# Exploring Earthquake Forecasting using Machine Learning

Jingyi Luo (jl6zh)

## Identify the Problem

Earthquake happens unexpected most times. When it occurs, earthquake can cause significant economic and social damages. Currently, earthquake forecasting is far from applicable. The United States Geological Survey (USGS) publish 50-100-year hazard maps (https://earthquake.usgs.gov/hazards/hazmaps/) that provide insights on long term (more than 1 year) estimate of future earthquake distribution. Short-term earthquake forecast is limited mostly due to our limited knowledge about the sophisticated physical processes related to earthquake processes. However, after decades of research and development, USGS and similar agencies around the globe have many types of observations and measurements associated with tens of thousands of earthquakes (https://earthquake.usgs.gov/earthquakes/). Among these data, earthquake parameters such as location (latitude, longitude, and depth), origin time, and magnitude are widely available. The data can be used to build machine learning models that may help us forecast earthquakes. Using machine learning techniques may allow us to build useful models and avoid complicated and hard-to-model earthquake physics. Those models are generally unavailable with other approaches. If useful models are obtained, earthquake centers and seismologists will be interested in these models. The short-term earthquake forecast could help us for the earthquake related disaster preparation and building code adjustments. Previous studies (e.g. Morales-Esteban et al., 2013; Reyes et al., 2013) only focus on earthquakes in a small area, I will explore ways to forecast earthquakes on the global scale.

## Define Objectives and Metrics

I will try to increase the earthquake forecast accuracy on the global scale using previously occurred seismic events. Comparing to published methods (e.g. Morales-Esteban et al., 2013; Reyes et al., 2013), I will need to convert the earthquake parameters into a fixed number of features and build measure learning models with these features. For each observation, the label will be whether an earthquake or earthquakes will occur a specific location. Since I will consider the features and labels on the global scale, both the features and labels are projected onto an evenly spaced grid. The data will be split to training and testing set. The Receiver Operating Characteristic (ROC) Area Under Curve (AUC) measured from the testing test will be used to validate and compare the machine learning models.

## Understand the State-of-the-Art

A few previous studies (e.g. Morales-Esteban et al., 2013; Reyes et al., 2013) have explored artificial neural networks for earthquake forecasting problem in local seismic zones. Martínez-Álvarez et al. (2013) investigated different indicators to forecast earthquakes. Moustra, Avraamides, and Christodoulou (2011) used seismic electric signals for earthquake forecasting and obtained promising results. Asencio-Cortés et al. (2016) studies sensitivity of different seismicity indicators and suggests choose the right features will affect prediction accuracy. Rhoades (2007) proposed that minor earthquakes can be used for forecasting large events. Asencio–Cortés et al. (2018) projected earthquake catalogs onto evenly spaced grid and use the resulting features to predict earthquake magnitudes in California. The published machine learning models are evaluated with testing data and different metrics were used. Giving these past investigations, the earthquake forecast problem is far from being solved. Global earthquake forecasting was not studied due to limited computational resources and varying number of earthquakes for a time window of interest.

## Define Hypotheses and Approach

My hypothesis is that future earthquake occurrence can be forecasted using previously occurred earthquake parameters on the global scale. If the hypothesis is true, the ROC AUC of the testing data should be larger than 0.5 (associated to random selection). If the hypothesis is incorrect, the ROC AUC will be close to or smaller than. Due to the computational considerations, I will only process earthquake catalog in just one year. In this way, the hypothesis can be tested in a finite time. I will use earthquake parameters such as latitude, longitude, depth, magnitude and origin time to compute machine learning models. Based on previous studies (e.g.  Asencio–Cortés et al. 2018), I will project both previous occurred earthquakes (features) and future earthquakes (labels) on to an evenly-spaced grid. The grid will convert the earthquake catalogs to features of the same size, which are much easier to use for machine learning algorithms. Possible data bias is that more grid points will be free of earthquakes.

I used both logistic regression and random forest technique to build machine learning models. Both methods are fast to compute and are suitable for the data size. I first select earthquakes occurred before a reference time and then search for earthquakes occurred after a reference time. These two earthquake catalogs are projected to a regular grid. For earthquakes before the reference time, earthquake parameters at each grid points are considered as features. Figure 1-4 show distribution of these earthquake parameters. Earthquakes occurred after the reference time are used as labels. If a specific grid point has earthquakes, the label is 1. If not, the label is 0. Besides the earthquake parameters, location (latitude and longitude) of each grid points are also used as features. To save computational resources, grid points without earthquake occurrence are not used in the calculations. Logistic regression assumes the relationship between the features and the labels are linear. My method is novel compared to published work. First, I study the earthquake forecast on the global scale. Secondly, feature calculation is adjustable and evaluated by testing data as well.

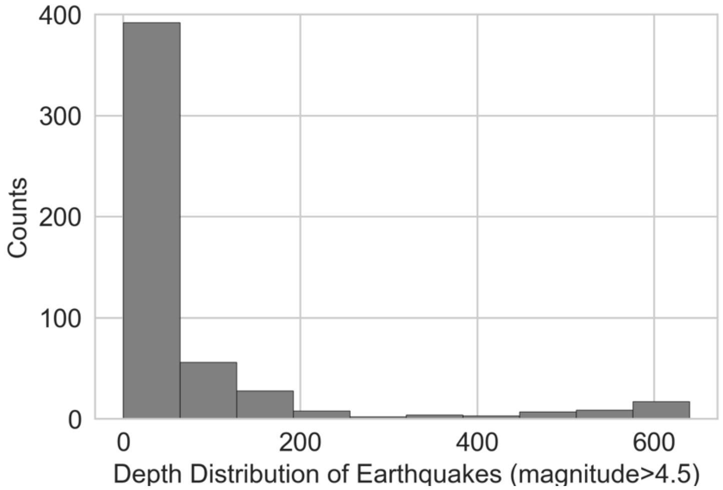

*Figure 1 Depth distribution of earthquake with magnitude larger than 4.5.*

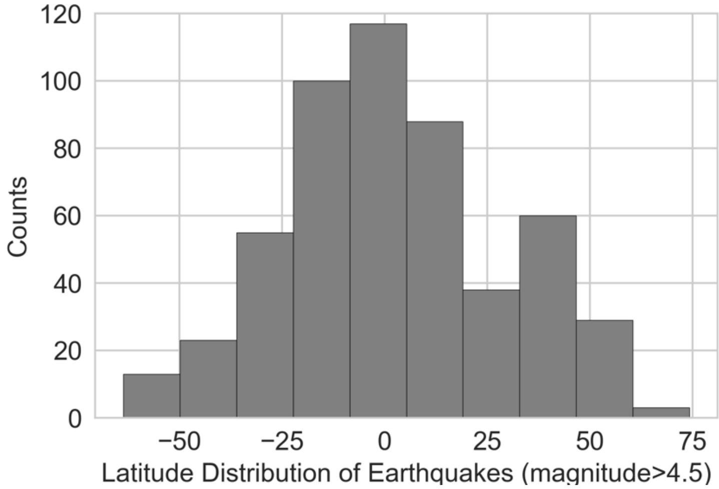

*Figure 2 Distribution of earthquake latitude.*

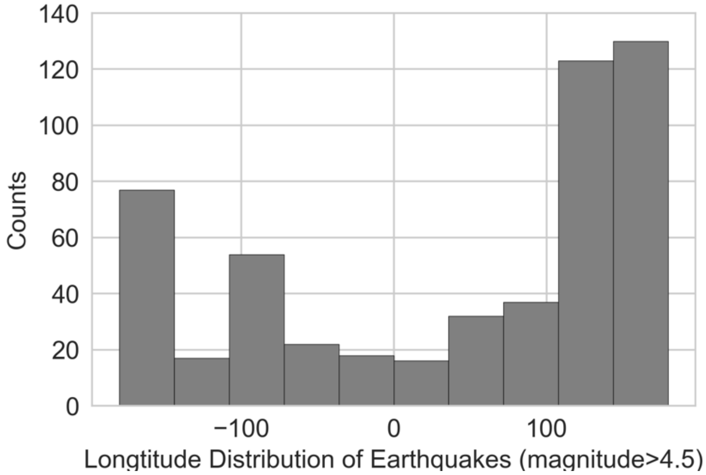

*Figure 3 Distribution of earthquake longitude*

If the maximum ROC AUC value of the machine learning models is close to 0.5, the results oppose my hypothesis. If the maximum ROC AUC value is larger than 0.5, the results support my hypothesis. The ROC AUC value evaluates the model performance faithfully since it is not significantly affected by imbalance class distribution.

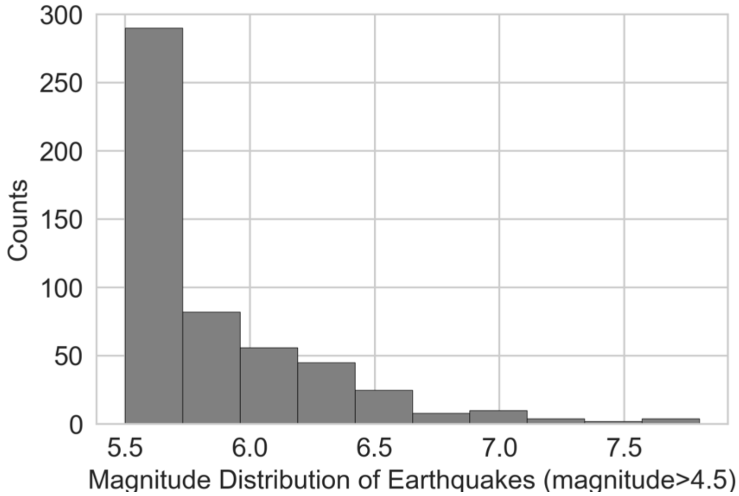

*Figure 4 Distribution of earthquake magnitude*

**Execute Approach**

I tested several parameters used in the feature calculations, such as the time span before of reference time, the time span before of reference time, the grid size, minimum earthquake magnitude to be included, and how to compute features at each grid point. When multiple earthquakes fall onto the same grid point, I could use minimum value, maximum value, average value, largest earthquake or earthquake counts for different magnitude ranges, and a combination of these measurements to represent the grid point.

The ROC AUC metrics were computed for both logistic regression models and random forest models using different preprocessing parameters. The results are summarized in Table 1. The ROC AUC values are changing in the range of 0.50 and 0.59. Since the maximum ROC AUC value is larger than 0.5, the results support the hypothesis. We can use previously occurred earthquakes to forecast future earthquakes on the global scale. However, the resulting machine learning models need to be improved since the error rates are large.

From the problem, I have learnt that previous earthquake parameters could help with earthquake forecasting though the accuracy needs to be improved. From the data and methods, I have learnt how to preprocess spatial and temporal variation measurements for machine learning algorithms. The evaluation process helped me to learn how to reliably measure model performance. My results support my hypothesis. My work shows how to prepare the earthquake parameters for machine learning techniques and tested a wide range of feature calculation procedures. The method and results could be a starting point for more sophisticated machine learning models.

*Table 1 ROC AUC values for logistic regression and random forest using different parameters. *Min represents using minimum values of each earthquake parameter; Max indicates maximum values are used; Avg means average values are used; Largest indicates parameters of the largest earthquakes are used to represent a grid point; Counts means earthquake count for different magnitude ranges are included.*

| Grid Size (degree) | Time Span after Reference time | Time Span before Reference time | Minimum Magnitude before reference time | Features Included* | ROC AUC Logistic Regression | ROC AUC Random Forest |
|---|---|---|---|---|---|---|
| 1x1 | 7 days | 30 days | 4 | Min, Max, Avg | 0.52 | 0.56 |
| 2x2 | 7 days | 30 days | 4 | Min, Max, Avg | 0.50 | 0.51 |
| 5x5 | 7 days | 30 days | 4 | Min, Max, Avg | 0.52 | 0.52 |
| 10x10 | 7 days | 30 days | 4 | Min, Max, Avg | 0.52 | 0.53 |
| 15x15 | 7 days | 30 days | 4 | Min, Max, Avg | 0.50 | 0.59 |
| 30x30 | 7 days | 30 days | 4 | Min, Max, Avg | 0.52 | 0.56 |
| 5x5 | 7 days | 30 days | 4 | Min, Max, Avg | 0.52 | 0.52 |
| 5x5 | 1 months | 30 days | 4 | Min, Max, Avg | 0.52 | 0.53 |
| 5x5 | 3 months | 30 days | 4 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 30 days | 4 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 4 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 6 months | 4 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 3.5 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 4 | Min, Max, Avg | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 4 | Min, Max, Avg, Largest | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 4 | Min, Max, Avg, Counts | 0.56 | 0.56 |
| 5x5 | 3 months | 3 months | 4 | Min, Max, Avg, Largest, Counts | 0.56 | 0.56 |

## References

Asencio-Cortés, G., F. Martínez-Álvarez, A. Morales-Esteban, and J. Reyes. 2016. “A Sensitivity Study of Seismicity Indicators in Supervised Learning to Improve Earthquake Prediction.” *Knowledge-Based Systems* 101 (June): 15–30. https://doi.org/10.1016/j.knosys.2016.02.014.

Asencio–Cortés, G., A. Morales–Esteban, X. Shang, and F. Martínez–Álvarez. 2018. “Earthquake Prediction in California Using Regression Algorithms and Cloud-Based Big Data Infrastructure.” *Computers & Geosciences* 115 (June): 198–210. https://doi.org/10.1016/j.cageo.2017.10.011.

Martínez-Álvarez, F., J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero. 2013. “Determining the Best Set of Seismicity Indicators to Predict Earthquakes. Two Case Studies: Chile and the Iberian Peninsula.” *Knowledge-Based Systems* 50 (September): 198–210. https://doi.org/10.1016/j.knosys.2013.06.011.

Morales-Esteban, A., F. Martínez-Álvarez, and J. Reyes. 2013. “Earthquake Prediction in Seismogenic Areas of the Iberian Peninsula Based on Computational Intelligence.” *Tectonophysics*. https://doi.org/10.1016/j.tecto.2013.02.036.

Moustra, Maria, Marios Avraamides, and Chris Christodoulou. 2011. “Artificial Neural Networks for Earthquake Prediction Using Time Series Magnitude Data or Seismic Electric Signals.” *Expert Systems with Applications* 38 (12): 15032–39. https://doi.org/10.1016/j.eswa.2011.05.043.

Reyes, J., A. Morales-Esteban, and F. Martínez-Álvarez. 2013. “Neural Networks to Predict Earthquakes in Chile.” *Applied Soft Computing Journal*. https://doi.org/10.1016/j.asoc.2012.10.014.

Rhoades, David A. 2007. “Application of the EEPAS Model to Forecasting Earthquakes of Moderate Magnitude in Southern California.” *Seismological Research Letters* 78 (1): 110–15. https://doi.org/10.1785/gssrl.78.1.110.