ANALYSIS SEISMIC WAVE DATA TO DETERMINE EARTHQUAKE EPICENTERS AND FOCAL DEPTHS USING DECISION TREES

[1] G. SRI LAKSHMI PRASANNA, [2] MRS. CH. DEEPTHI

[1] PG Scholar in the department of MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

[2] Assistant Professor in the department of CSE/MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

ABSTRACT

We develop a random forest (RF) model for rapid earthquake location with an aim to assist earthquake early warning (EEW) systems in fast decision making. This system exploits P-wave arrival times at the first five stations recording an earthquake and computes their respective arrival time differences relative to a reference station (i.e., the first recording station). These differential P-wave arrival times and station locations are classified in the RF model to estimate the epicentral location. We train and test the proposed algorithm with an earthquake catalog from Japan. The RF model predicts the earthquake locations with a high accuracy, achieving a Mean Absolute Error (MAE) of 2.88 km. As importantly, the proposed RF model can learn from a limited amount of data (i.e., 10% of the dataset) and much fewer (i.e., three) recording stations and still achieve satisfactory results (MAE<5 km).The algorithm is accurate, generalizable, and rapidly responding, thereby offering a powerful new tool for fast and reliable source-location prediction in EEW.

Index Terms—Earthquake Early Warning (EEW) system; Machine learning; Earthquake Location.

INTRODUCTION

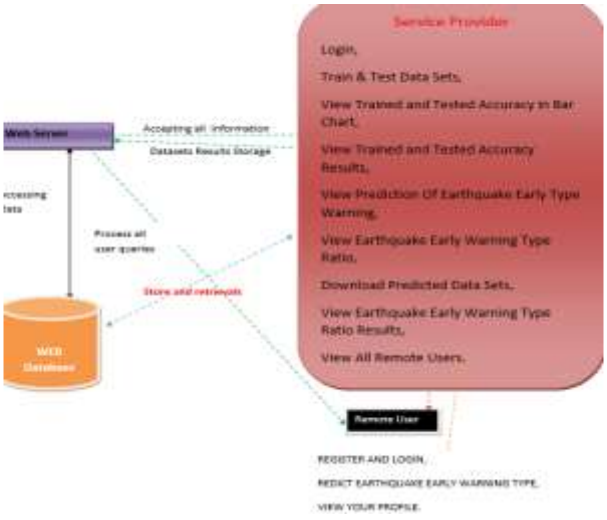
EARTHQUAKE hypocenter localization is essential in the field of seismology and plays a critical role in a variety of seismological applications such as tomography, source characterization, and hazard assessment. This underscores the importance of developing robust earthquake monitoring systems for accurately determining the event origin times and hypocenter locations. In addition, the rapid and reliable characterization of ongoing earthquakes is a crucial, yet challenging, task for developing seismic hazard mitigation tools like earthquake early warning (EEW) systems [1]. While classical methods have been widely adopted to design EEW systems, challenges remain to pinpoint hypocenter locations in real-time largely due to limited information in the early stage of earthquakes. Among various key aspects of EEW, timeliness is a crucial consideration and additional efforts are required to further improve the hypocenter location estimates with minimum data from 1) the first few seconds after the P-wave arrival and 2) the first few seismograph stations that are triggered by the

ground shaking.

The localization problem can be resolved using a sequence of detected waves (arrival times) and locations of seismograph stations that are triggered by ground shaking. Among various network architectures, the recurrent neural network (RNN) is capable of precisely extracting information from a sequence of input data, which is ideal for handling a group of seismic stations that are triggered sequentially following the propa- gation paths of seismic waves. This method has been inves- tigated to improve the performance of real-time earthquake detection [2] and classification of source characteristics. Other machine learning based strategies have also been proposed for earthquake monitoring. Comparisons between traditional machine learning methods, including the nearest neighbor, decision tree, and the support vector machine, have also been made for the earthquake detection problem [3]. However, a common issue in the aforementioned machine learning based frameworks is that the selection of input features often requires expert knowledge, which may affect the accuracy of these methods. Convolution neural networks-based clustering methods have been used to regionalize earthquake epicenters [4] or predict their precise hypocenter locations [5]. In the latter case, three-component waveforms from multiple stations are exploited to train the model for swarm event localization. In this study, we propose a RF-based method to locate earth- quakes using the differential P-wave arrival times and station locations (Figure 1). The proposed algorithm only relies on P- wave arrival times detected at the first few stations. Its prompt response to earthquake first arrivals is critical for rapidly disseminating EEW alerts. Our strategy implicitly considers the influence of the velocity structures by incorporating the source-station locations into the RF model. We evaluate the proposed algorithm using an extensive seismic catalog from Japan. Our test results show that the RF model is capable of determining the locations of earthquakes accurately with minimal information, which sheds new light on developing

efficient machine learning.

SYSTEM ARCHITECTURE


Service Provider
Login,
Train & Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
View Prediction Of Earthquake Early Type Warning,
View Earthquake Early Warning Type Ratio,
Download Predicted Data Sets,
View Earthquake Early Warning Type Ratio Results,
View All Remote Users.
Web Server
Accepting all information
Datasets Results Storage
ccessing
ata
Process all
user queries
Store and retrievals
WEB
Database
Remote User
REGISTER AND LOGIN,
REDICT EARTHQUAKE EARLY WARNING TYPE,
VIEW YOUR PROFILE.

METHODOLOGY

A. Differential travel timebased epicenter prediction

To estimate the epicenters of earthquakes, theRF model is trained via a supervised learning scheme. Two sets of properties, including the differential P-wave arrival times and station locations, are utilized as the model input, which can be expressed as

$$X = [T_i, Y_i, Z_i], \quad (1)$$

where T_i represents the P-wave travel time of the i^{th} station relative to that of a reference station, and Y_i and Z_i are the corresponding latitude and longitude of the target station. In this study, we set the P-wave arrival time at the first recording station as the reference (i.e., $T_i = t_i \; t_1$) and utilize five stations to locate the earthquakes. The input parameters consist of a total of 14 features that are defined as

$T_i = \{t_2 - t_1, t_3 - t_1, t_4 - t_1, t_5 - t_1\}$,

$Y_i = \{y_1, y_2, y_3, y_4, y_5\}$,

$Z_i = \{z_1, z_2, z_3, z_4, z_5\}$.

The combination of these features enables the network to determine the relative location (i.e., the latitude/longitude difference) between the earthquake and the reference station.

B. Random Forest (RF)

The final output of a RF model is obtained by averaging the predictions from K trees as

$$\overline{H}(X) = \bar{H}(X) = \left(\frac{1}{K}\right) \sum_{k}^{K} H(X;\vartheta_k),$$

where O denotes the latitude and longitude difference between the event and the reference station and N represents the number of training earthquakes. We tune two hyperparameters,the maximum number of trees (*mtree*) and the maximum depth of each tree (*mdep*), during the training process. The training of each tree is conducted by randomly drawing M records [6] from the training earthquakes, with a sampling ratio (MS) that varies between 0 and 1. Each node in a decision tree (except for the leaf node) is split into morebranches while considering a random subset of features, with the number of features represented by MF . The training process is performed through the following steps:

A) Growing the number of trees to *mtree*.

B) Picking M random records according to the MS factor.

C) Randomly splitting each tree into *mdep* levels.

D) Randomly selecting the MF at each splitting node.

E) Obtaining the averaging of the *mtree* trees outputsaccording to Equation. 3.

F) Obtaining the loss function according to Equation. 4.

G) Repeating steps B-F until the loss function converges.

C. The Architecture of RF

The robust performance of the RF model relies on well- designed network architecture. We tune its parameters by a trial-and-error approach. Firstly, we test the number of trees (*mtree*) from 500 to 10000 with an interval of 500, and the depth of each tree (*mdep*) from 10 to 200 with an interval of 5. The optimal values of *mtree* and *mdep* are 1000 and 100, respectively. Secondly, we increase the *MS* factor from

0.1 to 1 with an interval of 0.1. A *MS* factor of 1 achieves the optimal result, which indicates that all records contribute positively to the training process. Finally, we test a series of *MF* , e.g, 2, 3, 5, 7, 8, 9, 11, and 13, and the optimal value is determined to be 8. The architecture of the proposed algorithm is shown in Figure 1.

ALGORITHM

Decision tree classifier : Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the **set Si.**

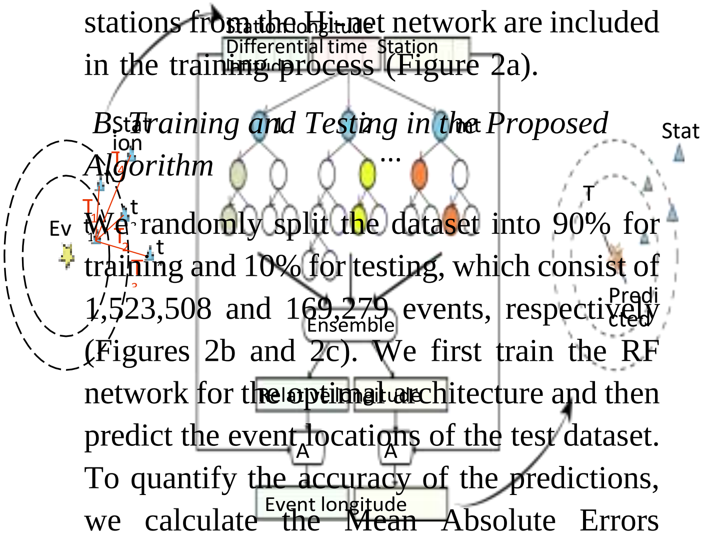
RESULTS ANALYSIS

A. *Dataset and Model Inputs*

We apply the proposed network to an earthquake detection problem in Japan (Figure 2a). catalog reported by the National Research Institute for Earth Science and Disaster Resilience, the Japan Meteorological Agency, and various institutions. This large catalog includes 2,235,159 regional seismic events recorded by the Hi-net seismic network between January 1st, 2009 and November, 11th 2020. For each event, we extract the source parameters including arrival times, magnitudes, depths, latitudes, and longitudes, as well as the locations of recording stations. We define qualified events for further analysis as those satisfying the following criteria: A) P-wave arrivals are detected at a minimum of five stations, B) Epicenter distances are less

than 1˚ (≈112 km), and C) Magnitudes of the events are

earthquakes while ensuring relatively reliable predictions. The final catalog, which contains a total of 1,692,787 qualified events, is characterized by a broad distribution of source parameters and offers an ideal dataset to train and test the proposed algorithm. In this catalog, longitude varies between 121.86° and 146.48° and the latitude between 23.42° and 46.22°. Event magnitude ranges from 0.10 M_L to 7.59 M_L, and depths from 0 to 440.78 km. Note that the intermediate (80-300km) and deep (300+km) events in the training dataset only affect the location accuracy marginally according to some tests we have performed. To train the network, we set the earliest arrival time of a group of P waves as the reference time (t_1), and determine its differential travel times relative to later P phases recorded at the other stations (T_i). The latitudes (Y_i) and the longitudes (Z_i) of the recording stations are also used as input parameters for the RF model (Figure 1). Finally, a total of 1,541 stations from the Hi-net network are included in the training process (Figure 2a).



B. Training and Testing in the Proposed Algorithm

We randomly split the dataset into 90% for training and 10% for testing, which consist of 1,523,508 and 169,279 events, respectively (Figures 2b and 2c). We first train the RF network for the optimal architecture and then predict the event locations of the test dataset. To quantify the accuracy of the predictions, we calculate the Mean Absolute Errors (MAEs) of the latitude and longitude between the reported values from the catalog and those estimated by the RF model. We achieve MAEs of 0.015° (≈ 1.625 km) and 0.023° (≈ 2.553 km) for the latitude (Figure 2d) and longitude (Figure 2e), respectively, with corresponding standard deviations of 0.033° and 0.052°. These location errors lead to a distance MAE value of 2.879 km (Figure 2f). The resulting R^2 score reaches 0.9998, which suggests highly consistent values between the predicted and the catalog locations. We define the events with a distance error below 0.1° (≈ 11.2 km) as true positive (TP), otherwise are false positive (FP), and calculate the accuracy ($\frac{TP}{TP+FP}$). The resulting accuracy rate is 94.39%. We further examine spatial variation in the accuracy across the study region (Figure 3a).

To better illustrate our location results, we select a subset of events in central Japan (Figures 3b), where

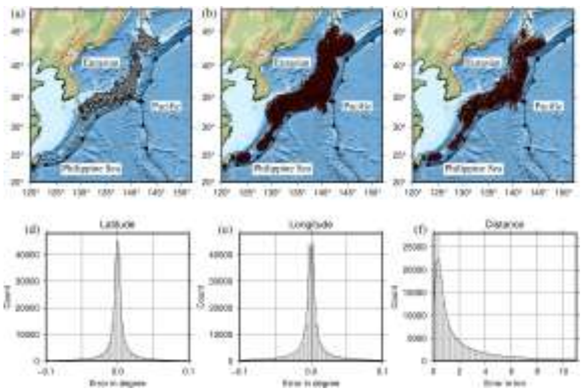


Fig. 2: (a) The distribution of the Japanese seismic stations. (b) The distribution of the training events. (c) The distribution of the testing events. (d) The error distribution between the catalog latitude and the estimated latitude corresponding to the proposed algorithm. (e) The error distribution between the catalog longitude and the estimated longitude corresponding to the proposed algorithm. (f) The error distribution between the catalog location and the estimated location corresponding to the proposed algorithm we observe relatively small errors in predicted locations if the first recording station is closer to the event. The distance errors of the estimated locations for all testing events show an overall small uncertainty of less than 0.1° (11.2 km), with slightly larger errors observed near the coastal regions (Figure 3c). This pattern is primarily caused by the varying station density and azimuthal station coverage, which is relatively sparse in the offshore region and limits the accuracy of location prediction. In the future, we will investigate how azimuthal distribution of stations will affect the accuracy of localization. Figure 3d shows the spatial distribution of ray density. The ray path density of each cell in a 250 250 grid is calculated by counting the number of rays intersecting that cell. Comparing Figures 3c and 3d, it is clear that the lowest ray-density area (e.g., around the coast) has the largest location-prediction error. For those high ray-density areas, e.g., the interior of Japan, the prediction errors are generally small.

CONCLUSION

We can pinpoint the specific site of the quake continuously by contrasting the hours of appearance of the P-waves with the areas of the seismic stations. One potential answer for this relapse issue is to utilize random forests (RF), with the result being the distinction in longitude and scope between the quake and the seismic stations. The contextual investigation of the Japanese seismic locale shows that it functions admirably and might be conveyed at this moment. We retrieve all occurrences from neighboring seismic stations that have a minimum of five P-wave arrival periods. After that, in order to build a machine learning model, we divided the extracted events into two datasets: one for training and one for testing. The recommended approach is adequately versatile to deal with continuous quake checking in additional troublesome districts; moreover, it can train using only three seismic sensors and 10% of the information, all while maintaining promising performance. Even though many networks are sparsely distributed, making it hard to train an efficient model using the random forest technique, one might make up for the absence of beam pathways in an objective district brought about by lacking index and station scattering by utilizing a few engineered datasets.

REFERENCES

[1] Q. Kong, R. M. Allen, L. Schreier, and Y.-W. Kwon, “Myshake: A smartphone seismic network for earthquake early warning and beyond,” *Science advances*, vol. 2, no. 2, p. e1501055, 2016.

[2] T.-L. Chin, K.-Y. Chen, D.-Y. Chen, and D.-E. Lin, “Intelligent real-time earthquake detection by recurrent neural networks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5440–5449, 2020.

[3] T.-L. Chin, C.-Y. Huang, S.-H. Shen, Y.-C. Tsai, Y. H. Hu, and Y.-M. Wu, “Learn to detect: Improving the accuracy of earthquake detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8867–8878, 2019.

[4] O. M. Saad, A. G. Hafez, and M. S. Soliman, “Deep learning approach for earthquake parameters classification in earthquake early warning system,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[5] X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, andW. Li, “Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method,” *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[6] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza, “Earthquake transformer an attentive deep-learning model for simultaneous earthquake detection and phase picking,” *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020

[8] S. M. Mousavi and G. C. Beroza, “A Machine- Learning Approach for Earthquake Magnitude Estima- tion,” *Geophysical Research Letters*, vol. 47, no. 1, p. e2019GL085976, 2020.