

## Machine Learning-Based Method for Earthquake Analysis in Japan: A Preliminary Study

Phongwit SRIKOKCHAROEN, Shota SUZUKI, Kantapisit LIMSUPAPUTTIKUN  
Advisor: Lin ZHU

*National Institute of Technology, Sendai College, Hirose Campus  
4-16-1 Ayashi-Chuo, Aoba-ku, Sendai-shi, Miyagi 989-3128, Japan  
E-mail: s2001123@sendai-nct.jp (Thai language supportable)*

In Japan, earthquakes frequently occur and cause devastating damage to human lives and infrastructure. This preliminary study aims to analyze earthquake data in Japan and utilize a machine learning method to predict the magnitude and depth of earthquake based on the earthquake location and timestamp.

We visualized data in three parts: locations of earthquakes, the relationship between the number of occurrences and time, and the relationship between the number of occurrences and magnitude (Gutenberg-Richter Law).

We then split data into the train set and the test set (80:20) in the Relative Intensity model, and into the train set, the test set, and the validation set (80:10:10) in the machine learning model. We use Relative Intensity (RI), receiver operating characteristic (ROC), and Pierce's skill score to find the locations with high risk for earthquake occurrences. Then, we use the machine learning-based model to find the relationship between time, locations, magnitude, and depth.

The result demonstrated that earthquakes in Japan can be explained using both conventional and proposed models. Also, while both of our models have their advantages and drawbacks, they can analyze the data to a certain extent.

**Keywords:** Gutenberg-Richter Law, Neural Network, Relative Intensity, Receiver Operating Characteristic, Pierce Skill Score

## 1. Introduction

In Japan, earthquakes frequently occur and cause devastating damage to infrastructure and human lives.

This preliminary study aims to utilize a machine learning-based method to analyze earthquake data in Japan. We also aim to prove the validity for machine learning method in earthquake analysis.

## 2. Framework (Data and Method)

### Part 1: Data Format and Data Visualization

In this study, an earthquake dataset, which contains the time, locations, magnitude, and depth of each earthquake in Japan from 1985-2022, was obtained from Japan Meteorological Agency.

We first converted the date and time to the timestamp for better usability. To further understand the dataset, we then visualized these data in 3 ways.

1. Locations of earthquake occurrences
2. The relationship between the number of occurrences and time
3. The relationship between the number of occurrences and magnitude (Gutenberg-Richter Law) is then also used to find the Gutenberg-Richter relationship in earthquakes.

We calculated the logarithm base 10 of occurrences and approximate the relationship using linear regression.

Then, we split the data for training each model. For the Relative Intensity Method (Conventional), we divided 70% of the data into the train set and 30% into the test set (time-sequential). For neural network, we normalized these data. Then, we separated data into X (timestamp, latitude, longitude) and Y (magnitude, depth). After that, we divided 80% of the data into the train set, 10% of the data into the test set, and 10% of the data into the validation set at random state=42.

### Part 2: Relative Intensity Method (Conventional)

The RI model measures past seismic activities in areas and is used to determine the potential for a new earthquake. We then define the study area (ranges from 30°N 128°E to 46°N 146°E) as a point with even spacing. The RI score, an index (ranges from 0 to 1) used in the RI method, is a min-max normalized Cumulative Benioff Strain of a point in the study area. The formula for Cumulative Benioff Strain at a time  $B_{xy}(t)$  is as follows.

$$B_{xy}(t) = \sum_{i=1}^{N_{xy}(t)} \sqrt{E_{xy}^{(i)}}$$

Where  $E_{xy}^{(i)}$  is Seismic Energy released from  $I^{\text{th}}$  earthquake,  $xy$  is the location of the point, and  $N_{xy}(t)$  being all earthquake within a certain radius of the  $xy$  point with magnitude larger than M5.0 up until time  $t$ . The Seismic Energy released in an earthquake is approximated using magnitude with the following formula.

$$E_{xy}^{(i)} = e^{5.24+1.44M}$$

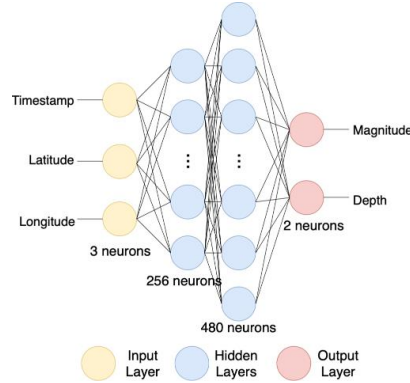
After calculating the RI score using the train set. We then find a threshold for the RI score that performs best on the test data; Pierce's Skill Score is used to measure performance. The formula for Pierce's Skill Score is as follows (equation 3), where a, b, c, and d in the formula are elements in the confusion matrix respectively.

$$\text{Pierce's Skill Score} = \frac{(ad - bc)}{[(a + c) \cdot (b + d)]}$$

We also tried setting study area spacing as 10 km. and 25 km. and compare both RI model performances.

### Part 3: Machine Learning-Based Model

We built a neural network to find the relationship between the time, locations, magnitude, and depth of the earthquake. Figure 1 shows the neural network used in this study, which consists of four layers. The first layer is the input layer. The second layer is 256 neurons and rectified linear unit (ReLU) as activation function. The third layer is 480 neurons and rectified linear unit (ReLU) as activation function. And the last layer is the output layer. We wanted two outputs, so we used 2 neurons and linear as activation function in the last layer.



**Figure 1 Neural network model for regression**

We trained this model with Adam (learning rate=0.05) optimizer, loss function = mean squared error (MSE), batch size=10,000, and 2,000 epochs. We then plotted the loss function of the train set and the validation set. After all, we evaluated this model with the test set and calculated the mean squared error. All mean squared error (MSE) is calculated using the following formula.

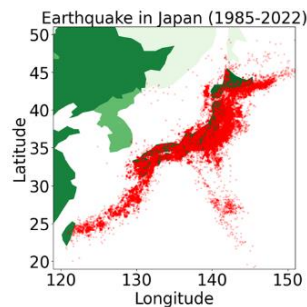
$$MSE = \frac{1}{b} \sum_{i=1}^b \frac{(M_{pred} - M_{true})^2 + (D_{pred} - D_{true})^2}{2}$$

Where  $b$  is batch size,  $M_{pred}$  is predicted magnitude,  $M_{true}$  is true magnitude,  $D_{pred}$  is predicted depth, and  $D_{true}$  is true depth.

### 3. Finding and Discussion

#### Part 1: Visualization

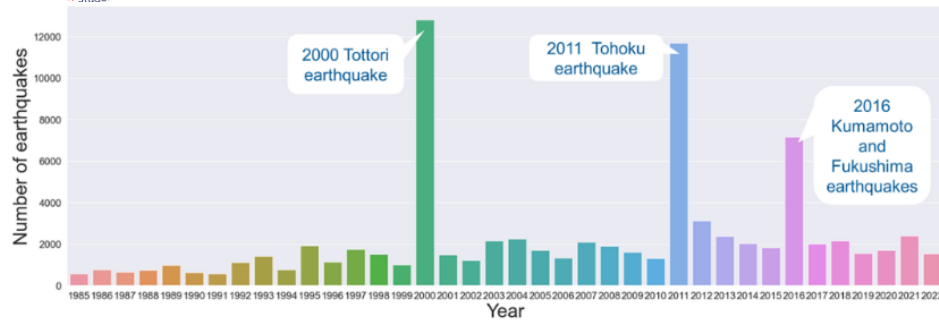
At first, we plotted all occurrences of the earthquake in Japan from 1985 to 2022 as figure 2 below.



**Figure 2 All earthquake occurrences in Japan during 1985-2022**

As figure 2 shows, Japan has an earthquake on the side of the Pacific Ocean because the location of Japan is on the Ring of Fire.

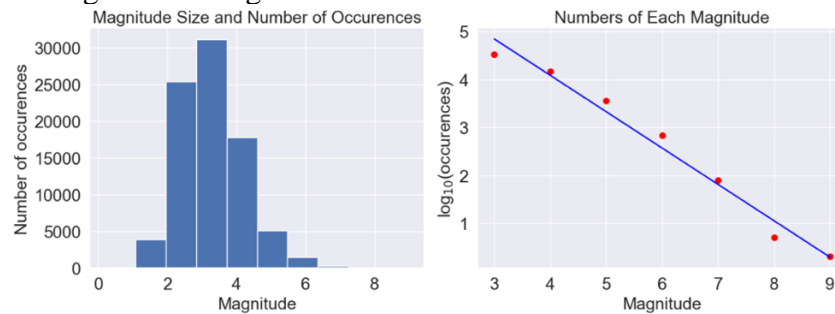
After plotting all the occurrences, we separated the data annually and find the relationship between the duration of peak occurrences (more than average occurrences) as figure 3 below.



**Figure 3 The relationship between the number of occurrences and time**

As figure 3 shows, all earthquakes must accumulate energy before occurring. From the graph, the years with a high earthquake occurrence rate are 2000, 2011, and 2016.

From the past database, we plotted the number of occurrences of each magnitude. The graph is shown in figure 4 and figure 5 below.



**Figure 4 and 5 The relationship between the number of occurrences and magnitude**

We calculate the logarithm with base 10 of the number of occurrences and plot the graph with magnitude. After using linear regression, we found the relationship between the number of occurrences (N) and magnitude (M) as written in the following equation.

$$N = 10^{-0.758737M+7.1147147}$$

We can prove that earthquake in Japan can be explained by Gutenberg-Richter law.

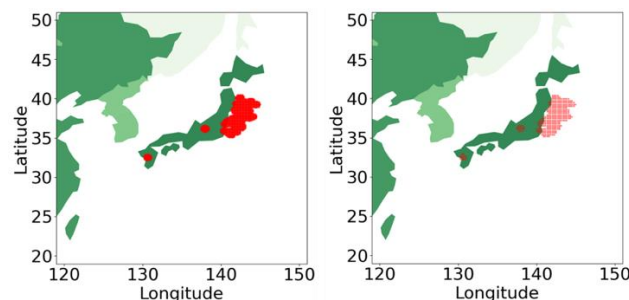
## Part 2: Relative Intensity (RI) with Receiver Operating Characteristic (ROC) and Pierce's Skill Score

Following the method discussed in the framework section above, we used the RI model on earthquake data and obtain the following results.

**Table 1 RI Model performances with different settings**

Study Area Spacing	Radius	RI value Threshold	Pierce's Skill Score
10 km.	50 km.	0.10090	0.69377
25 km.	50 km.	0.10672	0.68784

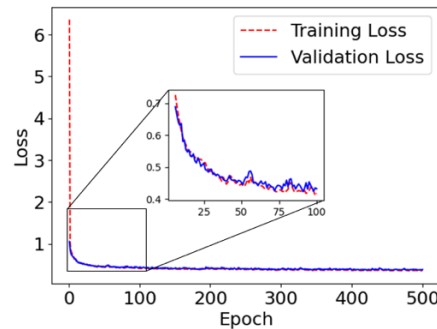
From table 1, the study area with 10 km. spacing performs slightly better than the study area with 25 km. spacing.



**Figure 6 Predicted result from RI model (left: 10 km. spacing, right: 25 km. spacing)**

From figure 6, the area that the RI model predicted is well-known in Japan for high earthquake occurrences. The area is also on the side of the Pacific Ocean, which aligns with the conclusion we arrived at earlier in the visualization section.

### Part 3: Machine Learning-Based Model



**Figure 7 Training loss and validation loss of model**

We used the test set to evaluate this model and got  $MSE=0.394$  as the result. From training, we got  $MSE=0.349$  for the train set and  $MSE=0.377$  for the validation set shown in figure 7.

### 4. Conclusion

By using various statistical methods and visualization, we can obtain the properties of Japan's earthquake data. For example, we found that every large earthquake needs time to accumulate energy. We also found that a small number of earthquakes occurred on the Sea of Japan side, almost all of the earthquakes in Japan occurred on the line of the Ring of Fire.

By the conventional model (relative intensity model), we can find the locations that have high risk for large-scale earthquakes with high accuracy. However, we cannot predict the precise timing of the earthquake.

By the proposed model (machine learning-based model), we can use this regression to find the relationship between timestamps, locations, magnitude, and depth. The model's performance is excellent. However, the model assumes all the inputs to be earthquakes. Hence, this model can only be used for data analysis and regression, but not for prediction.

### 5. Acknowledgement

We would like to express our greatest appreciation to Japan Meteorological Agency, National Research Institute for Earth Science, and Disaster Resilience for providing the earthquake data in Japan.

We also would like to express our greatest appreciation to Dr. ZHANG Xiaoyong from Sendai KOSEN for giving us some advice about machine learning (neural networks).

At last, we would like to express our greatest appreciation to Prof. Santi Pailoplee from Geology, Chulalongkorn University for giving us some advice about seismic hazard analysis, and statistical seismology.

### 6. References

- [1] B. Gutenberg & C. F. Richter (1944) *Frequency of Earthquakes in California*. Pasadena, California.: California Institute of Technology.
- [2] Bruce A. Bolt (2022) *Earthquake Geology*. <http://www.britannica.com/science/earthquake-geology>. [accessed 9/23/2022]
- [3] Japan Meteorological Agency. 2022. Earthquake Database in Japan. <http://www.data.jma.go.jp/svd/eqdb/data/shindo/index.html>. [accessed 9/23/2022]
- [4] Kristy French Tiampo and Robert Shcherbakov. (2012) *Seismicity-based earthquake forecasting techniques: Ten years of progress*. [http://www.researchgate.net/publication/251496912\\_Seismicity-based\\_earthquake\\_forecasting\\_techniques\\_Ten\\_years\\_of\\_progress](http://www.researchgate.net/publication/251496912_Seismicity-based_earthquake_forecasting_techniques_Ten_years_of_progress). [accessed 9/23/2022]