Journal of Data Science,18(1). P. 44 – 55,2020
DOI:10.6339/JDS.202001_18(1).0002

# Application of Statistical Control Charts to detect unusual frequency of earthquake in the world

Fariha Taskin, Mohammad Shahed Masud

*Institute of Statistical Research and Training (ISRT),*

*Dhaka University, Dhaka, Bangladesh.*

## ABSTRACT

Earthquake in recent years has increased tremendously. This paper outlines an evaluation of Cumulative Sum ($CUSUM$) and Exponentially Weighted Moving Average ($EWMA$) charting technique to determine if the frequency of earthquake in the world is unusual. The frequency of earthquake in the world is considered from the period 1973 to 2016. As our data is auto correlated we cannot use the regular control chart like Shewhart control chart to detect unusual earthquake frequency. An approach that has proved useful in dealing with auto correlated data is to directly model time series model such as Autoregressive Integrated Moving Average ($ARIMA$), and apply control charts to the residuals. The $EWMA$ control chart and the $CUSUM$ control chart have detected unusual frequencies of earthquake in the year 2012 and 2013 which are state of statistically out of control.

**Keywords**: CUSUM chart, EWMA chart, ARIMA.

# 1. Introduction

Earthquake is one of the most horrific and devastating natural phenomena in the world. It is the sudden, rapid shaking of the earth, caused by the breaking and shifting of subterranean rock as it releases strain that has accumulated over a long time. Earthquakes may damage household items, building to move off foundations or collapse, damage roads, bridges and dams, cause fires and explosions. They may also trigger landslides, avalanches and tsunamis.

The origin of earthquake is as old as the origin of the Earth, but due to lack of knowledge and scientific instruments it was tremendous challenge for ancient scientists to collect and analysis the earthquake data. Now-a-days the advancement of science and technology makes it easy to collect and analysis the earthquake data. Chen et al. (2010) analysed earthquake data with the help of frequency time analysis. Machado and Lopes (2013) analysed global earthquake data covering the period from 1962 up to 2011. Gupta and Gupta (2016) analysed earthquake data with the help of Big Data technology and visualized with the help of Tableau in India from 1800 to 2014. A multivariate non-parametric hazard model (Ata and Ozel, 2011) was used to analyse 111 destructive earthquakes having magnitude greater than 5 between the years 1903 to 2009 in Turkey. Reyes (2013) studied the spatial distribution of cluster associated to the aftershocks of the megathrust Maule earthquake for the year 2010. Besides the analysis of earthquake data, there are some studies that predict the occurrences of large scale earthquakes. For example, Amei et al. (2012) predicted a total number of 12 large scale earthquakes based on the worldwide earthquake data during 1986 to 2009. Alam (2015) forecasted the earthquake behaviour in Indonesia based on the earthquake data during the year 1980 to 2007. Last et al. (2016) predicted the magnitude of the earthquakes in the area of Israel and its neighbouring countries using the earthquake data from the year 1983 to 2010.

Recently, much research in the literature has focused on whether there is an increase in the frequency of earthquake occurrences. Any increase in the frequency of earthquakes is due to climate change in the world (Yi˘giter, 2012). There is considerable debate on whether climate change really does increase the frequency of natural disasters such as earthquakes and volcano eruptions. In many studies, it is emphasized that there is serious concern about impact of climate change on the frequencies of hazardous events (Lindsey, 2007; Mandeville, 2007 etc.).

The earth has experienced several earthquakes starting from the end of 19th century. Although this amount has been increased a little higher than the previous years, statistical evidence is thus required to determine whether this recent number of earthquake is just a random phenomenon or a genuine shift. Statistical control chart is a technique that can be used to determine the shift in number of earthquake. This technique was used (Justin et al. 2012) to detect climate change in Masvingo city in Zimbabwe.

The standard assumptions that are used to apply control charts are that the data are normally and independently distributed. When these assumptions are satisfied, one may apply conventional control charts and draw conclusions. The conventional control charts do not work well and give misleading results if the data exhibit even low levels of correlation over time (Bisgaard and Kulahci 2005). An approach that has proved useful in dealing with auto correlated data is to directly model time series model such as Autoregressive Integrated Moving Average (ARIMA), and apply control charts to the residuals (Montgomery and Mastrangelo, 1991). The purpose of this study is to determine whether the recent number of

earthquake is just a random phenomenon or a genuine shift using statistical control chart techniques. The earthquake data is an auto correlated data. So to detect shift in number of earthquake, we identify an appropriate $ARIMA$ model to the earthquake data and then apply the residuals of this $ARIMA$ model to Cumulative Sum Control Chart (Page, 1961) and Exponentially Weighted Moving Average (Roberts, 1959).

## 2. Methodology

### 2.1 ARIMA time series analysis

In statistics and econometrics and in particular time series analysis, an autoregressive integrated moving average ($ARIMA$) model is a generalization of an autoregressive moving average ($ARMA$) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).

ARIMA model is denoted by $ARMA\ (p, q)$ where $p$ represents autoregressive terms and $q$ represents moving average terms. Combine both $p$ autoregressive terms and $q$ moving average terms, $ARMA\ (p, q)$ with mean $\mu$ can be written as:

$$X_t = \mu + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \qquad (1)$$

where $X_t$ is the original series, for every t. The $\phi_i$are the parameters of the autoregressive part of the model and $\theta_i$ are the parameters of the moving average part. Assumed that $t$ is independent of $X_{t-i};\ i = 1, \ldots, p$. If the process is non-stationary then the model will be written as $ARIMA\ (p, d, q)$.

Modelling an $ARMA\ (p, q)$ process requires stationarity. A stationary process has a mean and variance that do not change over time and the process does not have trends. Stationarity test can be performed in many ways. In this study we perform the graphical test and the Augmented Dickey-Fuller (ADF) test.

In graphical procedure, observations are plotted against the time. If there is any trend of increasing or decreasing, then it will be interpreted as that the process is non-stationary and then an appropriate procedure has to be taken to achieve stationarity. However, if there is no trend, then the process will be interpreted as stationary.

The most widely used test to check stationarity is the Dickey-Fuller (DF) test. Assuming an $AR(1)$ model as below:

$$\Delta X_t = \gamma^* X_{t-1} + e_t. \qquad (2)$$

we test the null hypothesis that $\gamma^* = 0$, meaning that the time series $X_t$ is not stationary. In addition to the model (2) above, a drift $\mu$ and additional lags of the dependent variable can be added:

$$\Delta X_t = \mu + \gamma^* X_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta X_{t-j} + \epsilon_t. \qquad (3)$$

The augmented Dickey-Fuller test evaluates the null hypothesis that $\gamma^* = 0$. The model (3) will be non-stationary if $\gamma^* = 0$. The model with a time trend can be considered as:

$$\Delta X_t = \mu + \beta t + \gamma^* X_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta X_{t-j} + \epsilon_t. \quad (4)$$

Then we test the hypothesis that $\beta = 0$ and $\gamma^* = 0$. Again, the model will be nonstationary if $\gamma^* = 0$.

After checking stationarity, model identification is required. In this paper we use two procedures for model identification. One is graphical procedure (ACF and PACF plot). And another is Akaike Information Criterion (AIC).

An ARIMA model can be chosen upon inspection of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). This approach relies on the following facts: (i) the ACF of a stationary AR process of order p goes to zero at an exponential rate, while the PACF becomes zero after lag p. (ii) for an MA process of order q the theoretical ACF and PACF exhibit the reverse behaviour (the ACF truncates after lag q and the PACF goes to zero relatively quickly). It is usually clear to detect the order of an AR or MA model. However, with processes that include both an AR and MA part the lag at which they are truncated may be blurred because both the ACF and PACF will decay to zero. One way to proceed is to fit first an AR or MA model (the one that seems more clear in the ACF and PACF) of low order. Then, if there is some further structure it will show up in the residuals, so the ACF and PACF of the residuals is checked to determine if additional AR or MA terms are necessary.

The Akaike information criterion was developed by Hirotugu Akaike (Akaike, 1974), originally under the name "an information criterion". The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. It is based on information theory, but a heuristic way to think about it is as a criterion that seeks a model that has a good fit to the truth but few parameters. It is defined as:

$$AIC = -2(ln(likelihood)) + 2K. \quad (5)$$

Where likelihood is the probability of the data given a model and K is the number of free parameters in the model. AIC scores are often shown as $\Delta AIC$ scores, or difference between the best model (smallest AIC) and each model (so the best model has a $\Delta AIC$ of zero).

For further use of the model, first we have to check whether the fitted model is appropriate. Whether the parameters of the model are significant or not is need to be checked as well as if the residuals are white noise or not. Whether the parameters are significant, can be examined by p-values and whether the residuals are white noise, can be checked by Ljung-Box statistic and the ACF plot of residuals. If the plot shows that the residuals stay within the limits, then it is said that residuals are white noise.

The Ljung-Box test is a type of statistical test of whether any of a group of autocorrelations of a time series is different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. Here, the null hypothesis is of data that are independently distributed and the alternative hypothesis is of data that are not independently distributed; they exhibit serial and the test statistic is:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k{}^2}{n-k}, \quad (6)$$

where, n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under the null hypothesis that the series is white noise (data are independently distributed), Q has a limiting $\chi^2$ distribution with p degrees of freedom.

## 2.2 Cumulative Sum(CUSUM) chart

The CUSUM chart directly incorporates all the information in the sequence of sample values by plotting the cumulative sums of observations of the deviations of the sample values from a target value. Let $y_i$ be the $i^{th}$ observation, which has a normal distribution with $\mu_0$ and standard deviation $\sigma$. Now if $\mu_0$ is the target, the CUSUM control chart is formed by plotting the quantity $C_i = \sum_{j=1}^{i}(y_i - \mu_0)$ against the sample number $i$. $C_i$ is called the cumulative sum up to and including the $i^{th}$ sample. The tabular CUSUM works by accumulating derivations from $\mu_0$ that are above target with one statistic $C^+$ and accumulating derivations from $\mu_0$ that are below target with another statistic $C^-$. The statistics $C^+$ and $C^-$ are called one-sided upper and lower CUSUMs, respectively. They are

$$C_i^+ = max[0,\ y_i - (\mu_0 + K) + C_{i-1}^+], \quad (7)$$

$$C_i^- = max[0,\ (\mu_0 - K) - y_i + C_{i-1}^-], \quad (8)$$

where the starting values are $C_0^+ = C_0^- = 0$. K is usually called the reference value and it is often chosen about halfway between the target $\mu_0$ and the out-of-control value of the mean $\mu_1$, and it can be calculated by

$$K = \frac{\delta}{2}\sigma = \frac{|\mu_1 - \mu_0|}{2} \quad (9)$$

$C_0^+$ and $C_0^-$ both are greater than K and if either $C_0^+$ or $C_0^-$ exceed the decision interval H, the process is considered to be out of control. We define $H = h\sigma$ and $K = k\sigma$, where $\sigma$ is the standard deviation of the sample variable used in forming the CUSUM. Using $h$ = 4 or $h$ = 5 and $k = \frac{1}{2}$ will generally provide a CUSUM that has good ARL properties against a shift of about $1\sigma$ in the process mean.

## 2.3 Exponentially Weighted Moving Average (EWMA) Control Chart

The EWMA control chart is approximately the same as CUSUM control chart and in some cases it is easier to set up and operate. This chart considers all previous points using a weighting factor that makes the outcome more influenced by recent points.

The EWMA $z_i$ is computed sequentially as a linear interpolation between the present observation $y_i$ and $z_{i-1}$, the previous EWMA. So the exponentially weighted moving average is defined as

$$z_i = \lambda y_i + (1 - \lambda)z_{i-1}. \quad (10)$$

Where, $0 < \lambda < 1$ is a weighted constant and the starting value is the process target, i.e. $z_0 = \mu_0$. Sometimes the average of preliminary data is used as the starting value of EWMA, so that $z_0 = \bar{y}$.

The EWMA is sometimes called a geometric moving average (GMA). The EWMA is used widely in time series modelling and forecasting. Since the EWMA can be viewed as a weighted average of all past and current observations, it is very insensitive to the normality assumption. It is therefore an ideal control chart to use with individual observations.

If the observations $y_i$ are independent random variables with variance $\sigma^2$, then the variance of $z_i$ is

$$\sigma_{z_i}^2 = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]. \tag{11}$$

By plotting $z_i$ versus the sample number $i$ (or time), the EWMA control chart would be constructed. The UCL, CL and LCL for the EWMA control chart are as follows:

$$UCL = \mu_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]}, \tag{12}$$

$$CL = \mu_0, \tag{13}$$

$$LCL = \mu_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]}, \tag{14}$$

where factor L is the width of the control limits. $L = 3$ (the usual three-sigma limits) works reasonably well.

## 3. Description and analysis of earthquake data

### 3.1 Data description

United States Geological Survey (USGS) is one of the providers of earthquake data from all over the world. The data provided by USGS for year 1900 to 2016 contains too many variables. Our variables of interest are year of occurrence and magnitude. Figure 1(a) displays the total number of earthquake for different magnitudes and Figure 1(b) displays the total number of earthquake for different years.

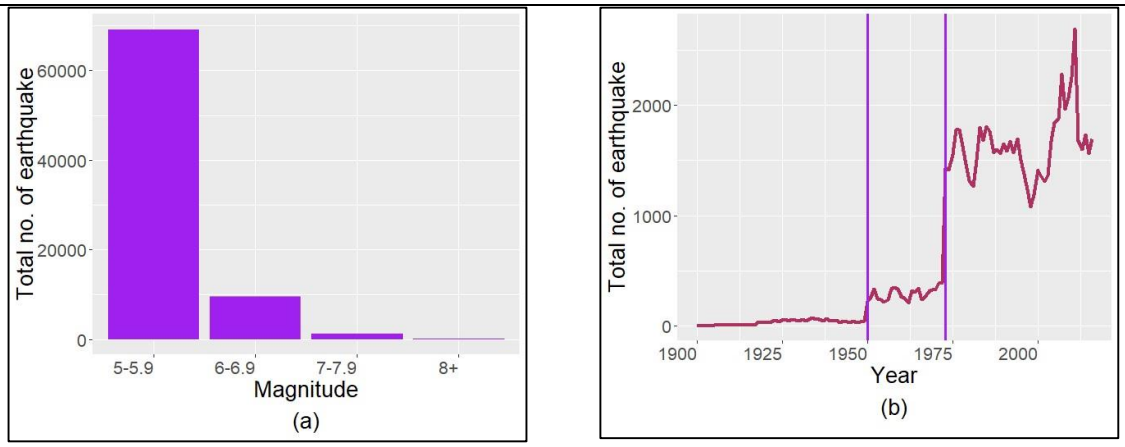


Figure 1: (a) Total number of earthquake for different magnitudes and (b) Total number of earthquake for different year.

The time series plot for the earthquake data (Figure 1(b)) indicates three significant shifts of number of occurrences such as 1900 to 1949, 1950 to 1972 and 1973 to 2016. We observe that the frequency of earthquake from 1900 to 1949 is very low and at the year 1950 there is a sudden shift. This shift continued for year 1950 to 1972. At year 1973 there is a very large shift in the frequency. Thus from 1973 the new shift has continued until present year.

Though from the graphical method it is quite clear that there are two shifts in the mean, however we can check this by performing global mean test (ANOVA). The test of hypothesis is: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a$: at least two are not equal. Where $\mu_1$, $\mu_2$, $\mu_3$are the mean of the number of earthquake for year 1900-1949, 1950-1972, 1973-2016 respectively.

Table 1: Analysis of variance table

| Frequency Group | DF | Sum sq | Mean sq | F-value | p-value |
|---|---|---|---|---|---|
| Factor effect | 2 | 64370424 | 32185212 | 914.27 | < 0.001 |
| Residuals | 114 | 4013175 | 35203 | | |

From Table 1 we can say that, as the p-value is less than the significant level $\alpha = 0.05$, we reject the null hypothesis. Thus we conclude that there is mean difference in year 1900-1949, 1950-1972 and 1973-2016 for total. As the number of occurrences is higher during the period 1973 to 2016, in this study we analyse this period.

### 3.2 Analysis of the data for 1973 to 2016

To analyse the time series earthquake data first we check for stationarity of the data. Figure 2 shows the time against frequency of the earthquake data. From the figure we find that sometimes the frequency of earthquake is increasing and sometimes decreasing and the mean and variance are not constant. And from ADF test (Table 2), we observe that for zero difference the p-value is insignificant. So we cannot reject the null hypothesis. That means, both techniques provide the same results, that the earthquake data is nonstationary.

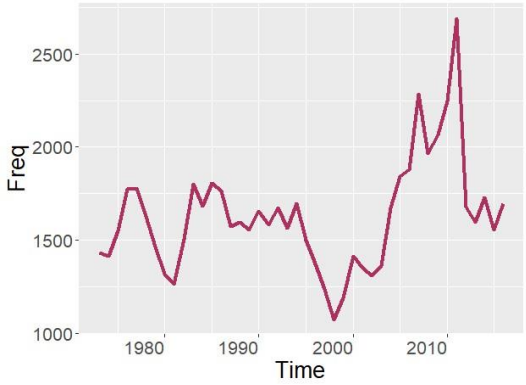


Figure 2. Plot of Frequency vs time

Table 2: Results of Augmented Dickey-Fuller (ADF) test

| p-value | | |
|---|---|---|
| Difference 0 | Difference 1 | Difference 2 |
| 0.4376 | 0.0794 | 0.01 |

For difference 1, the data also remain nonstationary. However, the earthquake data become stationary at difference 2. So, the integrated part of the $ARIMA$ model is 2.

As the integrated part i.e. d of the models for the earthquake data is known, to identify the value of $p$ and $q$, we use two techniques, graphical procedure and AIC value. Figure 3 shows the ACF and PACF of second difference of the data. These figures suggested an $AR$ (2) and $MA$ (0) model is operating. The minimum value of AIC for the data is also for $ARIMA$ (2,2,0) (Table 3). So for the earthquake data, we have identified an $ARIMA$ (2,2,0) model.

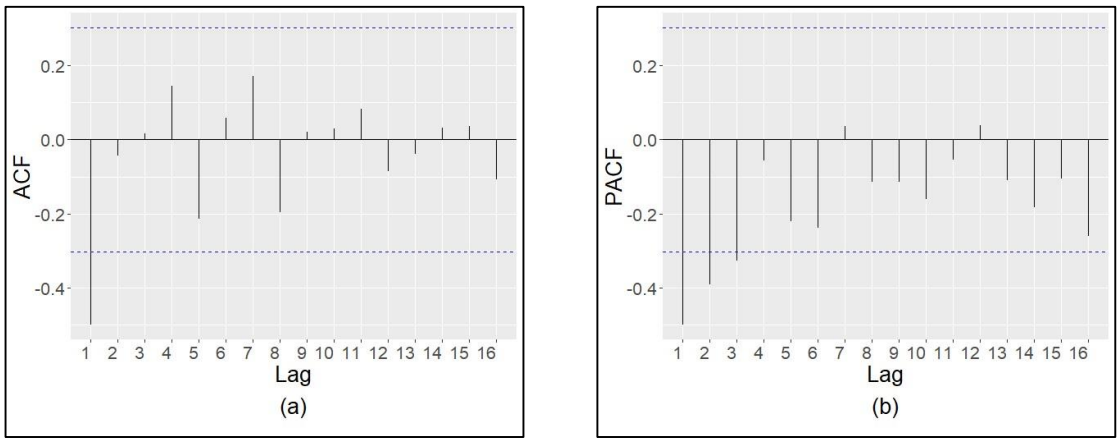


Figure 3: (a) ACF and (b) PACF plots for second difference

Table 3: AIC values for different $ARIMA(p, d, q)$ models for earthquake data:

| AIC values | | |
|---|---|---|
| $ARIMA(0,2,0)$ | $ARIMA(1,2,0)$ | $ARIMA(2,2,0)$ |
| 613.8245 | 603.7525 | 598.8686 |

After model identification, we have to check the model adequacy i.e. if the residuals are white noise or not. The ACF plot (Figure 4) of the residuals from the $ARIMA$ (2,2,0) model shows all correlations within the threshold limits indicating that the residuals are behaving like white noise. A Ljung-Box test (Table 4) returns a large p-value, also suggesting the residuals are white noise i.e. residuals are independently distributed. And the selected models can be fitted for further procedure.

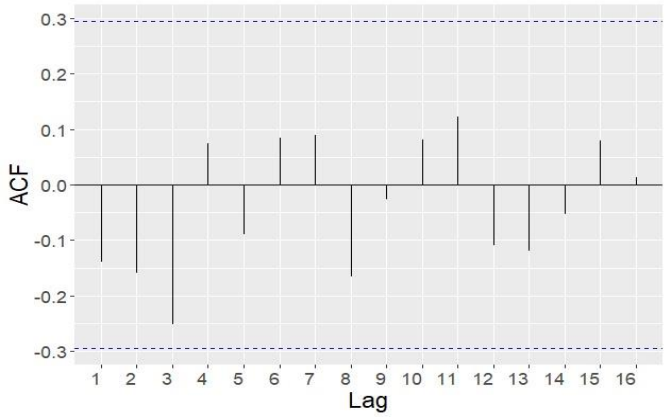


Figure 4: ACF plot for the residuals

Table 4: Box-Ljung test

| $\chi^2$ | DF | p-value |
|---|---|---|
| 17.832 | 20 | 0.5985 |

Now the residuals of the $ARMA$ (2,2,0) model are plotted to two control charts such as $CUSUM$ and $EWMA$ to identify whether the number of earthquake is statistical control or not. For constructing $CUSUM$ control chart, we use $h$ = 4 and $k = \frac{1}{2}$. Here, we define for UCL, $H = h\sigma$ and for LCL, $H = -h\sigma$. Figure 5 shows that most of the $CUSUM$ points lie on the central line 0, however there are two $CUSUM$ points (year 2012 and 2013) which are beyond the LCL line. So there is abnormality in the frequency of earthquake data and we may conclude that it is out of statistical control.

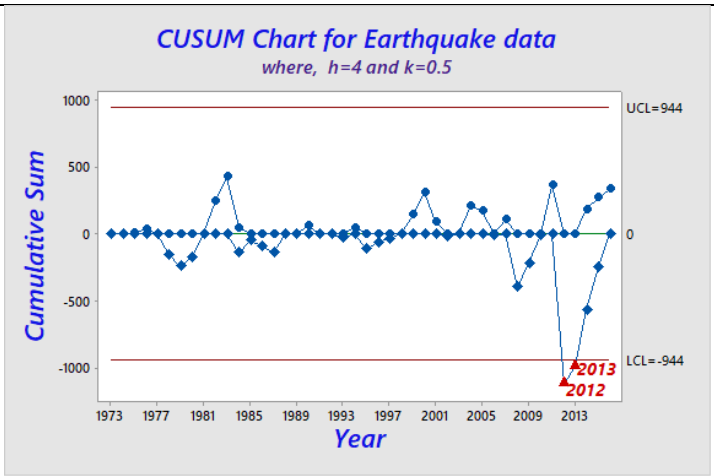


Figure 5: Cumulative sum control chart

For $EWMA$ control chart, we use the value of $\lambda$ equals to 0.4 for our earthquake data. Figure 6 shows that, the starting residual is zero. Then the residuals gradually increase and decrease over time. And for the year 2012, we find that the state of the process is out of control.

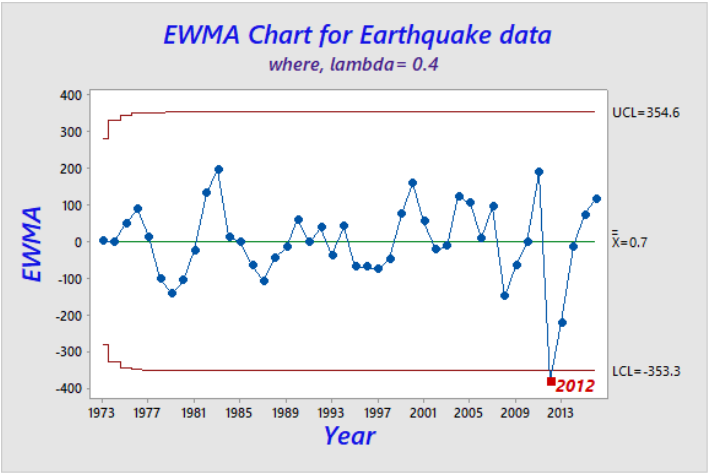


Figure 6: Exponentially Weighted Moving Average control chart

## 4. Conclusion

Now-a-days with the help of science and technology, it is possible to collect and analyse earthquake data. Earthquake is a natural disaster and increase of the frequency of earthquake indicates the change of world climate which is a big issue now-a-days. In this paper we are trying to identify whether the number of earthquake in the world is under state of statistical control or not.

In this article it is observed that there are two shifts in the mean of the number of earthquakes. For the year 1900-1949, the mean was very small. Then, for the year 1950-1972 the mean had increased slightly. However, we witnessed a tremendous shift in mean at 1973. There can be several reasons for this significant difference. It is possible that at the beginning of the 1900s, the measurement and detecting equipments of earthquakes were so ancient that those could not detect the earthquakes correctly. Back then science and technology was not as advanced as now. But at the end of the century, the situation has changed. Better and advanced

tools have invented that can detect the magnitude correctly. Yet, may be in the 1900s, earthquake did not appear as frequent as now.

In recent year, world has experienced several earthquakes. At the beginning of the study we had hypothesised that may be there are some abnormalities for these consequences. Though we have found most of the points are in the control zone, but some of them are in the state of out of control. So we may conclude that the predominant number of earthquake is not a random phenomenon, moreover it is a genuine shift that we should be concerned about and perform necessary research to find the real reasons behind those earthquakes.

## References

[1] Alam, M. M. (2015). Time series modelling for forecasting the earthquake behaviour in Indonesia. *Water and Geoscience,* 174-179.

[2] Amei, A.; Fu, W. and Ho, C. H. (2012). Time series analysis for predicting the occurrences of large scale earthquakes. *International Journal of Applied Science and Technology,* Vol. 2, No. 7: 64-75.

[3] Ata, N. and Ozel, G. (2011). A multivariate non-parametric hazard model for earthquake occurrences in Turkey. *Journal of Data Science,* 9, 513-528.

[4] Bisgaard, S. and Kulahci, M. (2005). Quality quandaries: the effect of autocorrelation on statistical process control procedures. *Quality Engineering*, 17(3):481–489.

[5] Chen, C.-K., Ho, C., Correa, C., Ma, K.-L., and Elgamal, A. (2011). Visualizing 3d earthquake simulation data. *Computing in Science & Engineering*, 13(6):52–63.

[6] Gujarati, D. N. (2009). Basic econometrics. Tata McGraw-Hill Education.

[7] Gupta, G. and Gupta, I. S. (2016). Earthquake data analysis and visualization using big data tool. In Proceedings of the 10th INDIACom; INDIACom-2016 2016 3rd International Conference on "Computing for Sustainable Global Development", 16th–18th March, 2016 Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi.

[8] Justin, C., Tinashe, C. P., Jonas, Z. R., Jonathan, M., and Marx, D. (2012). Application of statistical control charts to climate change detection in masvingo city, Zimbabwe. *Journal of Environmental Research and Development,* Vol, 7(2).

[9] Last, M.; Rabinowitz, N. and Leonard, G. (2016). Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. *PLoS ONE,* 11(1): e0146101. doi:10.1371/journal. pone.0146101.

[10] Lindsey, H. (2007). Natural' catastrophes increasing worldwide, world net daily web page.

[11] Machado, J. A. T. and Lopes, A. M. (2013). Analysis and visualization of seismic data using mutual information. *Entropy*, 15(9):3892–3909.

[12] Mandeville, M. (2007). Eight charts which prove that chandler's wobble causes earthquakes, volcanism, el nino and global warming.

[13] Montgomery, D. C. and Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23(3):179–193.

[14] Page, E. S. (1961). Cumulative sum charts. *Technometrics*, 3(1):1–9.

[15] Reyes, J. E. C. (2013). Nonparametric assessment of aftershock clusters of the Maule earthquake Mw = 8.8. *Journal of Data Science*, 11, 623-638.

[16] United States geological survey, USGS. https://earthquake.usgs.gov/ earthquakes/search/. Accessed: 2017-03-30.

[17] Yigˇiter, A. (2012). Change point analysis in earthquake data. In Earthquake Research and Analysis-Statistical Studies, *Observations and Planning*. InTech.