Statistical Modeling and Forecasting of Seismic Events

Iris Wang
Scarsdale High School,
Scarsdale, New York

Abstract

Understanding earthquakes remains a challenge in our world which is strongly affected by natural forces. This study analyzes earthquake data across various seismic regions, focusing on the frequency and intensity of seismic activity. Monthly maximum magnitudes are modeled using extreme value distribution theory. We use nonparametric methods, such as locally fitted regressions and splines, alongside parametric ARIMA models to assess temporal patterns. Machine learning techniques are incorporated for anomaly detection, and earthquake occurrences are modeled using a Poisson process based on interarrival times. These methods provide insights into earthquake dynamics and may improve risk assessments.

Keywords: Earthquake forecast, Extreme Value theory, thin plate smoothing spline, locally estimated scatterplot smoothing, time series models, anomaly detection, Poisson process

1 Introduction

1.1 Background

Understanding the extreme magnitudes of natural disasters, particularly earthquakes is vital for effective risk assessment and preparedness. Statistical modeling is crucial in quantifying the likelihood of rare, high-magnitude events that can have devastating impacts. Using historical earthquake data and probabilistic techniques, statisticians develop models that estimate these events' frequency, intensity, and geographic distribution. Such models are indispensable not only for seismologists but also for urban planners, engineers, and policymakers who rely on accurate predictions to mitigate potential risks. In this article, we explore the principles of statistical modeling applied to extreme earthquake magnitudes, the challenges of capturing rare events, and how advances in data science enhance our capacity to forecast and manage earthquake risks.

This article examines the frequencies and magnitudes of earthquakes across 11 seismic zones worldwide. They are (in alphabetical order): California (United States), Chile, China, Tibet, India, Iran, Japan, Mexico, Philippines, Taiwan, and Turkey. Seismographs classify

these zones (depicted in Figure 1 below) based on the dynamics of tectonic plate movements, which fall into four distinct categories: subduction zones, collision zones, strike-slip faulting, and complex tectonic regions.

• Subduction Zones: Among the most powerful earthquakes are those that occur in subduction zones, where one tectonic plate is forced beneath another. Seismic zones in Chile, Japan, Mexico, and the Philippines exemplify this phenomenon. Chile's location along the Peru-Chile Trench, where the Nazca Plate subducts beneath the South American Plate, led to the 1960 Valdivia earthquake with a magnitude of 9.5, the most powerful earthquake ever recorded. In Japan, the Nankai Trough experiences frequent megathrust earthquakes as the Philippine Sea Plate subducts beneath the Eurasian Plate, often resulting in tsunamis. Other regions, such as Mexico and the Philippines, also face significant seismic risks from subduction. Mexico's seismic activity arises from the Cocos Plate's subduction beneath the North American Plate along the Middle America Trench, while the Philippine Sea Plate's subduction along the Philippine Trench often triggers earthquakes and volcanic activity in the Philippines.

• Collision Zones: In Central China, Tibet, and India ongoing collisions between the Indian Plate and the Eurasian Plate have formed the Himalayan mountain range and the Tibetan Plateau, both known for intense seismic activity. Earthquakes in these areas result from thrust faulting as the Indian Plate is forced beneath the Eurasian Plate. While both regions involve compressional tectonics, they differ in seismic hazard scale: Central China experiences moderate to large earthquakes, while the Himalayas face a higher risk of catastrophic seismic events due to more significant tectonic forces.

• Strike-Slip Faulting: California and Turkey are characterized by strike-slip faulting, where two tectonic plates slide past each other horizontally. The San Andreas Fault in California and the North Anatolian Fault in Turkey are instances of this type of fault system, capable of producing large earthquakes with seismicity concentrated along the fault lines. In California, the movement occurs between the Pacific and North American Plates, while the North Anatolian Fault involves the westward movement of the Anatolian Plate relative to the Eurasian Plate.

• Complex Tectonic Regions: Iran and Taiwan present highly complex tectonic environments. Iran is situated at the convergence of the Arabian and Eurasian Plates, where both thrust and strike-slip faulting contribute to its status as one of the most seismically active regions in the Middle East. Likewise, Taiwan lies at the intersection of the Eurasian and Philippine Sea Plates, featuring both strike-slip and compressional faulting.

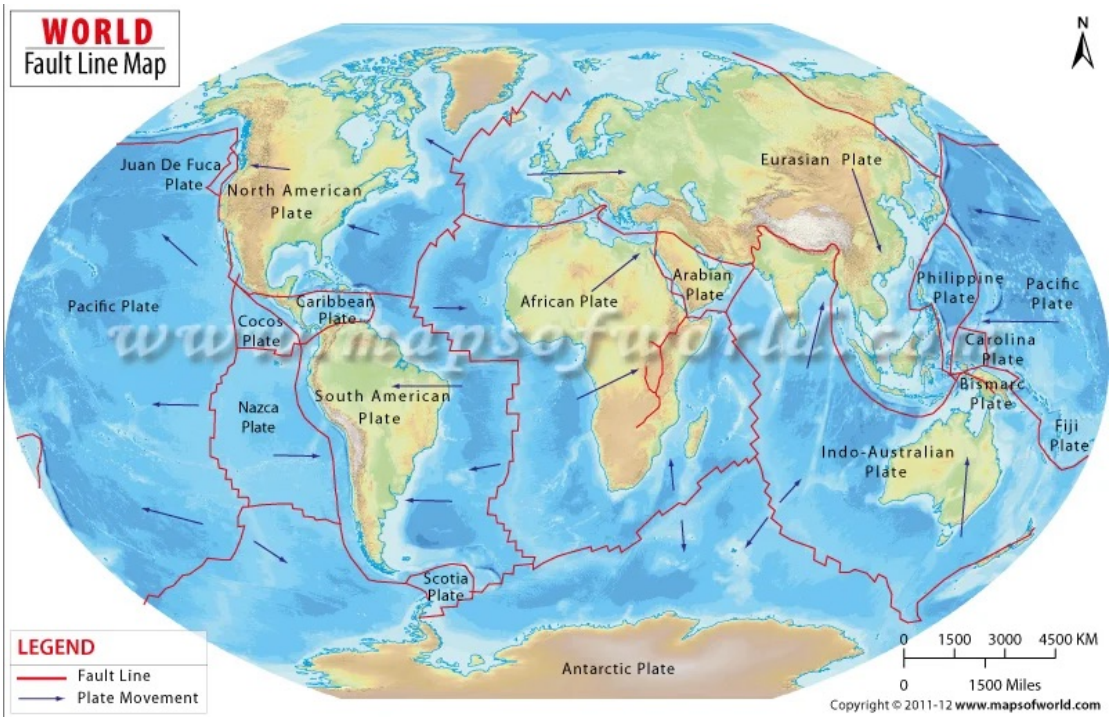

Figure 1: Map of Earthquake Faults Around the World. *Source:* MapsofWorld (`https://www.mapsofworld.com/`)

1.2 Literature Review

There has been significant effort in the past to understand and interpret the physical precursors of seismic events. Physical precursors include foreshocks, ground deformation, changes in groundwater levels, and variations in radon gas emissions among others. Nevertheless, it is a daunting task to predict earthquakes because they are characterized by varying magnitudes and unpredictable patterns [1,2].

In terms of geological approaches, one fundamental concept is the elastic rebound theory, which explains how stress accumulates along faults over time until it exceeds the frictional strength of rocks, leading to an earthquake. Rate-and-state friction laws provide insight into how frictional resistance on faults evolves, helping researchers model aftershocks and future seismic events based on prior slip behavior. Geological studies also play a crucial role in predicting earthquakes. By examining tectonic plate movements and studying paleoseismology, researchers can assess regions at risk for future quakes. Paleoseismology involves analyzing geological records of past earthquakes to identify patterns and recurrence intervals. Geodetic measurements, such as the Global Positioning System (GPS) and the Interferometric Synthetic Aperture Radar (InSAR), allow scientists to monitor ground deformation and stress accumulation along faults with high precision. Strain gauges placed near fault lines provide real-time data on ground movement, helping to identify sudden changes that may signal an impending earthquake.

Probabilistic Seismic Hazard Analysis (PSHA) remains central in the field. It attempts to estimate the probability of different categories of seismic events occurring over time in a given area based on historical earthquake data and tectonic models. Although it's useful for long-term risk assessment, it lacks the temporal precision required for short-term prediction [3,4].

Statistical and probabilistic models have grown in popularity in recent decades as tools to predict earthquakes. These models often rely on historical earthquake data to estimate the likelihood of future events. For example, the Poisson process is a model frequently used in seismology, assuming that earthquakes occur independently over time. However, this assumption has come under criticism, leading to the development of more complex models such as the Epidemic Type Aftershock Sequence (ETAS) model [5]. This model is an extension of the Poisson process to allow for aftershocks, which represents the fact that if an earthquake occurs in one geographic region, further events are more likely in the same region. This model, however, is not very useful in predicting impending main shocks [6]. A third statistical approach is Bayesian inference, which integrates prior knowledge and rapidly updates predictions as new data become available, making it more effective for real-time forecasting [7].

Recent advances in machine learning and artificial intelligence have introduced new prospects in earthquake prediction. Only a few highly complex machine learning algorithms can sift through seismic data to pick out patterns and correlations that are not evident in other forms of research. The machine learning models employ random forests, support vector machines, and neural networks to predict future earthquakes. A major advantage of these models is their capability to deal with any sort of complexity in a dataset. Deep learning, which is a subcategory of machine learning, also shows some potential for earthquake prediction. Zhu and Beroza [8] developed a deep-learning model that could detect low-frequency earthquakes more accurately than traditional methods. Their model uses convolutional neural networks to process seismic waveforms, allowing them to identify subtle patterns that could represent precursors to larger seismic events.

Despite these developments, the use of machine learning for earthquake prediction remains in its early stages. One of the greatest challenges is the "black box" nature of many machine learning models: it is hard to understand what physical processes the models are capturing. Additionally, research is now being conducted to test how well such models will do in generalizing across different seismic regions and conditions [9]. Currently, prediction through geospatial technologies like remote sensing and Geographic Information Systems (GIS) is playing a central role. Satellite-based remote sensing allows the monitoring of both ground deformation and movements of faults among other geophysical phenomena in order to provide earlier warning signs of seismic activity. In addition, GIS technologies integrate and analyze spatial data originating from multiple sources, which enables seismic hazards to be mapped out and areas with high risks to be recognized [10].

Due to the complexity of earthquake prediction, integration of multiple disciplinary approaches became essential. By using each model's advantages, multi-disciplinary models

using physical, statistical, and machine-learning methods increase the accuracy of prediction. For example, Jordan et al. suggest a multi-tiered approach where seismic risk might be assessed using geological surveys, statistical models, and real-time analysis of seismic data [11]. On the other hand, with regard to standardized testing and evaluation of models for earthquake forecasting, the Collaboratory for the Study of Earthquake Predictability (CSEP) has been very much at the forefront. The efforts of CSEP underline the real requirement for rigorous testing of models and open sharing of data in the seismological community [12].

Despite all the advances successfully made in this field, earthquake prediction still faces many challenges. The difficulty comes from the fact that earthquakes are unpredictable, the modeling is crude, and data availability is poor. Moreover, the ethical and societal implications that go with earthquake prediction, such as the possibility of false alarms leading to public unrest, make the challenge even bigger within this field [13]. Future research is likely to focus on improving the accuracy and interpretability of machine learning models, as well as the integration of diverse data sources, including seismic, geodetic, and geophysical data. Advances in computational power and data processing techniques will also play a critical role in enabling more sophisticated analyses and real-time predictions [14].

1.3 Data Overview

This study utilizes data spanning from June 15, 1980, to June 15, 2024, encompassing eleven global locations in alphabetical order: California (United States), Chile, China, Tibet, India, Iran, Japan, Mexico, the Philippines, Taiwan, and Turkey. The data are sourced from the United States Geological Survey Earthquake Hazards Program <https://www.usgs.gov/programs/earthquake-hazards>, which monitors earthquakes, assesses their impacts, and conducts research under the National Earthquake Hazards Reduction Program, a collaborative effort of four federal agencies established by Congress.

For our analysis, we concentrated on the time stamp (date + time), and magnitude recorded on the Richter scale, filtering for magnitudes of 3.0 and above. A snippet of the data for Tibet is presented in Table 1 below.

1.4 Article Layout

This article is organized as follows: For each seismic region, we first model the frequency and intensity of monthly maximum magnitudes. This analysis is based on the theoretical framework of extreme value distributions. Next, we focus on modeling all earthquake occurrences with a minimum of 3.0 magnitudes through time series analysis. We examine the temporal behavior of magnitudes using nonparametric methods such as locally fitted linear regressions and splines, while also fitting various parametric time series models, building up to an autoregressive integrated moving average model. We then perform anomaly detection analysis using machine learning techniques. Finally, we analyze the distribution of interarrival times to model earthquake occurrences as a Poisson process.

Time	Magnitude
2024-06-14T23:24:18.637Z	4.6
2024-06-09T16:01:23.107Z	4.4
2024-06-05T12:04:59.113Z	4.3
2024-06-03T18:54:06.025Z	4.6
2024-06-03T18:53:48.588Z	4.2
2024-06-01T10:59:01.251Z	4.3
2024-06-01T03:00:02.003Z	4.4
2024-06-01T00:46:37.593Z	5.6
2024-05-28T01:13:06.793Z	4.3
2024-05-26T23:02:53.725Z	4.4

Table 1: Tibet Earthquake Data Snippet

2 Extreme Value Modeling

2.1 Theoretical Framework

According to the Fisher–Tippett–Gnedenko theorem, when properly normalized, the maxima of earthquake magnitudes converge to a Gumbel distribution as the sample size increases. To express this mathematically, consider a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$ and let $M_n = \max(X_1, X_2, \ldots, X_n)$ represent their maximum value. To examine the limiting behavior of M_n, we define random variables $Z_n = \dfrac{M_n - b_n}{a_n}$ where $a_n > 0$ and b_n are sequences that depend on n and are chosen to ensure convergence. The theorem asserts that as n approaches infinity, the distribution of Z_n converges to one of three types: Gumbel, Fréchet, or Weibull, depending on the characteristics of the original distribution of X_i's. Notably, it is established that the maxima of earthquake magnitudes follow a Gumbel distribution in the limit (**reference is needed**). The cumulative distribution function of the Gumbel distribution is given by

$$F(x; \mu, \beta) \, = \, exp\Big(- \, e^{-\frac{x-\mu}{\beta}} \Big), \; x > 0,$$

where μ is the location parameter, and $\beta > 0$ is the scale parameter.
When fitting a Gumbel distribution to data using R, such as through the 'eva::fit()' function, the code estimates the location and scale parameters of the distribution using Maximum Likelihood Estimation (MLE). Specifically, this function takes a dataset as input and provides estimates for the parameters that describe the Gumbel distribution, which is characterized by its ability to model the distribution of maximum values. The normalizing constants a_n and b_n play a crucial role in Extreme Value Theory (EVT) by rescaling the sequence of maxima $M_n = \max(X_1, \ldots, X_n)$ of independent and identically distributed (i.i.d.) random variables X_i. As n approaches infinity, the normalized form $(M_n - b_n)/a_n$ converges to a Gumbel distribution, allowing for effective modeling of extreme events. Thus, a_n adjusts the spread of the maxima, while b_n shifts the distribution, ensuring that the rescaled maxima conform to the limiting Gumbel distribution in the context of EVT.
In practice, a block maxima approach is employed to model extreme values. This method

involves analyzing the maximum values over specific intervals, or "blocks." Statistical techniques, such as maximum likelihood estimation, are used to estimate the parameters of the Gumbel distribution based on the observed maxima.

2.2 Applications

In this analysis, we used months as blocks and calculated the maximum earthquake magnitudes for each month. These maxima were then normalized according to extreme value theory, and we fitted the Gumbel distribution using the maximum likelihood estimation method, supported by numerical optimization. The same process was applied to data from each of the 11 seismic zones individually. We calculated asymptotic confidence intervals for the Gumbel parameters and produced diagnostic plots such as Q-Q plots, extreme value analysis (EVA) plots, and histograms. We demonstrate our analysis for the Tibet region, with Figure 2 below showing the corresponding graphs.

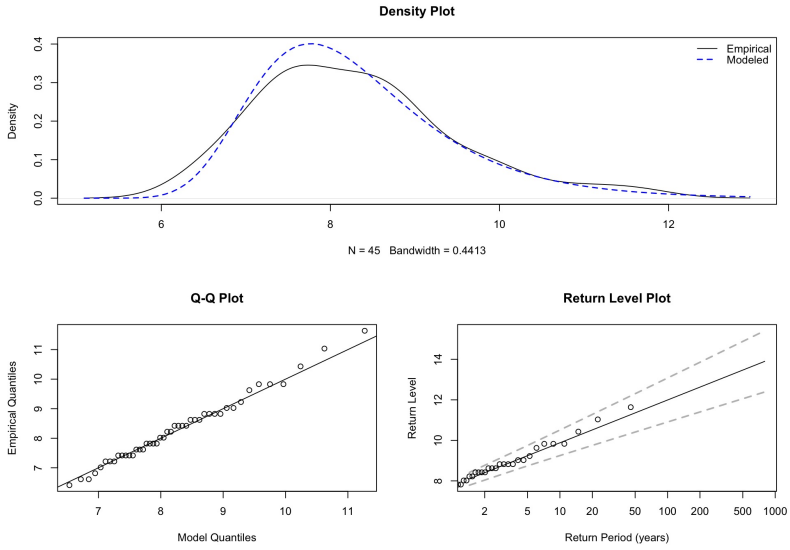

Figure 2: Tibet regions: Q-Q plot and EVA plots

For instance, in the case of the Tibet region, the location parameter was estimated at 7.77 with a 95% confidence interval (CI) of [7.48, 8.05], and the scale parameter was estimated at 0.92 with a 95% CI of [0.71, 1.13]. The mean was calculated as 8.30 with a 95% CI of [7.95, 8.64]. A chi-squared goodness-of-fit test yielded a p-value of about 0.30, indicating that the Gumbel distribution fits the data well. To visually represent this fit, a histogram was overlaid with the Gumbel curve (see Figure 3 below).

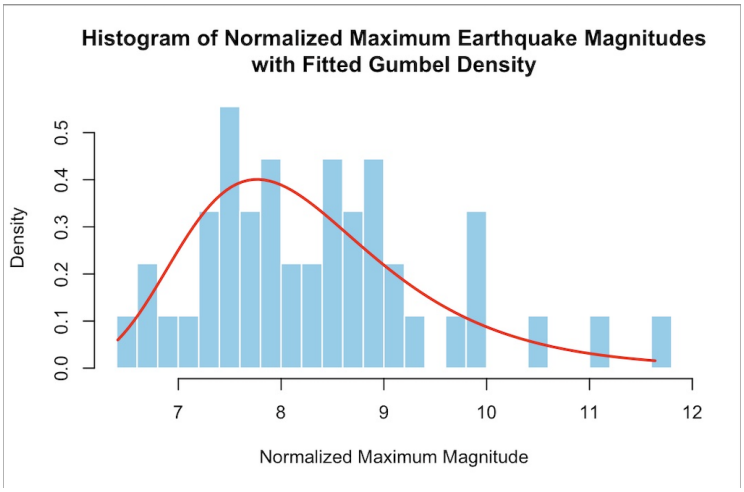

Figure 3: Histogram of Normalized Maximum Earthquake Magnitudes with Fitted Gumbel Density for Tibet

For each of the eleven sites, when performing the chi-squared test, a degree of freedom of 9 is used (12 categories, minus 1 for the total and 2 for the estimated parameters). The Gumbel distribution fits well the data from eight out of eleven sites. The results for all the sites are summarized in Table 2 that follows.

Region	$\hat{\mu}$	95%LCL	95%UCL	$\hat{\beta}$	95%LCL	95%UCL	χ^2	p-value	Gumbel?
China	7.1843	6.8852	7.4833	0.9738	0.7455	1.2020	22.1646	0.0084	No
Tibet	7.7667	7.4841	8.0493	0.9177	0.7085	1.1268	10.7223	0.2952	Yes
India	7.0592	6.7236	7.3947	1.0906	0.8435	1.3377	16.6770	0.0540	Yes
Iran	7.8013	7.4843	8.1183	1.0276	0.7942	1.2610	14.0177	0.1220	Yes
Japan	9.0775	8.7986	9.3564	0.9012	0.7053	1.0971	9.3298	0.4070	Yes
Mexico	8.9628	8.6157	9.3100	1.1205	0.8848	1.3562	12.7932	0.1720	Yes
Philippines	9.1371	8.8277	9.4464	0.9979	0.7794	1.2165	17.8582	0.0369	No
Chile	7.8357	7.5352	8.1361	0.9822	0.7486	1.2159	13.9092	0.1256	Yes
Taiwan	7.4839	7.1535	7.8142	1.0702	0.8309	1.3095	7.4460	0.5908	Yes
Turkey	7.1988	6.8913	7.5062	1.0024	0.7625	1.2422	17.9831	0.0354	No
US California	6.8259	6.4724	7.1793	1.1467	0.8831	1.4102	11.5445	0.2402	Yes

Table 2: ML Estimates and Chi-squared Test Results for 11 Regions with 95% CI for both Location and Scale Parameters

2.3 Frequency analysis

For the eleven sites, the time differences between consecutive monthly maxima are determined and used to estimate the date and magnitude of the next significant earthquake.

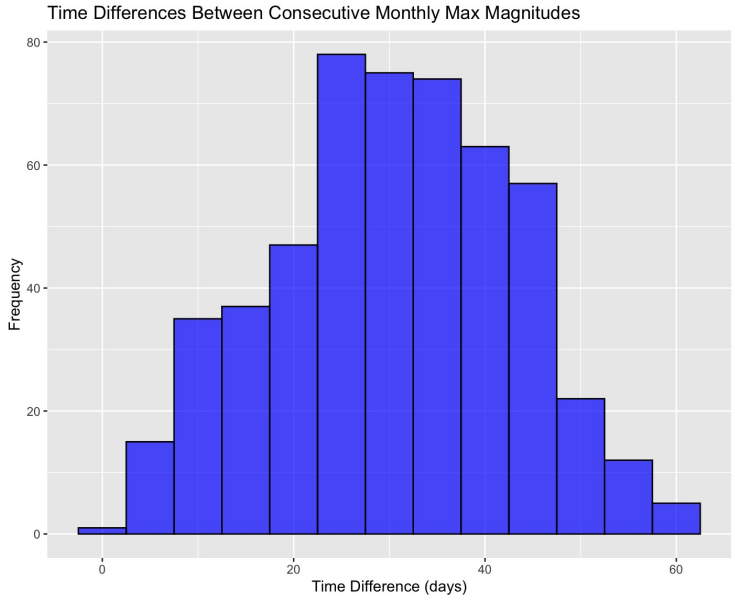

Figure 4: Histogram of Time Difference between Two Consecutive Monthly Maxima for Tibet

With information including the mean, median, and mode of the amount of time between monthly maxima, predictions are formed. For example, for Tibet, the time difference variable had a median of 31 days, a mean of 30.55 days, and a mode of 27 days. The CI was [7, 53]. For the magnitude variable, there was a median of 5, a mean of 5.077, and a mode of 5. The 95% CI for the mean was [4.2, 6.4]. The prediction was made that from June 30th, 2024 to July 3rd, 2024, there would be an earthquake in Tibet with a magnitude between 5.032 and 5.123, but neither the magnitude nor the date was correctly predicted. Using monthly maxima to predict the next significant earthquake did not prove to be useful, as only four out of the eight sites had correct date predictions, and none had correct magnitude predictions.

Region	China	Tibet	India	Iran	Japan	Mexico	Philippines	Chile	Taiwan	Turkey	US California
					Time Difference						
Median	31	31	31	30	31	30	31	30	30	30	31
Mean	30.89	30.55	31.18	30.53	30.39	30.4	30.42	30.63	31.21	30.58	30.4
Mode	31	27	33	21	29	31	36	23	26	21	28
95%LCL	6	7	7	6	7	7	7.175	7.575	9	5.725	6
95%UCL	55.05	53	56	54	54	54	53	54.425	56	56	54.825
					Max Magnitude						
Median	4.8	5	4.8	5.1	5.6	5.4	5.5	5	4.8	4.7	4.435
Mean	4.878	5.077	4.866	5.18	5.655	5.462	5.628	5.138	4.9	4.872	4.597
Mode	4.8	5	4.6	5	5.4	5.2	5.2	4.7	4.5	4.5	4.2
95%LCL	4	4.2	4.1	4.5	4.8	4.414	4.9	4.2	4	4	3.6
95%UCL	6.2	6.4	6.2	6.5	7.1	7.1	7.1	6.6425	6.4	6.3275	6.1
					Prediction						
Magnitude Prediction	4.8779	5.0774	4.8656	5.1805	5.6551	5.4623	5.6282	5.1375	4.9002	4.8720	4.5974
95%LCL (Magnitude)	4.8281	5.0319	4.8173	5.1352	5.6041	5.4034	5.5796	5.0815	4.8449	4.8211	4.5395
95%UCL (Magnitude)	4.9276	5.1228	4.9138	5.2258	5.7061	5.5212	5.6768	5.1935	4.9555	4.9228	4.6554
Date Prediction	2024-07-08	2024-07-02	2024-07-09	2024-07-14	2024-07-02	2024-07-05	2024-07-11	2024-07-08	2024-07-03	2024-07-10	2024-07-08
95%LCL (Date)	2024-07-07	2024-06-30	2024-07-08	2024-07-12	2024-07-01	2024-07-04	2024-07-10	2024-07-07	2024-07-02	2024-07-08	2024-07-07
95%UCL (Date)	2024-07-09	2024-07-03	2024-07-10	2024-07-15	2024-07-03	2024-07-06	2024-07-12	2024-07-09	2024-07-04	2024-07-11	2024-07-10
Correct? (Magnitude)	No	No	No	No	No	No	No	No	No	No	No
Correct? (Date)	No	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No

Table 3: Time Difference, Max Magnitude, and Prediction Analysis for Each Region

3 Nonparametric Modeling

3.1 Locally Estimated Scatterplot Smoothing Method

3.1.1 Theoretical Framework

The locally estimated scatterplot smoothing (LOESS) is a nonparametric method for modeling a series of data with no assumptions about the distribution of the data. The general form of modeled relation is defined by the formula

$$y = f(x, \ldots, x_k) + \varepsilon \tag{1}$$

where f is an unknown function, and ε represents the error terms that are independently and identically distributed with a zero mean and constant variance. The LOESS method evaluates the function f at each data point and plots on the scatterplot a curve that connects the fitted points by straight lines.
The estimation of f at each point in the data is done by fitting a weighted polynomial in the local neighborhood of each point. The smoothing parameter p/n represents the fraction of all points that are captured by each local neighborhood $\mathcal{N}_p(\mathbf{x}^0)$ with the center at a fixed point $\mathbf{x}^0 = (x_1^0 \ldots, x_k^0)$. A weighted linear regression of the form

$$l(\mathbf{x}) = l(x_1, \ldots, x_k) = \beta_0 + \beta_1(x_1 - x_1^0) + \cdots + \beta(x_k - x_k^0)$$

is fitted through the points $\mathbf{x} = (x_1, ..., x_k)$ in $\mathcal{N}_p(\mathbf{x}^0)$. The line is weighted because it is chosen in such a way that it minimizes the sum

$$\sum_{\mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}^0)} \Big(y_i - l(\mathbf{x}_i)\Big)^2 w\Big(\frac{\|\mathbf{x}_i - \mathbf{x}^0\|}{r(\mathbf{x}^0)}\Big)$$

where $\mathbf{x}_i = (x_{1i}, \ldots, x_{ki})$, $i = 1, \ldots, p$. Here $\|\cdot\|$ denotes the standard Euclidean distance, that is,

$$\|\mathbf{x}_i - \mathbf{x}^0\| = \sqrt{(x_{1i} - x_1^0)^2 + \cdots + (x_{ki} - x_k^0)^2},$$

$w(\cdot)$ is the weight function, and $r(\mathbf{x}^0) = \max\limits_{\mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}^0)} \|\mathbf{x}_i - \mathbf{x}^0\|$ is the radius of the neighborhood. The weight function used in the analysis is the normalized tricube function $w(x) = \dfrac{32}{5}(1 - \|x^3\|)^3$, if $\|x\| \leq 1$, and 0, otherwise.

3.2 Thin-plate Smoothing Spline Method

3.2.1 Theoretical Framework

The thin-plate smoothing spline (TPSS) method is another nonparametric method that estimates the function f in (1) by minimizing the sum of squares of the residuals

$$\sum_{i=1}^{n} \big(y_i - f(x_{1i}, x_{2i}, \ldots, x_{ki}\big)^2 + \lambda\, J_m(f)$$

where λ is a smoothing parameter responsible for the smoothness of the fitted surface, and the roughness penalty term J_m is defined as

$$J_m(f) = \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum \frac{m!}{\alpha_1! \alpha_2! \ldots \alpha_k!} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_k^{\alpha_k}} \right)^2 dx_1 \cdots dx_k$$

with $\alpha_1 + \cdots + \alpha_k = m$. The quantity m is the degree of smoothness of the function f. A common approach is to consider functions f of the following algebraic form:

$$f(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_n \mathbf{x}_n + \sum_{i=1}^{n} w_i \|\mathbf{x} - \mathbf{x}_i\|^2 \ln\left(\|\mathbf{x} - \mathbf{x}_i\|\right)$$

where $\beta_0, \ldots, \beta_n$ are real-valued coefficients, $\mathbf{x}_i = (x_{1i}, \ldots, x_{ki})$, $i = 1, \ldots, n$ are the data points, and w_i's are real-valued weights for each data point.

3.3 Application

We fit LOESS and TPSS to the data for all eleven sites. The smoothing parameter for LOESS is chosen as 0.05, whereas for TPSS it is 0.000005. Graphs for Tibet data are presented in Figures 5 and 6 below. From the graphs, despite highly dispersed data points, the fitted LOESS and TPSS curves reveal periodic fluctuations, indicating an underlying trend amidst the noise.

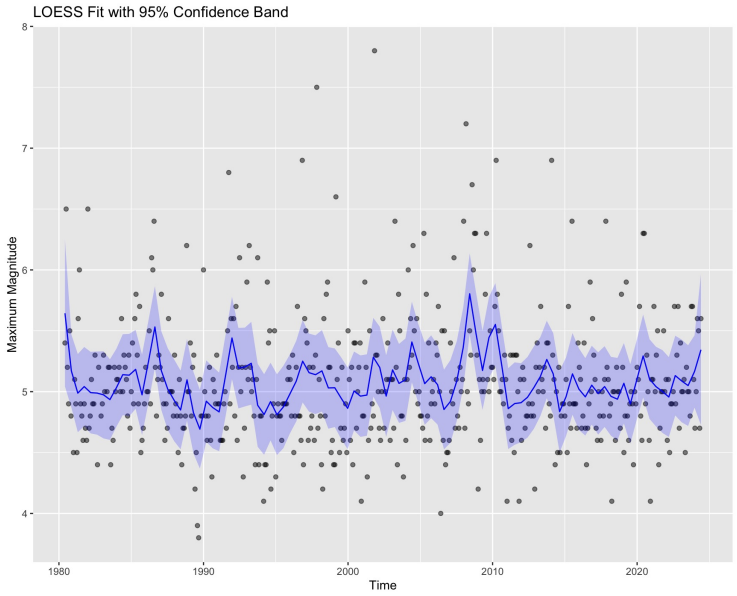

Figure 5: Fitted LOESS Curve with a 95% Confidence Band for Tibet

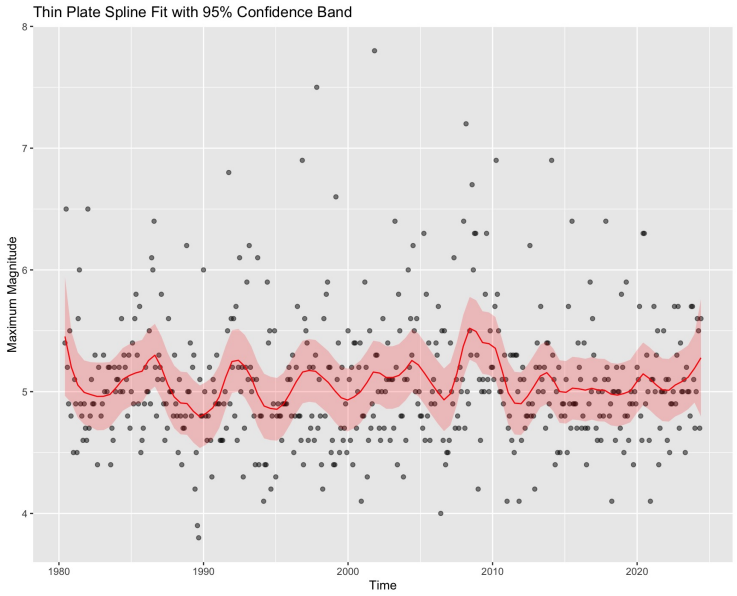

Figure 6: Fitted TPSS Curve with a 95% Confidence Band for Tibet

4 Time Series Models

4.1 Theoretical Framework

Time series models are crucial for analyzing data collected over time. A commonly used model is the autoregressive (AR) model. For a time series y_t, the AR(p) model is defined as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where $\phi_1, \ldots, \phi_p$ are the autoregressive parameters, and ε_t is a white noise error term with a mean of zero and constant variance σ^2. The autoregressive structure implies that the current value y_t is regressed on its past values up to lag p. Similarly, a moving average (MA) model of order q is written as:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\theta_1, \ldots, \theta_q$ are the moving average coefficients, and ε_t is a random process with mean zero and variance σ^2. The MA model expresses the current value y_t as a linear combination of past errors. The autoregressive moving average (ARMA) model, which combines both AR(p) and MA(q) models, is given by:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

This model captures both the dependence on past values and past errors, making it suitable for stationary time series. To ensure that a time series is stationary, we often employ the Augmented Dickey-Fuller (ADF) test. The ADF test examines whether the data have a unit root (non-stationary) or is stationary. The ADF regression equation is:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_p \Delta y_{t-p} + \varepsilon_t$$

where $\Delta y_t = y_t - y_{t-1}$. The test statistic is calculated as:

$$\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}.$$

If the test statistic τ is smaller than the critical value, we reject the null hypothesis (non-stationarity) and conclude that the process is stationary. The Partial Autocorrelation Function (PACF) helps in identifying the appropriate lag p for AR models by measuring the correlation between y_t and y_{t-k}, accounting for the effect of intermediate lags $1, \ldots, k-1$. The PACF at lag k is the correlation between y_t and y_{t-k} after removing the effect of the lags between 1 and $k-1$. Significant spikes in the PACF plot indicate the lags that should be included in the model.

To compare the model fits, the Akaike Information Criterion (AIC) is widely used as a metric for model selection. It is given by $AIC = -2lnL + 2k$ where L is the maximum value of the likelihood function of the model, and k is the total number of parameters the model utilizes. A lower AIC value indicates a better-fitting model, as it achieves a better trade-off between goodness of fit and simplicity.

4.2 Application

We refined the data by selecting only events with magnitudes greater than 6.0 and applied a range of time series models to each site. The optimal model for each location was determined by comparing Akaike Information Criterion (AIC) values, with the model having the lowest AIC being chosen as the best fit. Figure 7 below illustrates both the fitted model predictions and the actual observed data for the Tibet region. In this case, the AR(1) model provided the best fit for the region's time series data.

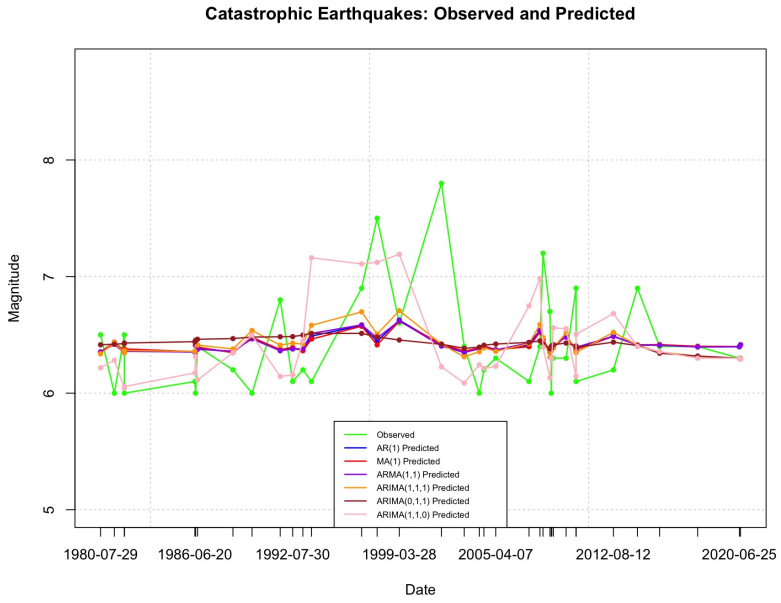

Figure 7: Fitted Time Series Models for Tibet's Data

Table 4 below presents the AIC values for each fitted model across all eleven regions. The model with the lowest AIC is selected as the best fit.

Model	China	Tibet	India	Iran	Japan	Mexico	Philippines	Chile	Taiwan	Turkey	US California
AR(1)	43.67524	45.71278	36.44550	62.99334	194.1014	179.6045	192.7338	130.1584	44.97536	53.1627	24.76077
MA(1)	43.40414	45.76629	35.45929	62.99220	194.0690	179.7224	192.8627	130.1316	44.93794	53.19481	24.24802
ARMA(1,1)	40.64334	47.68394	36.52129	64.99003	195.9143	181.3718	193.4757	131.7950	45.34316	52.74383	24.62957
ARIMA(1,1,1)	47.32254	48.88182	39.71896	66.50585	198.9591	184.0309	197.3524	134.1322	49.04869	56.27383	28.24141
ARIMA(0,1,1)	46.10187	48.09877	39.60975	64.84639	197.3393	183.0446	196.0604	132.2398	47.23685	54.37357	26.86149
ARIMA(1,1,0)	58.05595	55.67496	47.38252	81.21313	277.3341	220.9284	243.2171	170.7055	62.85192	61.87926	35.11636
Best Model	ARMA(1,1)	AR(1)	AR(1)	MA(1)	MA(1)	AR(1)	AR(1)	MA(1)	MA(1)	ARMA(1,1)	MA(1)

Table 4: AIC Values and Model Comparison for Different Regions

5 Anomaly Detection

5.1 Theoretical Framework

Anomalies in time series data are often identified by examining the residuals after accounting for trends and seasonality. This is done through a process called de-trending, where the trend and seasonal elements are removed, leaving behind residuals that represent the core fluctuations in the data. These residuals are then analyzed for deviations, with significant outliers standing out from the expected range.

Outliers are typically determined using the interquartile range (IQR). Values are flagged as outliers if they fall below $Q_1 - 3 \cdot IQR$ or exceed $Q_3 + 3 \cdot IQR$. The first quartile (Q_1) represents the point below which 25% of the data lies, while the third quartile (Q_3) represents the value below which 75% of the data falls. The IQR, calculated as $IQR = Q_3 - Q1$ captures the range of the middle 50% of the data, making it a robust indicator of variability that is less sensitive to extreme outliers.

The default threshold of $3 \cdot IQR$ is not fixed. In R, outliers are identified as values that fall $0.15/\alpha \cdot IQR$ beyond the quartiles. The default setting of $\alpha = 0.05$ results in a multiplicative factor of 3. Increasing α causes more observations to be considered outliers, while decreasing α results in fewer observations being classified as outliers.

To examine further the identified anomalies, a bootstrap method is applied, and a bootstrap confidence interval is computed. This resampling technique involves drawing a large number of samples (typically over 1,000) with replacements from the original dataset, ensuring that the resample sizes match the original sample size. The mean or other statistic of interest is then calculated for each bootstrap sample. The distribution of these statistics, often visualized as a histogram, forms an empirical distribution. Confidence intervals are then derived from this distribution, typically by taking percentiles of the resampled statistics, such as the 2.5th and 97.5th percentiles for a 95% confidence interval. Such a confidence interval is known as an Efron percentile confidence interval.

5.2 Application

We detect anomalies separately for each region, selecting only events with magnitudes greater than 5.0 and adjusting the value of α to have ten to fifteen anomalies showing. For instance, Tibet has an alpha value of 0.04, producing 14 anomalies. If anomalies are less than 30 days apart, they are regarded as aftershocks and do not count separately. Figure 8 visualizes the anomalies for the Tibet data set.

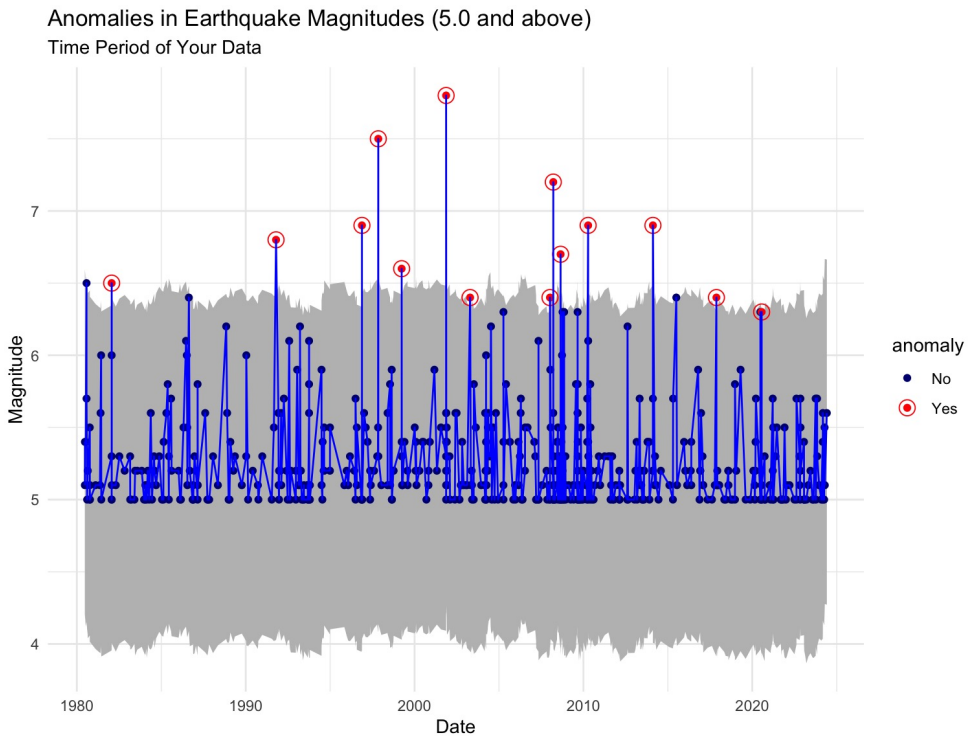

Figure 8: Anomalies for Tibet Earthquake Data

To continue our analysis of anomalies, we concentrate on the anomalies specific to each region and apply the bootstrap method to calculate the point estimate of the mean magnitude, along with the Efron percentile confidence interval for that mean (see Table 5 below). By excluding the last anomaly, we can determine whether this data point is covered by the confidence interval produced by our analysis. For the region of Tibet, the mean magnitude is calculated as 6.84, with a 95% confidence interval of [6.58, 7.2]. The last anomaly, with a severity of 6.6, is captured by this confidence interval, indicating a reasonable prediction for this region.

Metric	China	Tibet	India	Iran	Japan	Mexico	Philippines	Chile	Taiwan	Turkey	US California
Alpha	0.06	0.04	0.05	0.045	0.033	0.042	0.03	0.045	0.08	0.06	0.1
# Anomalies	11	14	10	15	12	11	13	15	14	10	13
Severity											
Mean	6.84	6.846	6.778	6.864	7.273	7.46	7.4	7.136	6.723	6.944	6.667
95%LCL	6.58	6.623	6.567	6.671	7.173	7.28	7.292	6.85	6.515	6.767	6.467
95%UCL	7.20	7.061	7.055	7.064	7.373	7.66	7.517	7.479	6.961	7.133	6.883
Median	6.6	6.8	6.7	6.7	7.3	7.4	7.4	7.0	6.7	6.8	6.65
95%LCL	6.5	6.5	6.5	6.6	7.1	7.2	7.25	6.7	6.4	6.7	6.45
95%UCL	7.25	6.9	6.9	7.1	7.4	7.7	7.6	7.2	6.9	7.2	7.0
Frequency											
Mean	1662.11	1090.17	1421.63	894.77	1448.10	1569.00	1054.55	1066.31	1090.92	1689.38	1229.82
95%LCL	869	612	762	442.5	562	826	599	448	811	822	532
95%UCL	2543	1689	2110	1400.5	2457	2306	1647	1768	1348	2723	2242
Median	1613	779	1137	600	587	1200	770	949	1191	1528.5	694
95%LCL	292	429.5	543	308	307	588	410	115.5	780	280	316
95%UCL	2373	1564.5	2539	1384.8	2730	2823	1695	1234	1451	2713	1528
Last Frequency	241	978	1315	1671	656	818	3700	455	563	829	419
Last Severity	6.6	6.3	7.8	7.3	7.5	7.6	7.6	6.7	7.4	7.8	6.0
Frequency in CI?	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
Severity in CI?	Yes	No	No	No	No	Yes	Yes	Yes	No	No	No

Table 5: Anomaly Severity and Frequency Analysis Across Regions

6 Interarrival Time Analysis

6.1 Theoretical Framework

In earthquake analysis, events are often modeled using a Poisson process, which has three key assumptions: events are independent, occur at a constant rate, and the probability of a future event is unaffected by past events. The intervals between consecutive events, referred to as interarrival times, are modeled as exponentially distributed random variables. For a Poisson process with a rate λ, the interarrival times T_i follow an exponential distribution characterized by the probability density function:

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

The mean interarrival time is $1/\lambda$, representing the expected time between events. This model implies that shorter intervals between earthquakes are more likely, which is often observed in aftershock sequences.

The assumption of exponentially distributed interarrival times is widely used in seismology to model random earthquake occurrences. Estimating λ from historical data allows researchers to model future events. Deviations from this model may indicate non-random patterns, such as earthquake clustering or changes in seismic activity.

To determine if the interarrival times adhere to an exponential distribution, a standard chi-squared goodness-of-fit test is conducted.

6.2 Application

We analyze data from all eleven regions, focusing on earthquakes with magnitudes of 3.0 or 5.0 to keep the number of events manageable. We compute the interarrival times and plot a histogram for each site. The chi-squared test is performed, with the parameter λ estimated as the reciprocal of the sample mean of the interarrival times. Figure 9 shows a histogram of the interarrival times for the Tibet data, alongside the fitted exponential distribution

curve. The chi-squared test statistic is 8.982936, yielding a p-value of 0.3437375, which exceeds 0.05. This suggests that the interarrival times follow an exponential distribution. Similarly, the analysis confirms that the interarrival times for all eleven regions fit an exponential distribution (see Table 6 below).

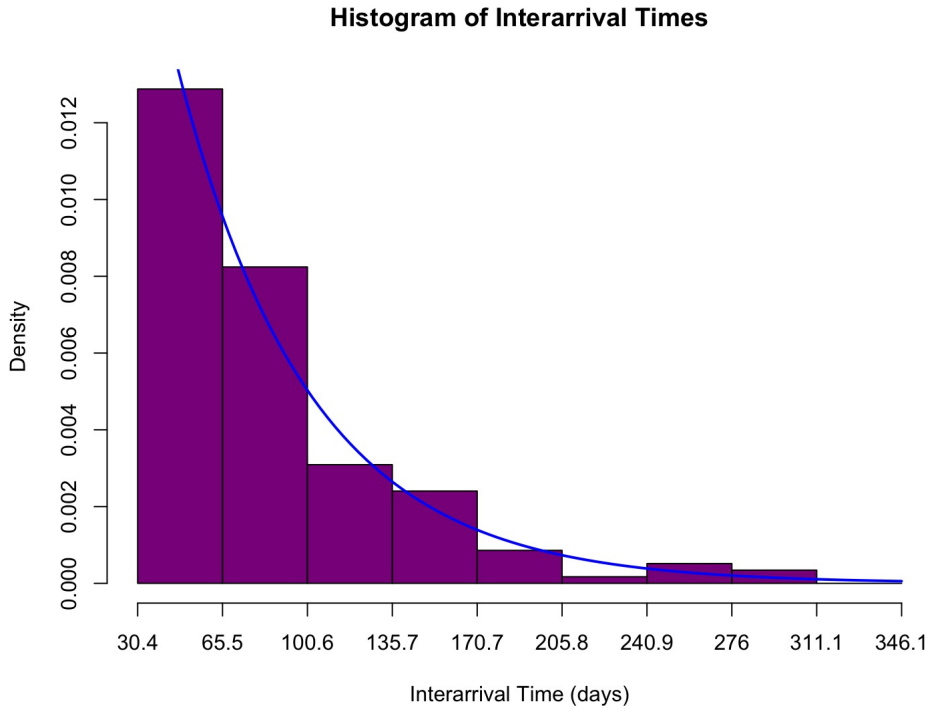

Figure 9: Interarrival Times for Tibet Overlaid with Exponential Fit

Metric	China	Tibet	India	Iran	Japan	Mexico	Philippines	Chile	Taiwan	Turkey	US California
Aftershock Max.	30	30	25	20	30	25	20	30	30	30	30
Min Mag	5	5	3	3	5	5	5	3	3	3	4
Chi-squared Stat.	8.9829	7.1373	10.7192	9.9317	11.7866	8.0913	3.8408	5.6301	12.2221	7.8538	5.0669
p-value	0.3437	0.5219	0.2181	0.2699	0.1610	0.4246	0.8712	0.6886	0.1416	0.4479	0.7504

Table 6: Interarrival Time Distributions Across Regions

7 Summary and Discussion

In this study, we explored various models to analyze earthquake data across different seismic regions. The investigation included modeling the frequency and intensity of monthly maximum magnitudes using extreme value distributions, followed by a time series analysis of earthquake occurrences We applied both nonparametric methods, such as locally fitted regressions and splines, and parametric models, including autoregressive integrated moving average (ARIMA) models, to examine temporal behavior. Additionally, we employed machine learning techniques for anomaly detection and modeled earthquake occurrences

using a Poisson process based on interarrival times.

Despite our comprehensive approach, none of the models provided consistently strong performance across all regions or earthquake magnitudes. Further research is necessary to develop more robust models that can better handle the complexity and unpredictability of earthquake data. Possible directions include incorporating more advanced machine learning techniques such as deep learning or considering physical factors like tectonic movements to improve prediction accuracy.

Supplemental Materials

All the data sets, codes, and plots can be accessed here: GitHub Repository: Earthquake-ML

Acknowledgments

Thank you to everyone who provided support throughout this journey, including my family, friends, and Areteem Institute for introducing me to this opportunity. I am especially grateful to Dr. Olga Korosteleva for her constant guidance and encouragement.

References

1. Scholz, Christopher H. *The Mechanics of Earthquakes and Faulting.* Cambridge University Press, 2002.

2. Vere-Jones, David. "Stochastic Models for Earthquake Occurrence." Journal of the royal statistical society series b-methodological 32 (1970): 1-45.

3. Cornell, C. Allin. "Engineering Seismic Risk Analysis." *Bulletin of the Seismological Society of America*, vol. 58, no. 5, 1968, pp. 1583-1606.

4. McGuire, Robin K. *Seismic Hazard and Risk Analysis.* Earthquake Engineering Research Institute, 2004.

5. Ogata, Yosihiko. "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association*, vol. 83, no. 401, 1988, pp. 9-27.

6. Alejandro Veen and Frederic P. Schoenberg. "Estimation of Space–Time Branching Process Models in Seismology Using an EM–Type Algorithm." *Journal of the American Statistical Association*, vol. 103, no. 482, 2008, pp. 614-624.

7. Marzocchi, Warner & Zechar, J. & Jordan, Thomas. (2012). *Bayesian Forecast Evaluation and Ensemble Earthquake Forecasting.* The Bulletin of the Seismological Society of America. 102. 2574-2584. 10.1785/0120110327.

8. Zhu, Weiqiang, and Gregory C. Beroza. "PhaseNet: A Deep-Neural-Network-Based Seismic Arrival-Time Picking Method." *Geophysical Journal International*, vol. 216, no. 1, 2019, pp. 261-273.

9. DeVries, Phoebe M R et al. "Deep learning of aftershock patterns following large earthquakes." Nature vol. 560,7720 (2018): 632-634. doi:10.1038/s41586-018-0438-y

10. Rouet-Leduc, Bertrand Hulbert, Claudia Lubbers, Nicholas Barros, Kipton Humphreys, Colin Johnson, Paul. (2017). Machine Learning Predicts Laboratory Earthquakes. Geophysical Research Letters. 44. 10.1002/2017gl074677.

11. Gahalaut, Vineet K., et al. "Earthquake Hazard in Northeast India: A Synthesis." *Journal of Seismology*, vol. 20, no. 2, 2016, pp. 517-532.

12. Jordan, Thomas H., et al. "Operational Earthquake Forecasting: State of Knowledge and Guidelines for Utilization." *Annals of Geophysics*, vol. 54, no. 4, 2011, pp. 315-391.

13. Jordan, Thomas Hutt, New Liukis, Masha Maechling, Philip Schorlemmer, D. ETH, Zurich, Switzerland, Zechar, J. Collaboration, CSEP. (2006). Collaboratory for the Study of Earthquake Predictability.

14. Clucas, Nathaniel. *Rare Events with High Risks: Extreme Value Theory with Applications in R.* 2019. Master's thesis, California State University, Long Beach.