Indonesian Journal of Electrical Engineering and Computer Science
Vol. 36, No. 3, December 2024, pp. 1594~1604
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v36.i3.pp1594-1604

Identification and characterisation of earthquake clusters from seismic historical data

Karmenova Markhaba[1], Tlebaldinova Aizhan[2], Alibekkyzy Karlygash[2], Zhantassova Zheniskul[1], Karymsakova Indira[3]
[1]Department of Computer Modeling and Information Technology, Higher School of IT and Natural Sciences, Sarsen Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan
[2]School of Digital Technology and artificial intelligence, Daulet Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan
[3]Department of Automation, Information Systems and Urban Planning, Shakarim University, Semey, Kazakhstan

Article Info

Article history:

Received Jan 10, 2024
Revised Aug 8, 2024
Accepted Aug 26, 2024

Keywords:

Algorithms
Clustering
Data analysis
Earthquake data
K-means
K-medoids
Seismic events

ABSTRACT

New approaches and methods based on machine learning technologies make it possible to identify not only the spread of earthquakes, but also to establish hidden patterns that allow further assessment of any risks associated with their occurrence. In this article, the clustering algorithms of K-means and K-medoids are applied for the analysis of seismic data recorded on the territory of the Republic of Kazakhstan. Using the Elbow and Silhouette methods, the optimal value of K clusters was determined, which was later used in classifying a data set using cluster analysis methods. The results of seismic data classification by clustering algorithms are in line with expectations. However, when measuring the quality of clustering, the accuracy of the model by the K-means method exceeded the accuracy of the K-medoids model, and the scoring value by the K-means method is ahead of the value by the K-medoids method. In addition, the presented results of descriptive statistics allowed to carry out a more in-depth analysis of the characteristics of each cluster.

Corresponding Author:

Karmenova Markhaba
Department of Computer Modeling and Information Technology
Higher School of IT and Natural Sciences, Sarsen Amanzholov East Kazakhstan University
Kazakhstan Street, 55, Ust-Kamenogorsk 070004, Kazakhstan
Email: mmm058246@gmail.com

1. INTRODUCTION

In history and in the memory of all mankind, many cases have been remembered, generated by various types of disasters, including natural ones, which also entail man-made disasters. Earthquakes are one of the most unpredictable and dangerous natural disasters. As you know, earthquakes bring total losses, in the first place–this is the loss of human lives. In order to ensure the safety of life and health of people, as well as to eliminate technical, economic and any other significant damage from the consequences of earthquakes, active scientific research is being carried out in this direction [1]. A recent example of such a catastrophic earthquake is the earthquake that struck southern Turkey on the border with Syria on February 6, 2023. According to scientists [2], one third of the territory of the Republic of Kazakhstan is seismically hazardous. It is where about half of the country's population lives, over 400 cities and towns are located, and 40% of the industrial potential is concentrated. Therefore, tracking and in-depth study of ongoing seismic events is very important for their further analysis and minimization of possible consequences.

In recent years, there has been a sharp increase in the number of scientific publications covering the main issues and problems of using machine learning methods and data mining technologies to solve problems in the field of seismology [3]. Most of these works are devoted to the development of approaches to predicting seismic events, namely, solving the problems of warning about long-term, medium-term, and short-term seismic hazards [4]-[7]. Another equally important category of tasks is related to the processing of seismic historical data for their in-depth study and identification of hidden patterns [8], [9]. For example, based on seismic data, a spatial-statistical study was carried out, which revealed clusters of earthquakes and their spatial relationship between tectonic and other geological processes [9]. To solve the problems of classifying seismic signals, various approaches are proposed. One of which is based on supervised machine learning methods and deep learning methods.

Seismic signals recorded in the local seismic network of Agadir (Morocco) [10] were classified using a multilayer perceptron neural network. The work was carried out in two stages: feature extraction and classification. The accuracy of the classifier proposed by the authors has reached more than 94%. Random forest classifier allows detection of mass movements, including seismic signals of earthquakes [11]. The classifier was tested on the basis of two data sets. Ways to improve the performance of the classifier are proposed, and the problem of lack of training data is solved. The developed model differs from other models in its simplicity in terms of setting parameters, and also does not require significant computational costs. Effective earthquake detection is implemented based on several machine learning methods with further comparative analysis: support vector machine, decision trees, random forest and linear regression [12]. As for deep learning methods [13], the developed module includes a transfer learning method, where the trained network uses a Bayesian approach, which results in a reduction in data labeling time and a quick generalization of the model for new data even with intense seismic activity. A new structure for automatic classification of earthquake magnitudes is described in [14]. This structure is based on a convolutional recurrent neural network, which uses a new approach to feature extraction. The peculiarity of the proposed structure is that it is able to classify both minor and strong earthquakes by magnitude.

Thus, we observe that a significant number of studies on this problem are based on supervised machine learning methods. It is known that supervised machine learning methods are aimed at classifying a predefined class of known signals, i.e., it assumes the presence of a complete set of labeled data for training the model at all stages of its construction. It is also worth noting that obtaining a high-quality labeled dataset, especially large-volume datasets, is a complex and routine task, although today there are different markup methodologies [15]. In this regard, in seismology, where most of the recorded data are unmarked, the relevance of using unsupervised learning algorithms to identify and study new classes of seismic events is growing and developing [16].

For a comprehensive understanding of global seismic activity in [17], a methodology for analyzing seismic data is presented. The authors' methodology uses exploratory data analysis (EDA), time dynamics research, spatial patterns analysis and cluster analysis. The result of the study demonstrated the global distribution of earthquakes, where cluster analysis showed certain hot spots susceptible to seismic activity. Iaccarino and Picozzi [18], the cluster analysis method is used. As a result of this analysis, the presence of two clearly distinguishable cluster groups of seismicity was found. These cluster groups belong to the preparatory stage, which has a higher percentage of events within it during the day before the main tremors compared to a random sample. Zaccagnino *et al.* [19], a cluster analysis of the seismicity of Turkey is also carried out, where an updated version of the Turkish Homogenized Earthquake Catalog (TURHEC) is used. The authors found that globally clustered and locally Poisson seismic activity has been observed in areas exposed to major seismic events over the past century.

In previous studies, seismic events have been studied by various approaches and machine learning methods, focusing on seismic signal processing and prediction methods, seismic hazard classification, and cluster group detection. However, these early works clearly lack in-depth studies on the structure of the detected cluster groups and their changes during the clustering process. The research issue of this article is to study and identify characteristic changes in the cluster ability process based on the intelligent analysis of seismic events recorded on the territory of the Republic of Kazakhstan for a deeper understanding of the nature of seismicity of this territory. The article presents a methodology for analyzing the seismic data set recorded by seismic stations on the territory of the Republic of Kazakhstan. The research methodology uses exploratory data analysis (EDA), cluster analysis and clustering assessment methods. The K-means method is used as a cluster analysis method, as well as the K-medoids method to compare the performance of clustering methods and verify the accuracy of the results. The analysis of statistical conclusions and a comprehensive assessment of the results obtained were carried out using descriptive statistics. The assessment of the quality of the constructed models was carried out by applying the "Silhouette" method, one of the indicative measures for validating the obtained clusters. Finally, the results of the evaluation of the quality of the constructed models were confirmed using external clustering quality assessment measures. In section 2 presents a methodology with a description of the data, with a description of the proposed approach and research methods. In section 3 contains the results of the study with the justification of the proposed concept. The conclusions and future directions of the study are summarized in section 3.

2. METHOD

2.1. Data description

Kazakhstan is located at the junction of two continents-Europe and Asia, bordering China, Kyrgyzstan, Turkmenistan, Uzbekistan and Russia. As it was noted earlier, one third of the territory of the Republic of Kazakhstan is earthquake-prone. According to the map of the general seismic zoning of Kazakhstan, as well as the presented spatial and temporal distribution of seismic events of recent years, it is shown that the most earthquake-prone regions are the borders with China and Kyrgyzstan. The work uses data on recorded seismic events on the territory of the Republic of Kazakhstan and its adjacent areas over the past 11 years, which are shown in Figure 1.

The data are taken from an open source of the LLP "seismological experimental and methodological expedition (SEME)" the purpose of which is to organize and conduct comprehensive studies in earthquake-prone areas of the Republic of Kazakhstan in order to predict earthquakes. Observations carried out in a wide frequency and dynamic ranges in real time are provided by a set of instruments including digital seismic stations (DM24, Q730, DAS6102; with seismometers with different registration periods from 360 seconds. up to 0.02 sec.) the Volcano seismotelemetric system and digital installations of strong ETNA movements. Strong earthquakes at a distance of several thousand kilometers are recorded using long-period instruments (periods from 360 to 10 seconds), and earthquakes in the near zone are recorded with highly sensitive short-period equipment (periods from 10 to 0.02 seconds).

The structure of the dataset includes attributes such as date, time, latitude, longitude, depth, magnitude and energy class, where date is the date of registration of seismic events, time is the time of registration of seismic events in GMT, latitude is latitude (north latitude), longitude is longitude (east longitude), depth is depth (km), energy_cl-energy class (Kr), magnitude-magnitude (MPVA). A fragment of the data set is presented in Table 1. The values of recorded seismic events vary in magnitude from 3.1 to 6.9, in depth from 0 to 220, in energy class from 7.1 to 14.9.

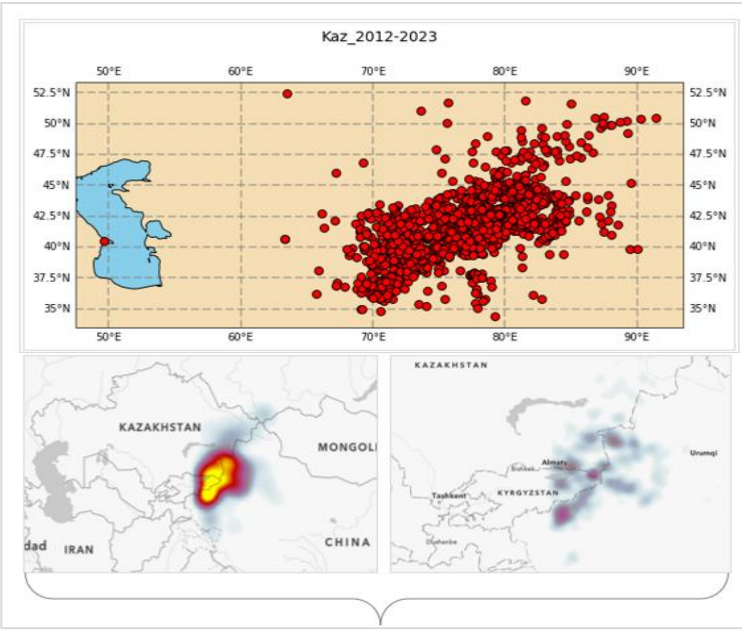


Figure 1. Earthquakes distribution during 2012-2023 in the study area

2.2. Approach

The general structure of the proposed approach is shown in Figure 2. The approach includes such main components as:

- preparation of data for further processing and analysis;
- conducting a comparative analysis of K-means and K-medoids clustering algorithms;

– conducting internal and external validation of clusters with the best quality indicators.

At the stage of data preparation for further processing and analysis, work has been carried out to transform data types and, depending on them, sets of procedures have been prepared for processing missing values by replacing them with an average value or modes. The clustering algorithms are implemented in the Python programming environment.

Table 1. Earthquake data in Kazakhstan

Date	Time	Latitude	Longitude	Depth	Energy_cl	Magnitude
02.04.2020	9:10:59	44.75	80.73	30	9.8	4.1
29.03.2020	14:28:27	39.97	77.42	10	9.3	4.1
29.03.2020	14:22:22	42.93	84.82	15	9.8	4.3
29.03.2020	4:51:41	41.36	72.9	10	9.8	4.3
29.03.2020	4:41:09	41.31	72.9	10	11.5	4.9
26.03.2020	4:48:45	40.87	70.66	5	11.5	4.8
22.03.2020	19:21:31	41.74	81.2	10	12.5	5
21.03.2020	13:08:15	42.67	84.87	5	9.7	4.3
20.03.2020	16:18:42	42.04	81.5	15	9.9	4.2

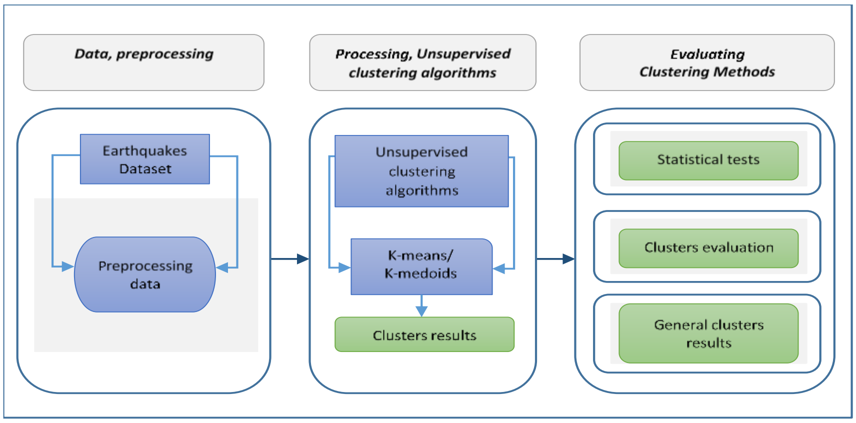


Figure 2. Block diagram of the developed method

2.3. Survey unsupervised clustering algorithms

Clustering is a data mining technology that can group data into different clusters. This technology is currently the most suitable for studying seismic data, in addition, it is becoming increasingly popular due to their convincing ability to identify discrete groups. The research methods in the work are the methods of machine learning without a teacher K-means and K-Medoids.

2.3.1. K-means clustering algorithm

As one of the most widely used clustering algorithms, the K-means method is characterized by its practicality, efficiency and good clustering effect [20]–[22]. Let be a data set X= $\{x_1, x_2, \ldots, x_m\}$, where M are objects in Euclidean space. The main goal of the K-means method is to divide the set of observations X into K clusters, $K_1, K_2, \ldots, K_k$, i.e., $K_i \subset X$, $|K_i| \geq 1$, and $K_i \cap K_j = \emptyset$ for $1 \leq i, j \leq k$, $i = j$. Each cluster must contain at least one object. The method classifies identical objects into several groups so that objects in the same cluster are as similar as possible, and objects from different clusters are different, i.e., not similar. Each cluster in K-means clustering is represented by its center (centroid). The cluster centroid is the average value of the objects that are assigned to the cluster. The difference between the object $l \in K_i$ and k_i, a representative of the cluster, is measured using the Euclidean distance dist (l, k_i), where dist (l, k_i) is the Euclidean distance between the points l and k_i. The quality of the K_i cluster can be measured by the intracluster variation, which is defined as (1):

$$A = \sum_{i=1}^{k} \sum_{l \in K_i} dist(l, k_i)^2 \quad (1)$$

where A is the sum of squared errors for all objects in the dataset; l is a point in space representing this object; k_i -is the centroid of the cluster K_i. In other words, the distance from each object to the cluster center is squared and these distances are summed.

2.3.2. K-medoids clustering algorithm

Unlike the K-means method, the K-medoids method works with representative values for each cluster. Instead of averages, the K-medoids method selects actual data points from clusters as their centers, i.e., it selects the medians of clusters as centers. Cluster formation is carried out by selecting K representative points, and then repeatedly moving to the best representatives of the cluster. All possible combinations of representative and non-representative points are analyzed, and the quality of the resulting clustering is calculated for each pair. The original representative point is replaced by a new point that causes the greatest reduction in the distortion function. At each iteration, the set of best points for each cluster generates new corresponding methods [23]-[26]. Figure 3 represents the pseudocode of the K-medoids algorithm. In the Figure 3, D means objects (1, 2, …, n). S_{ij} square error function and dis (D_i, D_j) distance between i and j objects.

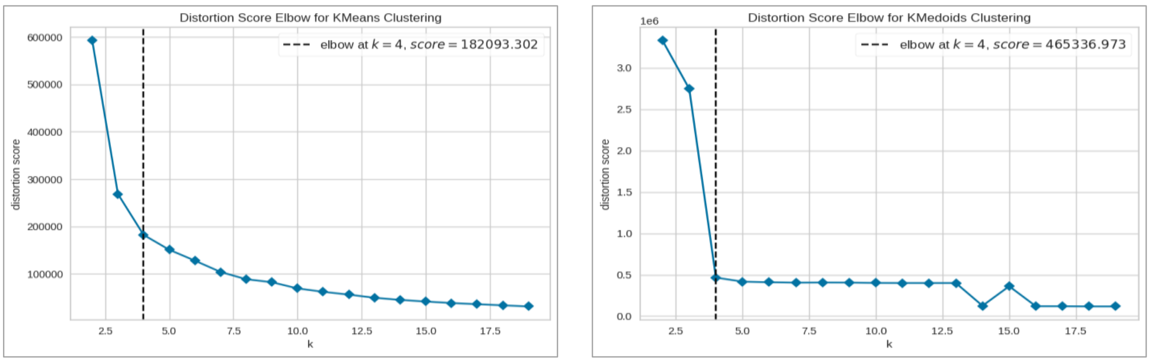
```
K-Medoids algorithm
Input:
- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.
arbitrarily choose k objects in D;
repeat
      for each Dj not a medoid do
            for each medoid Di do
                  compute square error;
function Sij;
find i,j where Sij is the smallest
if   Sij<0 then
replace medoid Di with Dj
until Sij>=0
for each Di∈D do
assign Di to kn where dis(Di, Dj) is the smallest over all medoids
```

Figure 3. K-medoids algorithm pseudocode

3. RESULTS AND DISCUSSION

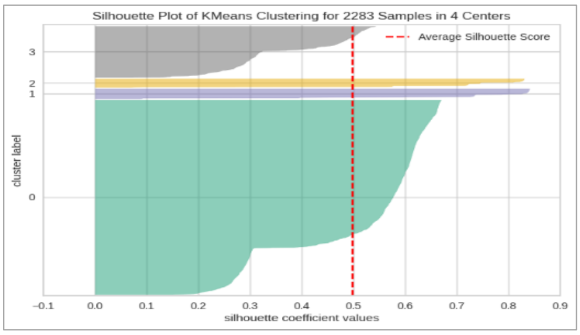
The implementation of cluster analysis began with the selection of the optimal number of clusters for the analyzed data, since the clustering methods used in this study work on the basis of an algorithm for dividing a vector space into a predetermined number of clusters K. There are various measures for determining the optimal number of clusters, most of which are based on the calculation of intracluster and intercluster distances within a single partition. In this paper, the determination of the optimal parameter K for the K-means and K-medoids methods was carried out using the elbow and Silhouette methods. The main idea of the elbow method is that it is possible to minimize the sum of the square inside the cluster (within-cluster sum-of-squares (WCSS)). The overall variation in the WCSS system measures the effectiveness of classification. Thus, the lower the WCSS value, the better the classification result is considered. Figure 4 is the result of the elbow method for choosing the optimal parameter K. As can be seen from Figures 4(a) and 4(b), the elbow method calculated the value K= 4 as the optimal parameter K for K-means that K-medoids. As the k value increases, the position where the improvement effect of the distortion degree decreases the most is the k value corresponding to the elbow [27].



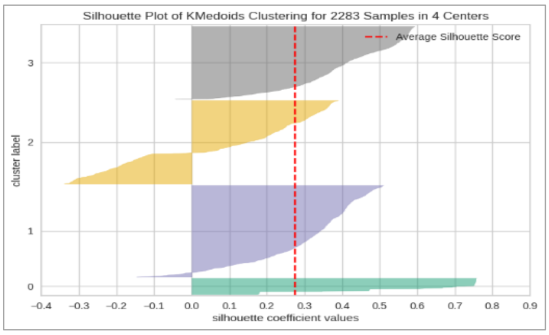
(a) (b)

Figure 4. Choosing the optimal number of clusters by the elbow method (a) for K-means and (b) for K-medoids

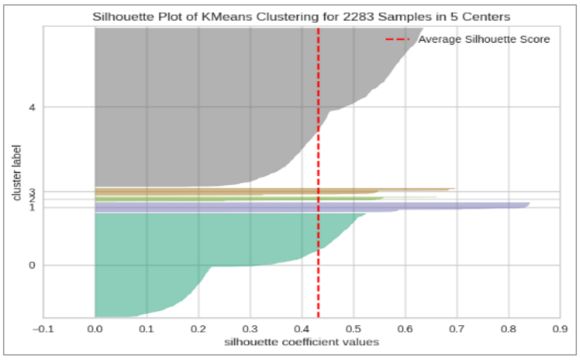
In the Silhouette method, the Silhouette coefficient is calculated using the average intracluster distance and the average distance to the nearest cluster for each object. The highest value of the “Silhouette” coefficient allows you to clearly identify clusters. Thus, based on the Silhouette scoring values, you can find the best clustering settings. The results of applying the Silhouette method to assess the quality of the constructed cluster model are visually presented in Figure 5. Figures 5(a), 5(c), 5(e) reflect sets of Silhouette coefficients at values K=4, K=5, K=6 for the K-means method. Figures 5(b), 5(d), 5(f) show sets of Silhouette coefficients at the same values K=4, K=5, K=6 for the K-medoids method. The scoring values for K=4, K=5, K=6 for the K-means and K-medoids methods are presented in Table 2. As can be seen from Table 2, at K=4, the scoring value for the K-means method reached a higher value of 0.50 than for the K-medoids method, which is 0.27.



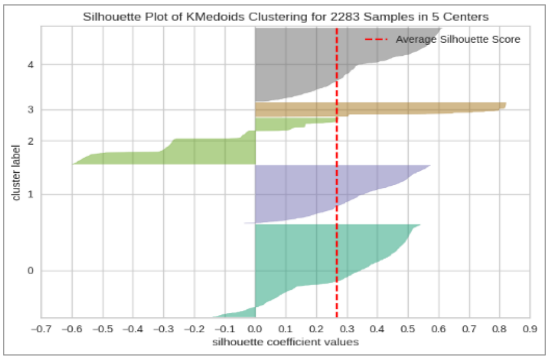
(a)



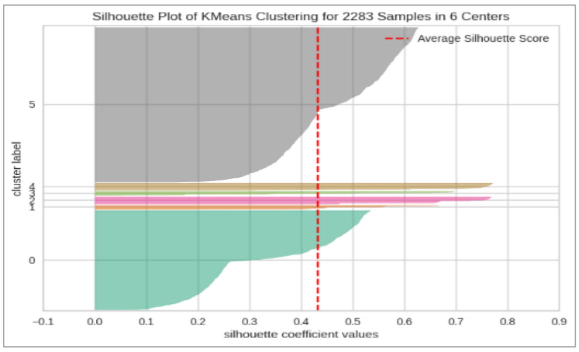
(b)



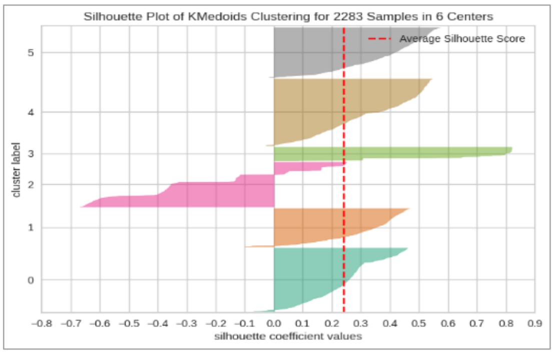
(c)



(d)



(e)



(f)

Figure 5. The value of the Silhouette for K-means method ((a) K=4; (c) K=5; (e) K=6) and K-medoids method ((b) K=4; (d) K=5; (f) K=6)

Table 2. The value of the Silhouette for K-means and K-medoids methods

Number of clusters	K-means	K-medoids
4	0.50	0.27
5	0.43	0.24
6	0.41	0.21

In the next step of the study, Figure 6 presents the results of applying clustering methods. As shown in Figures 6(a) and 6(b), the results of clustering by K-means and K-medoids demonstrate 4 clusters that are well separated from each other with clear boundaries and their centroids. Figures 6(c) and 6(d) show 3D visualizations of these 4 clusters. However, a comparison of the Silhouette coefficients of both methods shows that the scoring value by the K-means method is much greater than by the K-medoids method. In this regard, further investigation of the identified cluster groups was carried out on the basis of the results obtained using the K-means method. In addition, as a result of measuring the quality of clustering, the accuracy of the model using this method K-means exceeded the accuracy of the model K-medoids. In addition, as a result of the clustering quality measurement, the accuracy of the K-means model exceeded the accuracy of the K-medoids model. Also, the external clustering quality assessment measures presented in Table 3 confirm the results obtained using internal clustering quality assessment measures. In this connection, further research of the identified cluster groups was carried out on the results obtained by the K-means method. Our study suggests that higher values for the external measures of clustering quality using the K-means method demonstrate clearly delineated cluster groups of earthquakes. Thus, Figure 7 represents the results of classification by the K-means method: Cluster 0, Cluster 1, Cluster 2, and Cluster 3.

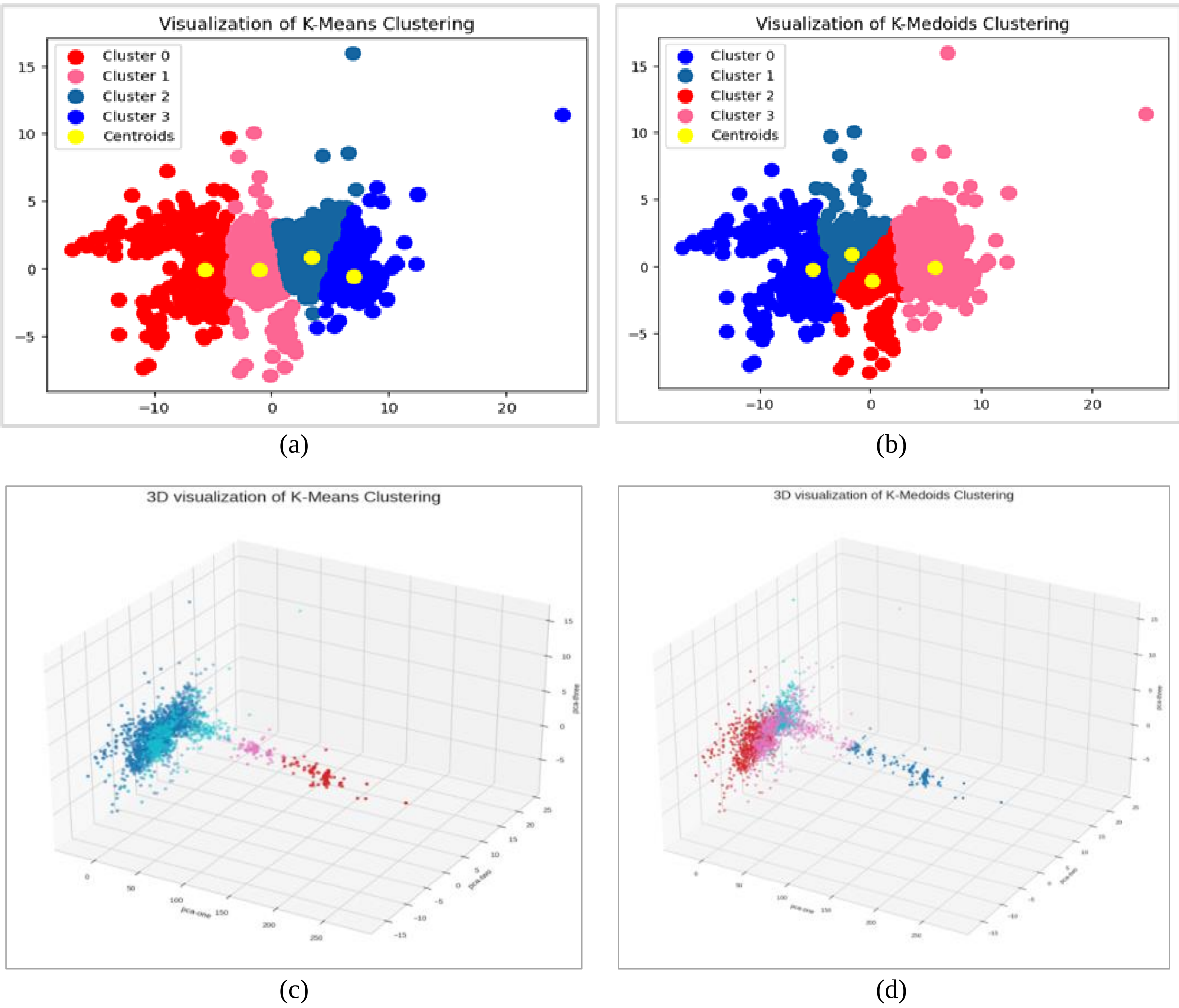


Figure 6. Plot of a four cluster data set from different view angles (3D and 2D): (a), (c) K-means method; (b), (d) K-medoids

Table 3. External measures for evaluating the results of clustering

Methods	ARI	NMI	Homogeneity	Completeness	V-measure
K-means	0.793	0.854	0.868	0.847	0.839
K-medoids	0.754	0.836	0.859	0.843	0.835

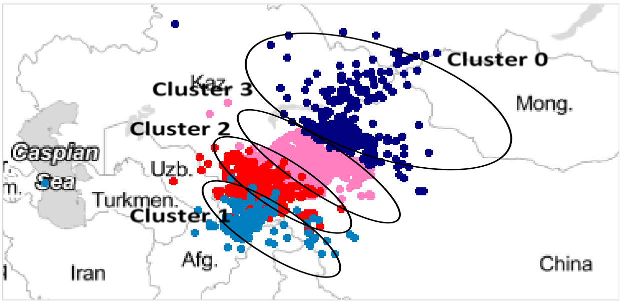


Figure 7. Cluster groups

The analysis of statistical conclusions and a comprehensive assessment of the results obtained were carried out using descriptive statistics. Table 4 presents the results of descriptive statistics for conducting a more detailed analysis of the clusters obtained and identifying the features of each of them.

Table 4. Statistical description of the earthquakes data set in each cluster

	Statistical description	Lat	Long	Depth	Energy_cl	Mag
Cluster 0	Count	1677	1677	1677	1677	1677
	Mean	41.752	77.446	8.863	10.147	4.485
	Std	2.462	4.117	4.259	1.099	0.544
	Min	34.39	63.31	0.0	6.3	2.9
	Max	52.48	91.42	15.0	15.1	7.3
Cluster 1	Count	89	89	89	89	89
	Mean	37.088	71.049	192.921	11.574	5.37
	Std	0.571	0.902	23.352	1.261	0.679
	Min	34.96	67.22	150.0	9.5	3.8
	Max	39.15	74.43	300.0	15.2	7.2
Cluster 2	Count	75	75	75	75	75
	Mean	37.156	71.299	102.066	11.197	5.186
	Std	0.594	1.024	16.564	0.968	0.551
	Min	35.9	67.29	65.0	9.6	4.2
	Max	39.44	75.12	145.0	13.9	6.7
Cluster 3	Count	442	442	442	442	442
	Mean	41.493	76.769	25.305	10.229	4.606
	Std	2.795	4.595	7.987	1.286	0.644
	Min	34.79	49.61	20.0	6.1	3.1
	Max	50.40	90.31	60.0	16.0	7.0

It can be seen from Table 4 that in Cluster 0 the frequency of occurrence of earthquakes is more frequent (within 1677) compared to other cluster groups. Seismic events are grouped in Cluster 0, where the average magnitude is ≈4.5 points, the energy class is ≈10.15 and the depth is ≈8.7 km. And in Cluster 1, the values of these indicators for magnitude and energy class are slightly higher than the values of Cluster 0, ≈5.4 points and ≈11.6, respectively, and the depth value is much higher than ≈192.92 km. As for the values of Cluster 2 indicators, they are closer to the values of Cluster 1: magnitude ≈5.2 points, energy class ≈11.2 and depth ≈102.07 km. The second cluster in terms of concentration of seismic event foci (within 442) is Cluster 3. Thus, as a result of descriptive statistics, we were able to identify the largest foci of earthquakes over the past 10 years.

The study of earthquake foci by Cluster 0 shows that mainly earthquakes partially cover the southeastern territories of the Republic of Kazakhstan with adjacent borders of the territories of the countries of China, Kyrgyzstan, and Uzbekistan. Most of the earthquake foci correspond to the territories of Kyrgyzstan and China. The location of the Cluster 0 centroid covers the average values of latitude 41.752 and longitude 77.446. According to this cluster, the main earthquake foci on the territory of the Republic of Kazakhstan are determined by the Zhetysui, Abai, East Kazakhstan regions. Also, this cluster determines single earthquake foci in Akmola, Karaganda and Pavlodar regions. The study of the Cluster 3 structure shows a denser concentration of earthquake foci in the southern territory of the Republic of Kazakhstan and a partial accumulation in the southeastern part. According to Cluster 3, South Kazakhstan, Almaty, Zhambyl, Turkestan regions are defined. Single earthquake foci correspond to the Ulytau region. The location of the centroid for Cluster 3 covers the average values of latitude 41.493 and longitude 76.769. The centroids of both clusters Cluster 0 and Cluster 3 are shifted to the territory of Kyrgyzstan. The study of Cluster 1 and

Cluster 2 showed that the distribution of earthquake foci and the location of their centroids mainly belong to the territory of Tajikistan, and therefore, a more detailed analysis of these clusters was not carried out in the work.

The study of each cluster made it possible to study the seismogenerating zones of the Republic of Kazakhstan. As it is known, according to historical data on the earthquakes that have occurred, the territories of the Republic of Kazakhstan are subject to the threat of destructive earthquakes. Recently, the research of Kazakhstani scientists and researchers on existing earthquake-prone regions also includes Atyrau, Aktobe, West Kazakhstan, Kyzylorda, Mangistau regions of the Republic of Kazakhstan. The appearance of new seismogenerating zones and faults is primarily associated with the geology and geomorphological structure of the earth, with the movement and creep of shear faults and plates [28], [29]. Thus, comparing the results of the study in the form of the obtained cluster groups of earthquakes with the seismogenerating map of the Republic of Kazakhstan and with the map of earthquake epicenters from ancient times to 2015, we see that the appearance and accumulation of earthquakes has not changed dramatically. In the course of the study, as previously noted, the identified cluster groups of earthquakes and their clusters are also associated with the main passing Junggar North Tian Shan fault. The regular occurrence of earthquakes in the adjacent areas with the main fault shows the frequency of occurrence of seismic events with different magnitudes. Thus, the obtained results will be used in the future for the spatial and temporal analysis of seismic datasets with the integration of new technologies [30] in modeling and forecasting seismic risks.

4. CONCLUSION

Earthquakes on the territory of the Republic of Kazakhstan and in its border zones with neighboring countries mainly occur in the southeastern part of the country, where Cluster 0 was identified, as well as in the southern and southeastern parts of the country as Cluster 3. As can be seen, the study of the earthquake dataset is based on the approach of data mining and clustering algorithms, such as K-means and K-medoids. In determining the optimal number of clusters for both clustering methods, two methods were used, such as the elbow method and Silhouette method, where the optimal value is K=4. Also, in assessing the quality of clustering, external measures were used to assess the quality of clustering, according to which the K-means method showed better results for identified cluster groups of earthquakes. The study of the features of cluster groups of earthquakes allowed us to obtain a modern understanding of the seismicity of the zones of the Republic of Kazakhstan. In the experimental part of the article, we studied in detail the structures of cluster groups, delved into the characteristic features of seismic events based on the magnitude, depth and energy class of earthquakes.

Our findings provide strong evidence that the studied seismic dataset based on the clustering method clearly shows stable earthquake cluster groups that have a complex spatial structure with characteristic features. Our study demonstrates that the complex spatial structure of cluster groups is closely related to the seismic activity of the country’s territory and adjacent regions of other countries. These results have profound implications for the prediction of earthquakes in the territory of the Republic of Kazakhstan, as well as for the assessment of seismic hazards in general. The results obtained highlight the importance of regular monitoring and research to better understand the seismic activity of our country. Future studies may consider the change and influence of time series on the formation of earthquake cluster groups, as well as show the change in the spatial structure of seismic events in the territory of the Republic of Kazakhstan. Further research is also needed to study the interaction of seismic events and external factors such as climate change, the emergence of new tectonic faults. Similar research will serve to improve existing spatial analysis methods, integrate new technologies to develop more sophisticated visualization tools and interactive platforms in real time, providing researchers and stakeholders with the necessary analysis and information.

ACKNOWLEDGEMENTS

We are grateful to LLP “SOME” and JSC “National Center for Seismological Observations and Research” of the Republic of Kazakhstan for providing information on earthquakes in an open Internet source. The article was carried out within the framework of the project of the Ministry of Education and Science of the Republic of Kazakhstan for grant financing of fundamental and applied scientific research of young postdoctoral scientists under the Zhas Galym project for 2022-2024. IRN: AP14972524 - “Development of VLC technologies in the management of unmanned vehicles.”

REFERENCES

[1] M. Karmenova, A. Nugumanova, A. Tlebaldinova, A. Beldeubaev, G. Popova, and A. Sedchenko, "Seismic assessment of urban buildings using data mining methods," in *ACM International Conference Proceeding Series*, Apr. 2020, pp. 154–159, doi: 10.1145/3397125.3397152.

[2] Ministry of Ecology and Natural Resources of the Republic of Kazakhstan, "National report on the state of the environment and on the use of natural resources of the Republic of Kazakhstan for 2022," 2023, [Online]. Available: https://www.gov.kz/memleket/entities/ecogeo/documents/details/518069?lang=ru.

[3] G. C. Beroza, M. Segou, and S. M. Mousavi, "Machine learning and earthquake forecasting—next steps," *Nature Communications*, vol. 12, no. 1, p. 4761, Aug. 2021, doi: 10.1038/s41467-021-24952-6.

[4] T. Bhandarkar, V. K, N. Satish, S. Sridhar, R. Sivakumar, and S. Ghosh, "Earthquake trend prediction using long short-term memory RNN," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, p. 1304, Apr. 2019, doi: 10.11591/ijece.v9i2.pp1304-1312.

[5] P. Xiong *et al.*, "Towards advancing the earthquake forecasting by machine learning of satellite data," *Science of The Total Environment*, vol. 771, p. 145256, Jun. 2021, doi: 10.1016/j.scitotenv.2021.145256.

[6] P. Debnath *et al.*, "Analysis of earthquake forecasting in India using supervised machine learning classifiers," *Sustainability (Switzerland)*, vol. 13, no. 2, pp. 1–13, Jan. 2021, doi: 10.3390/su13020971.

[7] P. A. Johnson *et al.*, "Laboratory earthquake forecasting: a machine learning competition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 5, Feb. 2021, doi: 10.1073/pnas.2011362118.

[8] M. Karmenova *et al.*, "An approach for clustering of seismic events using unsupervised machine learning," *Acta Polytechnica Hungarica*, vol. 19, no. 5, pp. 7–22, 2022, doi: 10.12700/APH.19.5.2022.5.1.

[9] B. Aslam and F. Naseer, "A statistical analysis of the spatial existence of earthquakes in Balochistan: clusters of seismicity," *Environmental Earth Sciences*, vol. 79, no. 1, p. 41, Jan. 2020, doi: 10.1007/s12665-019-8790-2.

[10] E. H. A. Laasri, E.-S. Akhouayri, D. Agliz, and A. Atmani, "Seismic signal classification using multi-layer perceptron neural network," *International Journal of Computer Applications*, vol. 79, no. 15, pp. 35–43, Oct. 2013, doi: 10.5120/13821-1950.

[11] M. Wenner, C. Hibert, A. Van Herwijnen, L. Meier, and F. Walter, "Near-real-time automated classification of seismic signals of slope failures with continuous random forests," *Natural Hazards and Earth System Sciences*, vol. 21, no. 1, pp. 339–361, 2021, doi: 10.5194/nhess-21-339-2021.

[12] W. Li, N. Narvekar, N. Nakshatra, N. Raut, B. Sirkeci, and J. Gao, "Seismic data classification using machine learning," in *Proceedings - IEEE 4th International Conference on Big Data Computing Service and Applications, BigDataService 2018*, 2018, pp. 56–63, doi: 10.1109/BigDataService.2018.00017.

[13] A. Bueno *et al.*, "PICOSS: Python interface for the classification of seismic signals," *Computers and Geosciences*, vol. 142, p. 104531, Sep. 2020, doi: 10.1016/j.cageo.2020.104531.

[14] M. Shakeel, K. Itoyama, K. Nishida, and K. Nakadai, "EMC: earthquake magnitudes classification on seismic signals via convolutional recurrent networks," in *2021 IEEE/SICE International Symposium on System Integration, SII 2021*, Jan. 2021, pp. 388–393, doi: 10.1109/IEEECONF49454.2021.9382696.

[15] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: a survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, p. 100336, Nov. 2021, doi: 10.1016/j.patter.2021.100336.

[16] L. Seydoux, R. Balestriero, P. Poli, M. de Hoop, M. Campillo, and R. Baraniuk, "Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning," *Nature Communications*, vol. 11, no. 1, p. 3972, Aug. 2020, doi: 10.1038/s41467-020-17841-x.

[17] M. Ibrahim and B. Al-Bander, "An integrated approach for understanding global earthquake patterns and enhancing seismic risk assessment," *International Journal of Information Technology (Singapore)*, vol. 16, no. 4, pp. 2001–2014, Apr. 2024, doi: 10.1007/s41870-024-01778-1.

[18] A. G. Iaccarino and M. Picozzi, "Detecting the preparatory phase of induced earthquakes at the geysers (California) using K-means clustering," *Journal of Geophysical Research: Solid Earth*, vol. 128, no. 10, Oct. 2023, doi: 10.1029/2023JB026429.

[19] D. Zaccagnino, L. Telesca, O. Tan, and C. Doglioni, "Clustering analysis of seismicity in the anatolian region with implications for seismic hazard," *Entropy*, vol. 25, no. 6, p. 835, May 2023, doi: 10.3390/e25060835.

[20] Y. Zhao, X. Huang, Y. Zhao, X. Liu, and R. Zhou, "The application of landscape character classification for spatial zoning management in mountainous protected areas – A case study of Laoshan national park, China," *Heliyon*, vol. 9, no. 3, p. e13996, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13996.

[21] G. S. Palupi and M. N. Fakhruzzaman, "Indonesian pharmacy retailer segmentation using recency frequency monetary-location model and ant K-means algorithm," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 6, pp. 6132–6139, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6132-6139.

[22] N. M. Mahfuz, M. Yusoff, and Z. Idrus, "Clustering heterogeneous categorical data using enhanced mini batch K-means with entropy distance measure," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1048–1059, 2023, doi: 10.11591/ijece.v13i1.pp1048-1059.

[23] M. Lee, S. Lee, J. Park, and S. Seo, "Clustering and characterization of the lactation curves of dairy cows using K-medoids clustering algorithm," *Animals*, vol. 10, no. 8, pp. 1–14, Aug. 2020, doi: 10.3390/ani10081348.

[24] A. V. Ushakov and I. Vasilyev, "Near-optimal large-scale k-medoids clustering," *Information Sciences*, vol. 545, pp. 344–362, Feb. 2021, doi: 10.1016/j.ins.2020.08.121.

[25] J. Deng, J. Guo, and Y. Wang, "A novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering," *Knowledge-Based Systems*, vol. 175, pp. 96–106, Jul. 2019, doi: 10.1016/j.knosys.2019.03.009.

[26] D. Wang, C. Jin, B. Xiao, Z. Li, and X. He, "Proof-of-activity consensus algorithm based on K-medoids clustering," *Big Data Research*, vol. 26, p. 100266, Nov. 2021, doi: 10.1016/j.bdr.2021.100266.

[27] Mengyao Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Accounting, Auditing and Finance*, vol. 3, pp. 9–16, 2020, doi: 10.23977/geors.2020.030102.

[28] D. Mackenzie *et al.*, "A creeping intracontinental thrust fault: past and present slip-rates on the Northern edge of the Tien Shan, Kazakhstan," *Geophysical Journal International*, vol. 215, no. 2, pp. 1148–1170, Nov. 2018, doi: 10.1093/GJI/GGY339.

[29] W. Wojcik, M. Karmenova, S. Smailova, A. Tlebaldinova, and A. Belbeubaev, "Development of data-mining technique for seismic vulnerability assessment," *International Journal of Electronics and Telecommunications*, vol. 67, no. 2, pp. 261–266, 2021, doi: 10.24425/ijet.2021.135974.

[30] K. Alibekkyzy, W. Wojcik, V. Kornev, and S. Belginova, "Robust data transfer paradigm based on VLC technologies," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 3, pp. 552–568, 2021, doi: 10.52209/1609-1825_2023_2_397.

BIOGRAPHIES OF AUTHORS

Karmenova Markhaba holds a M.Sc. in Computer science from East Kazakhstan State University, Kazakhstan. She has a Ph.D. in Information Systems (by industry) from The Daulet Serikbayev East Kazakhstan Technical University (EKTU), Kazakhstan. She is currently an Associate Professor IT in Natural Sciences School of the Sarsen Amanzholov East Kazakhstan University. She has authored or coauthored more than 15 papers. Her field of interest includes the applications of artificial intelligence, spatial data analysis, machine learning, IT for Education and Training. She can be contacted at email: mmm058246@gmail.com.

Tlebaldinova Aizhan received her B.Sc. in Applied Mathematics from the East Kazakhstan State University, Kazakhstan, and her MSc in the same University. In 2016, she was awarded her Ph.D. in Information Systems (by industry) from the Daulet Serikbayev East Kazakhstan State Technical University (EKTU), Kazakhstan. She is an Associate Professor of the School of Digital Technology and artificial intelligence of the Daulet Serikbayev EKTU, Kazakhstan. Her primary research interests are in the field of big data, data mining, machine learning, image processing, pattern recognition. She can be contacted at email: tlebaldinovazhan@gmail.com.

Alibekkyzy Karlygash received an academic degree of master of technical sciences in the specialty “Instrument making” and a degree of candidate of technical sciences. In the specialty “Automation and Control” from the East Kazakhstan State Technical University Daulet Serikbayeva (EKTU), Ust-Kamenogorsk, Kazakhstan, in 2014 and 2022, respectively. Currently, he is an associate professor, Ph.D. of the Faculty of Information Technologies and Intelligent Systems of the East Kazakhstan Technical University. Serikbayeva D., Ust-Kamenogorsk, Kazakhstan. Her is the author or co-author of more than 15 publications. Her research interests include LED systems, data communication, self-driving cars. She can be contacted at email: karlygash.eleusizova@gmail.com.

Zhantassova Zheniskul is associate professor of the Computer Modeling and Information Technologies Department at the East Kazakhstan University named after S. Amanzholov, Kazakhstan. She Holds a Candidate of Technical Sciences degree in System Analysis, Management and Information processing. Her research areas are design of databases in information systems, cryptographic methods of cybersecurity. She is a recipient of different national and regional awards such as Best University Teacher Award in Kazakhstan Republic, Diploma of head of the East Kazakhstan region “For significant contribution to the development of the university, achievements in the upbringing of the younger generation”, Honorary certificate of the city government of Ust-Kamenogorsk “For contribution to the development of Ust-Kamenogorsk city etc. Her research interests include design information systems, risk management and cryptographic algorithms. She can be contacted at email: zhantasova@vku.edu.kz.

Karymsakova Indira received an academic master’s degree in Information Systems from Shakarim Semipalatinsk State University (SSU) in Semey, Kazakhstan in 2007 and a PhD in Information Systems from Daulet Serikbayev East Kazakhstan Technical University (EKTU) in Ust-Kamenogorsk, Kazakhstan in 2022. Currently, she is an Acting Associate Professor, Ph.D. of the Department of Automation, Information Systems and Urban Planning of the Shakarim Semey University, Semey, Kazakhstan. She is an author or co-author of more than 35 publications. Her research interests include mathematical and computer modeling, robotics, artificial intelligence. She can be contacted at email: indira.karymsakova@gmail.com.