Randomness in the global earthquake distribution

Prithvi Thakur

1. Introduction

Earthquake prediction is one of the most fundamental challenges in the earth science community. The reason for lack of progress in this direction is attributed to sparsity of data and the lack of direct observations since earthquakes happen at a depth. But a more fundamental question is whether the distribution of earthquakes is random? Any evidence of non-randomness in the observed earthquake catalogue would give us hope that earthquake prediction, to a certain extent, might be possible. On the other hand, an inherent randomness in the earthquake catalogue would suggest that there is no basis for prediction. I test the temporal randomness in the global earthquake catalogue by comparing it to a series of synthetic random process, and see whether the observations are significantly deviating from a synthetic realization of an equivalent random process. This study is based on a research paper by Shearer and Stark (2012).

I use global earthquake data obtained from USGS (https://earthquake.usgs.gov/earthquakes) as our observations. Fig. 1, 2 show the earthquake magnitude distribution through time for global earthquake magnitudes greater than or equal to 7. The figures clearly show a lack of big earthquakes from 1960 to 2000. The time period from 1950 to 1965 and 2004-present shows elevated seismicity rates for very large earthquakes compared to the other time intervals. These seismicity trends are almost absent for the smaller earthquakes. So, the question is that are these clustering of big earthquakes statistically significant, or are they just one realization of a random distribution. For these clustering to be statistically significant, there has to be a physical

mechanism that allows the big earthquakes to be clustered without affecting the distribution of smaller earthquakes. We model the earthquake catalogue as a homogenous Poisson's process in time, and look at the significance of the apparent anomalies mentioned above.

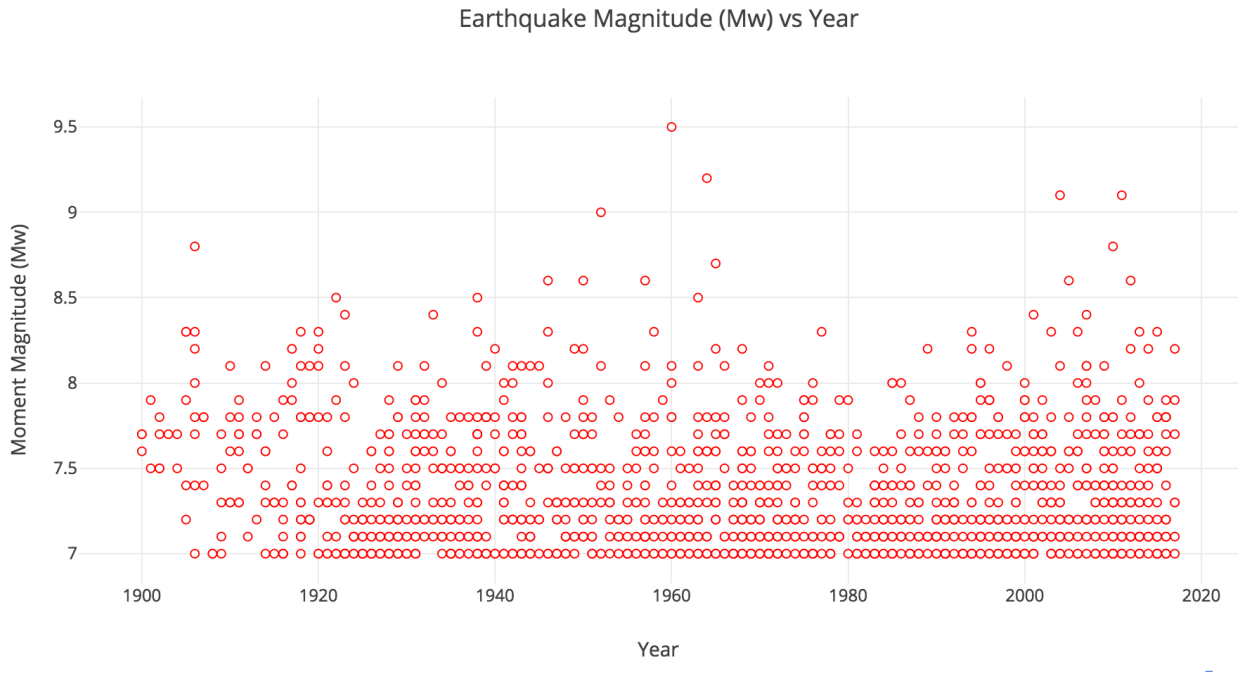


Fig.1. Global distribution of earthquake magnitudes since 1900. Data obtained from USGS.

2. An overview of Poisson's process

In probability and statistics, a measure of points located randomly in any mathematical space is called a Poisson process. This process is commonly defined in Euclidean space and used as a mathematical model to simulate natural processes which are apparently random. Mathematically, the number of events in a finite interval of time *(t)* obeys the Poisson distribution, or the probability of a Poisson random variable *N* being equal to *n* is given by the Poisson distribution:

$$P\{N(t) = n\} = \frac{\lambda}{n!} e^{-\lambda} \quad (1)$$

where, P is the probability, and λ is the Poisson parameter used to define the distribution. If this parameter λ is constant, then the Poisson process is said to be homogeneous. The parameter

λ can be interpreted as average number of points per unit of time, in this context. Given that a set of points distributed in space is a homogeneous Poisson process, the points would be uniformly distributed and independent of each other. The first online payment system using the blockchain technology, bitcoin, uses a mathematical model based on homogeneous Poisson point process (Nakamoto, 2008).

An inhomogeneous Poisson process has a domain varying parameter λ. For a Euclidean space, λ can be a function of spatial coordinates, and integrating it over the region of interest gives us the expected number of points of the Poisson process in the bounding region.

Distribution of earthquakes is spatially inhomogeneous, temporally homogeneous (SITHP) Poisson process (Gardner and Knopoff, 1974; Shearer and Stark, 2012). In the present study, we will only be looking at the temporal distribution of earthquakes and compare it to a homogeneous Poisson's process.

3. Declustering

Some of the early studies of global earthquake catalogues (Aki, 1956) and local catalogues (Knopoff, 1964) led to the conclusion that the sequence of earthquake distribution in space is non-Poissonian. We know that earthquakes do not happen randomly in space, and there are a series of smaller earthquakes following a larger earthquake known as aftershocks. If the aftershock happens to be bigger than the previous earthquake, then the previous earthquake is termed as foreshock, and the aftershock becomes the mainshock. In order to study whether the observed clustering in the catalogue is random, and get meaningful interpretations, we need to

remove any known systematic patterns found in the catalogue. Therefore, any inference about the distribution of earthquake is dependent on how well are we able to remove the aftershocks, and this process is known as declustering (van Stiphout et al., 2012). It involves identifying potential clusters in the observed data, and removing them while retaining only the biggest earthquake. Gardner and Knopoff (1974) came up with a relatively simple method, known as the mainshock window method, for declustering the earthquake catalogue. It involves punching a hole in the catalogue after each mainshock, and remove the earthquakes in that hole. Another common method used for declustering is the linked-window method, wherein an event is selected to belong to a particular cluster iff that event falls within the window of more than one event in that cluster (Reasenberg, 1985). More sophisticated declustering methods have been developed in the recent times, which use stochastic selection rules to decide the position of a particular event within the given cluster (Zhuang et al., 2002).

Shearer and stark (2012) use a simple approach of declustering by removing events occurring within 3 years and 1000km of a larger event. This approach can be classified as the mainshock-window method. I use a similar approach as that of Shearer and Stark (2012), and remove the smaller events within the specified window of 3 years' time and 1000km distance. This choice of threshold is pretty conservative, and it may remove a lot of events that would not be classified as aftershocks otherwise. This is done because we only want to look at the correlation between distant earthquakes, and not focus on clustering in the regional level.

Following the above guidelines, I wrote a declustering algorithm described as follows:

Let us assume we have a total of N earthquakes within the time period of 1900-2017. In the USGS catalogue used in this study, N = 1353.

1. For each event N_i, where i = 0 to N, we filter the group of earthquakes that are within 3 years after the given earthquake, and within a distance of 1000 km. This filtering domain is hereafter referred to as a window. We can neglect the depth dimension since the spatial window is very large. For N data points, we would have N windows.
2. We retain only the largest magnitude event in each window. Therefore, we would have N largest events.
3. Note that we calculate the window for each earthquake, thus we would have a number of repeating largest events in the above step. We merge the repeating events. This is the first iteration of declustering. Now the new number of events would be $N` \subseteq N$.
4. We repeat the above three steps for a few iterations until the new N` is the same as N at the previous step.

I ran the declustering algorithm 10 times, and the number of data points in each iteration is summarized in Table 1. This number saturates to 737 after 5 iterations. Thus, from our original catalogue of 1353 events, we have only 737 events in our declustered catalogue. The declustered earthquake distribution with time is shown in Fig. 2.

Table 1. Number of earthquakes in the catalogue after each declustering iteration

Iterations	Number of earthquakes
0 (Original catalogue)	1353
1	844
2	757
3	742
4	738
5	737
6	737
7	737
8	737
9	737
10	737

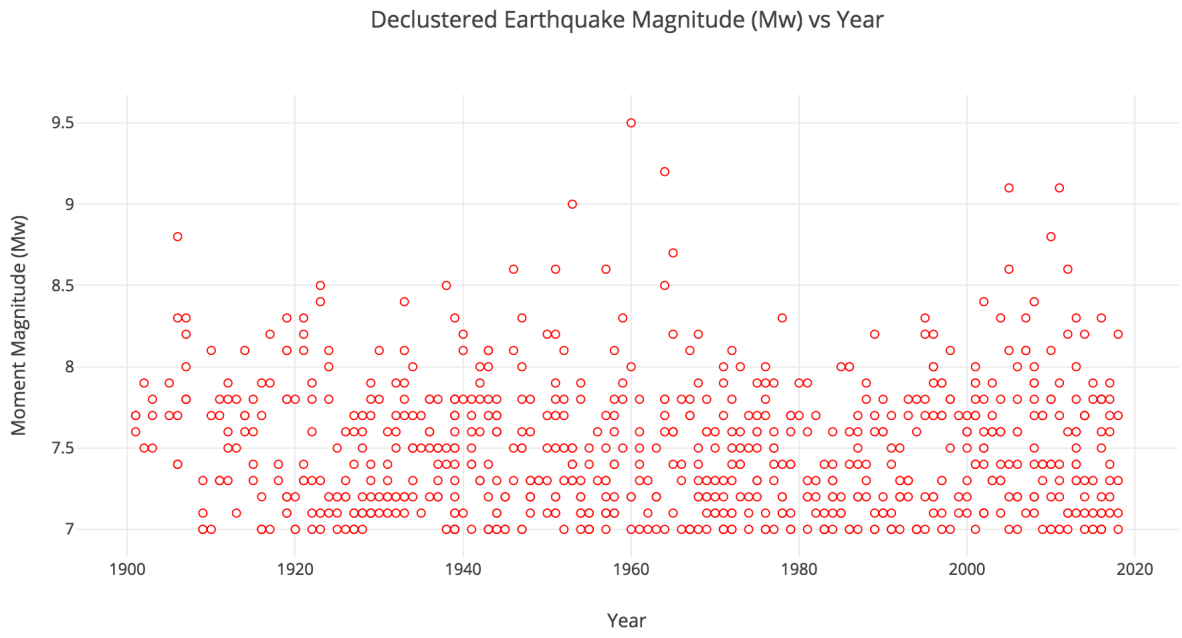


Fig. 2. Temporal distribution of earthquakes with respect to their moment magnitudes after declustering. The total number of declustered earthquakes = 737.

Using three magnitude cutoff values of 7.0, 7.5, and 8.0, I plot the number of earthquakes each year for the three cutoff values. Thus, we filter the earthquakes with magnitudes greater than 7.0, 7.5, and 8.0 respectively and count the number of earthquakes per year. This is shown in Fig. 3 as annual earthquake rates.

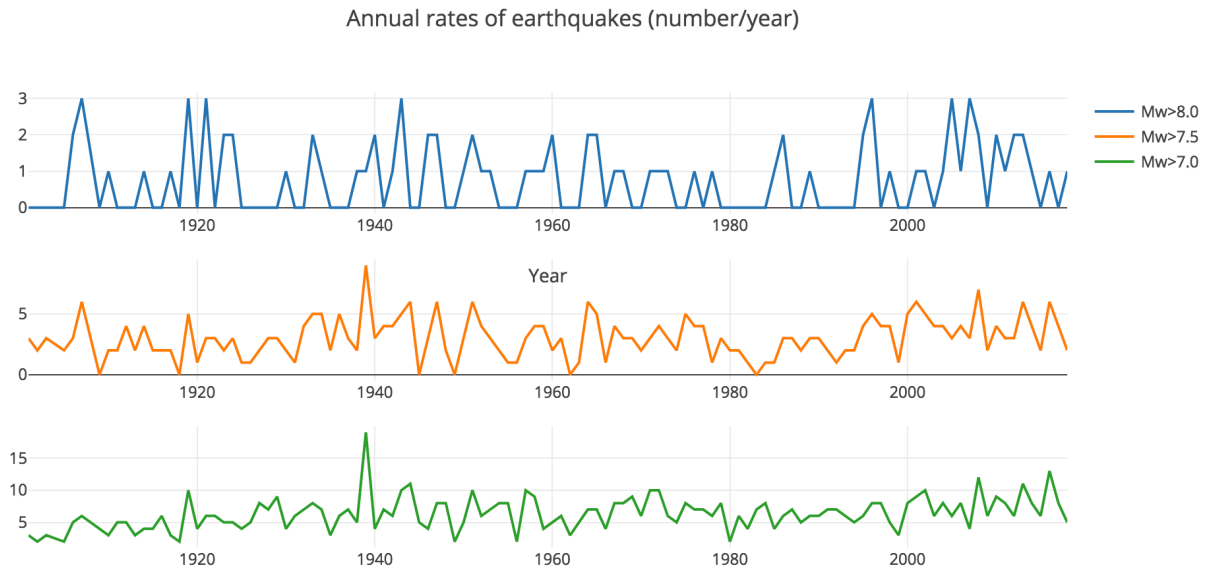


Fig. 3. Annual earthquake rates for three magnitude cutoff values of M≥7.0, M≥7.5, and M≥8.0. The y axis shows the number of events in that year plotted against time in years.

4. Monte Carlo simulations

One of the important properties of Poisson process is that the events at given times are independent, identically distributed (iid) events. This implies that each random variable has the same probability distribution as the others, and all the events are mutually independent. One way of generating a Poisson process is to draw a fixed number of events from a Poisson distribution (eq. 1), and draw the corresponding time of each event from the uniform distribution.

In order to test the statistical significance of the apparent clustering in the catalogue, as described in section 1, we use Monte Carlo simulations to generate synthetic earthquake catalogue, having exactly the same number of events as observed in the declustered catalogue. These synthetic catalogues are modeled as a homogeneous Poisson's point process in time. In doing this, we are making an implicit assumption that the observed earthquake distribution is a homogeneous Poisson process in time, and this is referred to as the null hypothesis. If the anomalies observed in the catalogue are also observed in the synthetic catalogues at some significance level, then the null hypothesis cannot be rejected.

The earthquake distribution follows the Gutenberg-Richter relationship between the magnitude and the frequency. The relation is given by:

$$N_{EQ} = 10^{(a-bM)} \tag{2}$$

where N_{EQ} is the number of earthquakes for a given magnitude M. a and b are constant parameters. The parameter a is the measure of total seismicity of the region. The value of 10^a gives the total number of earthquakes. Therefore, the value 10^{-bM} is the sum of probability of occurrence of earthquakes for each magnitude. For a global catalogue, I have assumed the 'b' value to be 1.3 (Shearer and Stark, 2012). This exponential relation between the earthquake magnitude and frequency tells us that there will be much more lower magnitude earthquakes, and this can be confirmed from looking at the observations in Fig. 1 and 2.

Now we have all the ingredients to simulate earthquake catalogues as Poisson process. Our observations consist of earthquakes from the time period 1900 to 2017. When converted to days, we have a catalogue with temporal interval of 0 to 42,470 days. A caveat here is that I have

assumed each month to contain 30 days, and ignored the leap year days when converting the time period into days. This might introduce small changes in the catalogue, but should not affect the overall results and interpretation. The observed, declustered catalogue consists of 737 events within a period of 42,470 days. In our simulated catalogue, we generate exactly the same number of events within the same period. The steps for simulating are elucidated below:

1. We randomly sample the time for each earthquake from a uniform distribution.
2. For the magnitude, we define a probability function of its occurrence. This probability function is the frequency of a particular magnitude computed in accordance with the Gutenberg-Richter relation given by eq. 2, divided by the theoretical maximum frequency corresponding to a magnitude 7, the lowest magnitude in our catalogue.
3. We generate a uniformly distributed random number between 7 and 9.5 (the two extreme values of magnitude in our catalogue), and select it if the probability obtained in step 2 is greater than a random number between 0 and 1.
4. We repeat the step 3 until we have 737 magnitudes in our catalogue.

This process of Monte Carlo simulation generates one catalogue of magnitude and time of earthquakes. We simulate 10,000 catalogues in this manner, under the null hypothesis assumption that the earthquake distribution follows a Poisson process in time. We estimate the probability of an observed anomaly as the fraction of the 10,000 catalogues that have atleast the observed number of events within an interval of specified length.

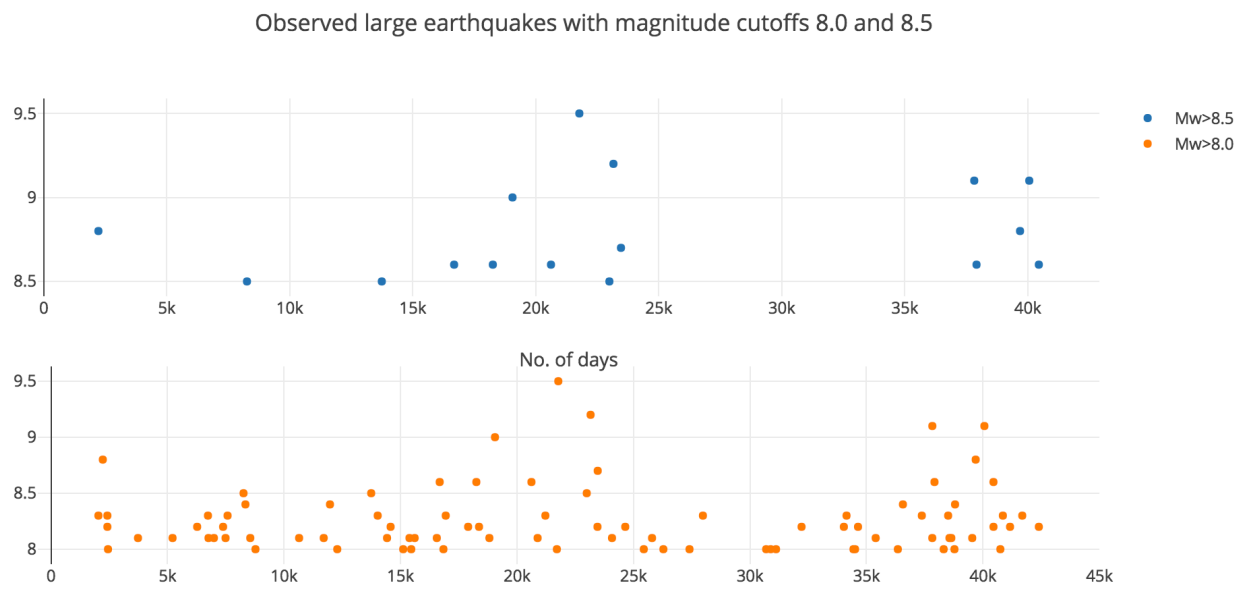


Fig. 4. Observed distribution of earthquakes with minimum magnitude of 8.5 and 8.0 respectively. Notice the clustering of large earthquakes in the recent period and the gap preceding it.

First, we look at the observed anomalies in the declustered catalogue. Looking at the first figure in Fig. 3, there are 5 events within the time period of 37,826 days to 40,451 days (counting from 1900). Due to the approximations used in converting time, I cannot pinpoint the exact earthquakes and their dates, but the 37,826th day earthquake corresponds to the 2004 Sumatra-Andaman earthquake, and 40,451 corresponds to the 2012 Sumatra earthquake (Duputel et al, 2012). Thus, there are 5 events out of the 16, with Mw≥8.5, that occur within 2626 days of each other. Looking at the recent seismicity elevation in the second graph of Fig. 3, we have 22 out of the 87 earthquakes with Mw≥8.0 occurring within the time period of 6053 days.

We look for similar anomalies as observed in the simulated catalogue, i.e, we look for the fraction of the 10,000 simulated catalogues that contain atleast 5 events of Mw≥8.5 occurring within a time period of 2626 days, and compute the probability. Out of the 10,000 simulations,

2912 simulations had more than 5 events of Mw≥8.5 within a time period of 2626 days, implying that there is 29.12 percent chance that under null hypothesis, we can have the seismicity distribution equivalent to the observed seismicity clustering. Similarly, there were 9508 out of 10,000 simulations with more than 22 earthquakes of Mw≥8.0 within a 6053 days interval. Thus, there is a 95% chance that the observations follow a null hypothesis for the Poisson process.

5. Kolmogorov-Smirnov (KS) test

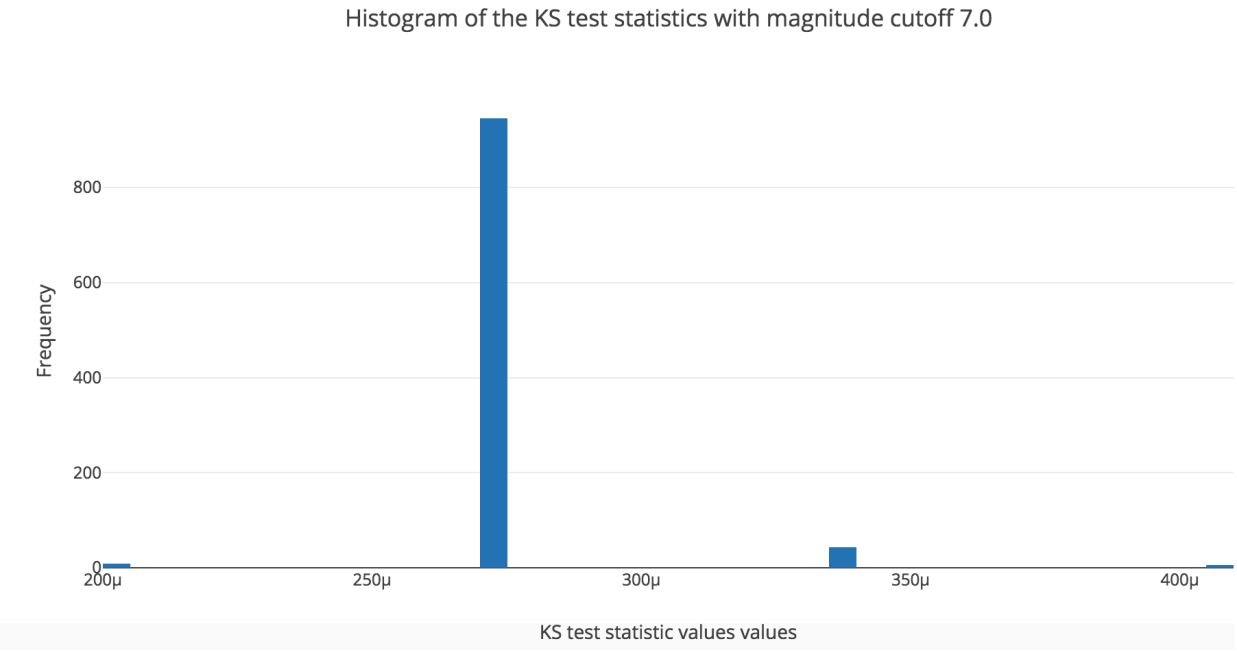
So far, we have looked at the observations, and empirically compared it to the Monte Carlo simulations. These results tell reinforce our confidence that the null hypothesis might be true and the given observations might not be significantly anomalous compared to a homogeneous Poisson process. But in order to quantify our confidence in the null hypothesis, we need to do a hypothesis testing using a test statistic. The Kolmogorov-Smirnov (KS) test is used to determine if two given distributions differ from each other (two-sample test), or to determine whether the given sample follows a reference probability distribution function (one-sample test). KS test is a non-parametric test and does not assume anything about the prior distribution of the variables. We perform a KS test of the hypothesis that the times in the declustered catalogue at different thresholds (see Fig. 3) are a sample of independent, identically distributed uniform random variables. We are testing the observations against the simulated earthquake catalogues, therefore we will use two-sample KS test.

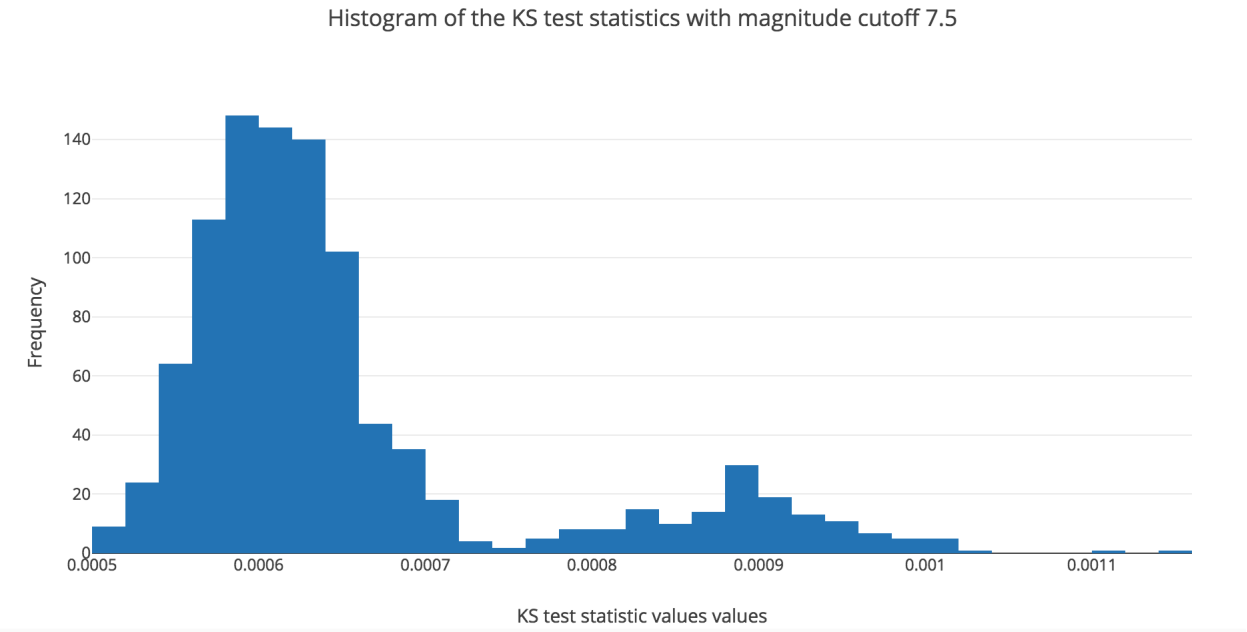
The KS test statistic (D) for determining whether the two given distributions differ, is defined as the largest difference between the two distributions.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \quad (3)$$

where $F_{1,n}(x)$ and $F_{2,m}(x)$ are the empirical distributions of the two samples, and *sup* is the supremum function, defined as the least element in the distribution of the difference greater than or equal to the all the elements of the distribution.

We calculate the KS statistic (D value) for the 10,000 simulated distributions with respect to the observed earthquake distribution. In order to do that, we divide the simulated and the observed catalogues into magnitude cutoffs shown in Fig. 3, and create a histogram from the discrete values. The D value is given by the maximum difference between the two histogram distributions. Fig, 5 shows a histogram of the 10,000 computed D values with respect to the observations. A lower D value implies that the two distributions under consideration are more likely to be equal.





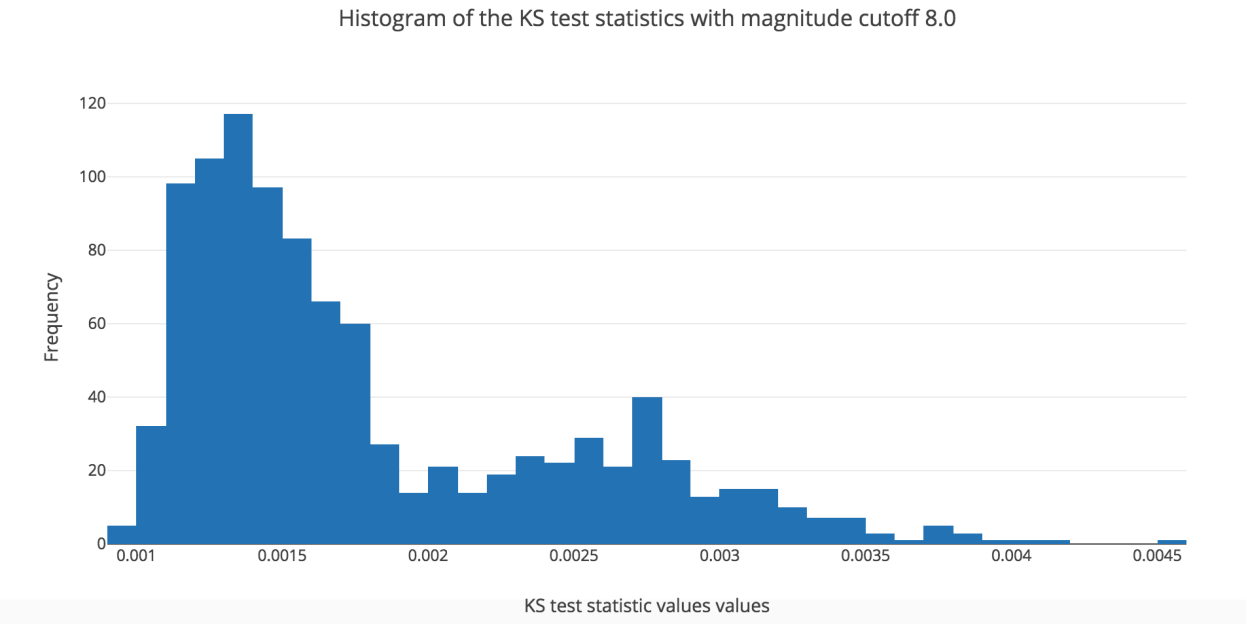


Fig. 5. The D values for different magnitude cutoff ranges. A lower value represents more probability of the two distributions under consideration belonging to the same Poisson process.

The null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{4}$$

where n, m are the number of values in the distribution and $c(\alpha)$ is a function of α for different confidence intervals. For the α values of 0.05, i.e., a confidence of 95%, the $c(\alpha)$ value is 1.36 (Source: Wikipedia: KS test). Since the value of the square root in the equation 4 is always around 0.707 as the sample sizes are comparable, therefore the D values in our results is always less than the critical level. Thus, we can never reject the null hypothesis that the observed distribution follows Poisson process.

The next step is to calculate the p-values for the KS test. The computation for p-value of the KS test statistic is quite complicated, and it involves computing infinite sum of an inverse exponential expression. The reason for this complicated expression is that the KS test is a nonparametric test, therefore there is no assumption about a specific distribution. The test statistic is simply the largest difference between any two given distributions. I have omitted the calculations of p-values due to this reason in this study.

6. **Conclusions**

In this study, I have tried to study the questions posed by Shearer and Stark (2012), and worked out a methodology following their guidelines. From a data analysis point of view, we have looked at declustering, its implications and a simple method for doing so. We have also looked at the definition of a Poisson process, its importance in modeling natural processes that are seemingly random, and how to model it using Monte Carlo simulations. We then move on to model estimation, where we test the hypothesis whether two given distributions are equal. This is done using a Kolmogorov-Smirnov test, and it is widely used in testing whether any observed distribution follows a Poisson process. There are two other important tests for testing the Poisson

hypothesis, namely the Poisson dispersion test and the multinomial chi-squared test. The Poisson dispersion test statistic is just the normalized variance within the specified interval. We divide the time domain into a fixed number of intervals and compute the dispersion test statistic for each interval. The dispersion for a typical Poisson distribution is 1, since it has equal variance and mean. A greater than 1 value means overdispersion and implies the existence of clusters. A lower than 1 value means underdispersion and indicates patterns in observations that are more regular than the randomness of a Poisson process. The multinomial chi-squared test also divides the time into a number of intervals, and tests how well the number of events in each interval agrees with the expected number of events under the assumption of a homogeneous Poisson process. I have not included these two tests in the analysis due to time constraints, and my lack of understanding of the implementation of these tests to the earthquake catalogue. This can be a subject of future work.

Acknowledgements

All the data analysis and plotting is carried out in python 3 using Jupyter Notebooks, and plotly distribution. The explanation of any concept not referenced is the courtesy of Wikipedia and Stackexchange.

References:

Aki, K. (1956). Some problems in statistical seismology. Zisin 8, 205-228.

Gardner, J.K. and Knopoff, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? Bull. Seis. Soc. Am., 64(15), 1363–1367.

Knopoff, L. (1964). Statistics of earthquakes in Southern California, Bull. Seis. Soc. Am., 55, 753-797.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Consulted*, 1(2012):28.

Reasenberg, P., A. (1985). Second-order moment of central California seismicity, 1969-1982. J. Geophys. Res., 90(B7), 5479–5495.

Shearer, P., M., and Stark, P., B. (2012). The global risk of big earthquakes has not recently increased. Proc. Nat. Acad. Sci., 109(3), 717–721.

van Stiphout, T., Zhuang, J., and Marsan, D. (2012). Seismicity declustering, Community Online Resource for Statistical Seismicity Analysis, doi:10.5078/corssa- 52382934.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. J. Am. Stat. Assoc., 97 (458), 369–380.