ambient air quality measurements in the district/city. The sampling locations represent industrial, residential transportation, and office areas using the manual passive sampler method. Meanwhile, the land cover quality index is measured based on the land cover area. The LCQI was calculated by comparing the forest and administrative area (DIC, 2019; Wear et al., 1998; Zhang et al., 2015).

Based on the Law of the Republic of Indonesia number 41 of 1999, each province has a minimum forest area of about thirty per cent of the total area (IGRL RI, 1999). The LCQI calculation assumes that regions with a forest area of thirty per cent of their administrative area are assigned a value of 50. At the same time, those with the highest LCQI value (100) are areas that have an area of 84.3 per cent of their administrative scope. The overall data in this study consists of 34 objects (provinces), each of which observes three environmental quality indicators.

Methods

In this study, we compare the block-based k-medoids algorithm (Block-KM) and Ward's method. We use Hartigan index to obtain the optimal number of clusters. Then we describe profile of each group based on the Multivariate Analysis of Variance (MANOVA).

A Block-based K-Medoids Partitioning Methods

Kariyam et al. 2022 have proposed a block-based k-medoids (Block-KM) algorithm to partition a dataset. The outline of the combined two stages in the Block-KM algorithm is as follows,

(i) For each object, i, $(i = 1, 2, \cdots, n)$, calculated two indicators, namely, (i) sum up of values on p variables,

$$w_i = \sum_{l=1}^{p} x_{il}, \tag{2}$$

where x_{il} can be either standardized or unstandardized data, and (ii) standard deviation on p variables,

$$u_i = \sqrt{\frac{\sum_{l=1}^{p}(x_{il} - \bar{x}_i)^2}{p-1}}, \tag{3}$$

where $\bar{x}_i = w_i/p$; $i = 1, 2, \cdots, n$; $l = 1, 2, \cdots, p$.

(ii) Arrange object i, $(i = 1, 2, \cdots, n)$, based on the standard deviation (u_i) and the sum (w_i) in ascending order. Select the initial medoids, namely, representative object of the first k blocks of a combination of u_i and w_i. Use it to get the initial groups.

(iii) For each cluster, update the current medoid based on the object that minimizes the average distance. Suppose, n_g is the number of group members gth, then the average distance for object ith, $\overline{D}_i$, to other things in its group, such as follows,

$$\overline{D}_i = \frac{1}{n_g} \sum_{j=1}^{n_g} d_{ij}, \tag{4}$$

where d_{ij} is distance of two objects i and j. In this paper, we implement the Euclidean distance.

(iv) Assigning each object to the nearest medoid and calculating the sum of the total distance within a cluster, $TD(k)$, as follows,

$$TD(k) = \sum_{g=1}^{k} \sum_{i=1}^{n_g} \sum_{l=1}^{p} \left|x_{gil} - m_{gl}\right| \quad (5)$$

where x_{gil} is object i-th for variable l-th in the cluster g-th; and m_{gl} is medoid cluster g-th for variable l-th.

(v) Repeat steps (iii) and (iv) until a pre-determined number of iterations is met.

Ward's Hierarchical Clustering Method

Ward's hierarchical clustering method considers minimizing the 'loss' of information from joining two groups. This method applies to any data, especially to small data. Ward's method is usually implemented with loss of information taken as an increase in an error sum of square criterion, ESS. First, for a given cluster, g, let $\mathrm{ESS_g}$ be the sum of the squared deviations of every item in the cluster mean (centroid). If there are currently k clusters, define ESS as the sum of the $\mathrm{ESS_g}$ or $\mathrm{ESS} = \mathrm{ESS_1} + \mathrm{ESS_2} + \ldots + \mathrm{ESS_k}$. At each step in the analysis, the union of every possible pair of clusters is considered, and the two groups whose combination results in the smallest increase in ESS are joined. Initially, each cluster consists of a single item, and if there are n items, $\mathrm{ESS_g} = 0, \mathrm{k} = 1, 2, \ldots, \mathrm{n}$ so $\mathrm{ESS} = 0$. At the other extreme, when all the clusters are combined in a single group, the value of ESS is given by (Kaufman and Rousseeuw, 1990; Everitt, 2011),

$$\mathrm{ESS} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^{\mathrm{t}} (\mathrm{x_i} - \bar{\mathrm{x}}), \quad (6)$$

where $\mathrm{x_i}$ is the multivariate measurement associated with the ith and $\bar{\mathrm{x}}$ is the mean of all objects. The results of Ward's method can be displayed as a dendrogram. The vertical axis is given the value of ESS at which the mergers occur.

Hartigan's Index

Hartigan, 1975 proposed an index to determine the number of clusters (also known as the Hartigan index). This index is calculated based on the trace of homogeneity within groups, $tr(W_k)$, such as follows,

$$H(k) = \left(\frac{tr(W_k)}{tr(W_{k+1})} - 1\right)(n - k - 1). \quad (7)$$

The number of clusters is the smallest of k $(k \geq 1)$, so Hartigan's index is less than ten $(H(k) \leq 10)$. In this case, W_k is the matrix of the squared deviation of all objects to the centroid (mean), which is defined as follows,

$$W_k = \sum_{g=1}^{k} \sum_{i=1}^{n_g} \left(\underline{x}_{gi} - \underline{\bar{x}}_g\right)^t \left(\underline{x}_{gi} - \underline{\bar{x}}_g\right), \quad (8)$$

where:

$\boldsymbol{p}$ = the number of variables

n_g = the number of group gth $(g = 1,2,\ldots k)$

x_{g_i} = object ith, $(i = 1,2,\ldots,n_g)$, on the group gth

$\bar{x}_g$ = average of variable in the group g

Multivariate Analysis of Variance

Multivariate analysis of variance (MANOVA) is a statistical technique used to analyze the relationship between two or more dependent variables and one or more independent variables (Johnson & Wichern, 2009). It is an extension of the Analysis of Variance (ANOVA). MANOVA tests the hypothesis that the means of two or more dependent variables are equal across levels of tne or more independent variables.

In this paper, we use MANOVA to compare several mean vectors arranged according to some (more than two) groups. It provides a way to evaluate the overall effect of the independent variable on all the groups simultaneously rather than evaluating the effect on each dependent variable individually.

RESULT AND DISCUSSION

The data used to group provinces in Indonesia are a quality index of water, air, and land cover. The data standardized to a hundred, where this number indicates the maximum value of the environmental quality index. The clustering methods are the Ward and block-based k-medoids partitioning methods. Implementing the Hartigan Index to these data concludes that the number of groups is four. This amount is based on the smallest group resulting in a Hartigan Index below ten, as shown in Figure 3.

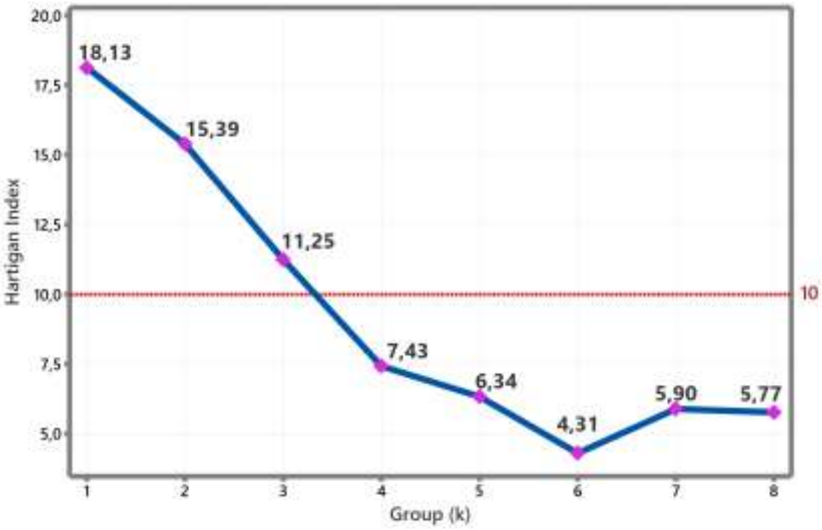

Figure 3 The plot of the Hartigan Index on Several Group Sizes

The dendrogram plot in Figure 4 was carried out simultaneously using three environmental quality indicators and the Ward method. The first group consists of seven provinces, the second group eight provinces, the third group ten provinces, and nine provinces in the fourth group, as shown in Figure 4.

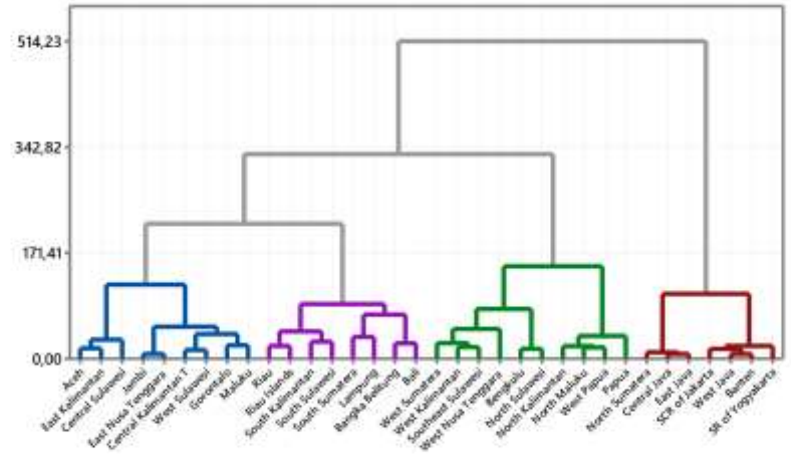

Figure 4 The Plot of Dendrogram with Ward method

The profile of each group is executed based on three indicators of the environment quality index, as shown in Fig. 5. The group with the best environmental quality consists of nine provinces marked in dark green. This group includes the provinces of Aceh, Jambi, East Nusa Tenggara, Central Kalimantan, East Kalimantan, Central Sulawesi, Gorontalo, West Sulawesi, and Maluku. Provinces with the worst environmental quality are in the first group, characterized by the lightest colour. This group's members are dominated by Java provinces, namely East Java, Central Java, West Java, Yogyakarta Special Region (SR), Special Capital Region (SRC) of Jakarta, Banten, and one province on the island of Sumatera, North Sumatera. Eight provinces with better environmental quality than those on the Java Islands, including Riau, Riau Islands, South Sulawesi, Lampung, Bangka Belitung, and Bali. Meanwhile, ten provinces, namely West Sumatera, West Kalimantan, Southeast Sulawesi, West Nusa Tenggara, Bengkulu, North Sulawesi, North Kalimantan, North Maluku, West Papua, and Papua, are the provinces with the best second environmental quality.

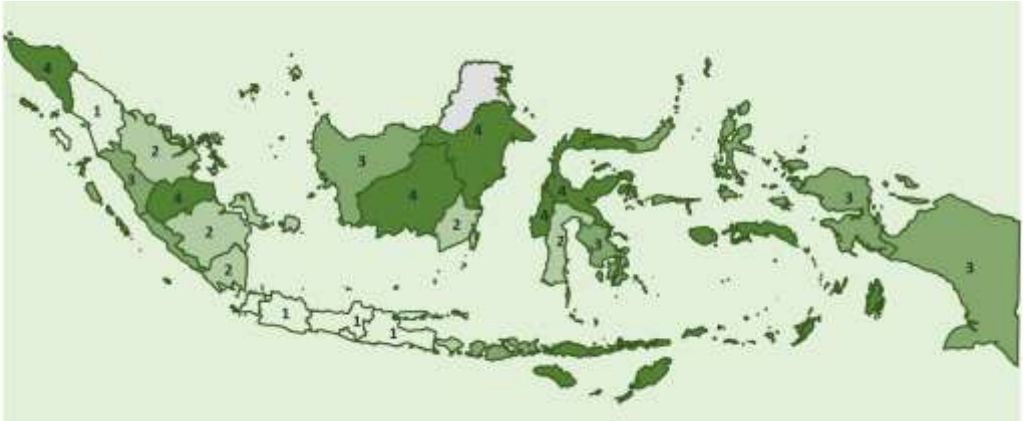

Figure 5 The profile of provinces in Indonesia based on the Environment Quality Index with Ward's method

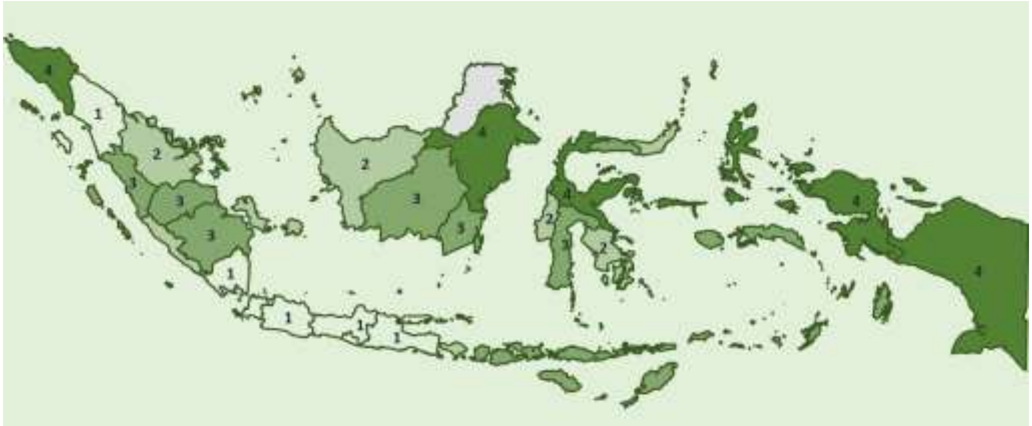

Figure 6 The profile of provinces in Indonesia based on the Environment Quality Index with Block-KM

The application of Block-KM for finding four groups of provinces in Indonesia based on the environmental quality index is as in Figure 6. The Block-KM method classifies the provinces of Aceh, East Kalimantan, North Kalimantan, Central Sulawesi, North Maluku, West Papua, and Papua in the same group (fourth group) with the best environmental quality. The seven provinces (first group) with the worst environmental quality according to the Block-KM method were the same as the results of the Ward method and added one region, namely Lampung. According to the Block-KM process, the ten provinces (third group) with the best second environmental quality are the Jambi, East Nusa Tenggara, South Sumatera, Central Kalimantan, South Kalimantan, South Sulawesi, Gorontalo, Maluku, West Nusa Tenggara, and West Sumatera. The other nine provinces fall into cluster two. There are 16 out of 34 or 47.1% of regions in the same quality order resulting from Ward's method and Block-KM algorithm.

To see the profile of each group by multiple comparison tests between groups produced by Ward's method, as shown in Table 1.

Table 1 The Multiple Comparison Test on the clustering result of Ward's Method

Indicators		Conclusion
1.	Air Quality Index	$\mu_1 < \mu_4 = \mu_2 = \mu_3$
2.	Water Quality Index	$\mu_1 = \mu_3 < \mu_4 = \mu_2$
3.	Land Cover Quality Index	$\mu_1 = \mu_2 < \mu_3 = \mu_4$

Using the 95% confidence level, the average air quality index of the seven provinces in the first group is statistically lower than the other groups. The average index is the same in the other three groups. The average water quality index for the first group is statistically the same as the average for group three and is more diminutive than for groups two and four. Meanwhile, the average land cover quality index of the first group was statistically the same as the second group and smaller than groups three and four.

The multiple comparison tests between groups produced by the Block-KM method are as in Table 2.

Table 2 The Multiple Comparison Test on the clustering result of the Block-KM Method

Indicators		Conclusion
1.	Air Quality Index	$\mu_1 < \mu_3 = \mu_2 = \mu_4$
2.	Water Quality Index	$\mu_1 < \mu_2 = \mu_4 < \mu_3$
3.	Land Cover Quality Index	$\mu_1 < \mu_2 = \mu_3 < \mu_4$

In descending order, statistically, the average air quality index in groups four, two and three have the same mean. The water quality index for group three has a higher average than the other groups. For the land cover quality index, group four has a higher mean than the other groups. The average for all indicators in group one is significantly lower than in other groups. These clustering results using the Block-KM method concluded that the eight provinces in the first group must seriously develop strategies to improve environmental quality.

CONCLUSION

In this paper, we have discussed the application of a block-based k-medoids algorithm and the Ward method to group provinces in Indonesia based on data from the water quality index, air quality index, and land cover quality index. Both ways produce four groups. Application MANOVA to compare the vector mean of the four groups concluded that the three areas with the best environmental quality index are Aceh, East Kalimantan, and Central Sulawesi provinces. The region that needs serious attention to improve the quality of their environment is Nort Sumatera, Lampung, Special Capital Region Jakarta, Banten, West Java, Special Region Yogyakarta, Central Java, and East Java.

REFERENCES

Berlemann, A. (2013). Using a water quality index to determine and compare creek water quality. Journal American Water Works Association. E291-E298. http://dx.doi.org/10.5942/jawwa.2013.105.0059

Data and Information Center, Secretariat General of the Ministry of Environment and Forestry, Indonesia. (2019). Indonesia's Environmental Quality Index 2019. Ministry of Environment and Forestry, Indonesia. https://www.menlhk.go.id/site/post/124

Everitt, B.S., Landau, S. Leese, M., and Stahl, D. (2011). Cluster analysis, 5th edn., John Wiley & Sons., Ltd., Publication. ISBN: 978-0-470-97844-3. pp. 49-50.

Hartigan, J. (1975). Clustering Algorithms. Wiley-Interscience, New York. ISBN: 0-471-35645-X. pp. 90-91.

Indonesian Government Regulation Law of the Republic of Indonesia number 41 of 1999. https://www.elaw.org/es/content/indonesia-number-41-1999-stipulation-act-forestry

Indonesian Government Regulation Number 82 of 2001. Water Quality Management and Water Pollution Control. https://www.informea.org/en/legislation/government-regulation-no-822001-management-water-quality-and-control-over-water

Inhaber, H. (1976). Environmental quality: outline for a national index for Canada. Ekistics, 41(243): 102-108. https://www.jstor.org/stable/43619060

Johnson, R.A., Wichern, D.W. (2009). Applied Multivariate Statistical Analysis, John Wiley & Sons. ISBN: 0-13-187715-1. pp. 301-302, 692-693.

Kariyam, Abdurakhman, Subanar, and Herni, U. (2022). The Initialization of Flexible K-Medoids Partitioning Methods Using a Combination of Deviation and Sum of Variable Values. Mathematics and Statistics, Volume 10 Number 5, pp: 895-908. https://doi.org10.13189/ms.2022.100501

Kariyam, Abdurkhman, and Effendie, A.R. (2022). The Use of Distance Blocks Representative To Avoid Empty Groups Due To Non-Unique Medoids, Eduves-Journal of Universal Studies, Volume 2 Number 10, pp: 2218-2228. https://doi.org/10.36418/eduvest.v2i10.625

Kariyam, Abdurakhman, Subanar, Utami, H., Effendie, A.R. (2022). Block-based K-medoids partitioning method with standardized data to improve clustering accuracy. Mathematical Modelling of Engineering Problems, Vol. 9, No. 6, pp. 1613-1621. https://doi.org/10.18280/mmep.090622

Kaufman, L., Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis, John Wiley & Sons, Inc., New YorkISBN: 0-471-73578-7. pp. 28-38, 68-71, 102-104.

Wear, D.N., Turner, M.G., R.J. Naiman. (1998). Land cover along an urban-rural gradient: implications for water quality. Ecological Applications, 8(3): 619-630. https://about.jstor.org/stable/2641254

Wolf, M.J., Emerson, J.W., Esty, D.C., de Sherbinin, A., Wendling, Z.A., et al. (2022). 2022 Environmental Performance Index, New Haven, CT: Yale Center for Environmental Law & Policy. epi.yale.edu. https://epi.yale.edu/downloads/epi2022report06062022.pdf

Zhang, T.Q., Zheng, Z.M., Lal, R., Lin, Z.Q., Sharpley, A.N., Shober, A.L., Smith, D., Tan, C.S., P. Van Cappellen (2018). Environmental Indicator Principium with Case References to Agricultural Soil, Water, and Air Quality and Model-Derived Indicators. Journal of Environmental Quality, 47: 191-202. https://doi.org/10.2134/jeq2017.10.0398

Zhang, Y., Liu, J., Wan, L., Qi, S. (2015). Land cover/use classification based on feature selection. Journal of Coastal Research, SI(73): 380-385. https://doi.org/10.2112/SI73-067.1