Data Mining Approach for Estimation of Earthquake Hazard in Shillong Region

Manisha Singh¹, Radhakrishnan Rambola²

¹Master of Technology Student, Department of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, INDIA
²Associate Professor, Department of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, INDIA

ABSTRACT

North India is one of the earthquake active zone in Indian subcontinent. In this area, a series of medium to large magnitude earthquake can be seen. The expected earthquake of intensity up to IX can be observed. Therefore, for reducing the loss of life and property, it is important to analyze the earthquake risk in the considered region, properly. In this paper data mining approach have been applied. This paper contains the calculation of the largest annual earthquake magnitude, its probability of occurrence and the return period of largest earthquake for the earthquake prone Shillong region of North East India using Gumbel’s type I distribution which is based on the extreme value theory has been estimated. Most probable annual maxima have also been calculated for the considered region. The line of expected extreme which is based on 46 years (1963-2008) of seismicity data of yearly extreme values of earthquakes for the region have been plotted. The medium to large size earthquakes which is expected to occur in this region has been predicted.

Keywords— Gumble’s Theory, Line of expected stream MM intensity.

I. INTRODUCTION

To analyze the seismic risk has been a major problem for earthquake engineering. Time of occurrence of earthquake, location and magnitude of future earthquake cannot be predicted precisely because of incomplete understanding of geophysical phenomenon and present infrastructure available. Many statistical approaches have been proposed for the estimation of occurrence of largest earthquake (Kaila et al., 1972; Yegulap and kuo 1974; Lomnitz, 1974; Shanker and Singh, 1997; Papazachos, 1989). These statistical methods are being used for risk analysis and probability estimation.

Extreme value theory given by Gumbel (1958) has been used by number of researchers for earthquake predictions of different region of the world. In this study calculations based on this model have been carried out for estimating the largest annual earthquake, its return period, and annual maxima in the Shillong region.

II. METHODOLOGY

The methodology used in this approach is Gumbel’s Type I Distribution. Initially, the Gumbel’s theory was proposed for the estimation of flood data analysis and now being frequently used for the earthquake data. It is based on the Poisson random variable function. The shape of Gumble model does not depends on the distribution parameters. It is a probability model which gives an adequate basis for making predictions concerning the occurrence of largest annual earthquake magnitudes. This theory requires p independent observations which are collected continuously from long years. This should be divisible into N data sets each having extremes (N). The largest magnitude from each earthquake data must be taken and should be arranged in ascending order of their magnitude i.e. m₁ < m₂ < m₃… < mp. Where p=N denotes total no of intervals of the data sets. The frequency of each (mk) in the order set of extremes is represented as

\[ G(m) = \exp(-\alpha \exp(-\beta m)), m \geq 0 \]  

Where K is the rank for each mk

Gumbel’s distribution of largest value is given by

\[ G(m) = \exp(-\alpha \exp(-\beta m)), m \geq 0 \]  

From this assumptions it follows that \( m \), largest annual earthquake magnitude, is distributed with cumulative distribution function \( G (m) \). This model is given by Epstein and Lomnitz (1966) for the occurrence of largest earthquake.

The Cumulative distribution function (2) is called the “Gumbel type I” distribution of largest value. The parameter \( \alpha \) and \( \beta \) are the average number of earthquake with magnitude greater than \( m \) per year, \( \beta \) is the inverse of the average magnitude of earthquake for the shillong region and \( m \) is the maximum annual earthquake magnitude. The parameter \( \alpha \) and \( \beta \) found from the least square fit to the equation

\[
\ln [- \ln G (m)] = \ln \alpha - \beta m
\]

Equation (3) is the derivation of equation (2). \( \beta \) is considered a measure of concentration about the mode. The table 2 defines the estimation of parameters \( \alpha \) and \( \beta \).

The return period of earthquake of magnitude \( m \) is given by

\[
T_m = 1/N_m
\]

Where \( N_m = \alpha e^{\beta m} \) is the number of expected earthquake per year exceeding magnitude \( m \).

\( T_m \) is also called average reoccurrence period of earthquake. Table 3 defines the average recurrence period for the given earthquakes within shillong region.

Earthquake Hazard \( H_t (m) \) is the probability of reoccurrence of an earthquake of magnitude \( m \) within a period of \( T \) years is given by

\[
H_t (m) = 1 - \exp (-\alpha T e^{\beta m})
\]

where \( P, B \) and \( M \) are sets of items (i.e., attributes), classified by two measures: support \( s \) and confidence \( c \). The support of a rule \( P \Rightarrow Q \) defines the probability that a database event contains both \( P \) and \( Q \), whereas the confidence of the rule defines the conditional probability that a database event containing \( P \) also contains \( Q \).

For an example, if we apply association rule on seismological data then it would be of the following type:

Location in \( L \) \( \land \) depth \( \geq 100 \) Km \( \Rightarrow \) magnitude \( \geq 5R \) \( [1\% , 50\%] \) which is express as follows: whenever an earthquake occurs in location \( L \) at a depth of over 100Km its magnitude is likely to be greater than 5R with a probability of 50%; this combination occurred in 1% of all recorded events.

In case of temporal association rule mining (sequencing), which detects correlations between events with time as in the following example:

Area in \( A_1 \) \( \land \) magnitude \( \geq 7R \) \( \Rightarrow \) Area in \( A_2 \) within [0, 30 days] \( [0.1\% , 30\%] \)

Which is express as follows: whenever an earthquake occurs in area \( A_1 \) with a magnitude greater than 7R it is likely that another earthquake occurs in area \( A_2 \) within a month after the first event with a probability of 30%; this combination occurred in 0.1% of all recorded events.

By identifying and analyzing event sequences (seismic sequences) seismologists can be aid in studying this type of earthquake behavior.

**Clustering:** Clustering of data (Kaufman & Rousseeuw, 1990; Jain et al., 1999) is the unsupervised process of grouping together sets of items into classes in accordance with a similarity measure. Hence, it is the behavior of items rather than that of individual events that is detected. Applications on seismic data maybe for the purpose of finding densely populated regions (according to the Euclidean distance) between the epicenters, and, hence, locating regions of high seismic frequency or dividing the area of a country into zones according to seismicity criteria (e.g., low/medium/high seismic load) as illustrated in Figure 1.
Various clustering approaches have been proposed in the literature. Local spatial-temporal clusters of low magnitude events can be extracted (Dzwinel et al., 2003) using multi-dimensional correlations. Correlations between the clusters and the earthquakes are also recognized. In case of multidimensional Signal (Sheikholeslami et al., 2000) for spatial data, processing techniques can be applied. A clustering method associated with wavelet transform can identify clusters by finding dense regions in the transformed data. Finally, hybrid methodologies have been proposed (Guo et al., 2003) where spatial clustering is combined with high-dimensional clustering.

Classification of Data: Data classification is one of the most common supervised learning techniques. Its first objective is to analyze a (labeled) training set and, with the help of this procedure, build a model for labeling new data entries (Han & Kamber, 2000). A classification model is built in first step using a training data set consisting of database records which are associated with certain class and a proper supervised learning method, e.g. decision trees or neural networks. For decision trees, the model consists of a tree of “if” statements focal to a label which denote the record of the class it belongs. In the second step, the built model which is not included in the training set is used for the classification of records. Several methods have been developed for classification, including decision tree induction, neural networks and Bayesian networks (Fayad et al., 1996).

For the example, the (hypothetical) decision tree illustrated in Figure 2 tries to “predict” the macro seismic intensity at a site given the depth and the magnitude of an earthquake, the geographic area and the local geology. Such implication uncovers correlations between the attributes of the seismological database and decision trees. Such a type is already used by local authorities to process actions for response and relief of the population after a strong earthquake.

IV. OUR APPROACH

Regression Analysis: The approach used here is statistical approach of linear regression. For implementation of this model, data form earthquake database was used, making a previous error data by perceptive the string data. For the best results which are obtained in the tests, data sets (numerical) were created, that generate a pattern relevant to this data. Some important relationship between variables are identified: magnitude calculation based on their latitude and longitude, number of death according to the earthquake magnitude, injuries per year, daily losses, losses per year and depth according to the length.

The graph between reduced statistics and design magnitude shows the equation of linear regression which is same like Gutenberg and Richter (1956)

\[ \log N = a - b (m) \]  

Where a, b is a regression coefficient.

V. DATA USED

In order to study the earthquake risk, probability of occurrence of the earthquake data distributed over 46 years (1963-2008) are divide into one year time interval such that at least one earthquake in each duration is observed which is necessary conditions for the validity of approach. Following catalogue shows the northeast India
and their adjoining region and the table gives the calculation of largest annual earthquake magnitude.

Figure 3: Seismicity of northeast India region (Singh et al., 2005). The structural features and the four major tectonic zones (I-IV) are also shown. Zone III, Shilong region has been taken for the present study.

The data given in table 1 is the data for the sillong region i.e for Zone III for which have calculated the frequency of the earthquake, its return period and the probability of the occurrence.

Table 1: Calculation for Gumbel’s Annual Maximum Distribution & Estimation of α and β.

<table>
<thead>
<tr>
<th>Extremes</th>
<th>Rank</th>
<th>Plotting Position G(m)</th>
<th>Reduced Statics Ln[-ln(G(m))]</th>
<th>Extremes</th>
<th>Rank</th>
<th>Plotting position G(m)</th>
<th>Reduced Statics Ln[-ln(G(m))]</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>1</td>
<td>0.021277</td>
<td>1.348111</td>
<td>4.9</td>
<td>24</td>
<td>0.510628</td>
<td>-0.30736</td>
</tr>
<tr>
<td>4.4</td>
<td>2</td>
<td>0.042553</td>
<td>1.149612</td>
<td>4.9</td>
<td>25</td>
<td>0.513195</td>
<td>-0.46002</td>
</tr>
<tr>
<td>4.5</td>
<td>3</td>
<td>0.06383</td>
<td>1.012159</td>
<td>4.9</td>
<td>26</td>
<td>0.551591</td>
<td>-0.52416</td>
</tr>
<tr>
<td>4.6</td>
<td>4</td>
<td>0.08506</td>
<td>0.901776</td>
<td>4.9</td>
<td>27</td>
<td>0.574468</td>
<td>-0.59003</td>
</tr>
<tr>
<td>4.5</td>
<td>5</td>
<td>0.10638</td>
<td>0.806793</td>
<td>4.9</td>
<td>28</td>
<td>0.595745</td>
<td>-0.65789</td>
</tr>
<tr>
<td>4.5</td>
<td>6</td>
<td>0.12766</td>
<td>0.721923</td>
<td>4.9</td>
<td>29</td>
<td>0.617021</td>
<td>-0.72605</td>
</tr>
<tr>
<td>4.6</td>
<td>7</td>
<td>0.14896</td>
<td>0.644082</td>
<td>5.2</td>
<td>30</td>
<td>0.638298</td>
<td>-0.80304</td>
</tr>
<tr>
<td>4.6</td>
<td>8</td>
<td>0.17021</td>
<td>0.573718</td>
<td>5.2</td>
<td>31</td>
<td>0.659574</td>
<td>-0.87668</td>
</tr>
<tr>
<td>4.9</td>
<td>9</td>
<td>0.19469</td>
<td>0.502545</td>
<td>5.2</td>
<td>32</td>
<td>0.680851</td>
<td>-0.95004</td>
</tr>
<tr>
<td>4.7</td>
<td>10</td>
<td>0.212766</td>
<td>0.436681</td>
<td>5.2</td>
<td>33</td>
<td>0.702128</td>
<td>-1.03948</td>
</tr>
<tr>
<td>4.7</td>
<td>11</td>
<td>0.234043</td>
<td>0.373116</td>
<td>5.2</td>
<td>34</td>
<td>0.723404</td>
<td>-1.12767</td>
</tr>
<tr>
<td>4.7</td>
<td>12</td>
<td>0.255319</td>
<td>0.311331</td>
<td>5.2</td>
<td>35</td>
<td>0.746481</td>
<td>-1.21246</td>
</tr>
<tr>
<td>4.8</td>
<td>13</td>
<td>0.276596</td>
<td>0.250913</td>
<td>5.2</td>
<td>36</td>
<td>0.769667</td>
<td>-1.31229</td>
</tr>
<tr>
<td>4.8</td>
<td>14</td>
<td>0.297872</td>
<td>0.191532</td>
<td>5.2</td>
<td>37</td>
<td>0.787234</td>
<td>-1.43033</td>
</tr>
<tr>
<td>4.8</td>
<td>15</td>
<td>0.319149</td>
<td>0.132866</td>
<td>5.2</td>
<td>38</td>
<td>0.806911</td>
<td>-1.54512</td>
</tr>
<tr>
<td>4.9</td>
<td>16</td>
<td>0.340426</td>
<td>0.079698</td>
<td>5.2</td>
<td>39</td>
<td>0.829787</td>
<td>-1.67868</td>
</tr>
<tr>
<td>4.7</td>
<td>17</td>
<td>0.361702</td>
<td>0.026792</td>
<td>5.2</td>
<td>40</td>
<td>0.851064</td>
<td>-1.82469</td>
</tr>
<tr>
<td>4.8</td>
<td>18</td>
<td>0.382979</td>
<td>-0.04106</td>
<td>5.2</td>
<td>41</td>
<td>0.872372</td>
<td>-1.99038</td>
</tr>
<tr>
<td>4.9</td>
<td>19</td>
<td>0.404255</td>
<td>-0.09304</td>
<td>5.2</td>
<td>42</td>
<td>0.893617</td>
<td>-2.185</td>
</tr>
<tr>
<td>4.9</td>
<td>20</td>
<td>0.425532</td>
<td>-0.15734</td>
<td>5.2</td>
<td>43</td>
<td>0.914894</td>
<td>-2.41971</td>
</tr>
<tr>
<td>4.9</td>
<td>21</td>
<td>0.446809</td>
<td>-0.21614</td>
<td>5.2</td>
<td>44</td>
<td>0.936177</td>
<td>-2.71874</td>
</tr>
<tr>
<td>4.9</td>
<td>22</td>
<td>0.468085</td>
<td>-0.27516</td>
<td>5.2</td>
<td>45</td>
<td>0.957447</td>
<td>-3.19354</td>
</tr>
<tr>
<td>4.9</td>
<td>23</td>
<td>0.489362</td>
<td>-0.33596</td>
<td>5.2</td>
<td>46</td>
<td>0.978723</td>
<td>-3.59411</td>
</tr>
</tbody>
</table>

Y = -3.384x + 16.13 this is called gumbel’s equation which is same as regression analysis equation.

Table 2. Estimated Gumbel’s Parameters α and β

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope(-β)</td>
<td>-3.384</td>
</tr>
<tr>
<td>β</td>
<td>3.384</td>
</tr>
<tr>
<td>Intercept(ln(α))</td>
<td>16.13</td>
</tr>
<tr>
<td>α</td>
<td>10119754.878</td>
</tr>
</tbody>
</table>

Here are the graphs for the Gumbel’s equation of table 1 for the estimation of α and β.

Fig 4. Graph between maximum magnitude and reduced statistics showing regression analysis equation.

Fig 5. Graph showing Probability of occurrence with design magnitude for t= 10, 20, 30 years.
Table 3: Predicted Annual number of earthquakes and its return period

<table>
<thead>
<tr>
<th>Magnitude (m)</th>
<th>$N_e$</th>
<th>$T_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>13.38</td>
<td>27.29 days</td>
</tr>
<tr>
<td>4.1</td>
<td>9.54</td>
<td>30.28 days</td>
</tr>
<tr>
<td>4.2</td>
<td>6.80</td>
<td>53.70 days</td>
</tr>
<tr>
<td>4.3</td>
<td>4.93</td>
<td>75.23 days</td>
</tr>
<tr>
<td>4.4</td>
<td>3.46</td>
<td>105.66 days</td>
</tr>
<tr>
<td>4.5</td>
<td>2.47</td>
<td>148.20 days</td>
</tr>
<tr>
<td>4.6</td>
<td>1.76</td>
<td>207.08 days</td>
</tr>
<tr>
<td>4.7</td>
<td>1.25</td>
<td>291.60 days</td>
</tr>
<tr>
<td>4.8</td>
<td>0.89</td>
<td>1.12 years</td>
</tr>
<tr>
<td>4.9</td>
<td>0.64</td>
<td>1.57 years</td>
</tr>
<tr>
<td>5</td>
<td>0.45</td>
<td>2.20 years</td>
</tr>
<tr>
<td>5.1</td>
<td>0.32</td>
<td>3.09 years</td>
</tr>
<tr>
<td>5.2</td>
<td>0.23</td>
<td>4.34 years</td>
</tr>
<tr>
<td>5.3</td>
<td>0.16</td>
<td>6.08 years</td>
</tr>
<tr>
<td>5.4</td>
<td>0.14</td>
<td>8.53 years</td>
</tr>
<tr>
<td>5.5</td>
<td>0.08</td>
<td>11.96 years</td>
</tr>
<tr>
<td>5.6</td>
<td>0.06</td>
<td>16.78 years</td>
</tr>
<tr>
<td>5.7</td>
<td>0.04</td>
<td>23.54 years</td>
</tr>
<tr>
<td>5.8</td>
<td>0.03</td>
<td>33.02 years</td>
</tr>
<tr>
<td>5.9</td>
<td>0.02</td>
<td>46.32 years</td>
</tr>
<tr>
<td>6</td>
<td>0.02</td>
<td>64.97 years</td>
</tr>
<tr>
<td>6.5</td>
<td>0.00</td>
<td>352.83 years</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>1916.01 years</td>
</tr>
</tbody>
</table>

Table 4: Most probable largest earthquake hazard $H_t (m)$ for Different Magnitude and Time Periods ($t=10$, 20, 30) years

Table 5: Most probable largest earthquake hazard $H_t (m)$ for Different Magnitude and Time Periods ($t=50$, 75, 100) years

Table 6: Most probable largest earthquake hazard $H_t (m)$ for Different Magnitude and Time Periods ($t=500$, 1000) years

Fig 6. Graph showing Probability of occurrence with design magnitude for $t=50$, 75, 100 years
Fig 7. Graph showing probability of occurrence with design magnitude for t= 500, 1000 years

For an earthquake of magnitude 6, which has a return period of 64.97 years, the 89% probability recurrence period is calculated as and reported in Table 7.

\[ T_{89} = -T_m \ln (1-0.89) \]

Table 7. Recurrence Period with 89% probability

<table>
<thead>
<tr>
<th>Magnitude(m)</th>
<th>Return Period (years)</th>
<th>Recurrence period (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.2</td>
<td>4.856004809</td>
</tr>
<tr>
<td>5.5</td>
<td>11.96</td>
<td>26.39900796</td>
</tr>
<tr>
<td>5.8</td>
<td>33.02</td>
<td>72.88421763</td>
</tr>
<tr>
<td>6</td>
<td>64.97</td>
<td>143.4066511</td>
</tr>
<tr>
<td>6.5</td>
<td>352.83</td>
<td>778.7928076</td>
</tr>
<tr>
<td>7</td>
<td>1916.01</td>
<td>4229.160806</td>
</tr>
</tbody>
</table>

VI. DISCUSSION AND CONCLUSION

A most important utility of this work is to express the usefulness and authenticity of extreme value theory when applied to the estimation of earthquake occurrence. Seismic threat and related earthquake engineering purposes usually require evaluation of return periods or probabilities of exceedance of specific levels of design load criteria or extremal safety conditions. The seismicity of the North east India region displays significant change in magnitude and frequency of earthquake in space and time. This study advocates that the region is one of the most earthquake active regions. The present technique uses 46 years of seismicity data to analyses the largest annual earthquake magnitude. The Gumbel’s extreme value method proved very useful to estimate the future earthquake occurrences for the civil and scientific uses. It has attracted the attention of seismologist for regional seismicity variation and earthquake engineers for civil protection. Tables 1 and 3 shows the probability of occurrence and return periods for different magnitude and time periods.

The merit of the analysis over other statistical techniques lies in the fact that one can forecast the mean return period of different magnitudes over several years. Here the return periods for magnitude 6 earthquake have been estimated to be 64.97 years. However, general interpretation of 89% probability recurrence period result is that, within the Shillong region, there is 89% probability that, in any given 143.41 years periods, at least one earthquake of magnitude 6.0 or greater will occur (Table 7). The earthquake hazard parameters for the considered region have been reported in Table (4, 5 and 6) for different magnitudes over 10, 20, 30, 50, 75, 100, 500, 1000 years duration. This table illustrates that risk value decreases with increasing magnitudes. Probability curves (Figures 5, 6 and 7) indicates that there is non-zero probability no event occurring over any given period of time, since probability of any particular event never reaches 1.0. For example, the probability of an earthquake of 6.0 occurring in Shillong area within any 500 years, is 0.999 and for 6.5 in 1000 years (Figure 7). It means, it is certain that at least one such event will occur within that period of time.

In all statistical analyses generally, a huge amount of seismicity data is required to derive return periods. Although, the present analysis uses 46-years of data to estimate the seismic risk and return period, however, the analysis may be considered to be reliable, since the value of the parameters lnα and β (Table 8) which are used in estimating the return periods and risk do not change much for the use of short or long duration of seismicity data. Thus the estimates offered in this study may permit certain consideration for civil engineering constructions.

This paper analyses that regression analysis (Figure 4) performs well in comparison to other data mining techniques. To improve the performance of regression analysis, other statistical based features linear regression can be subsumed.

Table 8. Earthquake hazard parameters for shillong Region

<table>
<thead>
<tr>
<th>Most probable annual maxima</th>
<th>lnα</th>
<th>t</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.76</td>
<td>10119754.88</td>
<td>3.384</td>
<td></td>
</tr>
</tbody>
</table>
VII. ACKNOWLEDGMENT

The authors would like to express his sincere thanks to all authors’ whose references and research works helped us a lot for the preparation of this manuscript.

REFERENCES