MAKAUT

Seismic Data Analysis Using Machine Learning

Fitting Seismic belt using typical clustering method and classifying earthquake

Abstract: The devastating effects of earthquakes are known to have on human lives and properties due to its suddenness, destructiveness, and inscrutability. Understanding the distribution of seismic data can help us identify the various risks that are associated with these natural disasters. In this paper we tried to converge machine learning and GIS technology together to analyze the seismic data. First of all we used a GIS software (ArcGIS) to map the positional parameters in the seismic data to coordinate points on a two dimensional plane with a base layer of world ocean base map, after which we applied DBSCAN and K-means clustering methods to cluster them. In addition, we take the depth and magnitude properties of seismic data and used K-Means clustering to classify them with the proper number of K value getting from Elbow method. The experiment show that the DBSCAN algorithm has a better effect on fitting the seismic belt and the classification result of K-means is also meet the expectations.

1. Introduction

Since the earthquake is very sudden, unfathomable and catastrophic, it always has been most feared natural disaster which brings with it the greatest loss to the human society. It causes deformation and destruction of houses, bridges, furthermore leads to various secondary deserters like landslides, tsunamis and wildfires. In recent years mega-earthquakes occurred quite frequently. The magnitude-7.6 quake occurred on October 8, 2005, in the Pakistan-administered portion of the Kashmir region. At least 79,000 people were killed and more than 32,000 buildings collapsed in Kashmir. On May 12, 2008, an earthquake measuring 8.0 on the Richter scale happened in the Wenchuan County, Sichuan Province, China, killing 69,000 people and injuring 370,000 people [1]. A magnitude-9.0 earthquake struck Japan in March 2011, triggering a tsunami that affected the most of the eastern Pacific. On December 26, 2004, at 7:59 am local time, an undersea earthquake with a magnitude of 9.1 struck off the coast of the Indonesian island of Sumatra. The tsunami killed at least 225,000 people across a dozen countries.

The distribution of earthquakes has always attract the attention of researchers. In the era of big data, the researchers can clench the behaviour of earthquake disaster and can reduce the risk by the use of new technical means to analyze seismic data. At present, there are many informatization methods used in seismic data analysis. Yang JL et al. [3] used wavelet decomposition, short time Fourier transformation and source scanning algorithm to explore the pre-seismic gravity anomalies prior to the earthquakes. Han P et al. [4] provided a new seismomagnetic data processing method by combining principal component analysis (PCA) and geomagnetic diurnal variation analysis. Zhao YG et al. [5] applied linear regression to the data analysis of the main aftershocks to provide a reference for the determination of aftershocks. The research on seismic belts is mostly concentrated on the temporal and spatial distribution of seismic distribution. Yin LR et al. [6] presented the fractal characteristics of seismicity spatial and temporal in the circum-Pacific seismic belt based on the methods of capacity dimension and information dimension.

The contribution of this project is to apply the typical clustering algorithms to the analysis of seismic data. The application effects of K-means algorithm and DBSCAN algorithm in the study of seismic belt fitting with seismic dataset are compared, and the visual display is finally carried out. The results show that compared to the K-means algorithm, the DBSCAN algorithm has a better effect on fitting the seismic belt. This result has certain reference value for the study of seismic distribution. This paper also combines the magnitude and depth properties of seismic data, uses the Elbow method to find the best K value, and then classifies the dataset by K-means algorithm. We visualize the results and the distinction of each classification is clear.

Fig.1. Indonesia Earthquake aftermath [26 dec,2004]

Fig.2. Pakistan-Kashmir Earthquake aftermath [8 oct,2005]

2. Method

2.1 Data and Data Source

Data for this project is acquired from *https://earthquake.usgs.gov* which is an U.S based Geological Survey website and provides science about the natural hazards that threaten lives and livelihoods. The data we are collected from the website contains 206351 rows and each row has 22 individual attributes or columns of last six years (2015-2021) earthquake seismic data. Many of these attributes are not related to our project work such as “id”, “type”, “location source”, “status” and so on because they are not the characteristics of earthquake. After dropping them from the data set, we completed our necessary cleaning procedure. The data set is defined as df = {latitude, longitude, depth, mag, Time, day, month, year}.

latitude: The angular distance of a place north or south of the earth's equator, usually expressed in degrees and minutes.

longitude: the angular distance of a place east or west of the Greenwich meridian, usually expressed in degrees and minutes.

depth: In seismology, the depth at which an earthquake occurs is called depth of focus or focal depth. Earthquakes are labeled "shallow" if they occur at less than 50 kilometers of depth and they are labeled "deep" if they occur at 300-700 kilometers of depth. Earthquakes with depths from 50 kilometers to 300 kilometers are labeled “intermediate”.

mag: Earthquake magnitude is a measure of the “size,” or amplitude, of the seismic waves generated by an earthquake source and recorded by seismographs. It is measured by Richter scale.

Time: In which time the earthquake has started.

day, month, year: The date in which the earthquake is recorded.

2.2 GIS Technology and Machine Learning

The domain of GIS comprises of a system that is competent enough to capture, store as well as analyse and display geospatial data, spatial relationships and various patterns. GIS technology is of utmost importance in visualizing spatial data infrastructure. Remote sensing is converged with GIS to include satellite imagery for wide 95 applications.

2.2.1 Operations in GIS

There are multiple operations in GIS application to perform Geospatial analysis. Here we used two following operations.

- **Overlay Analysis:** It specifies an operation where two or more layers are overlaid [9]. Various points or polygons can be put together for performing geospatial analysis. Diverse layers generated during plotting of data points over a map are required to be overlaid and this operation is called overlay analysis. The data points can also be labelled during this operation to explain the data points overlaid on the map. Various types of overlay analysis exist such as weighted-overlay, map-overlay, and so on.

- **Heatmap Analysis:** The analysis of density of data points in a particular region is done by the generation of various heatmaps. The heatmap is created by using pseudo-colour codes to signify various regions as per density. Various colour intensities gradually decrease as the concentration of data points reduce [10]. Also, kernel density interpolation is a technique which uses various modal functions to generate a heatmap can be used.

2.2.2 Machine Learning

Machine learning is a sub-field of artificial intelligence where on providing input a specific problem we get a program as the output [11]. Various operations exist in machine learning such as regression, classification, association, clustering, ensembling, feature extraction, dimensionality reduction, principal component analysis, maximum likelihood estimation and so on. For our project we used K-means and DBSCAN clustering method to fit the seismic belts and for classification we also used K-means algorithm.

- **K-means clustering**

The spatial k-means clustering is an iterative approach to cluster various data points either based on the centroid values or on the Euclidean distance [12]. For data in raster format, centroids are chosen as the clustering parameter whereas for data in vector format, Euclidean distance is calculated for each data point to perform the spatial clustering.

The k-means algorithm divides a set of **N** samples **X** into **K** disjoint clusters **C**, each described by the mean $\boldsymbol{\mu_j}$ of the samples in the cluster. The means are commonly called the cluster "centroids"; note that they are not, in general, points from **X**, although they live in the same space.

The K-means algorithm aims to choose centroids that minimise the **inertia**, or **within-cluster sum-of-squares criterion**:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Inertia can be recognized as a measure of how internally coherent clusters are.

Algorithm:

- Take latitude and longitude in n seismic records to form a two-dimensional dataset D1={x_1,x_2,…., x_n}, where x_i is a sample in dataset D1. k samples { μ_1, μ_2 ,…., μ_k }are selected randomly among all samples in dataset D1 as the initial mean vector;
- Calculate the distance of each point [x_j($1\leq j \leq n$)] from each centroid [μ_i] respectively: d_{ij} = || x_j - μ_i $||_2$
- Assign each nearest data point [x_j] to its closest centroid, creating a cluster C_λ.
- Recalculate the position of the k centroids by the means of each independent dimensions of seismic data.
- Repeat the process 2 and 4, until the centroids are no longer move.

Elbow Method:

Here, we use the Elbow method to determine the optimal *k* value. The method selects the optimal *k* value by calculating the square sum of error. The formula is:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

Where C_i is the *i*-th cluster, *p* is the data element in C_i , m_i is the centroid of C_i , and the result of SSE represents the quality of the clustering result. The core idea of this method is that as the value of *k* increases, the number of clusters of samples increases, and the degree of aggregation of each cluster becomes higher. Then the value of SSE becomes smaller. When *k* is smaller than the actual number of clusters, the value of *k* increases, and the magnitude of SSE decreases greatly. When *k* is equal to the number of real clusters, if *k* increases again, the amplitude of SSE will decrease sharply, and the image tends to be stable.

- **DBSCAN Clustering**

  The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. The central component to the DBSCAN is the concept of *core samples*, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, min_samples and eps, which define formally what we mean when we say *dense*. Higher min_samples or lower eps indicate higher density necessary to form a cluster.

Algorithm:

- Initialize the core object collection $\Omega = Ø$;
- Traverse all seismic Data sample points to get the numbers of sample points in the ε - neighbourhood. If the number of one point is not less than MinPts, this sample point x_i is the core object;
- From any core point in the core object set, find out all the points with their density-reachable points to generate clusters. This process is iterated over and over until all core points are accessed.

  An algorithm is used here to determine the value of ε , along the following steps:

- For any point xi in the given dataset D, calculate the Euclidean distance d_{ij} to the remaining points, and sort it in ascending order to obtain the distance set M;
- Set the MinPts to k, and take the k-th distance from the distance set M as the k-distance of the point x_i;
- Calculate the k-distance of all points to form the k-distance set E;
- Sort the elements in E in ascending order, draw an image, to find out the k-distance value of the point x_ε where the change is the most intense, that is the desired ε value.

3. Data Analysis and Visualization

3.1 Overlay Analysis:

We took the last six years (2015-2021) of earthquake data which consists of 206351 earthquake points to visualize. So, we used an GIS software ArcGIS to visualize them with multiple layers. Firstly, we imported our csv file as shape file in the software then created our first layer which is base map of Ocean Base. After that we created second layer (Fig.3) with that shape file. We added two more additional layer (Fig.4) of seismic plat and active volcano to see the relationship with earthquake. We observed that earthquake density is higher around the seismic belt and active volcanos.

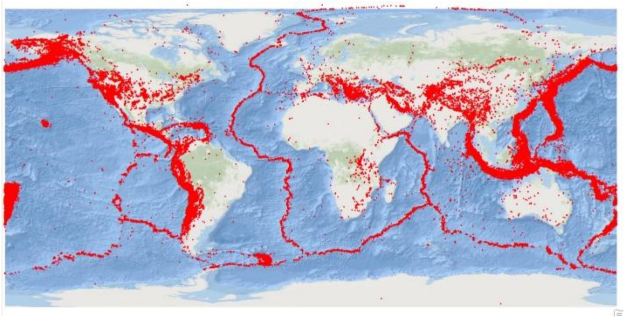


Fig.3 earthquake points (in red) on ocean base map

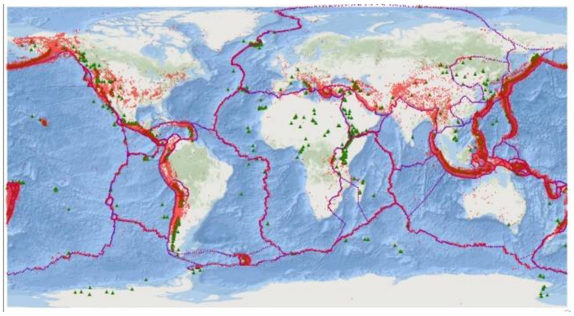
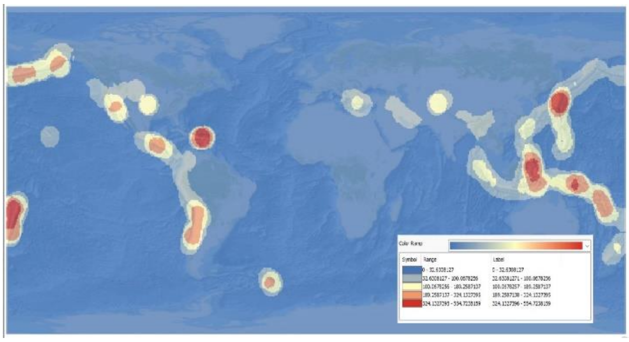


Fig.4 earthquake (in red), seismic belt (in blue) and volcano (in green) on ocean base map

3.2 Heatmap Analysis:

The analysis of density of earthquake data points in a particular region is done by a special heatmap tool in ArcGIS. In Fig.5 we can see highly dense area with red colour and lowly dense area with blue colour. In Fig.6 we can see highly dense area with black colour and lowly dense area with white colour. It is a Gray scale heatmap.



Symbol	Range	Label
🟦	0 - 32.6308127	0 - 32.6308127
▨	32.6308127 - 100.0678256	32.63081271 - 100.0678256
🟨	100.0678256 - 189.2587137	100.0678257 - 189.2587137
🟧	189.2587137 - 324.1327395	189.2587138 - 324.1327395
🟥	324.1327395 - 554.7238159	324.1327396 - 554.7238159

Fig.5 Heatmap (highly dense- red, lowly dense-blue)

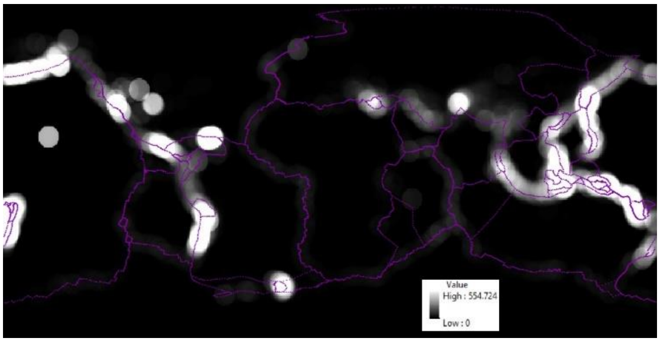


Fig.6 Gray-Scale Heatmap (highly dense- red, lowly dense-blue)

3.2 Basic Analysis

We used python for basic analysis of the seismic data. Using bar plot we can see that 2018 had the highest number of seismic records 44417 and 2017 had lowest number of seismic records 22722.

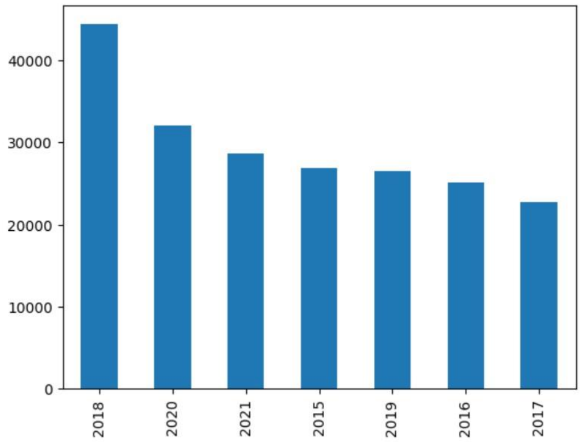


Fig.7 Bar plot (year VS number of earthquake)

Year	Seismic Records
2015	26853
2016	25116
2017	22722
2018	44417
2019	26506
2020	32079
2021	28658

Table.1 Year with seismic records

The next bar plot shows that month July had the highest number of seismic records 25742 and February had lowest number of seismic records 14496.

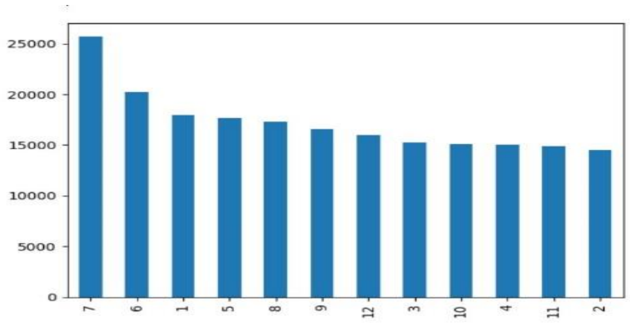
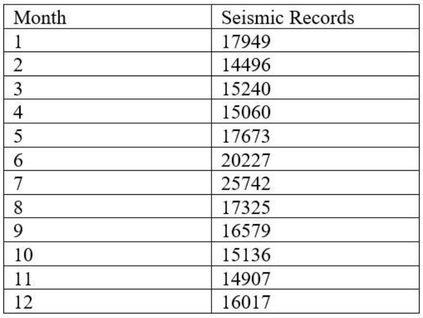


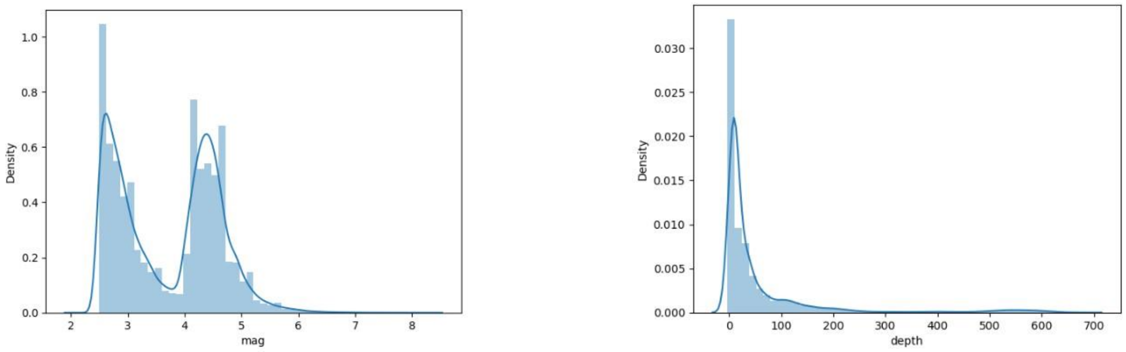
Fig.8 Bar plot (month VS earthquake)



Month	Seismic Records
1	17949
2	14496
3	15240
4	15060
5	17673
6	20227
7	25742
8	17325
9	16579
10	15136
11	14907
12	16017

Table.2 Month with seismic records

Now to see the distribution of magnitude and depth we ploted Fig.9 (a) magnitude against mag density and in (b) depth against depth density.



(a)

(b)

Fig.9 distribution plot: (a) mag against mag density, (b) depth against depth density

3.3 Fitting the seismic belt

After having published a first article in 1912, Alfred Wegener was making serious arguments for the idea of continental drift in the first edition of *The Origin of Continents and Oceans.* Afterward his idea give birth to tectonic plates movement. There are usually seven or eight "major" plates: African, Antarctic, Eurasian, North American, South American, Pacific, and Indo-Australian. The latter is sometimes subdivided into the Indian and Australian plates. Our work here is to fit the seismic plate boundaries with the help of some classic machine learning clustering method.

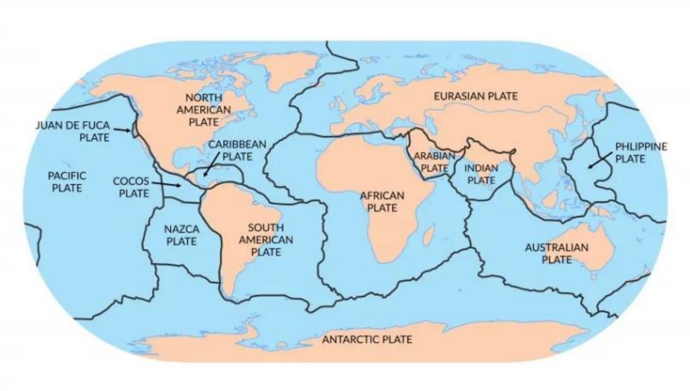


Fig.10 Alfred Wegener in Greenland in the winter of 1912–13.

Fig.11 Major tectonic plates

3.3.1 Fitting the seismic belt with the help of K-means Algorithm

For the first experiment we set the K value as default which is 3. The result can be seen in the Fig.12 that due to its algorithmic characteristics, the shape of the cluster is convex and cannot be divided into strips. The number of samples in the three cluster is 58034, 87331 and 60986 respectively. Their proportions are 28.12%, 42.32%, and 29.55%.

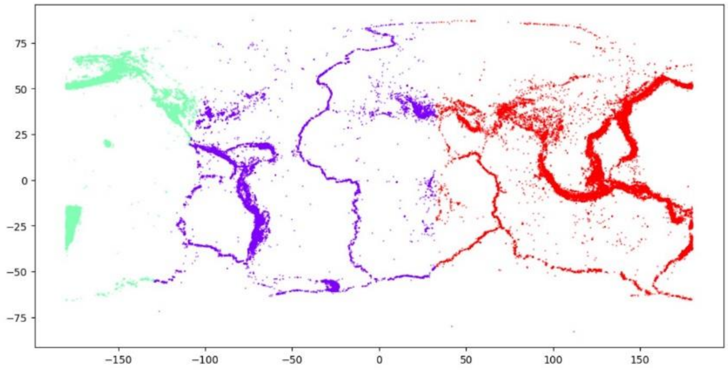
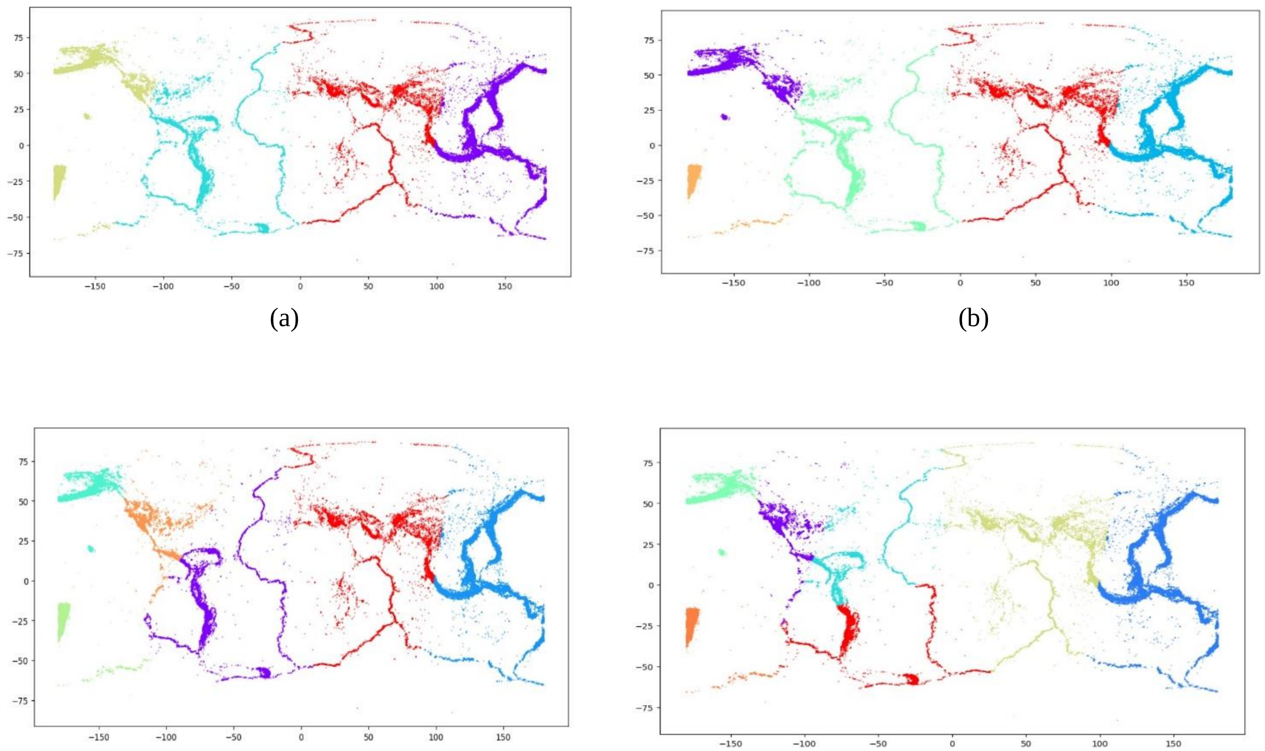


Fig.12 K-means Clustering result (k =3)

Combined with the division of seismic zones and the proportion of data in clustering, we think when the *k* value is 3, the results do not meet the expected effect. Considering that belts can still be divided in the three major seismic zones, the *k* values are set to 4, 5, 6, and 7 respectively to continue the experiment. The results are shown in Fig.13.



(a)

(b)

(c)

(d)

Fig.13 K-means clustering results (a) K = 4, (b) K = 5, (c) K = 6, (d) K = 7

It can be seen from the four subgraphs in Fig. 13 that as the value of *k* increases, the number of clusters increases. But the Western Pacific Seismic Belt is split, while the remaining clusters do not change. The East Pacific Seismic Belt and the Ridge Seismic Belt in the west are gathered as a cluster, the Eurasian Seismic Belt and the east of the Ridge Seismic Belt are gathered into a class. From the results, it does not meet the actual situation of seismic zone division, so the K-means clustering algorithm can’t fit the division of seismic zone well.

3.3.2 Fitting the seismic belt with the help of DBSCAN Algorithm

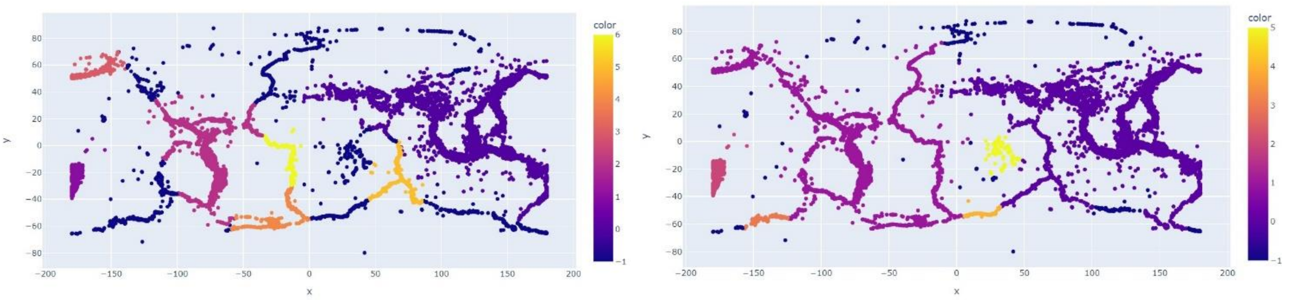
The ε values are obtained under the premise that the values of MinPts are 100, 200, 400, and 800 respectively by using the method in Section 2.2.2. The results are shown in the following table:

MinPts	ε	Number of cluster	Noise	SI
100	5	23	7.78	0.388
	10	6	1.26	-0.197
	15	2	0.304	0.299
	20	2	0.215	0.293
	40	1	0	
200	5	10	19.83	0.186
	10	10	6.511	0.421
	15	6	1.187	0.355
	20	2	0.415	0.267
	40	1	0	

400	5	9	26.18	0.154
	10	7	15.33	0.417
	15	7	6.99	0.431
	20	4	2.115	0.617
	40	1	0	
800	5	8	43.12	0.134
	10	5	26.06	0.468
	15	4	17.39	0.555
	20	5	10.36	0.422
	40	1	0	0.11

Table.3 Clustering results with different MinPts and ε

Based on the above clustering results, it can be found that for the same MinPts value, as the ε increases, the number of clusters decreases, the number of data identified as noise decreases. No matter the noise ratio is too large or too small, the clustering result is very poor for fitting the seismic belt. When the noise is too large, the amount of data participating in the clustering is small, and the result is incomplete. While the noise is too small, the number of clusters is too few and the accuracy is too low. When the value of MinPts is 400 or 800, since the minimum number of points in the neighborhood is too large, only regions with a particularly high density can be clustered for different ε values. At the same time, the proportion of noise in clustering changes drastically, so the effect of clustering is not ideal. Among the other two results, there are clusters with better fitting effects, as shown in Fig.14.



(a) (b)

Fig.14 DBSCAN clustering result: (a) MinPts = 400, ε =15 (b) MinPts = 200, ε =15

Fig. 14.(a) shows the clustering results for MinPts=400, and ε = 15. The East Pacific Seismic Belt is well restored, and the Western Pacific Seismic Belt is divided into three categories. Among them, the Eurasian Seismic Belt is represented by four categories, which is also in line with the actual situation. The Ridge Seismic Belt is relatively complete. Fig. 14.(b) shows the clustering results for MinPts=200, and ε =15, where the Western Pacific Seismic Belt has a good fitting. However, part of the Eurasian Seismic Belt is connected with the East Pacific Seismic Belt, and the range of which is too large. The Ridge Seismic Belt is still relatively complete. Since the dataset used in this paper does not have a label for the seismic belt to which it belongs, it can only be compared with the known seismic band division from the result graph. In summary, the DBSCAN algorithm can fit the distribution of the seismic zone well under the appropriate parameters, and the effect is in line with expectations.

3.5 Classification of Earthquake

First a scatter plot of depth and magnitude is drawn, as shown in Fig. 15. It can be seen from the figure that the seismic magnitude properties are very discrete. Under the same magnitude, the depth of the earthquake will also accumulate at some depth, and the density of it is high.

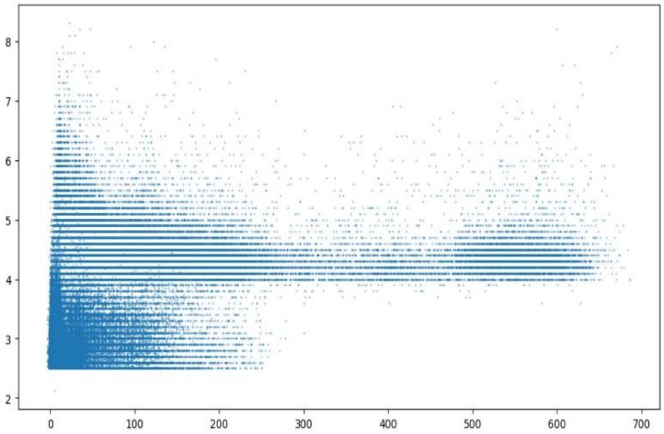


Fig.15 Depth-Magnitude scatter plot

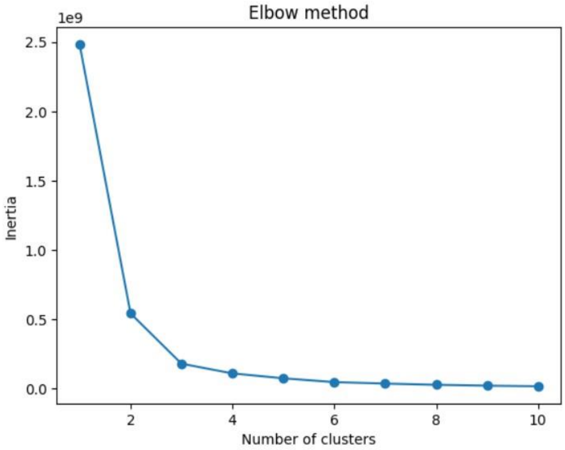


Fig.16 SSE-k diagram

The Elbow method mentioned in Section 2.2.2 is used here to determine the optimal *k* value. The data should be normalized by 0-means or the results will be affected. The relationship between SSE and *k* is shown in Fig.16. It can be seen from the figure that when the value of *k* is 1, 2, 3, the SSE drops sharply. And then *k* is 4, the line tends to be stable. Obviously, for the clustering of the dataset, 4 is the suitable *k* value. The clustering result is shown in Fig.17 and Table.4.

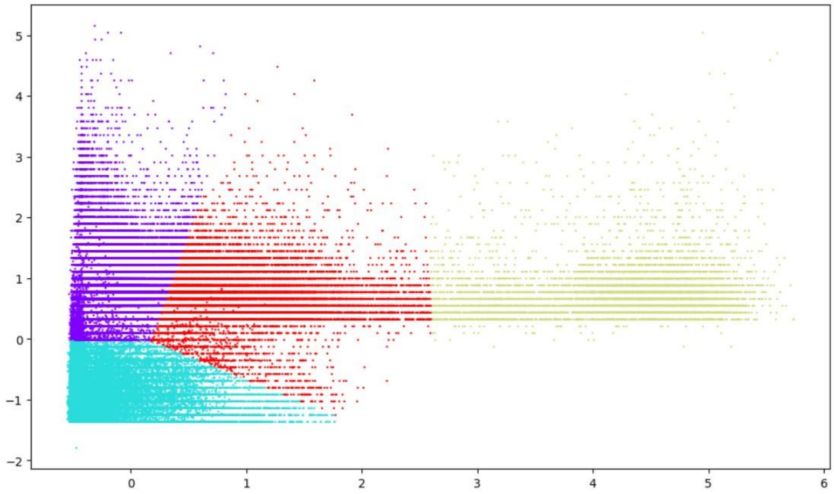
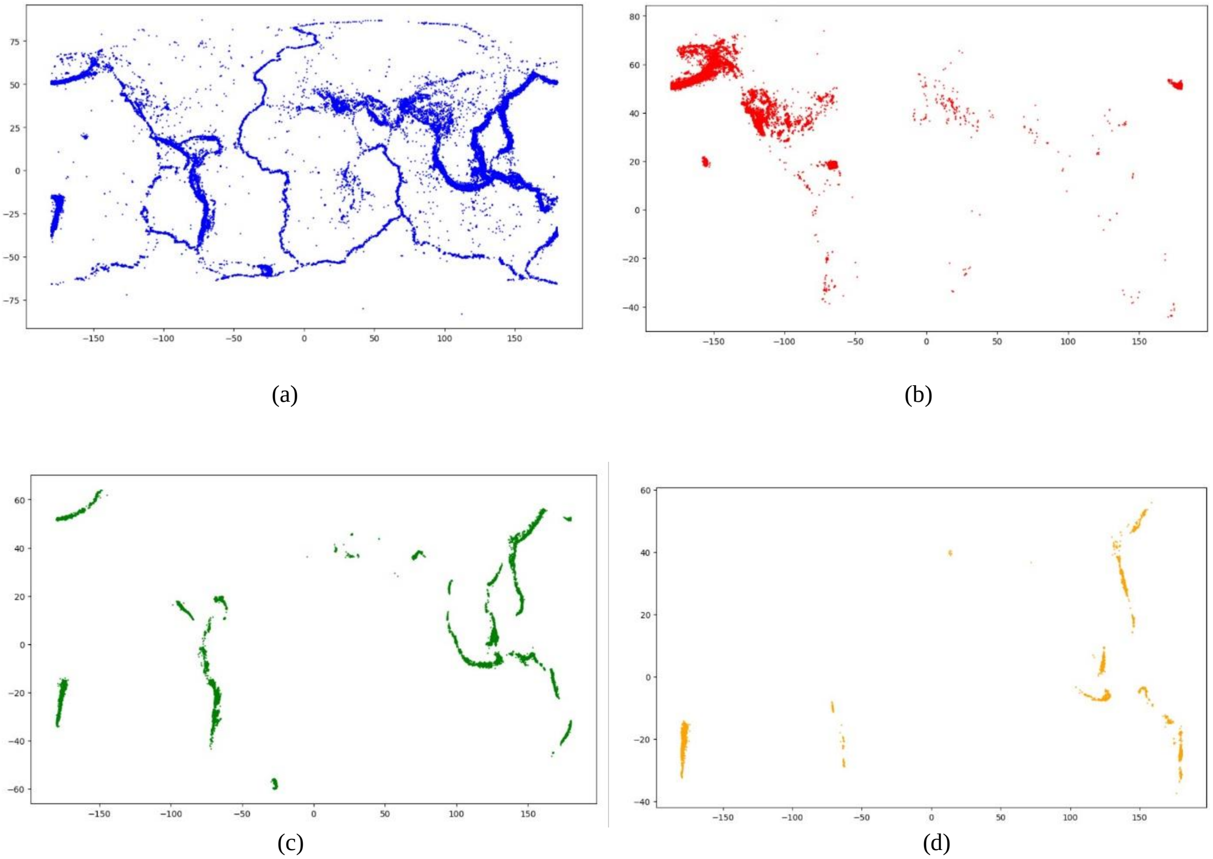


Fig.17 Depth-Magnitude clustering result

Cluster	Depth(standardized)	Magnitude(standardized)	Quantity
1	-0.29060718	0.91431649	80789
2	-0.32174948	-0.9641965	97257
3	0.9548029	0.68121476	19866
4	4.25067809	0.75420406	8439

Table.4 Depth-Magnitude clustering results

Among the results, the violate colour area is cluster 1, and the data ratio is about 39.1%. cyan colour area is cluster 2, whose data is about 47.13%, red colour is cluster 3, whose data is about 9.6%, yellow colour is cluster 4, and data of it is 4.08%. We can view the distribution of these clusters by setting them colours by their labels and redrawing the images, as shown in Fig.17.



(a)

(b)

(c)

(d)

Fig.18 Distribution of each Cluster results: (a) Cluster-1 (b) Cluster-2 (c) Cluster-3 (d) Cluster-4

It can be seen from Fig.17 and Fig.18 (a) and (b) that the depths of cluster centers in cluster 1 and cluster 2, that is, the mean values of depths are similar and low value, while the magnitudes of cluster centers are different, which is high for cluster 1 and low for cluster 2. The earthquakes in cluster 1 must have a clear distinction because of shallow depth and high magnitude. The location of the earthquake in this cluster-3 is highly concentrated and can be roughly divided into three parts. From the left the first part concentrated around Aleutian trench which is near the Gulf of Alaska. The Second part is concentrated around the Peru-Chile trench which is west side of the South Africa and the third part is concentrated around the Kurile-Japan trench, Mariana trench and Java trench. In cluster-3 the depth is shallow and average magnitude is high so that the seismic intensity is high in that area. The cluster-4 is highly concentrated around the Kurile-Japan trench, Mariana trench and Tonga trench. In summary, the effect is ideal by putting earthquakes and magnitude together in clustering analysis. This has certain significance for excavating seismic law and studying seismic distribution.

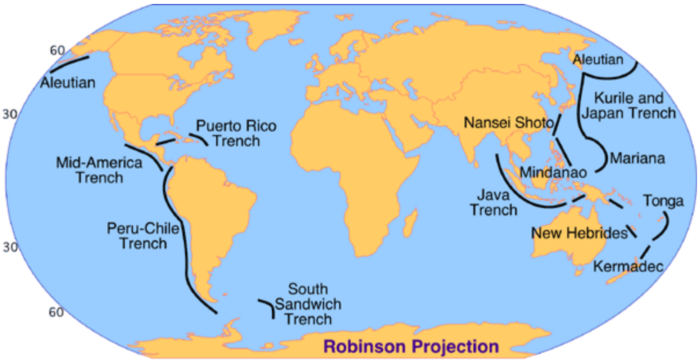


Fig.19 Trenches on world map

4. Conclusion

Earthquakes still cannot be predicted accurately at present, but the analysis of seismic data has been studied continuously. In this paper, two typical clustering algorithms K-means and DBSCAN are applied to the analysis of seismic data. We compare the application of two algorithms in seismic distribution research, fit and visualize the seismic belts. And then the earthquakes are clustered by focal depth and seismic magnitude, so that they are classified. The results show that it is completely feasible to apply the clustering algorithm to seismic data analysis. For the study of fitting seismic belts, the density-based DBSCAN algorithm is far superior to the K-means algorithm. At present, a large part of the prediction of earthquakes is based on the mining of existing earthquake data and the construction of models. Therefore, it is also practical to apply clustering algorithms to seismic distribution research. In the future, time series can be introduced into the research to extend the study of seismic distribution from the simulate surface to the space-time dimension. At the same time, we can introduce regression analysis method to partially predict the earthquake.