**Earthquake Nowcasting in Taiwan: Analysis of Interevent Count Distribution**

**Group 10 – Cox Group**

**Introduction**

**What is an Earthquake?**

An earthquake is a sudden shaking of the ground which occurs when two tectonics plates suddenly slide against each other. The space between the plates or the surface where they come into contact or slide is called the fault or the fault plane. The location below the surface where the earthquake begins is called the hypocenter and its corresponding point on the earth surface is called the epicenter. Sometimes an earthquake has fore-shocks, which are smaller earthquakes that occur in the same place as the larger earthquake that follows. The larger earthquake is called the ‘mainshock’. The mainshocks are usually followed by aftershocks. The size of the earthquake is called the magnitude which is measured on the Richter scale.

**Differences between Forecasting, Prediction and Nowcasting**

**Earthquake Prediction** is a branch of seismology for determining when and where an earthquake will occur along with how intense it will be. It is empirical in nature and usually relies on observations for a short-term prediction. An earthquake must define three elements, the date, the location and the intensity or magnitude of the earthquake. It determines the parameters for the next major earthquake. However, there is not enough scientific evidence to prove an earthquake.1

On the other hand, a **forecast** is a probabilistic statement which estimates the likelihood of an earthquake based on data factors such as the frequency of occurrence and the magnitude of earthquakes in a given area over many years. It usually employs timeseries analysis and we apply trend methods to forecast.

**Earthquake nowcasting** is a method to evaluate the current state of seismic hazard from large earthquakes.2 It uses the count of the number of small earthquakes since the last large earthquake in a defined region (in our case Taiwan) to estimate the current hazard level in the region. While forecasting is the calculation of probabilities for the future, nowcasting is the calculation of present state of the system.

We will be applying the method of Earthquake Nowcasting to the country of Taiwan in order to estimate the current state of seismic hazard from large earthquakes by calculating the **Earthquake Potential Score** (EPS).

**Is Taiwan prone to Earthquakes?**

Taiwan lies near the junction of two tectonic plates in the South China Sea on the circum-Pacific seismic zone. Earthquake occurrence here is quite common and strong earthquakes also occur frequently. According to the data provided by the Central Weather Bureau (CWB) seismic network, in the years till 1990 the average number of earthquakes was 2200 annually out of which 214 were felt. Geologists have identified 42 active faults on the island but most of the earthquakes detected in Taiwan are due to the convergence of the Philippine sea plate and the Eurasian Plate to the east of the island.3

**Preprocessing the Available Dataset**

The dataset that we utilized for this report basically constituted of the various earthquakes that have occurred in Taiwan with respect to the following conditions: **latitude 21oN-26oN, longitude 120oE-122oE and time period 1963-2020.** The software Python (Jupyter Notebook) has been utilized to filter out the unnecessary data and extract the useful data. First, the dataset had been imported to the software where all the large earthquakes (M≥6) were extracted from the dataset, as shown in Table 1.

Table

Description automatically generatedTable

Description automatically generatedThe interevent counts (Natural Time) for each large earthquake were found out and then added to this dataset. The **interevent counts** are described as the cumulative counts of “small” events (4.0≤M<6.0) between two successive “large” earthquakes (say, M≥6), as shown in Table 2.

**Table 1:** Large Earthquake Dataset **Table 2:** Dataset with Interevent Counts

Table

Description automatically generatedAll the large earthquakes that had a zero interevent count were discarded from the dataset. A final preprocessed dataset is given in Table 3 with all the relevant information required. The last value in the dataset (83 interevent counts) has null values because this is the intervent count for the upcoming earthquake which has not occurred yet.

**Table 3:** Dataset-Non-Zero Interevent Counts

After discarding the large earthquakes with zero interevent counts, it was observed that the region of Taiwan had undergone **37 cycles of large earthquakes** in the time period from 1963-2020.

The dataset in Table 3 is now modelled with 8 candidate distributions that include Weibull, Normal (Gaussian), Exponentiated Weibull, Lognormal, Exponential, Gamma, Inverse Gaussian and Inverse Weibull distributions. This is done to find out that distribution which best fits the given data using the various Goodness of Fit Tests such as KS (Kolmogorov-Smirnov) Test and AIC scores.

**Estimation of Parameters and Goodness of Fit Tests**

The two methods that we used to estimate the parameters were: Method of Moments (MoM) and Maximum Likelihood Estimation (MLE). The parameters of the various candidate distributions were estimated using the **actuar** and **fitdistrplus** libraries on **R**. Initially, we had tried to perform the estimation on Python, however certain statistical distributions were not available, and the estimations of the parameters were only limited to Maximum Likelihood Estimation method. The KS statistic, p-value and AIC scores were also obtained using the same software packages. The summary of the results is given in Tables 4 and 5.

**MLE**: Maximum likelihood estimation is used to make inferences about the population that is most likely to have generated the sample. The likelihood function is known as the joint density function of the observed data. Our goal is to maximize this likelihood function and obtain the optimum values of the parameters.

**MOM**: Method of Moments estimation starts by indicating the population moments as functions of the parameters to be estimated. Those expressions are then equated to the sample moments to obtain the parameters.

The parameters estimated using MLE are chosen over those estimated by MOM as MLE is known to provide better estimates. Also, due to certain limitations in the software packages in R only 7 of the 8 distributions’ parameters can be estimated by MOM (Only MLE estimates can be obtained for the Exponential Weibull Distribution).

**KS Test**: The KS test is a non-parametric goodness-of-fit test used to validate a continuous fully defined distribution for a sample of data given, or whether an underlying probability distribution differs from a hypothesized distribution.

**AIC**: The Akaike Information Criterion (AIC) is a mathematical method for evaluating how well a specific model fits the original data from which it was generated, i.e., given a number of models for a dataset, AIC helps us compare these models and estimate quality of each model, relative to each of the other. Thus, AIC can be used for model selection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Distribution** | **Estimated Parameters (MLE)** | **KS Statistic Value** | **KS P-value** | **AIC** |
| **Exp. Weibull** | Kappa= 0.8186113  Lambda=46.3922652  Alpha= 1.6679260 | 0.075206 | **0.9849** | 396.6046 |
| Log-Normal | Meanlog= 3.793244 sdlog= 1.124995 | 0.08184828 | 0.9653 | 398.4172 |
| Weibull | Shape= 1.06983 Scale= 74.80112 | 0.08388776 | 0.957 | 394.9955 |
| Gamma | Shape= 1.14987772 Rate = 0.01579841 | 0.08672685 | 0.9436 | 394.8383 |
| Exponential | Rate= 0.01373932 | 0.08928736 | 0.9296 | 393.2745 |
| Inv. Weibull | Shape= 0.7205068  Scale= 24.2254196 | 0.1649457 | 0.2665 | 415.1995 |
| Normal | Mean=72.78378, Sd=70.74550 | 0.1654269 | 0.2634 | 424.174 |
| Inv. Gauss | Mean= 72.71221 Shape 23.01678 | 0.2013781 | 0.09947 | 414.0056 |

The KS test is used as the measure of determining the best fit distribution for the Interevent Counts Data, rather than the AIC test. This is because the AIC scores do not vary much from one distribution to another, making it difficult for us to make a clear distinction between the fitting of the various candidate distributions, one of the reasons for this is the parametric nature of AIC tests where the result depend on the number of parameters while the KS test being non-parametric will have an upper hand. Not only that the KS Test is free from a priori distributional assumption and it performs well under a wide range of population distributions.4 The AIC score also provides better results in the case of a large sample size, however our sample size consists of merely **37** datapoints.

**Table 4**: Results of MLE estimated parameters for the candidate distributions and Goodness of Fit Test Results using R

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Distribution** | **Estimated Parameters (MoM)** | **KS Statistic Value** | **KS P-value** | **AIC** |
| Gamma | Shape= 1.05845300  Rate = 0.01454243 | 0.07893 | 0.9753 | 394.9938 |
| Weibull | Shape= 1.02893  Scale= 73.63527 | 0.081619 | 0.9662 | 395.0902 |
| Exponential | Rate = 0.01373932 | 0.089287 | 0.9296 | 393.2745 |
| Log-Normal | Meanlog= 3.954920  sdlog= 0.815565 | 0.15392 | 0.3447 | 409.471 |
| Normal | Mean=72.78378 Sd=70.74550 | 0.16543 | 0.2634 | 424.174 |
| Inv. Gauss | Mean= 72.71221  Shape= 23.01678 | 0.20138 | 0.09947 | 414.0056 |
| Inv. Weibull | Shape= 2.556288  Scale= 49.535455 | 0.35941 | 0.0001412 | 415.1995 |

**Table 5:** Results of MOM estimated parameters for the candidate distributions and Goodness of Fit Test using R

The results from the KS Test indicate that the Exponentiated Weibull Distribution is the best candidate fit for the given dataset, owing to the largest KS p-value and smallest KS distance. The five best candidate distributions that fit this data are Exp. Weibull, Log-Normal, Weibull, Gamma and Exponential (all having a KS p-value > 0.9). The normal distribution on the other hand which describes a lot of natural phenomena shows a **terrible fit** for the Interevent Counts Data (having a p-value of merely 0.2634).

Chart, histogram

Description automatically generatedFigure 1 is shown below plotting the data on a histogram plot and along with the top distributions’ density functions on Python.

The lognormal curve has been excluded owing to a few limitations of the Python libraries. Figure 1 gives us assurance that the KS Test results are accurate, and the best fit distributions obtained by the KS Test show a good fit on the histogram plotted data.

**Brief Analysis of the Earthquakes in Taiwan**

Chart, scatter chart

Description automatically generatedFigure 2 shows a basic scatter diagram of the earthquakes that have occurred in Taiwan since 1963. A majority of the earthquakes are small earthquakes which have magnitudes lying between 4 and 6. This scatter plot has been made using the Python library **geopandas** through a shape file of Taiwan’s map with respect to its latitude and longitude that has been obtained from **MIT GeoWeb**5. Figure 3 is a scatter diagram indicating the large magnitude earthquakes. These earthquakes are represented by red dots on the map of Taiwan. Majority of the country’s large earthquakes seemed to have occurred in the south-eastern and eastern parts of the country.

Chart, scatter chart

Description automatically generated

**Earthquake Potential Score (EPS)**

The EPS is a quantification of the current state of earthquake hazard for the chosen region and it indicates the current level of seismic progress through the regional earthquake cycle of recurring events in the region.

We know that when the EPS score is close to 1 (or close to 100 on a scale of 100), it indicates that there is an extremely high probability that a large earthquake is bound to occur, possibly even on the same day.

If n(t) is the number of small earthquakes since the last large earthquake, the EPS score is defined to be: **EPS = P{X ≤ n(t)}**, where P is the CDF of small earthquakes occurring between large earthquakes. EPS is a monotonic function whose value lies between [0,1] and resets to 0 at the onset of every large earthquake.4

EPS is therefore the current level of hazard and assigns a number between 0% and 100% to every region so defined, thus providing a unique measure. Physically, the EPS corresponds to an estimate of the level of progress through the earthquake cycle in the defined region at the current time. The nowcasting approach however yields no information regarding the future. This means that a large earthquake may take place at EPS scores of even as less as 30-40 for a particular country/ city.

To illustrate the nowcasting approach in the study region, we computed EPS values for the two seismically exposed megacities in Taiwan, Taipei and Tainan.

We have considered a radius of R = 100, 150 and 200 km around these cities and found all the earthquakes that are lying within these regions by using the latitude and longitudes given in the dataset for each earthquake**. The small earthquakes are considered to have magnitudes between [4,6) whereas for the large earthquakes M ≥ 6, as considered previously.**

The city center coordinates for the two cities are:

**Taipei City Center - 25.0330° N, 121.5654° E**

**Tainan City Center - 22.9997° N, 120.2270° E**

The EPS scores have been obtained through R by using the best fit distribution i.e., Exponentiated Weibull distribution with the parameters as stated in Table 4.

The data is shown for the various radii taken keeping the city center coordinates as the center of the circular region.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| City | Radius  (R) | Date of Last Large Event  (mm/dd/yy) | Magnitude of Last Large Event | Current Interevent Count | Number of Cycles | Earthquake Potential Score (EPS) |
| Taipei | 100 km | 5/15/2002 | 6.2 | 215 | 3 | **0.9505872** |
| 150 km | 4/18/2019 | 6.3 | 41 | 16 | **0.4206074** |
| 200 km | 4/18/2019 | 6.3 | 48 | 22 | **0.4779803** |
| Tainan | 100 km | 2/5/2016 | 6.2 | 45 | 4 | **0.4541065** |
| 150 km | 2/5/2016 | 6.2 | 119 | 20 | **0.8155418** |
| 200 km | 4/18/2019 | 6.3 | 61 | 32 | **0.5699152** |

Table 6: Nowcast Values for Taipei and Tainan with R = 100 km, 150 km and 200 km

The EPS scores for Taipei (R = 100 km) and Tainan (R = 150 km) indicate that the progress of the current earthquake cycle is almost complete suggesting a higher risk of imminent large scale earthquakes. The other EPS scores indicate roughly 50% progress of the earthquake cycle in the various regions surrounding the 2 cities. A bar graph in Figure 4 gives us a pictorial comparison of the intervent counts for different values of R.

Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generated

CDF plots are utilised to obtain the EPS information.2 Figure 5 shows the CDF plots that have been obtained from R for all the best fit curves with respect to the ECDF curve of the dataset of Interevent Counts.

The curves shown in the CDF Plot (Figure 5) indicate the KS Test Validity and give us assurance of an accurate EPS score. We will now use the Exponentiated Weibull distribution to obtain the EPS score for Taipei for R = 100 km in Figure 6.

Chart, scatter chart

Description automatically generated

For 215 Intervent Counts it can be observed that the EPS score from the Exponentiated Weibull Distribution is approximately 95 on a scale of 100.

**Data Driven Models based Analysis**

Motivation and Problem Statement being considered:

We wanted to verify our MLE + KS based estimation even further, to solve this problem we used a data-driven non-parametric based density estimation known as Kernel Density Estimation.

Methodology:

Kernel Density Estimation in crude terms can be explained as a modification of histogram-based density estimation where instead of stacking the blocks aligned with the *bins* of histogram, we stack the blocks aligned with the point they represent, thus placing a distribution per point.

The free parameters of kernel density estimation are the kernel,which specifies the shape of the distribution placed at each point, and the kernel bandwidth, which controls the size of the kernel at each point.

For our problem we used the Kernel as Gaussian distribution and to select the bandwidth parameter we did a cross-validation hyperparameter search which resulted in best bandwidth corresponding to our problem as 10.

Chart, histogram

Description automatically generated

As is evident from Figure 7, Kernel density estimation also provided an estimation which closely resembles our MLE+KS based estimation thus even further verifying our MLE+KS based estimation.

**Conclusion**

We have used Maximum Likelihood (MLE) and Method of Moments (MoM) for the estimation of parameters. Further we used two tests for finding the best-fit distribution: Kolmogorov-Smirnov test (K-S test) and Akaike Information Criterion (AIC test).

As mentioned earlier preferring parameters obtained via MLE and considering K-S test superior, we get Exp. Weibull, Log-normal, Weibull, Gamma and exponential as significant fits compared to others, where **Exp. Weibull** is the best fit with parameters: **Kappa= 0.8186113, Lambda=46.3922652 & Alpha= 1.667926**. Normal distribution cannot describe the Interevent Counts Data owing to a low KS p-value and is a terrible fit for the dataset.

Further to calculate the EPS scores in Taiwan, we used the obtained Exp. Weibull distribution with the corresponding parameters. Our cities of concern are Tainan and Taipei, and by using the CDF from above distribution we get the EPS values for both the cities within 100 km, 150 km and 200 km of the center point of each city. The results clearly indicate that the area under 100 KM radius for Taipei is at highest risk for an imminent large earthquake to occur as the seismic cycle has reached **95% towards completion**. For Tainan, the area under 150 km is at a higher risk of an imminent large earthquake due to **81.5% seismic cycle completion**. KDE is then utilized as a post-processing method to validate and strengthen our results, and the distribution obtained by KDE closely resembles our MLE+KS test result. (Compare **Figure 1** and **Figure 7**)

**Density Functions of the Candidate Distributions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Distribution** | **Density Function** | **Interpretation** | **Conditions** |
| Normal |  | *β– scale* | t > 0, -∞ < α < ∞,  *β* > 0 |
| Exponential |  | *α - scale* | t > 0, *α* > 0 |
| Weibull |  | *α – scale,*  *β– shape* | t > 0, *α* > 0, *β* > 0 |
| Gamma |  | *α – scale,*  *β– shape* | t > 0, *α* > 0, *β* > 0 |
| Log Normal |  | *α – log - scale,*  *β– shape* | t > 0, -∞ < α < ∞,  *β* > 0 |
| Inverse Gaussian |  | *β*/*α – shape* | t > 0, *α* > 0, *β* > 0 |
| Inverse Weibull |  | 1/*α – scale,*  *β– shape* | t > 0, *α* > 0, *β* > 0 |
| Exponentiated  Weibull |  | *α– scale, β–shape, γ- shape* | t > 0, *α* > 0, *β* > 0,  *γ* > 0 |

**References**

1. Wood, H. O., & Gutenberg, B. (1935). EARTHQUAKE PREDICTION. Science, 82(2123), 219–220. doi:10.1126/science.82.2123.219

2. John B Rundle, DL Turcotte, A Donnellan, L Grant Ludwig, Molly Luginbuhl, and G Gong. Nowcasting earthquakes. Earth and Space Science, 3(11):480–486, 2016.

3. Yu, S.-B., Chen, H.-Y., & Kuo, L.-C. (1997). Velocity field of GPS stations in the Taiwan area. Tectonophysics, 274(1-3), 41–59. doi:10.1016/s0040-1951(96)00297-1

4. Pasari, S., & Sharma, Y. (2020). Contemporary Earthquake Hazards in the West-Northwest Himalaya: A Statistical Perspective through Natural Times. Seismological Research Letters. doi:10.1785/0220200104

5. Hijmans, R. and University of California, Berkeley, Museum of Vertebrate Zoology. (2015). First-level Administrative Divisions, Taiwan, 2015. UC Berkeley, Museum of Vertebrate Zoology. Available at: <http://purl.stanford.edu/gm351ph7903>

**Codes and Datasets Used for the Report:** [**GitHub Link**](https://github.com/arsh-k/Natural-Time-Analysis-Taiwan)