Jurnal Varian
Vol. 5, No. 2, April 2022, Hal. 189–195
e-ISSN: 2581-2017

Algorithm K-Prototype for Clustering The Earthquake on Sulawesi Island

Suwardi Annas[1], Irwan[2], Rahmat H.S[3], Zulkifli Rais[4]
[1,3,4]Statistics Department, Universitas Negeri Makassar, Indonesia
[2]Mathematics Department, Universitas Negeri Makassar, Indonesia

Article Info

Article history:

Received : 04-30-2022
Revised : 04-28-2022
Accepted : 04-20-2022

Keywords:

Algorithm K-Prototype;
Clustering;
Earthquake;
Magnitude.

ABSTRACT

Natural disasters that had occurred in Indonesia consist of hydrometeorology: floods, droughts, and landslides, geophysical: volcanic earthquakes and volcanic eruptions, and biological: epidemics. Regarding the tectonic earthquake on Sulawesi Island, there are at least 2 earthquake disasters that became national disasters, namely in Central Sulawesi and West Sulawesi in the range of 2017 to 2021. This study aims to cluster tectonic earthquakes on Sulawesi Island, from 2017 to 2020, as the basis for formulating disaster mitigation plans. This study used tectonic earthquake data from 2017 to 2020 obtained from BMKG Gowa, Indonesia. The variables used are magnitude, depth, and distance category. Because they are mixed variables, this study used a k-prototype algorithm. There are four clusters in 2017, six clusters in 2018, five clusters in 2019, and six clusters in 2020 based results cluster on a ratio of within-cluster distance against between-cluster distance. It can be related to the active fault on Sulawesi Island. The characteristics of clusters form each year are the greater magnitude.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v5i2.1908

Corresponding Author:

Suwardi Annas
Department of Statistics, Universitas Negeri Makassar.
Email: suwardi_annas@unm.ac.id

A. INTRODUCTION

Natural disaster that had occurred in Indonesia consist of hydrometeorology: floods, droughts, and landslides, geophysical: tectonic earthquakes and volcanic eruptions, and biological: epidemics (Rencana Nasional Penanggulangan Bencana 2020-2024 - BNPB, 2019). Regarding to the tectonic earthquake on Sulawesi Island, there are at least 2 earthquake disasters that became national disasters, namely in Central Sulawesi and West Sulawesi in the range of 2017 to 2020. This should had been a concern in evaluating earthquake patterns from year to year to compile tectonic earthquake disaster mitigation plans.

Seeing that several earthquakes have occurred in the Sulawesi Island region, the composition of disaster mitigation efforts should be adjusted to the characteristic conditions of each region. These characteristics can consist of the depth and strength of the earthquake. The aim is to see existing patterns as the basis for the preparation of mitigation efforts so that the policies drawn up are appropriate (White et al., 2017).

It can be done by using clustering analysis. Cluster analysis is one of the topics of multivariate statistical analysis or statistical learning, which is also known as unsupervised learning (Ansori Mattjik and Sumertajaya, 2011). Cluster analysis is the process of collecting n objects into k groups with k less than n (Ji et al., 2012). Objects with similar characteristics to each other are grouped into a group, while other objects are collected in different clusters. The group formed hereinafter is called the cluster (Nooraeni et al., 2021).

The similarity between objects is obtained based on the variables that characterize the observed objects. To measure the similarity, it is conducted by using the concept of distance. Mathematically, the smaller the distance between objects, the more similar the objects are and vice versa. The concept of distance that is commonly used is Euclidean distance (Dinh et al., 2021).

According (Pham et al., 2011) Further mentioned that there are two main problems that need to be considered in non-hierarchical clustering, namely the number of clusters and the selection of cluster centre's because the clustering results depend on the selected centroids. Another challenge encountered is the type of variable that characterizes the objects (Li et al., 2019). Characteristics of objects consisting of numerical variables are measured by Euclid distance as in the k-means algorithm (Akramunnisa and Fajriani, 2020). Furthermore, the characteristics of objects consisting of categorical variables can be measured using the mode, the smaller the value of the mode, the more similar objects are and vice versa. This concept is used in the k-modes algorithm, where the mode is the centroid of a cluster (Mau and Huynh, 2021).

When the object characteristics consist of numeric and categorical variables, the concept of distance that can be used is a combination of the concepts of k-means and k-modes distances (Kuo et al., 2021) (Nooraeni et al., 2021). This was proposed by k-prototype method was proposed because the objects that are often encountered in real-world databases are mixed type objects between numeric and categorical (Kuo and Wang, 2022). Furthermore, this method can overcome the challenges of large-scale data compared to hierarchical-based methods (Pham et al., 2011). With the research on categorizing areas on the island of Sulawesi based on earthquake events that have occurred from 2017 to 2020, it is found that the strength and potential for regional earthquakes in South Sulawesi are grouped. So that the results of this study can be taken into consideration in the preparation of disaster mitigation policies.

B. LITERATURE REVIEW

The K-Prototype algorithm is one of the Clustering methods based on partitioning (Pham et al., 2011) (Iriawan et al., 2018). This algorithm is the result of the development of the K-Means algorithm (Mau and Huynh, 2021) (Ahmad and Dey, 2011). to handle clustering on data with mixed numeric and categorical type attributes (Dinh et al., 2021). The development carried out by Huang maintains the efficiency of the K-Means algorithm in dealing with large data and can be applied to numerical and categorical data (Kuo et al., 2021). The basic development of the K-Prototype algorithm is in measuring the similarity (similarity measure) between the object and its centroid (prototype) (Pham et al., 2011). In general, the K-Prototype algorithm is divided into three main stages (Sulastri et al., 2021), namely:

1. Initial initialization of the prototype. In this process, several k prototypes will be selected randomly from the X dataset according to the specified number of clusters.
2. Allocation of objects in X to the Cluster with the closest prototype. Measure the object distance to all prototypes and place the object in the closest cluster. At this stage the K-Prototype algorithm allocates all objects in the dataset to the cluster where the prototype of the cluster has the closest distance to the data object. Allocating all objects in Data set X to the cluster that has the closest prototype distance to the object being measured. For each time object X has been allocated, the next step will be to calculate (update) the related prototype cluster.
3. Reallocation of objects If there is a change in the prototype. After all objects in X have been allocated, the next step will be to re-measure the distance between all objects in X against all existing prototypes. If an object is found that is closer to another prototype, membership transfer will be carried out and then an update will be made on the old cluster prototype and the new cluster prototype. This process will continue until there are no more changes to the prototype or until the stopping criteria are met.

C. RESEARCH METHOD

Cluster analysis procedures using k-prototype are as follows (Dinh et al., 2021); (Mau and Huynh, 2021):

1. Data preparation, checking the tidy of data
2. Data exploration, identify the relationship of variables by visualization using scatterplot and boxplot
3. Data transformation, magnitude and depth earthquake is transformed as follows:

$$x* = \frac{x - \bar{x}}{s} \tag{1}$$

   where, $x*$ is the transformed variable, x is the original variable, and $\bar{x}$ is the average, and s is the standard deviation
4. Implementation of k-prototype clustering as follows (Mau and Huynh, 2021) (Sulastri et al., 2021):
   (a) Determining the centroid of the cluster as many as the k, where $k < n$, n is the number of samples as the starting point $C_1, C_2, \ldots, C_k$ on every variable $(X_1, X_2, \ldots, X_p)$;

DOI: https://doi.org/10.30812/varian.v5i2.1908

(b) Calculating the distance or similarity of data points on the data set against the centroid of the cluster, the data points are grouping into the cluster that has the closest distance to the centroid as follows:

$$d(X,Y) = \sum_{j=1}^{p}(x_{jn} - y_{jn})^2 + \gamma \sum_{j=p+1}^{m} d(x_{jc}, y_{jc}) = \begin{cases} 0 & , x_{jc} \neq y_{jc} \\ 1 & , x_{jc} = y_{jc} \end{cases} \tag{2}$$

$d(X,Y)$ is distance or similarity of object X and Y, p and m are the number of numerical variables and categorical variable respectively, j is the j^{th} variable, n and c is corresponding to numeric and category. The first term is Euclid distance for numerical characteristics and the second terms is frequency mismatch of level category for categorical characteristics where γ is a parameter that balances the variable scale difference.

(c) Calculating the new centroid of the cluster after all objects have been grouped into clusters, and then re-grouping all objects on the new centroids.

(d) The process would stop if there were not changing to the centroids, or it has been convergent.

5. The optimum cluster selection using diversity values. It is conducted by k optimum selection; Value of k is selecting by using ratio of variety of within-cluster distances (S_W) against variety of between-cluster distance (S_B) (Sulastri et al., 2021) (Sarma et al., 2013). The ratio is plotted against the number of clusters (k) and the selected k is whose greatest changing of ratio proposed S_W and S_B of numerical variable is obtained by using Equation 3 (Ji et al., 2013):

$$S_{W_n} = \frac{1}{K}\sum_{k=1}^{K} S_k, \quad S_{B_n} = \left[\frac{1}{K-1}\sum_{k=1}^{C}(\bar{x}_k - \bar{x})\right]^{\frac{1}{2}} \tag{3}$$

For categorical variable, (Mau and Huynh, 2021) proposed within and between sums of square are obtained by using Equation 4 as follows:

$$\begin{aligned} &S_{Wc} = (MSW)^{\frac{1}{2}}, \quad MSW = \frac{SSW}{(n-K)} = \frac{1}{(n-K)}\left[\frac{n}{2} - \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{m=1}^{M} n_{mk}^2\right] \\ &S_{Bc} = (MSB)^{\frac{1}{2}}, \quad MSB = \frac{SSB}{(K-1)} = \frac{1}{(K-1)}\left[\frac{1}{2}\left(\sum_{k=1}^{K}\frac{1}{n_k}\sum_{m=1}^{M} n_{mk}^2\right) - \frac{1}{2n}\sum_{m=1}^{M} n_m^2\right] \\ &n = \sum_{k=1}^{K} n_k = \sum_{m=1}^{M} n_m = \sum_{k=1}^{K}\sum_{m=1}^{M} n_{mk} \end{aligned} \tag{4}$$

6. Cluster description

D. RESULTS AND DISCUSSION

1. Data Description

On the data preparation, there are 34 earthquake events with zero magnitude. That are in 2017 and 2019, so that the records are removed. The reason why it is removed that earthquake with zero magnitude that it is not categorized as tectonic earthquake but as an impact of human activities. Hence, the remaining earthquake events are 6493 where 678 in 2017, 2029 in 2018, 2024 in 2019, and 1762 in 2020.

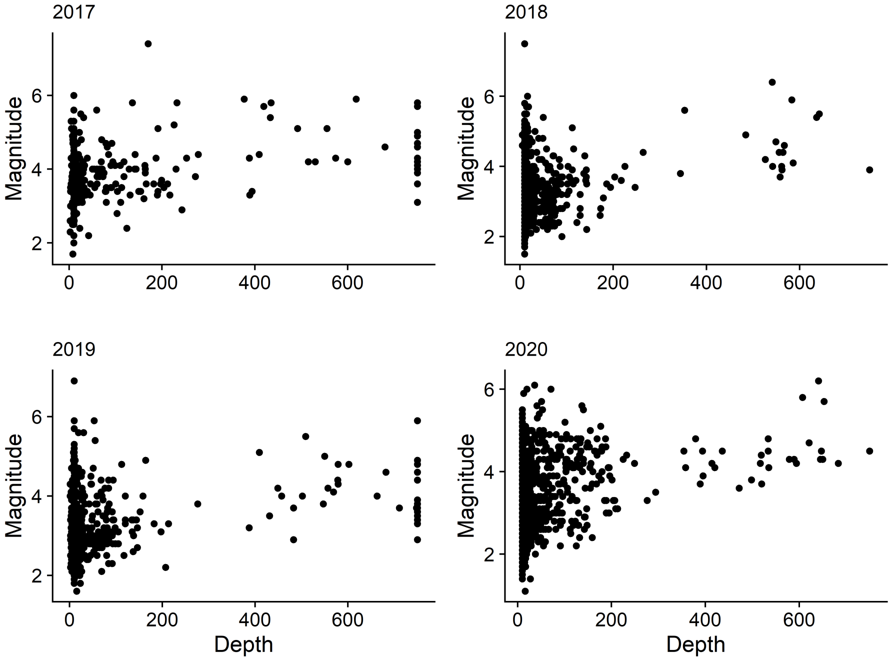

Figure 1. Relationship between magnitude and depth from 2017 until 2020

Relationship between earthquake magnitude and earthquake depth tends to directly weak relationship. It can be seen on Figure 1 which depicted that point patterns form positive pattern and spread enough for each year. The relationship between earthquake distance category and those numerical variables i.e., depth and magnitude are depicted on Figure 2. It seems like the relationship on Figure 1.

There is not significantly difference either magnitude against earthquake distance category or depth against earthquake distance category for each year. Hence, it can be conclude that there is not a significance relationship among used variables. Finally, the scale of magnitude and depth of earthquake is difference as depicted on Figure 1 and Figure 2, so they are transformed using Equation 1.

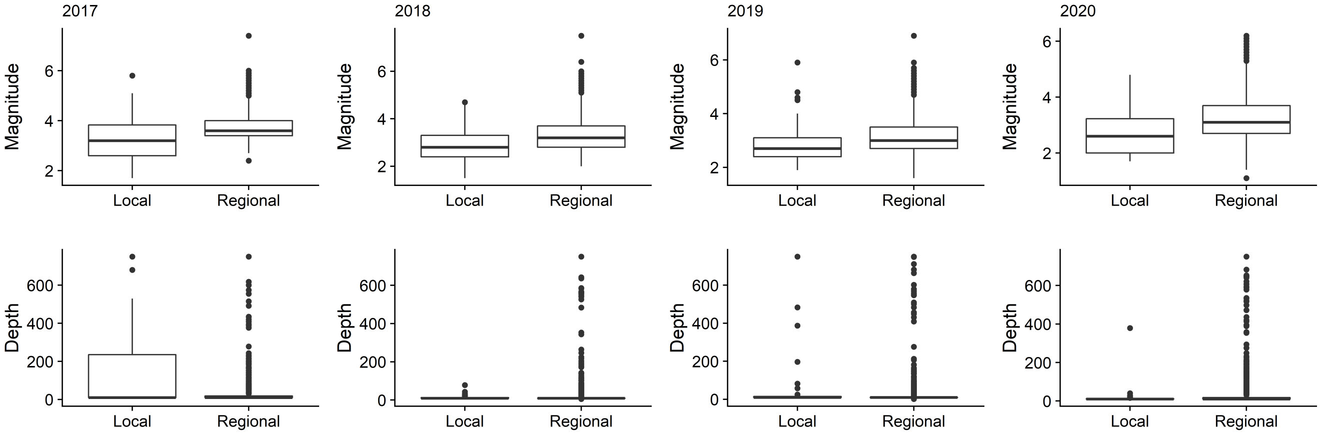

Figure 2. Magnitude and depth based on earthquake distance category from 2017 until 2020

2. The Cluster Optimum Selection

In this study, the grouping is done every year and the values of for each year are 7.60, 11.43, 14.19, and 14.90. By using these balancing parameters, the optimum k value for clustering tectonic earthquakes on Sulawesi Island is shown in Figure 3. In general, the value of the SW to SB ratio is fluctuating so that the k value is chosen based on the largest change in the ratio. The

DOI: https://doi.org/10.30812/varian.v5i2.1908

optimum k values are 4, 6, 5, and 6 for each year, respectively. The optimum number of clusters in each year is different because this is thought to be related to faults around the island of Sulawesi. The faults around the island of Sulawesi are the Kendari, Makassar, Luwuk, Matano, Palu Koro, Saddang, Walanae, and Poso faults (White et al., 2017).

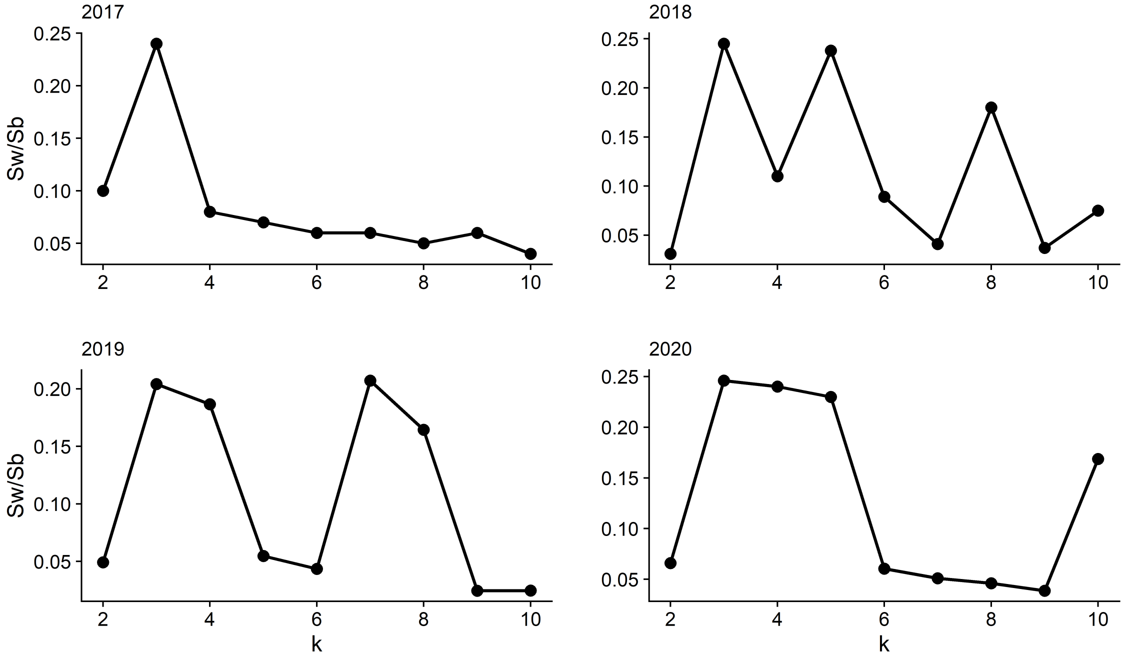

Figure 3. Ratio of S_W and S_B

3. Cluster Description

The number of elements of clusters for each year is shown on Table 1. Based on the number of elements of cluster, Cluster 3 is the cluster with the most elements for each year. Conversely, Cluster 6 tends to have the fewest elements for each year.

Table 1. The Number of Elements of Clusters

Cluster	2017	2018	2019	2020
1	33	673	210	317
2	341	18	349	652
3	215	446	717	668
4	89	242	38	18
5		557	710	41
6		93		30

DOI: https://doi.org/10.30812/varian.v5i2.1908

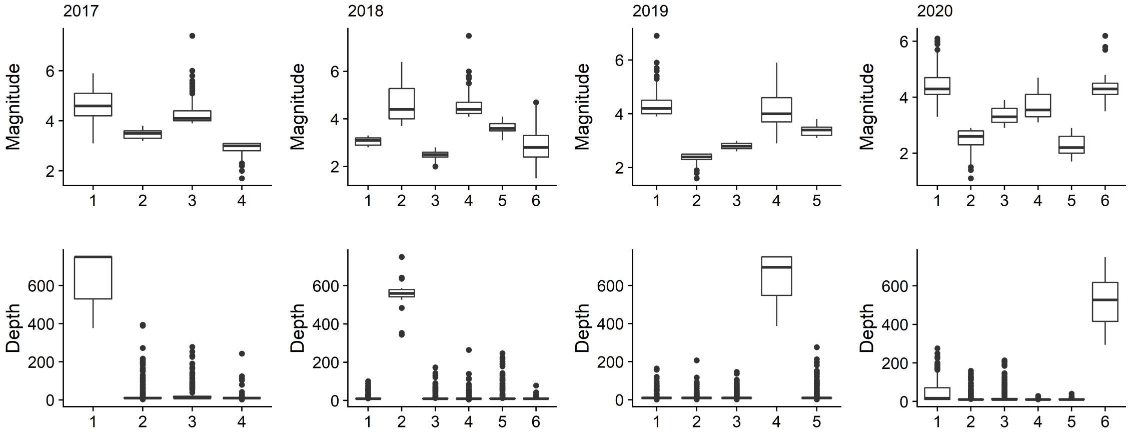

Figure 4. Magnitude and depth of clusters from 2017 until 2020

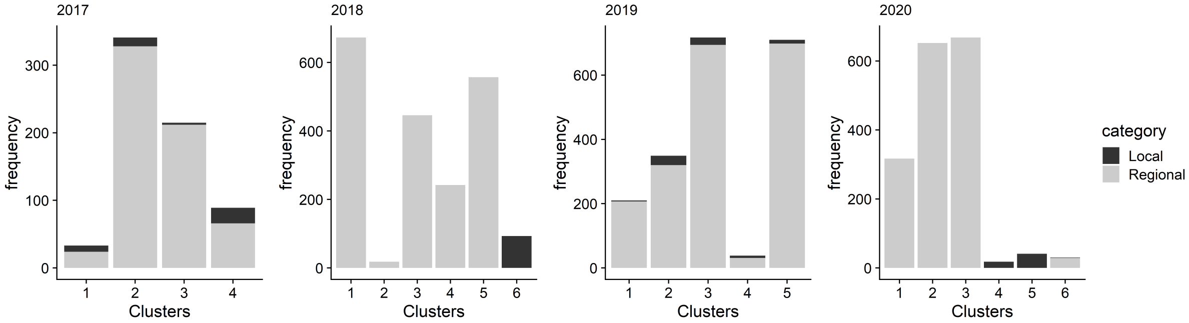

Figure 5. Earthquake distance category of clusters from 2017 until 2020

Figure 4 depicted magnitude and depth of earthquake each cluster. Cluster 1 and Cluster 2 contain greater magnitude of earthquake than others and Cluster 4 contains the least magnitude of earthquake in 2017. Nevertheless, Cluster 3 and Cluster 4 contain outliers. In 2018, Cluster 2 and Cluster 4 contain greater magnitude of earthquake than the others where the least magnitude of earthquake is Cluster 3. In 2019, Cluster 1 and Cluster 4 contain greater magnitude of earthquake than the others where the least magnitude of earthquake is Cluster 2. Finally, in 2020, Cluster 1 and Cluster 6 contain greater magnitude of earthquake than the others where the lowest magnitude of earthquake is Cluster 5.

E. CONCLUSION AND SUGGESTION

In term of depth of earthquake, there is only one cluster contains the deepest for each year. They are Cluster 1 in 2017, Cluster 2 in 2018, Cluster 4 in 2019, and Cluster 6 in 2020. Regarding to the distance category, most of earthquake from 2017 to 2020 is regional level. There is only a cluster contains local level that is in 2018. It can be related to the volcanic earthquake in Palu and Donggala on September 2018. From the findings above, the greater magnitude the deeper the earthquake is. In addition, the volcanic earthquakes are moderately regional earthquake.

REFERENCES

Ahmad, A. and Dey, L. (2011). A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters*, 32(7):1062–1069.

Akramunnisa, A. and Fajriani, F. (2020). K-Means Clustering Analysis pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan. *Jurnal Varian*, 3(2):103–112.

Ansori Mattjik, A. and Sumertajaya (2011). *Sidik Peubah Ganda Dengan menggunakan SAS*. IPB PRESS Edisi.

DOI: https://doi.org/10.30812/varian.v5i2.1908

Dinh, D.-T., Huynh, V.-N., and Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, 571:418–442.

Iriawan, N., Fithriasari, K., Ulama, B., Suryaningtyas, W., Susanto, I., and Pravitasari, A. (2018). Bayesian Bernoulli Mixture Regression Model for Bidikmisi Scholarship Classification. *Jurnal Ilmu Komputer dan Informasi*, 11(2).

Ji, J., Bai, T., Zhou, C., Ma, C., and Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, 120:590–596.

Ji, J., Pang, W., Zhou, C., Han, X., and Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 30:129–135.

Kuo, R.-J., Zheng, Y., and Nguyen, T. P. Q. (2021). Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Information Sciences*, 557:1–15.

Kuo, T. and Wang, K.-J. (2022). A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification. *Computers & Industrial Engineering*, page 108164.

Li, C., Wu, X., Cheng, X., Fan, C., Li, Z., Fang, H., and Shi, C. (2019). Identification and analysis of vulnerable populations for malaria based on K-prototypes clustering. *Environmental research*, 176:108568.

Mau, T. N. and Huynh, V.-N. (2021). An LSH-based k-representatives clustering method for large categorical data. *Neurocomputing*, 463:29–44.

Nooraeni, R., Arsa, M. I., and Projo, N. W. K. (2021). Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering. *Procedia Computer Science*, 179:677–684.

Pham, D.-T., Suarez-Alvarez, M. M., and Prostov, Y. I. (2011). Random search with k-prototypes algorithm for clustering mixed datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2132):2387–2403.

Sarma, T. H., Viswanath, P., and Reddy, B. E. (2013). Speeding-up the kernel k-means clustering method: A prototype based hybrid approach. *Pattern Recognition Letters*, 34(5):564–573.

Sulastri, S., Usman, L., and Syafitri, U. D. (2021). K-prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University. *Indonesian Journal of Statistics and Its Applications*, 5(2):228–242.

White, L. T., Hall, R., Armstrong, R. A., Barber, A. J., BouDagher Fadel, M., Baxter, A., Wakita, K., Manning, C., and Soesilo, J. (2017). The geological history of the Latimojong region of western Sulawesi, Indonesia. *Journal of Asian Earth Sciences*, 138:72–91.