EasyChair Preprint

№ 8371

Earthquakes in Turkey Situation, Linear Regression and Clustering Using K-Mean Algorithm

Hunaida Avvad and Yousef Al Barjakly

June 26, 2022

EARTHQUAKES IN TURKEY

SITUATION, LINEAR REGRESSION AND CLUSTERING USING K-MEAN ALGORITHM

Hunaida AVVAD [1], Yousef ALBARJAKLI[2]

[1] *Faculty of Economics and Administrative Sciences - MIS Department , Izmir Bakircay University , Izmir , Turkey*

[2] *Faculty of Economics and Administrative Sciences - MIS Department , Izmir Bakircay University , Izmir , Turkey*

EARTHQUAKES IN TURKEY

SITUATION, LINEAR REGRESSION AND CLUSTERING USING K-MEAN ALGORITHM

Abstract

Unlike other natural disasters, earthquakes happen so frequent, this makes them most terrifying, and precautions would have to made in every aspect of human's lives, constructions, business and individual. Earthquake's prediction has been every geologist, scientists and mathematicians target for centuries. The goal of this research is to cluster Turkish cities using k mean algorithm. Linear regression was used to extrapolate the seismic activity in Turkey. This preliminary research cover also a descriptive part of earthquakes in Turkey. The dataset used is owned by Bogazici University and the analysis was conducted using Tableau public. We identified that earthquake count case clustering showed better accuracy than maximum magnitude case clustering, also linear regression line equation and chart predicted that the earthquake counts in Turkey is expected to increase by 11% in the next 10 years.

Keywords: *Earthquakes, Clustering, K-mean Algorithm, Linear Regression*

1 Introduction

Turkey is considered to by seismically active (1). Hazard level, has been classified as high and a number of devastating earthquakes did occurs over the past decades. There is more than a 20% chance of potentially damaging earthquake shaking in any project area in the next 50 years. Based on this information, the impact of earthquakes should be taken into consideration in all aspects of any construction projects, on business and on individual. The positive point is that large earthquakes are rare events.

Past decades a number of sudden and devastating earthquakes occurred in Turkey, Twain, and India (2). Many questions were raised, when do earthquakes occurred?, are their triggers?, what happens during an earthquake?.

Earthquakes occur due to several types of seismic waves generated at the earth focus (3). The primary waves that arrive at recording station P are similar in nature to sound waves, the next waves to follow are the secondary waves S, there are shear transverse waves and travel on lower speed than the P waves. There are two basic types of earthquakes; volcanic that is most common in Pakistan and the tectonic earthquake such as the earthquakes in Turkey.

Statistical science and data mining techniques have been used to analyze, describe, and predict an earthquake and earthquake sequences (4) (5) (6) (7) (8; 9).

In this research we will conduct a descriptive analysis on earthquakes in Turkey that will include earthquakes occurrence frequency, compare Turkey with the surrounding region as far as earthquake

counts and intensity, calculate the regression for the Turkish region and preform a clustering analysis for Turkish cities. The rest of the paper is organized as this; Section1 discusses the literature survey, section2 describes our methodology, section3 contains analysis and results, and section4 contains the conclusion and future work.

2 Literature Survey

Linear regression have been used to predict and describe varies aspect of the earthquakes. (10) Used lines of best fit to estimate maximum magnitude from graphs comparing historical earthquake magnitudes and lengths of associated surface ruptures. In his work, he investigated the linear regression or correlation models for making statistical extrapolation from data on historical events. The regression equation was:

$$M\text{œ}(L) \left[\frac{1}{n} + 1 + \frac{(\log L - \overline{\log L})^2}{\sum_{i=1}^{n} (\log L_i - \overline{\log L})^2}\right] = M(L) + s$$

M is the maximum earthquake magnitude and L is the rapture length.

(11) Used logistic general linear regression models (GLM) to represent the statistical relationships between the factors controlling landslides (such as epicenter distance, rainfall during Hurricane Otto, altitude, and slope) for Costa Rico event in 2016.The statistical analysis supports the existence of coupled earthquake-hurricane dynamics with higher landslide densities close to the epicenter.

(12) Used a cluster analysis which is a multivariate method that searches for patterns in a dataset by grouping the observations into clusters. The goal of this method is to find an optimal grouping for which the observations or objects within each cluster are similar (homogeneous). In their research, he used K-mean clustering approach to cluster earthquakes epicenters within the Bengkulu province (Indonesia) and surrounding areas.

(13) research was conducted on Ecuador 2016 earthquake of magnitude 7.8. This earthquake had 4389 aftershocks and lasted from April 2016 until July 2017. The researcher used GEO-K-mean clustering approach to conduct clustering analysis for purpose of identification of the number of clusters based on spatial localization from a geophysical point of view, and identify the geodynamical sense of the performed clusters, i.e to cluster the post-earthquakes based on the dynamic forces within earth.

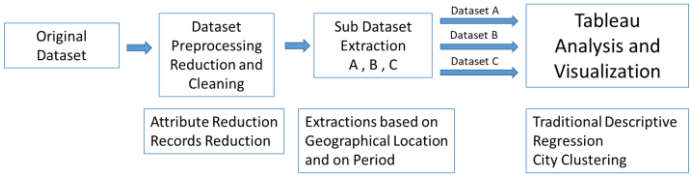
3 Methodology

In this quantitative research, we used a dataset from Kaggle.com. This dataset is owned and collected by Boğaziçi Üniversitesi Rektörlüğü and the covers the earthquakes from 1910-2017 for Turkey and the surrounding region.

We used Tableau (Public version). Tableau is a data visualization tool used for data analysis and business intelligence. Gartner's Magic Quadrant classified Tableau as a leader for analytics and business intelligence.

Figure 1. shows the work plan which consist of three stages, first stage of dataset preprocessing includes data cleaning and attribute reduction, the second stage of data preprocessing includes extracting 3 sub datasets from the dataset, detailes on this part are discussed in section 2.3 below, the third and final part is the analytical part, we used Tableau public to visuallize and analyze the sub datasets .

Figure 1. Work Plan



3.1 Dataset Description and Preprocessing

Boğaziçi Üniversitesi owns the dataset used, it can only be used for uncommercial issues, the dataset was downloaded from Kaggle.com. Table1. shows the original dataset summary.

Table 1. Original Dataset Summary.

Parameter	Description
Period	From 1910 to 2017

Covered regions	Turkey and all sounding countries
Number of Attributes	16 excluding earthquake ID
Number of records	24008

3.2 Attribute Reduction

Principal component analysis is the most widely known technique of attribute reduction by means of projection (PCA) (14). The purpose of this method is to obtain a projective transformation that replaces a subset of the original numerical attributes with a lower number of new attributes obtained as their linear combination, without this change causing a loss of information. The attributes that we kept are shown in Table 2. and these attributes will be using in the analysis and have the following properties:

- The removed attribute will not affect the dataset analysis results
- The kept attribute are the ones with no missing data

Table 2. Dataset Attributes.

Attribute	Remark
Date	Written as day-months-year from the right to left
Time	Written as 12:00 hours, all events were listed during the 12:00 hours
Country	/
City	City/Station
Region	Significant for Turkey
Xm	Earthquake magnitude expressed in Richter scale

3.3 Records Reduction and Sub Dataset extraction

The dataset covers years between 1910 and 2017l; the dataset includes records for all surrounding countries and all part of Turkey with a total of 24008 records. Therefore, for sake of this research we created three sub datasets. The sub datasets were created for varies reasons:

- To get a clearer picture when we want to explore the Turkish region alone.
- Reducing the period to 1970-2017 instead of 1910-2017 for sub dataset B and C was due to the believe that the accuracy and credibility of monitoring equipment might not be accurate during the early 20th century and specially during 1st and 2nd war

Table 3. shows the extracted 3 sub datasets from the original dataset and describes the differences between these three sub datasets and after attributes reduction.

Table 3. Sub Datasets Descriptions.

Attribute	Number of Records	Description and remarks
Sub Dataset A	11851	Only Turkish cities are included and covers the period of 1910-2017
Sub Dataset B	21558	All the regions are included and covers period 1970-2017.
Sub Dataset C	10837	Only Turkish cities are included and covers for period of 1970-2017

4 Analysis and Results

In this section, the analysis includes time earthquake count frequency, earthquake counts for the whole region covering all the period , Turkish cities earthquake counts and regional earthquake maximum magnitude over the period . Also, we applied linear regression analysis for Turkish region based on earthquake counts and in the last part of this section we used K-means algorithm to cluster Turkish cities based on two cases, case1 earthquake counts and case2 earthquake magnitude.

4.1 Descriptive Analysis

In the descriptive analysis we have worked on the frequency of earthquake count for Turkey between 1912-2017 using sub dataset A, regional earthquake counts between 1970-2017 using subset C, earthquake counts for Turkish cities between 1970-

2017, and regional magnitude from 1970-2017 for all the region using sub dataset B. Here are the details for the descriptive analysis part.

4.1.1 The frequency of earthquake over time

We used sub dataset A, the purpose of this analysis is to observe if the frequency of earthquakes is changing over the time.

Figure 2. Sub Dataset A -Earthquakes Count 1912-2017 Turkey

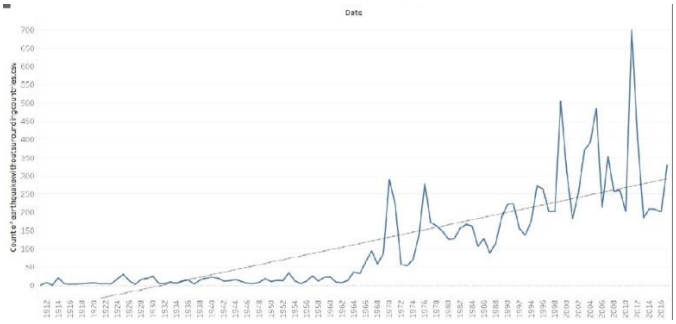


Figure 2. shows the earthquakes of magnitude over 3.5 count between 1912-217 for Turkey. The following can be observed from Figure 2.:

- Maximum count from the period was 699 in year 2011
- The extrapolated line for the graph shows that earthquakes counts are increasing over time.
- The trend line shows that the earthquakes count from 1912-the 60's were in the range from 1-34, for the rest of the period the range were 8-699 counts. This observation could possibly be because of poor earthquake recording and monitoring systems at early stages of the century. Also, it could be a true indication that earthquake counts are increasing.

4.1.2 Regional Earthquake counts 1970-2017

we used sub dataset C to check the regional earthquake counts for turkey and all surrounding areas. Figure 4. illustrates the earthquake counts in the period 1970 to 2017 in Turkey and all surrounding regions.

Figure 3. Sub Dataset C Earthquake Count 1970-2017 Turkish Regional Area

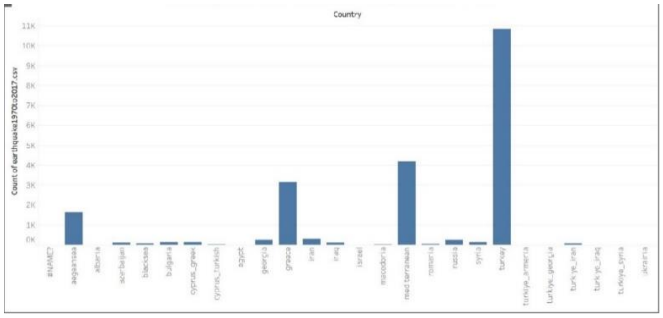


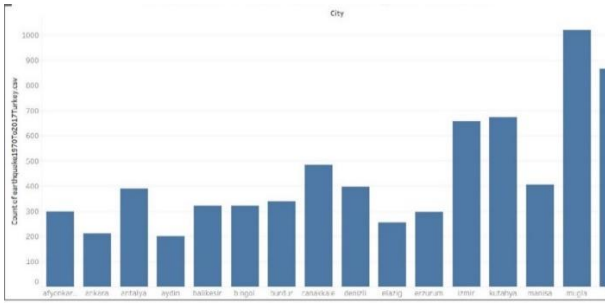
Figure 3. show that Turkey is most seismically active among the entire region followed by the Mediterranean and Greece. The rest of the area showed minor seismic activities. Two options as far as this observation:

- Possibly the data is not accurate as far as activities outside Turkey, it is possible to verify that by obtaining more sets of data from deferent sources.
- OR Turkey is considered as a more seismically active

4.1.3 Counts, Turkish Cities 1970-2017

we used sub dataset C to illustrate the Earthquakes counts in Turkish cities during the period 1970-2017. Figure 4. shows the Earthquakes counts in Turkish cities during the period 1970-2017, only cities with counts greater than 200 was filtered. The reasons are, to get a clearer visualization of the graph and main concern at this stage are the top 4 cities that are concerned as seismically active areas. We can see from Figure4 that cities with greatest counts are Mugla, Van, Katahya, Izmir, and Canakkale.

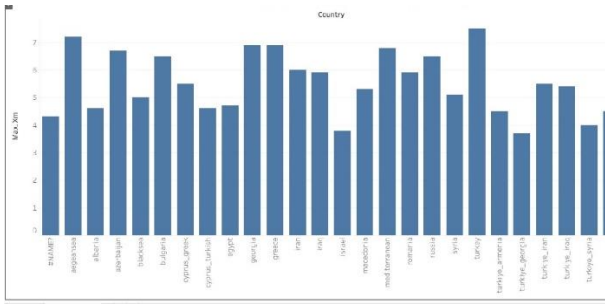
Figure 4. Sub Dataset C Earthquake Counts 1970-2017 for Turkish cities of counts greater than 200



4.1.4 Regional Magnitude 1970-2017 for all the region

We used sub dataset B in this step to check the magnitude for all regions together. Figure 5 shows that largest earthquakes magnitude occurred in the area between 1970-2017. Table 4. below shows greatest 6 countries with maximum magnitude occurred.

Figure 5. Sub Dataset B- Max Earthquake Magnitude 1970-2017 Region



From Figure 5.we can observe that Turkey encountered the maximum magnitude earthquake among all the regions and that the Further we get from Turkey; earthquakes magnitudes get less.

Table 4. Sub Dataset B Max Earthquake Occurred Between 1970-2017.

Attribute	Remark
Turkey	7.5
Aegean Sea	7.2
Greece	6.9
Georgia	6.9
Mediterranean	6.8
Azerbaijan	6.7

4.1 Linear Regression analysis

Linear regression has been used to predict and describe varies aspect of the earthquakes (10). The goal from this analysis is to calculate the regression line and to extrapolate the future earthquakes counts in Turkey. We used Tableau to calculate and plot the regression line.

The regression line for a simple linear equation is:

$$Y = w.X + b + E$$

In our case Y is Earthquakes counts, X is the year, and w are coefficient of regression and b is the line intercept with the y access.

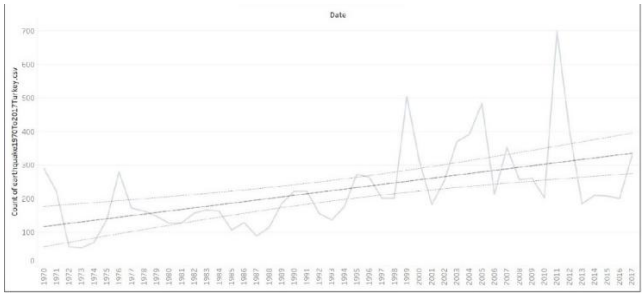
E is the error constant and have a different value for every point on the chart

$$E_i = Y_i - w.X_i - b$$

The equation was obtained from Tableau as:

$$Y = 4.64427 . X - 9032.6 + E$$

Figure 6. Sub Dataset C- Counts in Turkey 1970-2017 Regression and Trend Line



4.2 Clustering Analysis of Turkish Cities

We used sub dataset C for both below casesand K-mean algorithm as a clustering technique. The purpose is to detect and visualize the number and details of the clusters detected. The clustering analysis will be conducted for two cases:

- Case1: Clustering Cities in Turkey based on earthquake counts.
- Case2: Clustering Cities in Turkey based on earthquake maximum magnitude.

Case1:

Figure 7 illustrates the number of clusters detected by Tableau tool based on earthquake counts using Sub Dataset C. Table 5. shows the clustering summary and cluster contens for each cluster.

Figure 7. Sub Dataset C for City Clustering Based on Count – Turkey 1970-2017

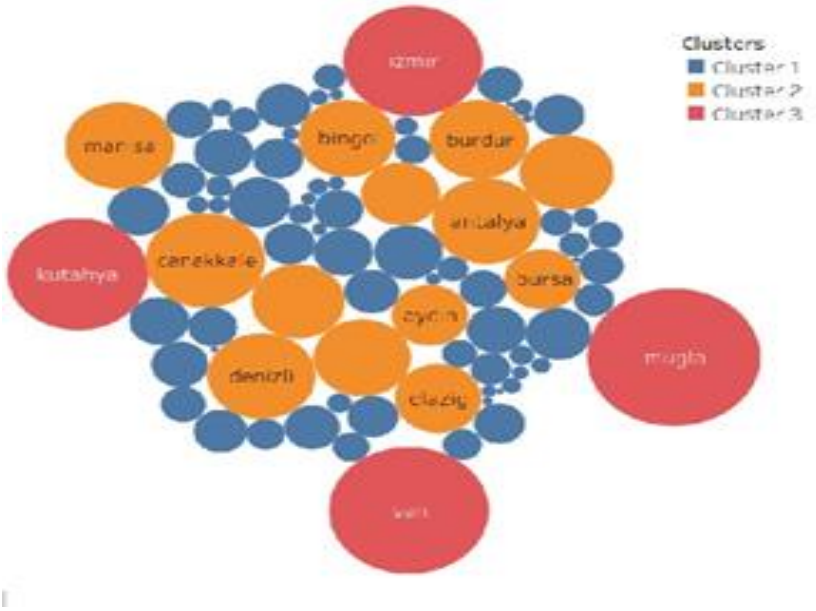


Table 5. Case1 Clusters Summary and Contents Based on Count.

Cluster	Count	Centers Count	Content
Cluster 1	67	52.284	Nigde, Mus, Mersin, Sakarya, Sivas, Kahramanmaras, Isparta, Sirnak, Zonguldak, Yozgat, Marmara-Denizi, Malatya, Konya, Kacaeli, Tekirdag, Sanliurfa, Tunceli, Tokat, Siirt, Osmaniye, Yalova, Usak, Hakkari, Nevsehir, Duzce, Gumeshane, Kars, Ordu, Karabuk, Karaman, Giresun, Amasya, Gaziantep, Aksaray, Agri, Igdir, Sinop, Erzincan, Mardin, Edirne, Corum, Cankiri, Kirsehir, Kirklareli, Kilis, Samsun, Kayseri, Trabzon, Kastamonu, Izmit, Rize, Istanbul, Hatay, Diyarbakir, Ardahan, Artivin, Bartin, Batman, Bayburt, Bilecik, Eskisehir, Bitlis, Bolu, Adana, Adiyaman.
Cluster 2	13	316.62	Aydin, Antalya, Denizli, Burdur, Bingol, Canakkale, afyonKarahisar, Erzurum, Elazig, Manisa, Bursa, Balikesir, Ankara.
Cluster 3	4	804.25	Van, Izmir, Kutahya, Mugla.

Case2:

Figure 8. illustrates the number of clusters detected by Tableau for Turkish cities based on earthquake magnitude using Sub Dataset C. Table 6. shows the clustering summary and cluster contens for each cluster.

Figure 8. Sub Dataset C Number of Clusters Based on Magnitude – Turkey 1970-2017

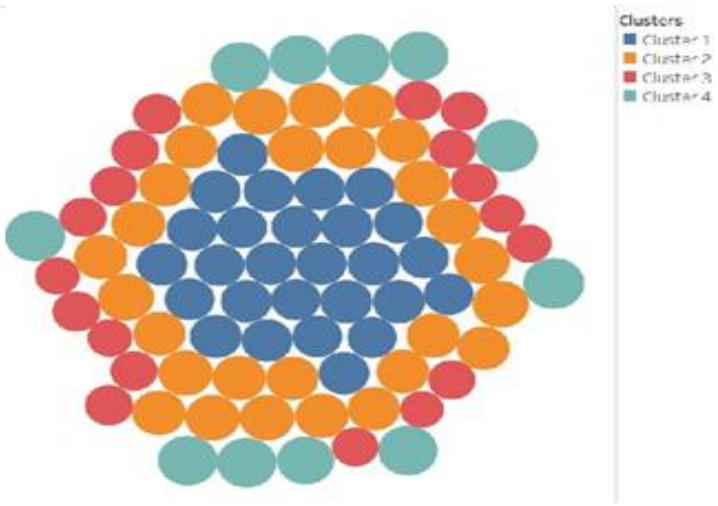


Table 6. Case2 Clusters Summary and Contents Based on Magnitude.

Cluster	Count	Centers Count	Content
Cluster 1	27	4.9815	Tokat, Tekirdag, Sivas, Yolova, Konya, Igdir, Isparta, Istanbul, Kayseri, Bitlis, Adiyaman, Gumushane, Kirsehir, Kahramanmaras, Bilecik, Bayburt, Bolu, Kirkkale, Rize, Bursa, Corum, Izmit, Mus, Mersin, Marmara-Denizi, Samsun, Zonguldak.
Cluster 2	28	5.6036	Manisa, Malatya, Izmir, Siirt, Gaziantep, Ardahan, Antalya, Hatay, Elazig, Denizli, Ankara, Sirnak, Amasya, Sanliurfa, Sakarya, Osmaniye, Eskisehir, Balikesir, Aydin, Usak, Tunceli, Hakkari, Cankiri, Burdur, Erzurum, Agri, Afyon Karahisar.
Cluster 3	18	4.2111	Sinop, Trabzon, Aksaray, Artivin, Kirklareli, Mardin, Giresun, Nevsehir, Yozgat, Nigde, Bartin, Batman, Edirne, Ordu, Kilis, Karabuk, Karaman, Kastamonu.
Cluster 4	11	6.8818	Mugla, Kocaeli, Diarbakir, Canakkale, Kutahya, Erzincan, Bingol, Adana, Kars, Duzce, Van.

5 Conclusion and Future Work

In the descriptive part of earthquake counts over time (regional), we noticed that there is a gradual and continues increase in earthquake frequency and specially in the last 40 years in Turkey and in the Turkish region. Furthermore, in the descriptive of counts and magnitude regional we noticed that Turkey is the most seismically active country among the whole region as far as earthquakes counts and magnitude. The third section of the descriptive covered counts and magnitude among Turkish cities; and here we saw that the maximum counts among Turkish cities are Van and Mugla. Although the earthquakes are distributed in most Turkey and areas with low or no seismic activity is very limited.

During our experiment to cluster the Turkish cities using k-mean algorithm, we got that the number of clusters detected based on earthquake counts was 3, but sum of squares is 2.8, as the number of clusters detected when using magnitude was 4 but sum of squares was 4.2, which is much higher than count case, which tells that using earthquake counts was more accurate method that using Magnitude i.e the degree of deviation was less when using earthquake counts.

Linear regression equation was calculated by Tableau to be Y=4.46X-9032+E and according to the regression line extrapolation and based on E=0, the earthquake counts are expected to increase by 11% every 10 years.

For future work, our current research showed many future research gabs that can be fulfilled by researchers; Time Series Pattern Recognition for earthquake sequences can be conducted using machine learning and chart pattern recognition to classify earthquake types (i.e. main shock, aftershock) for several locations. During our analysis we noticed a kind of relation between earthquake counts and earthquake magnitude, finding a correlation was outside the scope of this research, but it is important to pursue in future work.

The clustering process that was conducted in this research took the whole period (1970-2017), results was interesting, but it left us with some questions in respect to changes in clustering distribution over time.

References

1. Think Hazad. [Çevrimiçi] https://thinkhazard.org/en/report/249-turkey/EQ.
2. Hiroo Kanamori, Emily Brodsky. *The Physics of Earthquakes* . basım yeri bilinmiyor : American Institute of physics, 2001.
3. H.S.Virk, Vivek Walia. Earthquakes , cases , precursors and predictions. *The national Science magazine.* 1976.
4. *Aftershock and Earthquake Statstics.* Atsu, Tokuji. 1970, Journal of the factulty of science.
5. A.Vecchio, V.Carnone , L.Sorriso-valvo , C.De Rose , P.Harabaglia. Statstical Properties of Earthquake clustering . *Nonlinear Process in Geophysics* . 2008.
6. *Siesmic Data Classification using machine learning.* Wenrui Li, Nakshatra Narvekar , Nitisha Raut Birsen Sirkeci , Jerry Gao. 2018. IEEE Fourth International Conference for big data computing servises and applications .
7. *Clasification of long term very long eriod volcanic earthquakes at whakaari/white island volcano new zealand* . Iseul {Park, Arthus Jolly , Ivan Lokmer , Ben Kennedy. basım yeri bilinmiyor : Springer, 2020.
8. shumway, Robert. Time frequency clustering and discriminant analysis. *Statistics and probability letter.* 2003.
9. Aaron Bostrom, Anthony Bagnall. *A shapelete transform for multivariate time series classifcation.* basım yeri bilinmiyor : UK engineering and physical science research council, 2018.
10. *Application of linear statistical models of earthquake magnitude versus fault length in estimating maximum expectable earthquakes.* Mark, Robert K. 1977, The Geological society of Merica.
11. Quesada-Román, Adolfo , Berny Fallas-López , I Karina Hernández-Espinoza , Markus Stoffel. *Relationships between earthquakes, hurricanes,.* basım yeri bilinmiyor : Springer, 2019.
12. *K-Means cluster analysis in earthquake epicenter clustering.* Pepi Novianti, Dyah Setyorini , Ulfasari Rafflesia. 2017, International Journal of Advances in Intelligent Informatics.
13. —.Fernando Mato, Theofilos Toulkeridis. 2017, IEEE Symposium Series on Computational Intelligence (SSCI).
14. Vercellis, Carlo. *Business Intellegence.* basım yeri bilinmiyor : Wiley and sons ltd, 2009.