International Journal of Innovative Research and Scientific Studies, 6(2) 2023, pages: 242-249

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Environmental-based diseases of students in urban and rural areas, Lampung province, Indonesia

Subian Saidi[1], Netti Herawati[2*], Khoirin Nisa[3]

[1]*Mathematics and Natural Science, University of Lampung, 35141 Indonesia.*
[2,3]*Department of Mathematics, Faculty of Mathematics and Natural Science, University of Lampung, Indonesia.*

Corresponding author: Netti Herawati (*Email: netti.herawati@fimpa.unila.ac.id*)

Abstract

The aim of this study is to examine environmental-based diseases among students in Lampung Province, Indonesia. A total of 1,000 students from all districts in Lampung Province participated. Subsequently, the student disease distribution was grouped using the ROCK (Robust Clustering Using LinKs) method. The results indicated that many students were in a cluster often exposed to environmental-based diseases (cluster 4). The type of environmental-based disease that mostly affects the students in cluster 4 is skin disease while very few suffer from intestinal worms. The results of this study also demonstrated that students in rural areas suffer more from environmental-based diseases than those who live in urban areas. The majority of students in urban areas suffered from air-sensitive diseases (42%). Cluster 4 was the group of students who suffered from environmental-based diseases. Environmental-based diseases arise or recur due to environmental pollution such as acute respiratory infections (ARIs), diarrhea, intestinal worms, pulmonary tuberculosis and skin diseases. This may be due to the lack of awareness and knowledge of students in rural areas regarding the importance of protecting the environment in order to avoid environmental-based diseases.

Keywords: Cluster, Diseases, Pollution, ROCK method, Students.

DOI: 10.53894/ijirss.v6i2.1249
Funding: This study received no specific financial support.
History: Received: 10 November 2022/**Revised:** 30 December 2022/**Accepted:** 12 January 2023/**Published:** 24 January 2023

Authors' Contributions: All authors contributed equally to the conception and design of the study.
Competing Interests: The authors declare that they have no competing interests.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained.
Ethical Statement: This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Environmental issues have become a very interesting topic in modern times. At present, understanding the environment is considered crucial. Looking at various regions, many problems occur in the environment including the disposal of plastic waste and factory waste. Such habits threaten the survival of animals, plants and humans. The consequent impacts are various, one of which is environmental-based disease. Students at Hatay Mustafa Kemal University

have been surveyed and the results demonstrate that environmental problems were not deemed important enough by students at the university [1]. They also noted that environmental education was taught from pre-school to the highest level of education. Febriza, et al. [2] have explained that there exists a relationship between CLHB (clean and healthy living behaviors) and environmental-based diseases. In Lampung province, environmental issues must be considered. Therefore, we surveyed students in the province of Lampung on the topic of environmental-based diseases. To obtain the data, we used a proportional sampling technique, distributing the survey to 1000 samples in the province of Lampung which was considered representative of the population in the province of Lampung. Furthermore, the results may serve as a reference when developing policies to tackle environmental pollution.

Cluster analysis is a multivariate analysis technique for grouping observational data or variables into clusters. According to the factors used for clustering, each cluster is homogeneous [3]. In general, cluster analysis is only developed to process one type of data: categorical or numerical data. A problem occurs in cluster analysis with mixed-scale data. Data transformation is a method used to group mixed-scale data by transforming categorical data into numerical data and vice versa. The advantage of the transformation method is that it can reduce computational complexity. However, it is considered somewhat imprecise due to missing information [4].

Cluster analysis can be divided into two methods based on the determination of the number of clusters to be counted: hierarchical and non-hierarchical methods. A hierarchical method is used if the desired number of clusters is unknown while a non-hierarchical method is applied if the desired number of clusters has been determined [5]. The Robust Clustering Using Links (ROCK) method can be used for categorical data [6, 7]. The ROCK algorithm plays an important role in data mining, data analysis that can help to discover knowledge from large amounts of data. The ROCK algorithm was adapted from a hierarchical method by proposing the concept of neighbors and links in order to measure the similarity between data using links instead of distances when grouping. The general idea of the ROCK algorithm starts with entering data, taking random samples, grouping using links and labeling the data [8-10]. The grouping of data in this method uses the concept of distance as measured by the Jaccard coefficient in order to measure similarity. Research on the ROCK method has been carried out by several authors [11-13]. In this paper, we present a study based on an environmental-based disease survey including clustering of the data using the ROCK method. The results of the study are expected to be valuable to the provincial government as a reference for planning environmental-based disease problem management programs for each district.

2. Materials and Methods

2.1. Study Variables

A total of 1000 students from all districts and cities in Lampung Province, Indonesia were surveyed in the environmental-based diseases study in 2022. Questionnaires were distributed using a proportional sampling technique through junior and senior high schools in each district and city in Lampung Province. The completed questionnaires were collected in sealed envelopes. Previously, a preliminary survey was conducted on a small sample (n = 200) aimed at testing the feasibility of the survey questions and indicators using validity and reliability tests. The results of the two tests indicated that the survey questions and indicators are quite feasible for use in research but some questions were excluded from the study as they were not valid. The indicator variables are provided in Table 1.

Table 1.
Indicator variables.

Indicators	Variable
Have you ever had coughs and colds that lasted for approximately 14 days in the last 5 months?	X_1
Have you experienced an increased frequency of bowel movements and has the consistency of faeces been watery for less than 14 days in the last 5 months?	X_2
Have you had intestinal worms in the last 5 months?	X_3
Have you ever had a cough with phlegm for more than 2-3 weeks in the last 5 months?	X_4
Have you had any skin conditions in the last 5 months (Rashes, redness and itchy skin)?	X_5
Do you have a health condition that makes you more sensitive to poor air quality than other people?	X_6

Environmental-based disease clustering was carried out using the ROCK (Robust Clustering Using LinKs) algorithm based on the steps detailed in Section 2.2 with the help of the R software. As the output of this algorithm, the data are grouped into *k* clusters for each selected value where $0 < \theta < 1$. Performance measurement of the clustering results was determined by calculating the ratio of SW (standard deviation in groups) and SB (standard deviation between groups) values, the best clustering was selected with respect to the smallest ratio value. After determining the best *k* value, the survey data was grouped based on the obtained clusters. The collected data was then described and analyzed.

2.2. Cluster Analysis

Cluster analysis aims to group observational data or variables into clusters such that each cluster is homogeneous according to the factors used for clustering. To obtain a homogeneous cluster, the similarity of the analyzed scores is used as a basis. According to Bunkers, et al. [14], cluster analysis is a data technique for grouping a set of data into several clusters based on the similarity of the characteristics of the attributes possessed by object data such that data objects in the same cluster have similar characteristics to each other but are not similar to data objects in different clusters.

The dissimilarity between two objects i and j (d_{ij}) is a function that has the following properties: $d_{ij} \geq 0$, $d_{ii} = 0$, $d_{ij} = d_{ji}$, and $d_{ik} + d_{jk} \geq d_{ij}$, $\forall\, i, j,$ and k. The greater the value of the dissimilarity between two objects, the greater the difference between them and they will tend not to be in the same group [3]. According to Sudarmadi, et al. [15],one of the factors influencing the results of the cluster formed is the distance between the observed objects. The following methods for measuring the distance between objects are based on the characteristics of the grouped variables:

1. Distance Measurement Methods for Binary Categorical Variables

If the observed variable is a binary variable that has two different characters for example 0 and 1, then the observed variable can be formed into a contingency table as shown in Table 1. To calculate the size of the distance between variables x_i and x_j to measure binary data, the following values are presented in Table 2.

Table 2.
Binary data contingency.

Category x_i	**Category**		**Total**
	1	**0**	
1	A	b	a+b
0	C	d	c+d
Total	a+c	b+d	a+b+c+d

Then, the distance between x_i and x_j can be calculated using Jaccard's distance as follows:

$$JACCARD(x_i, x_j) = \frac{a}{a+b+c}.$$

2. Distance Measurement Method for Nominal Categorical Variables

Distance measurement for nominal variable data follows the same concept as the matching coefficient for dice where the categories can be more than two kinds. When the number of variables is m, the formula for measuring the nominal variable distance between x_i and x_j is:

$$sim(x_i, x_j) = \frac{1}{m}\sum_{l=1}^{m} S_{ijl},$$

Where $S_{ijl} = 1$ if $x_{il} = x_{jl}$ and $S_{ijl} = 0$ if $x_{il} \neq x_{jl}$.

3. Distance Measurement Method for Ordinal Categorical Variables

Measurement for ordinal variable data follows the same concept as for numerical data. One method that can be used for ordinal variables is the Manhattan distance. When the number of variables is m, the formula for measuring the distance between x_i and x_j (nominal variables) is:

$$sim(x_i, x_j) = \sum_{l=1}^{m} |x_{il} - x_{jl}|.$$

4. Distance Measurement Methods for Numerical Variables

For variables that have a numerical data type, the general distance used is the Euclidean distance. If there are two observations of m-dimensional variables, namely $x = [x_1, x_2, \ldots, x_m]^T$ and $y = [y_1, y_2, \ldots, y_m]^T$, the following formula is used to calculate the Euclidean distance between x and y:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2},$$
$$d(x, y) = \sqrt{(x - y)^T (x - y)}.$$

A cluster can be a "good cluster" if it has high homogeneity between members (within the cluster) and high heterogeneity when compared to another cluster (between clusters).

In general, cluster analysis only focuses on variables whose data types are numerical [16] but there are cases with categorical data and mixed numerical and categorical data. Cluster analysis of categorical data cannot be treated the same as numerical data. This is due to the special nature of categorical data which makes grouping categorical data more complicated than grouping numerical data [16].

Clustering of categorical data is carried out by using similarity or distance measures for categorical-scale data and can be accomplished using hierarchical or non-hierarchical methods. However, according to Guo, et al. [11], hierarchical and non-hierarchical clustering methods are considered inappropriate to use for categorical data. Therefore, the ROCK method was developed for clustering categorical data.

The ROCK algorithm uses the link concept as a measure of similarity to form clusters. It can handle outliers quite effectively: by trimming outliers, it makes it possible to remove unrelated data so that it does not participate in clustering. However, in some situations, outliers can form small clusters [11]. Clustering of categorical data using the ROCK algorithm is carried out in three steps as follows:

1. Calculating the similarity using the Jaccard formula [17]. The size of the similarity between the objects i and j is calculated as:

$$sim(X_i, X_j) = \left|\frac{X_i \cap X_j}{X_i \cup X_j}\right|, i \neq j.$$

2. Determining neighbors. Observations are determined to be neighbors if $sim(X_i, X_j) \geq \theta$.
3. Counting links between observation objects. The size of a link is influenced by the threshold (θ) value which is a user-defined parameter to control how close the relationship between objects is $(0 < \theta < 1)$. The size of a link is

calculated using the following formula: $link[C_i, C_j] = \sum_{X_i \in C_i, X_j \in C_j} link\,(X_i, X_j)$, which represents the number of links of all possible pairs of objects in C_i and C_j where n_i and n_j represents the number of members in groups i and j respectively and $f(\theta) = \frac{1-\theta}{1+\theta}$.

The ROCK method uses links as a measure of the similarity between objects. Let X_i, X_j, and X_k be the objects or observation. If X_i is a neighbor of X_j and X_j is a neighbor of X_k, then it is said that X_i has a link with X_k, even though X_i is not a neighbor of X_k. The links for all n possible pairs of objects can be calculated using a matrix of $n \times n$, where entry ij of the matrix has a value of 1 if X_i and X_j are similar and it has a value of 0 if X_i and X_j are not similar. The number of links between pairs X_i and X_j is obtained from the multiplication results between rows i and j of the matrix. If the number of links between X_i and X_j is high, then the possibility that X_i and X_j are in the same cluster will be greater [18]. Merging clusters with the ROCK algorithm is based on the goodness measure between clusters $g(C_i, C_j)$ which is the number of links divided by the possible links formed based on the cluster [19] calculated as follows:

$$g(C_i, C_j) = \frac{link\,[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \quad .$$

The performance measurement of grouping results (model testing) is a key step to determine the grouping validity. A good grouping has high homogeneity among members of a given group and high heterogeneity between groups [16]. According to Bunkers, et al. [14], the performance of grouping results for variables with numerical data scales can be determined through the ratio of SW and SB and by the sum of the total squares (SST) for categorical data. According to the above as well as the total number of squares in a group (SSW) and the sum of squares between groups (SSB), the following indices were calculated:

$$MSS = \frac{SST}{(n-1)}, ; \; MSW = \frac{SSW}{(n-C)}, \text{ ; and } MSB = \frac{SSB}{(C-1)}.$$

The standard deviation within groups (SW) and standard deviation between groups (SB) for categorical data can be formulated as follows:

$$S_W = [MSW]^{1/2} \text{ and } S_B = [MSB]^{1/2}.$$

With numerical data, the performance of a grouping method for categorical data is better if the ratio between SW and SB is smaller which indicates that there is maximum homogeneity within groups and maximum heterogeneity between groups [14].

3. Results and Discussion

When grouping categorical data using the ROCK method, the first thing to do is initialize each observation as a cluster with a single member in order to form a distance matrix between observations. The distances obtained from the 1000 observations were put into a matrix with dimensions of 1000×1000. Then, the value of the ratio for each θ was examined. Several values of θ were used namely: θ = 0.01, θ = 0.05, θ = 0.10, θ = 0.15, θ = 0.20, θ = 0.25, θ = 0.30 and θ = 0.35. The θ values were tested by adjusting the data and assessing the clustering results. The best clustering results were determined by the lowest ratio values of S_W and S_B. The lower the ratio, the better the grouping. The results for each θ are provided in Table 3.

Table 3.
The ratio of S_W to S_B for each considered θ.

$\boldsymbol{\Theta}$	Number of clusters	SW	SB	Ratio
0.01	3	0.455	11.209	0.041
0.05	3	0.449	11.086	0.041
0.10	2	0.845	38.523	0.022
0.15	2	0.845	38.523	0.022
0.20	2	0.845	38.523	0.021
0.25	4	0.625	18.975	0.032
0.30	4	0.766	20.667	0.037
0.35	5	0.875	19.299	0.045

Based on the table, it can be seen that when θ = 0.10, 0.15 and 0.20, there were only two clusters. However, clusters with more than two groups were considered. Based on the table above, θ = 0.25 was chosen as it led to more than two clusters and had a smaller ratio than the others. The ratios in Table 4 are presented graphically in Figure 1.

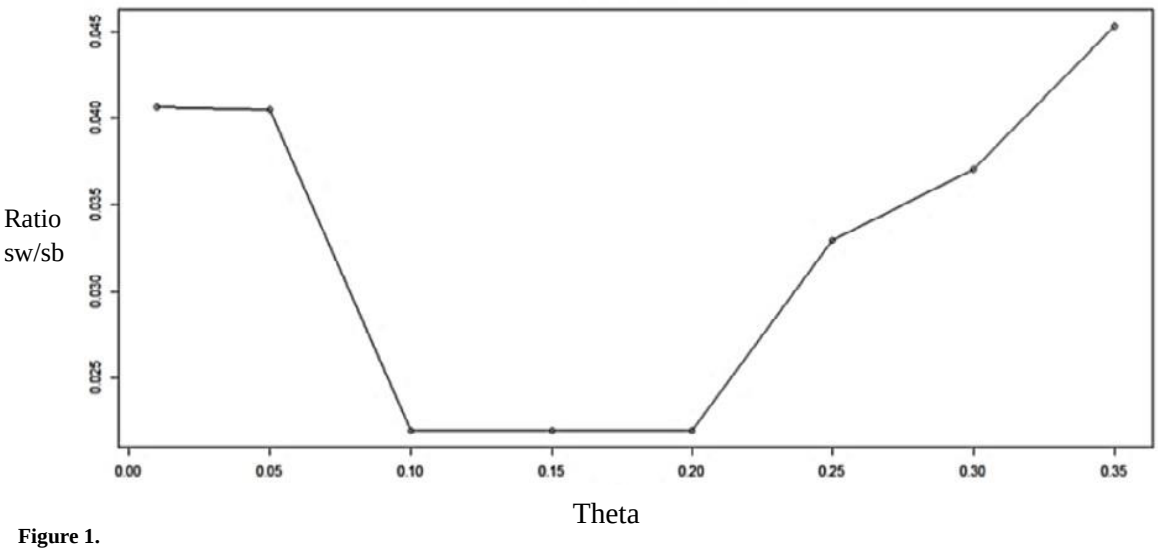

Figure 1.
The ratio of S_W to S_B.

Figure 1 shows that the smallest values obtained were θ = 0.10, 0.15 and θ = 0.20. However, it was necessary to check whether one of the three met the desired number of clusters. By using θ = 0.25, the data was divided into four data groups or clusters. The criteria for selecting this value were that the number of clusters should be greater than two but the ratio should also be minimized. The cases θ = 0.10, 0.15 and 0.20 indeed gave the smallest ratios but the number of clusters obtained with these values was not in accordance with our objectives. While θ = 0.01, 0.05, 0.30 and 0.35 provided the desired number of clusters, their ratio values were higher. Therefore, in the remaining case, θ = 0.25 gave the desired number of clusters (i.e., more than 2) and had the smallest ratio. Therefore, the optimal number of clusters based on θ = 0.25 was four.

Based on the optimum value of *k*, environmental-based disease grouping was carried out as shown in Figure 2. Figure 2 shows how the cases of environmental-based diseases (namely, ARI, diarrhea, worms, tuberculosis, skin diseases and air sensitivity) were divided into the four clusters. Cluster 1 was the group of people who had never experienced an environmental-based disease. Cluster 2 was a group of people who had suffered from some environmental-based diseases but most had not. Cluster 3 was a group of people with close ratios of patients to non-patients. Meanwhile, cluster 4 was the group of people who suffered mostly from environmental-based diseases. Each cluster based on environmental disease is shown in Figure 2.

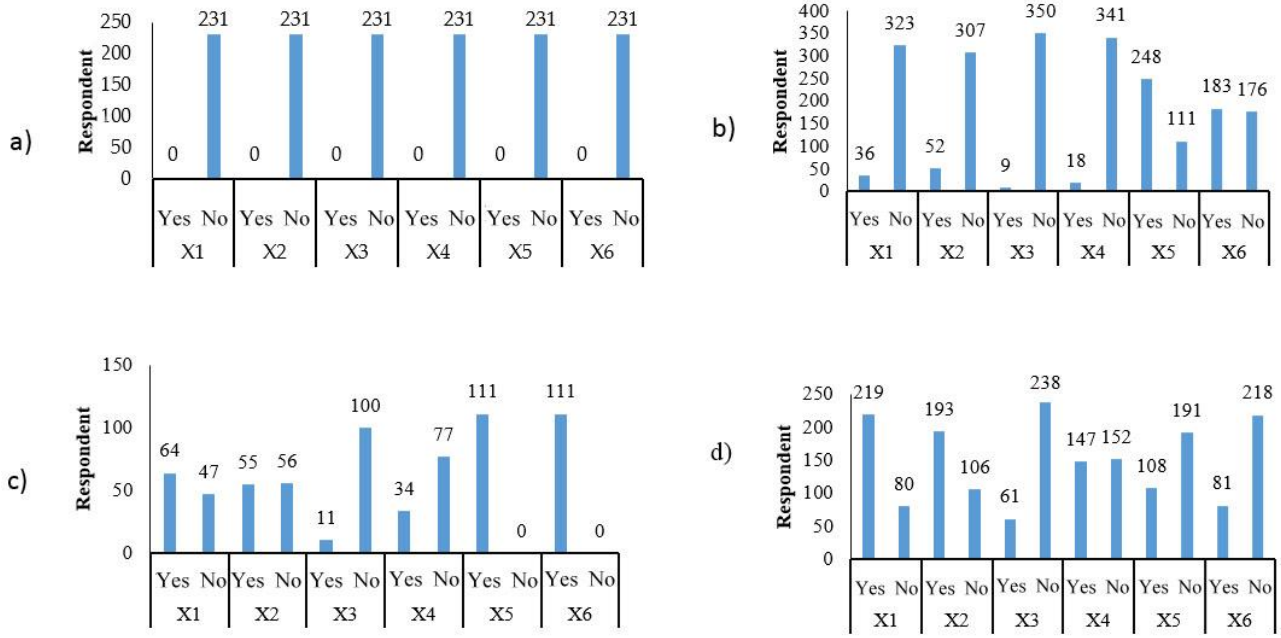

Figure 2.
Environmental-based disease characteristics of each cluster (a=Cluster 1, b= Cluster 2, c = Cluster 3, d = Cluster 4).
Note: X_1 = ARI; X_2 = Diarrhea; X_3 = Worms; X_4 = Tuberculosis; X_5 = Skin diseases; X_6 = Air sensitivity.

Figure 2 shows that the first cluster is a community cluster that did not suffer from any disease. This cluster consisted of the 23.1% of the 1000 people surveyed who had not a cough or cold, watery stools, helminthiasis, coughing with phlegm rashes, redness or itching of the skin. Thus, the first cluster can be said to be a healthy community cluster.

Cluster 2 consisted of 35.9% of the 1000 people surveyed having the largest number of participants. In cluster 2, there were 36 people who suffered from coughs and colds that lasted for approximately 14 days, compared to 323 who did not indicating that the majority of people in cluster 2 did not suffer from ARI. Meanwhile, there were 52 people who experienced an increased frequency of defecation and the consistency of loose or watery stools for approximately 14 days compared to 307 who did not. This indicates that only a small percentage of people in cluster 2 had diarrhea but the number of people suffering from diarrhea was higher than those with ARI. People suffering from intestinal worms were less as compared to other disease (only 9 people compared to 350 people who did not experience worms). As for pulmonary tuberculosis, there were 18 people who had experienced it compared to 341 people who had not (higher than those suffering from worms) but lower than those suffering from ARI and diarrhea.

The interesting thing in cluster 2 is the indication of skin disease and air sensitivity. The number of sufferers of these two diseases was higher than those who did not suffer. This was inversely proportional to the other four diseases in cluster 2. The number of people with skin diseases was 248 compared to 111 who did not suffer. They got rashes, redness and itching of the skin. This indicates the poor condition of the environment in which they live. Meanwhile, the number of sufferers who were sensitive to air was 183 compared to 176 who were not sensitive. The characteristics of the community in cluster 2 are people who are healthy. The majority of people in cluster 2 only suffer from mild diseases such as skin diseases. Cluster 3 had the smallest number of people with only 11.1% of the 1000 people surveyed. It can be seen that the majority of people were affected by ARI with a total of 64 compared to 47 who were not affected by the disease. The majority of people experienced cough and cold that lasted for approximately 14 days. Meanwhile, in terms of diarrheal disease, the number of those who suffered it and those who did not was balanced namely, 55 versus 56 respectively. For intestinal worms, there were 11 people who were affected compared to 100 people who were not affected. This indicates that the majority of people did not experience intestinal worms. The same thing occurred for pulmonary tuberculosis where the majority of people did not suffer from tuberculosis with a total of 77 people compared to 34 who had tuberculosis. This indicates that tuberculosis is not a common disease in this community. All people in cluster 3 suffered from skin diseases. Each individual had skin diseases. Thus, it can be concluded that cluster 3 includes a higher ratio of people suffering from environmental-based diseases compared to clusters 1 and 2.

Cluster 4 had the highest ratio of disease sufferers with 28.9% of the 1000 responses. A total of 219 of them suffered from ARI with symptoms of cough and runny nose lasting for approximately 14 days compared to 80 people who did not suffer from the disease. The same thing was observed for diarrheal disease where the majority experienced diarrheal disease (193) compared to those who did not (106). As for intestinal worms and pulmonary tuberculosis, the majority of people did not suffer from these two diseases. 289 people in cluster 4 did not suffer from these two diseases. Interesting results were obtained for skin diseases and sensitive to air. In the previous three clusters, these two diseases were linear or in line with the other results. In cluster 4, the majority of people did not experience these two diseases similarly to cluster one but rather opposite from clusters 2 and 3.

Further analysis was carried out to determine the distribution of students suffering from environmental-based diseases in each cluster. The results are provided in the following Table 4.

Table 4.
Percentage of students suffering from environmental-based diseases in each cluster.

Environmental-based diseases	Cluster			
	1	2	3	4
ARI	0%	11%	20%	69%
Diarrhea	0%	17%	18%	64%
Worms	0%	11%	14%	75%
Tuberculosis	0%	9%	17%	74%
Skin disease	0%	53%	24%	23%
Air sensitivity	0%	49%	30%	22%
Average	0%	25%	20%	54%

From the table, it can be seen that in cluster 1, all diseases had the same percentage (i.e., 0%). This indicates that the students in cluster 1 were a group of students who did not suffer from environmental-based diseases. Interestingly, in clusters 2, 3 and 4, all diseases presented different percentages. For ARI disease, students in cluster 2 presented 11% while cluster 3 had 20% and cluster 4 presented 69% as ARI patients. Thus, the majority of ARI patients were in cluster 4 with a very high percentage. In cluster 2, diarrheal disease was 17% which means that 17% of students rarely suffered from diarrheal disease. This percentage increased to 18% in cluster 3 while in cluster 4, the percentage of students was very high. 64% of students suffered from diarrheal diseases. Cases of diarrheal disease were almost the same as cases of ARI. In the case of intestinal worms in cluster 2, 11% of students suffered from intestinal worms. Meanwhile, in cluster 3, 14% of students often suffered from intestinal worms. In contrast, in cluster 4, 75% of students very often suffered from intestinal worms. For pulmonary tuberculosis disease, in clusters 2 and 3, 9% and 17%, had tuberculosis disease. Interestingly, in cluster 4, the percentage of students with tuberculosis disease was almost the same as those with worms where 74% of students very often suffered from tuberculosis disease. Notably, skin disease in cluster 2 had the largest

percentage among the other clusters with a percentage of 53%. In cluster 3, skin disease had 24% while in cluster 4 it had 23%. For air-sensitive diseases, cluster 2 again presented the largest percentage with 49% of people experiencing air sensitivity. In cluster 3, 30% of the respondents suffered from air sensitivity. Finally, in cluster 4, 22% of the community reported air-sensitive diseases. We also conducted an analysis of the distribution of environmental-based diseases among students living in urban and rural areas. Diseases in urban and rural areas are presented in the following figure:

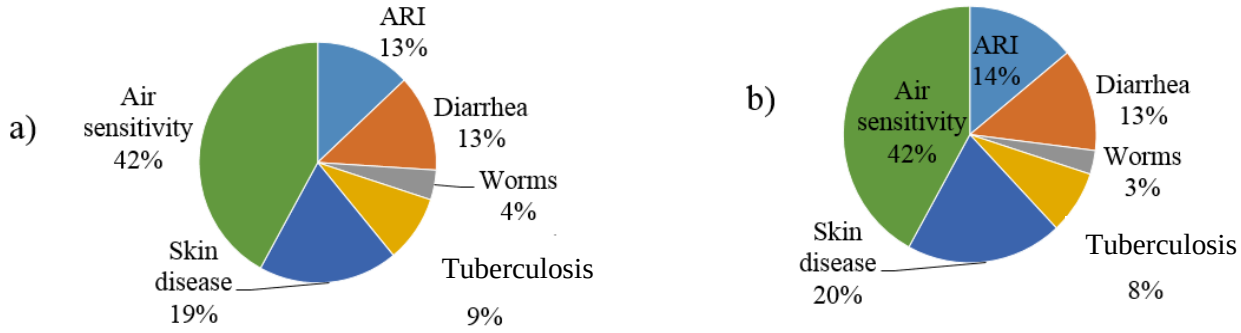

Figure 3.
Distribution of environmental-based diseases in urban areas (a) Distribution of environmental-based diseases in rural areas (b).

From Figure 3, we can see that the majority of students in urban areas suffered from air-sensitive diseases (42%) while the least-reported disease was intestinal worm disease (3%). Not much different from the result for urban students, the majority of students in rural areas also experienced air-sensitive diseases (42%). Similarly, the disease that was reported the least was intestinal worms (4%). The two figures abovementioned provide information that air-sensitive diseases are of particular concern not only in cities but also in villages. On the other hand, intestinal worms disease should not receive special attention as the number of sufferers of this disease is a very small percentage. A comparison of the percentage of students suffering from the diseases in urban and rural areas was also carried out. The results are presented in Table 5.

Table 5.
Percentage of students with environmental-based diseases in urban and rural areas.

Environmental-based diseases	Urban	Rural
ARI	45%	55%
Diarrhea	45%	55%
Worms	42%	58%
Tuberculosis	40%	60%
Skin disease	45%	55%
Air sensitivity	44%	56%
Average	44%	56%

Based on the table, it can be seen that the majority of ARI diseases occurred in rural areas with 55% compared to 45% in urban areas. The same phenomenon was observed for diarrheal diseases where students from rural areas were more likely to get diarrheal diseases than students from urban areas. As for helminthiasis, the percentage of students affected by helminthiasis was 58% in rural areas compared to 42% in urban areas again showing that rural areas are more prone to helminthiasis than urban areas. For tuberculosis, the percentage of students affected by the disease in rural areas was quite high (it was the largest) at 60% compared to 40% in urban areas. This shows that the hot rural areas are very prone to tuberculosis. Similarly, rural areas had the highest rates of skin disease and air sensitivity with percentages of 55% and 56% compared to 45% and 44% for urban areas. The overall comparison of the spread of disease in urban and rural areas is presented in Figure 4.

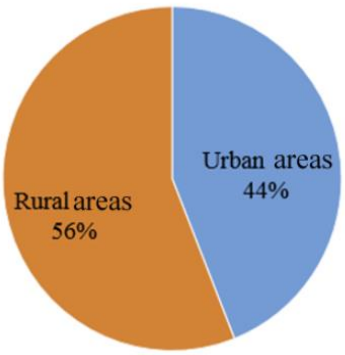

Figure 4.
Environmental-based disease distribution in urban and rural areas.

We found four clusters of students in Lampung Province based on the environmental-based diseases they had experienced using the ROCK method. Cluster 1 was a cluster of students who had never suffered from an environmental-based disease. In cluster 2, only a small number of students had suffered from environmental-based diseases. Cluster 3 is a balanced cluster where most students suffered from environmental-based diseases. Finally, in cluster 4, many students had suffered from environmental-based diseases. Through the clustering method based on the ROCK algorithm, we found that the spread of environmental-based diseases was dominated by students living in rural areas (56%). This is due to the lack of awareness and knowledge among students in rural areas regarding the importance of protecting the environment in order to avoid environmental-based diseases. This is in line with previous research [13, 15]. Future work will combine more variables and expand the data to determine the overall level of knowledge and awareness of the Indonesian population about environmental-based diseases.

4. Conclusion

The ROCK method was used to describe the distribution of environmental-based diseases among students in Lampung Province. We found four clusters among students based on their experience of environmental-based diseases in Lampung Province. Cluster 1 was a cluster of students who had never suffered from an environmental-based disease. In cluster 2 only a small number of students had suffered from environmental-based diseases. Cluster 3 is a balanced cluster where most students suffered from environmental-based diseases. Finally, in cluster 4, many students had suffered from environmental-based diseases. We also found that the spread of environmental-based diseases was dominated by students living in rural areas, 56% of diseases were experienced by rural students. This is assumed to be due to a lack of knowledge, perception and conscientiousness among students in rural areas regarding the importance of protecting the environment in order to avoid environmental-based diseases.

References

[1] E. Bozdogan, S. Sahinler, and E. Korkmaz, "Environmental awareness and attitudes in university students. an example from Hatay (Turkey)," *Oxidation Communications*, vol. 39, no. 1, pp. 661-672, 2016.

[2] N. Febriza, U. M. Tang, and E. Maryanti, "The Influence of Clean and Healthy Living Behavior (PHBS), Income and Sanitation on the Incidence of Diarrhea in Meranti Pandak Village, Rumbai Pesisir Pekanbaru," *Journal of Environmental Science*, vol. 9, no. 1, pp. 12-22, 2015.

[3] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, 6th ed. New Jersey: Pearson, Prentice Hall, 2007.

[4] R. R. Dewangan, L. K. Sharma, and A. K. Akasapu, "Fuzzy clustering technique for numerical and categorical dataset," *International Journal of Computational Science and Engineering*, pp. 75–80, 2010.

[5] M. d. P. Harb, L. Silva, T. Ayass, N. Vijaykumar, M. Silva, and C. R. Francês, "Dendrograms for clustering in multivariate analysis: Applications for COVID-19 vaccination infodemic data in Brazil," *Computation*, vol. 10, no. 9, p. 166, 2022, https://doi.org/10.3390/computation10090166.

[6] Y. Zhao, X. Liu, and W. Wang, "ROCK clustering algorithm based on the P system with active membranes," *WSEAS Transactions on Computers*, vol. 13, pp. 289-299, 2014.

[7] H. Sofyan, M. Iqbal, M. Marzuki, and M. Muhammad, "The comparison of k-modes clustering and ROCK clustering to the poverty indicator in Samadua Subdistrict , South Aceh," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1087, no. 012085, pp. 1-8.

[8] S. Alvionita and A. Suharsono, "Ensemble ROCK methods and ensemble SWFM methods for clustering of cross citrus accessions based on mixed numerical and categorical dataset," in *IOP Conference Series: Earth and Environmental Science*, 2017, vol. 58, no. 012029, pp. 1–10.

[9] M. A. Putri and S. Abdullah, "Clustering analysis of senior high school in West Java based on educational facilities," *In Journal of Physics: Conference Series. IOP Publishing*, vol. 1725, no. 1, p. 012032, 2021, https://doi.org/10.1088/1742-6596/1725/1/012032.

[10] P. M. Bhagat, P. S. Halgaonkar, and V. M. Wadhai, "Review of clustering algorithm for categorical data," *International Journal of Engineering and Advanced Technology*, vol. 3, no. 2, pp. 341-345, 2013.

[11] L. Guo, J. Liang, Y. Zhu, Y. Luo, L. Sun, and X. Zheng, "Collaborative filtering recommendation based on trust and emotion," *Journal of Intelligent Information Systems*, vol. 53, no. 1, pp. 113-135, 2019, https://doi.org/10.1007/s10844-018-0517-4.

[12] R. A. Wibowo, K. Nisa, H. Venelia, and W. Warsono, "Robust clustering of covid-19 pandemic worldwide," *BAREKENG: Journal of Mathematics and Applied Sciences*, vol. 16, no. 2, pp. 687-694, 2022, https://doi.org/10.30598/barekengvol16iss2pp687-694.

[13] W. Wang, Y. Liu, L. Zhang, L. Ran, S. Xiong, and X. Tan, "Associations between indoor environmental quality and infectious diseases knowledge, beliefs and practices of hotel workers in Wuhan, China," *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, p. 6367, 2021, https://doi.org/10.3390/ijerph18126367.

[14] M. J. Bunkers, J. R. Miller, and A. T. DeGaetano, "Definition of climate regions in the Northern Plains using an objective cluster modification technique," *Journal of Climate*, vol. 9, no. 1, pp. 130-146, 1996, https://doi.org/10.1175/1520-0442(1996)009%3C0130:docrit%3E2.0.co;2.

[15] S. Sudarmadi, S. Suzuki, T. Kawada, H. Netti, S. Soemantri, and A. Tri Tugaswati, "A survey of perception, knowledge, awareness, and attitude in regard to environmental problems in a sample of two different social groups in Jakarta, Indonesia," *Environment, Development and Sustainability*, vol. 3, no. 2, pp. 169-183, 2001, https://doi.org/10.29333/iji.2019.12323a.

[16] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis*, 7th ed. New Jersey: Prentice Hall, Inc, 2009.

[17] T. Alqurashi and W. Wang, "Clustering ensemble method," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 6, pp. 1227-1246, 2019.

[18] M. Dutta, A. K. Mahanta, and A. K. Pujari, "QROCK: A quick version of the ROCK algorithm for clustering of categorical data," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2364-2373, 2005, https://doi.org/10.1016/j.patrec.2005.04.008.

[19] A. Tyagi and S. Sharma, "Implementation of ROCK clustering algorithm for the optimization of query searching time," *International Journal on Computer Science and Engineering*, vol. 4, no. 5, p. 809, 2012.