MODELLING EARTHQUAKE INTER-EVENT TIME DISTRIBUTION

(APPLIED STATISTICAL METHODS)

BERNOULLI GROUP

Ankit Sonthalia
Nayan Khanna
Rohit K Bharadwaj
Jatin Singh
Aniruddha Jayant Karajgi
Rahul Jha
Subham Kumar Dash
Meet Kanani

November 22, 2019

Birla Institute of Technology and Science, Pilani

Contents

Abstract

Earthquakes have caused serious damages to life and property over the years. Due to the fact that they cannot be predicted with reliable precision, the threat they pose becomes even greater. However, ever since seismological agencies around the world have started maintaining records of earthquake data, researchers have been trying to use the same and model the behavior of the underlying probability distribution, provided such distribution exists, that governs the data. In the current analysis, we have attempted to fit earthquake inter-event time data corresponding to the Himalayas and the adjacent regions, to various probability distributions. We have then evaluated these models using two popular non-parametric tests. The gamma distribution was found to be the best fit among the applied models, with the shape parameter equal 0.8906 and a scale parameter equal to 635.324. Finally, we have discussed briefly on a potential procedure for forecasting earthquakes.

1 Introduction

A BRIEF WORD ON EARTHQUAKES Planet Earth is made up of four major layers: the **inner core**, the **outer core**, the **mantle** and the **crust**. The crust, along with some portion of the mantle, forms a **skin** on the surface. This skin, however, is made up of several pieces, called **tectonic plates**, the whole set of which largely resembles a giant **jigsaw puzzle**. These tectonic plates are constantly in **motion**, often **colliding** with each other. These movements give rise to **faults** or fractures, usually along the **plate boundaries**. The plate boundaries are **rough** along these faults. This causes the plates to often get **stuck** together at the boundaries, instead of **sliding past** one another. Meanwhile, the other parts of the plates keep moving, gradually building up a **force** that tries to drive the plates apart at the boundaries. Once this force is great enough, it overcomes the **friction** between the rough edges of the plates. This event releases an enormous amount of **energy**, and sends it outward in all directions, in the form of **seismic waves**. These waves carry all the energy that was being stored at the plate boundaries due to restricted movement. Once these waves hit the surface of the Earth, they unleash huge damages on life and property.

PREDICTION VS FORECASTING The terms **prediction** and **forecasting** are often used interchangeably, which can be misleading. **Prediction** is largely concerned with the specification of **time**, **magnitude**, and **approximate location** of the next major event, and it aims at being **precise** enough so as to able to issue a reliable **warning** in advance. **Forecasting**, on the other hand, is a fairly different branch of science of seismology that concerns itself with the general assessment of **seismic hazard** in a given area over a given period in future time.

WHY FORECAST EARTHQUAKES? While **evacuation** is the **safest** protective measure in preparation for an incoming earthquake, it is rarely the **wisest** one because predicting the exact time, location and magnitude of an earthquake is very difficult. Forecasting, however, often proves instrumental in the **probabilistic assessment of threat**. While it does not claim to give precise predictions, it provides a fairly good **general idea** as to which areas are at relatively higher risk. The availability of this **information** helps governments take **initiatives** towards securing these areas as much as possible, so as to **minimize the damages** to life and property, in the event that a major shock should actually hit the concerned areas[1].

2 Methodology

2.1 Acquiring the Dataset

The data required for fitting our earthquake forecast model was obtained from the ISC Bulletin. We have focused particularly on the area covering the **Himalayas** and adjacent regions. The ISC

Catalogue was searched for all the earthquakes that have occurred in the region in the period between 1900 and 2019. Additional search parameters included a **depth** parameter, whose value was set to be confined between **0** and **200** km, and **magnitude**, between **6** and **10**. A total of 86 such events were found.

2.2 Seismicity Declustering

NEED FOR DECLUSTERING The earthquake data obtained above, originally does not consist of **independent and identically distributed** (**i.i.d.**) random variables. Often a major disturbance is followed by shorter disturbances, called **aftershocks**[2]. Shorter, less significant events called **foreshocks** may sometimes lead up to a major mainshock. It has thus been observed that using the dataset as it is for estimating the model parameters, might and often does lead to poor results. Hence, there exists a need to promptly identify and remove all such dependent events so that the list of events that remains, is i.i.d. as desired.

DECLUSTERING ALGORITHM In our analysis, earthquakes with higher magnitude are assumed to be independent. It is also assumed that the influence of a given mainshock over a potential aftershock is a decreasing function of both, the spatial distance between them, and the time difference. As one additional assumption which, for our purposes, will not have any significant effect on the final results, while simplifying our analysis significantly, foreshocks have been treated as if they were aftershocks. With these assumptions, the list of events was arranged in decreasing order of magnitude, starting at the event with the highest magnitude. The algorithm starts at the earthquake with the highest magnitude, and constructs a window around it, taking its magnitude M as a parameter. The spatial and temporal radii, r and t, respectively, are calculated as follows[3]:

$$r = e^{-1.024+0.804M} \pm 15(km) \tag{1}$$

$$t = e^{-2.87+1.235M} \pm 60(days) \tag{2}$$

Once the values of r and t have been obtained, the algorithm iterates through the rest of the list and calculates, for each potential aftershock, its spatial and temporal distances from the mainshock. Under the condition that these distances fall within the constructed window, the corresponding event is removed from the list. With the remaining list, the algorithm then moves on to the earthquake with the next highest magnitude and repeats the process. When the algorithm terminates, i.e., when the only earthquakes that remain in the list are those that have already been considered, we have a list of i.i.d. events.

OBTAINING INTER-EVENT TIMES Finally, the declustered catalogue is arranged chronically (with the oldest event coming first). For each earthquake, except the very first one in the list, the time difference, in days, is calculated between the current earthquake and the one immediately preceding it. This operation results in the generation of a list of inter-event times. In all subsequent steps of our analysis, this list will serve as the data to which we will fit our models.

2.3 Candidate Models

Four different probability distributions were considered for the purposes of fitting our data. We shall be introducing each of them briefly.

2.3.1 Exponential Distribution

The exponential distribution is one of the simplest models that have been studied in the interest of earthquake forecasting. It is used to model the distribution of inter-event time between events that follow a Poisson process. Hence, an important assumption that this distribution requires is that each occurrence is independent of other occurrences.

Parameters The exponential distribution has a single parameter: β, also known as the **scale parameter**. Mathematically, it is the reciprocal of the rate parameter, λ, of the corresponding Poisson process.

Probability density function The probability density function of the exponential distribution is given by

$$f(x;\beta) = \begin{cases} \frac{1}{\beta}e^{\frac{-x}{\beta}} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3}$$

where β is the scale parameter.

2.3.2 Gamma Distribution

Another common distribution frequently used in the modelling of earthquake inter-event times[4] is the gamma distribution.

Parameters The gamma distribution has two parameters: the shape parameter, α, and the scale parameter, θ. If α is set to one, we obtain the exponential distribution as a special case of the gamma distribution.

Probability Density Function The probability density function of the gamma distribution is given as

$$f(x;\alpha,\theta) = \frac{x^{\alpha-1}e^{-x/\theta}}{\theta^{\alpha}\Gamma(\alpha)} \tag{4}$$

for $x > 0; \alpha, \theta > 0$, where $\Gamma(\alpha)$ is the Gamma function, given by

$$\Gamma(\alpha) = (\alpha - 1)! \tag{5}$$

2.3.3 Weibull Distribution

The Weibull distribution, named after the Swedish Mathematician of the same name, is used for modelling hazard rates of systems[5] [6], often as part of a quality checking routine. As a probabilistic model of the **time to failure** of a system, it forms a suitable candidate model for the purpose of earthquake forecasting.

Parameters The Weibull distribution has two parameters : the shape parameter, k, and the scale parameter, λ.

Probability Density Function The probability density function of the Weibull distribution is given as

$$f(x;k,\lambda) = \begin{cases} \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(\frac{x}{\lambda})^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{6}$$

2.3.4 Lognormal Distribution

While the normal distribution is a popular candidate model for data that shows a reasonable amount of symmetry, it is not applicable in this analysis due to lack of the same. However, another distribution, called the **lognormal distribution** comes into play here. It is based on the assumption that the logarithm of the random variable is normally distributed. Due to the shape and properties of the logarithmic function, this distribution takes care of the skewness in the data.

Parameters The lognormal distribution has two parameters: μ and σ, the mean and standard deviation, respectively, of the natural logarithm of the given continuous random variable.

Probability Distribution Function The probability distribution function of the lognormal distribution is given as

$$f(x;\mu,\sigma) = \frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(lnx-\mu)^2}{2\sigma^2}) \tag{7}$$

2.4 Parameter Estimation

For each of the four models discussed above, parameters were estimated so as to best fit the inter-event time data that we have analysed. Two methods of parameter estimation were used: **Maximum Likelihood Estimation** and **Method of Moments**.

2.5 Model Evaluation

Having fitted the models to our data, we applied two tests to evaluate the obtained distributions for **goodness of fit**. These two tests are the Kolmogorov-Smirnov (KS) Test and the Akaike Information Criterion (AIC) test. The model that returns the best results with respect to these tests will finally be considered for forecasting.

3 Results

The probability distributions were fitted to the data using the **fitdistr** function of the **fitdistrplus** package of the **R** programming language. This function contains built-in code for estimating the parameters of our distributions using both, the Maximum Likelihood Estimation method and the Method of Moments.

After feeding our data into the function, we obtained the plots corresponding to each of these distributions, along with the parameter estimations using both methods. The results of our analysis have been summarized in the tables and figures below.

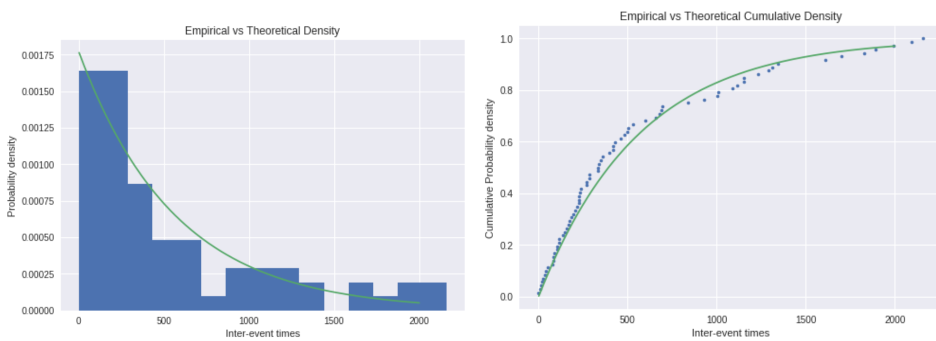
The figures are self-explanatory.

	Parameter 1	Parameter 2	AIC Values	KS_stat	KS_pval
Exponential Distribution	N/A	β = 565.93	**395.568**	0.0749	0.7866
Gamma Distribution	α = 0.8906	θ = 635.324	**396.911**	**0.0628**	0.9222
Weibull Distribution	k = 0.9328	λ = 548.354	1060.141	**0.0641**	0.9102
Lognormal Distribution	μ(log) = 5.68	σ(log) = 1.39	1074.17	0.084	0.6588

Figure 1: Results using MLE

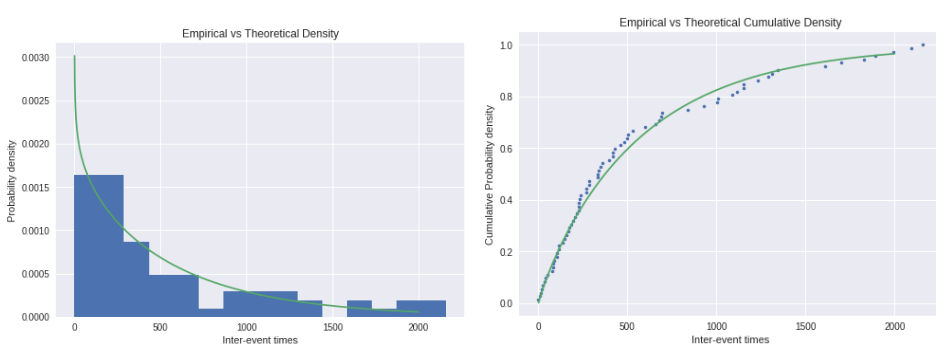
	Parameter 1	Parameter 2	AIC Values	KS_stat	KS_pval
Exponential Distribution	N/A	β = 565.93	1058.713	0.0749	0.7866
Gamma Distribution	α = 0.9844	θ = 575.043	1060.545	0.0717	0.8275
Weibull Distribution	k = 1.0087	λ = 572.577	1074.17	0.0802	0.7128
Lognormal Distribution	μ(log) = 5.98	σ(log) = 0.83	1137.914	0.1556	0.0546

Figure 2: Results using MoM



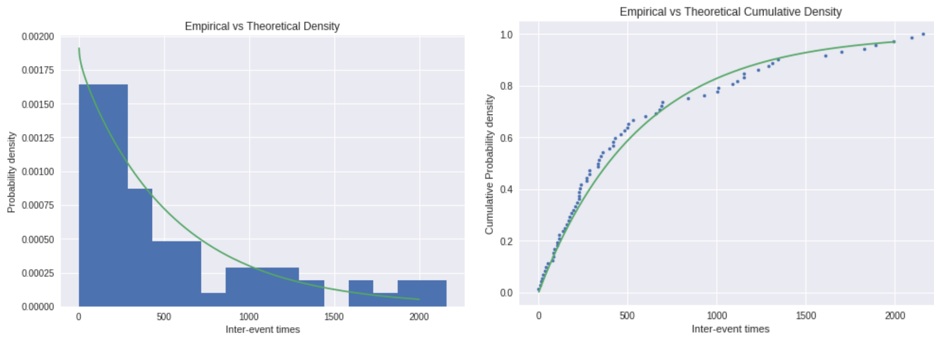
(a) Probability Distribution Function (b) Cumulative Distribution Function

Figure 3: Exponential Distribution (Using MLE)



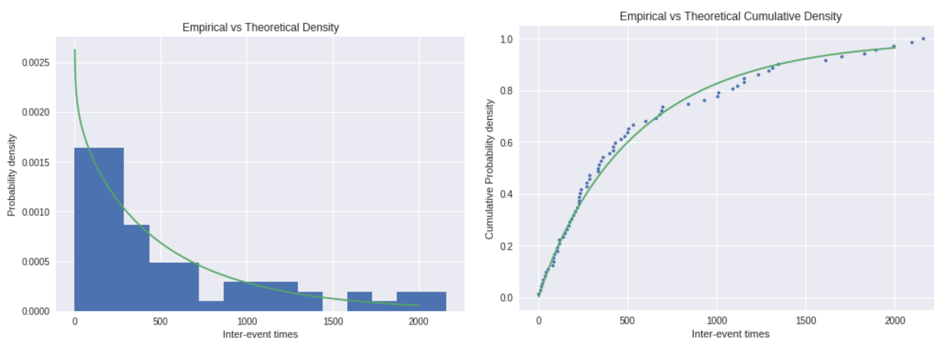
(a) Probability Distribution Function (b) Cumulative Distribution Function

Figure 4: Gamma Distribution (Using MLE)



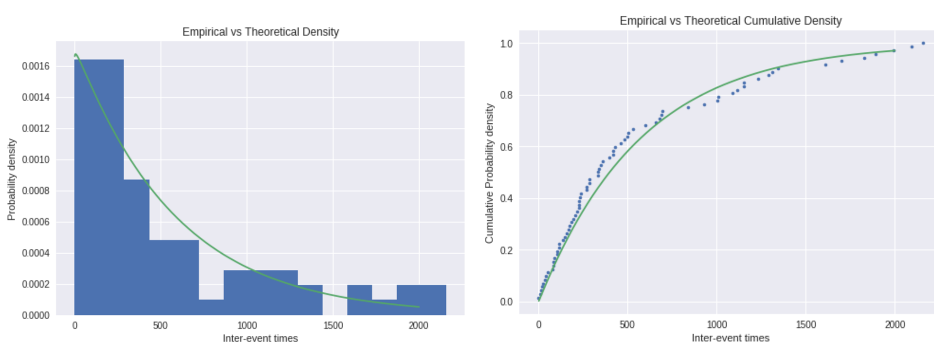
(a) Probability Distribution Function (b) Cumulative Distribution Function

Figure 5: Gamma Distribution (Using MoM)



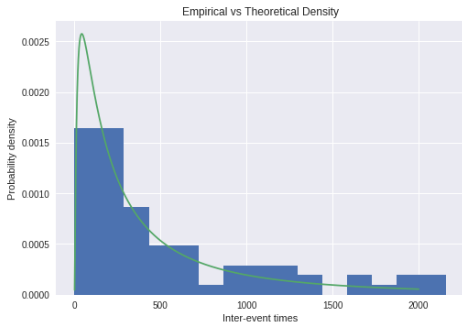
(a) Probability Distribution Function (b) Cumulative Distribution Function

Figure 6: Weibull Distribution (Using MLE)

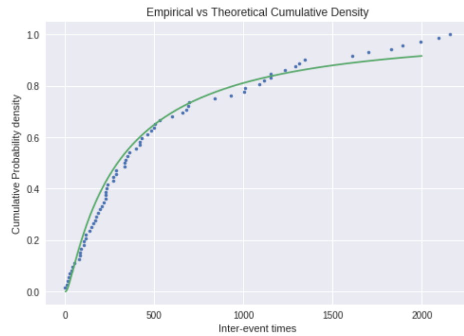


(a) Probability Distribution Function (b) Cumulative Distribution Function

Figure 7: Weibull Distribution (Using MoM)

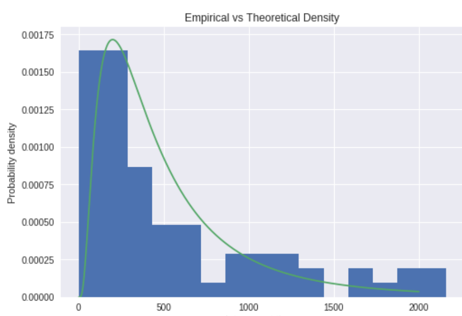


(a) Probability Distribution Function

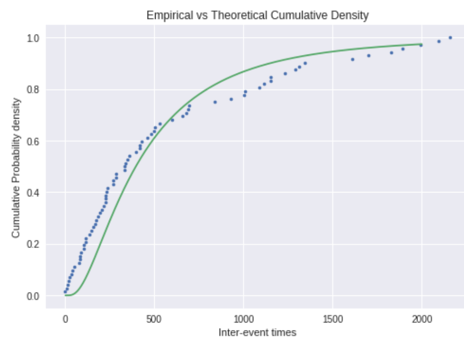


(b) Cumulative Distribution Function

Figure 8: Lognormal Distribution (Using MLE)



(a) Probability Distribution Function



(b) Cumulative Distribution Function

Figure 9: Lognormal Distribution (Using MoM)

4 Conclusions and Final Remarks

After completion of our analysis, the following significant results were noted:

1. The **AIC test** established the exponential distribution ($\beta = 565.93$) as the best fit model for our data. The gamma distribution ($\alpha = 0.8906, \theta = 635.324$) comes a close second with an AIC value exceeding that of the former by a mere 1.3 units.

2. The **KS test** established the gamma distribution ($\alpha = 0.8906, \theta = 635.324$) as the best fit model for our data, while the Weibull distribution ($k = 0.9328, \lambda = 548.354$) comes a close second, its value of the KS test statistic exceeding that of the former by a mere 0.013 units.

3. All of the distributions, except the lognormal model estimated using the method of moments, have passed the KS goodness of fit test. The lognormal distribution is generally not the best alternative for this purpose; the results of our analysis have proven this.

4. The models estimated using Maximum Likelihood Estimation have, in general, turned out to perform better than those estimated using Method of Moments.

These results make considerable sense because of the fact that the Gamma distribution and the Weibull distribution have proven successful in capturing the properties of the inter-event time distribution in several instances of earthquake forecasting attempts, across the world. Since the underlying processes that give rise to earthquakes, (and, in turn, the corresponding data that is subsequently used for analysis,) work in similar ways, under the same laws of physics, across the world, the data generated should be expected to follow a space-invariant pattern on a large scale.

While the Gamma and the Weibull distributions stand out as the best fitting models when evaluated using the KS test, they lose out to the exponential distribution (which is, essentially, a special case of the Gamma distribution), by a few points. Since the AIC test penalizes a model for an increase in its number of parameters, it assigns a slightly better score to the exponential distribution, due to the presence of just one parameter as opposed to two.

4.1 A Short Word on Earthquake Forecasting

Now that we have analysed our data, fitted probability distribution models and established their goodness of fit using non-parametric tests, we are in a position to start talking about earthquake forecasting. Since the **gamma distribution** estimated using MLE was found to be the best fit to our data using the KS test, we will be making use of the same in this final discussion. We will demonstrate how the Cumulative Density Function can be used to forecast earthquakes.

For a given time t (in days), it is now possible to comment on the probability that the next earthquake will occur within t days, beginning at the day that the most recent earthquake occurred. Using the gamma distribution with parameters $\alpha = 0.8906$ and $\theta = 635.324$, we can then calculate the cumulative probability at a few points. The results have been summarized below, for four data points. They have been compared with their cumulative frequency in our original data.

As can be observed, the forecasting cumulative densities are significantly close to the trend that has been observed in the past.

Time (in days) t	CDF P(T <= t)	Empirical cumulative Frequency
100	0.184	0.1666
365	0.491	0.541
730	0.725	0.736
1000	0.824	0.763

Figure 10: An example of earthquake forecasting and comparison with actual data

References

[1] Richard Andrews. Earthquake prediction and public policy: Recent experiences in california. In *Practical Approaches to Earthquake Prediction and Warning*, pages 675–680. Springer, 1985.

[2] Paul Reasenberg. Second-order moment of central california seismicity, 1969–1982. *Journal of Geophysical Research: Solid Earth*, 90(B7):5479–5495, 1985.

[3] Sumanta Pasari. Stochastic modelling of earthquake interoccurrence times in northwest himalaya and adjoining regions. *Geomatics, Natural Hazards and Risk*, 9(1):568–588, 2018.

[4] Cataldo Godano. A new expression for the earthquake interevent time distribution. *Geophysical Journal International*, 202(1):219–223, 2015.

[5] Sumanta Pasari and Onkar Dikshit. Impact of three-parameter weibull models in probabilistic assessment of earthquake hazards. *Pure and Applied Geophysics*, 171(7):1251–1281, 2014.

[6] Sumanta Pasari and Onkar Dikshit. Distribution of earthquake interevent times in northeast india and adjoining regions. *Pure and Applied Geophysics*, 172(10):2533–2544, 2015.