Time Series Analysis on Earthquakes Using EDA and Machine Learning

Muhammad Fakhrillah Abdul Azis*, Fariz Darari*†, Muhammad Rizqy Septyandy‡
*Faculty of Computer Science
Universitas Indonesia, Depok, Indonesia
Email: fpvariel@gmail.com, fariz@ui.ac.id
†Tokopedia-UI AI Center of Excellence
Universitas Indonesia, Depok, Indonesia
‡Geology Study Program, Faculty of Mathematics and Natural Science
Universitas Indonesia, Depok, Indonesia
Email: m.rizqy@sci.ui.ac.id

***Abstract*—An earthquake is a sudden, rapid shaking of the ground caused by the shifting of the Earth's tectonic plates. Earthquakes pose serious threats that cause economic losses and casualties. To mitigate such risks, it is crucial to better understand earthquakes through data-driven analysis. In this paper, we propose an approach to time series analysis over earthquake data, consisting of two steps: exploration and prediction. The exploration step relies on exploratory data analysis (EDA) comprising descriptive statistics and data visualization, whereas the prediction step focuses on how to predict the number of earthquakes for the following years. We perform our time series analysis using various machine learning techniques over a global earthquake dataset from 1965-2016 and report insights as well as lessons learned from the study.**

***Keywords*—Earthquake, Time Series Analysis, EDA, Machine Learning, Linear Regression, LSTM, Prophet**

I. Introduction

An earthquake is a natural phenomenon occurring when there is a sudden, rapid shaking of the ground. Earthquakes are usually caused by the breaking and displacement of rocks below Earth's crust [1]. The deadliest earthquake ever recorded happened in 1556, striking the province of Shaanxi, China, and claiming the lives of about 830,000 people at the time.[1] In Indonesia, over 200,000 people lost their lives in the 2004 Indian Ocean earthquake and tsunami. Without a doubt, earthquakes may bring tremendous (negative) impacts to human lives and also the economy.

In light of the (potential) damage of earthquakes, a number of studies have been done in order to better understand how, when, and where earthquakes may occur. Rouet-Leduc et al. [2] predicted so-called labquakes (that is, earthquakes in the laboratory settings) using machine learning (ML) techniques. The predictions were made based on the immediate characteristics of the acoustic signal without considering its history. In [3], Martínez-Álvarez et al. proposed the use of seismicity indicators to predict earthquakes in Chile and the Iberian Peninsula based on the application of artificial neural networks (ANNs). The work relied on Weka,[2] an off-the-shelf ML tool, and focused on forecasting earthquake occurrences in a reduced area with a temporal horizon of 5-7 days. Asim et al. [4] employed an earthquake prediction system based on Support Vector Regressor (SVR) and Hybrid Neural Network (HNN) on Hindukush, Chile, and Southern California regions. Anagnostopoulos and Moretti [5] proposed assessment criteria as to how earthquakes may affect the overall safety of a building based on the damage of its components. Their work may serve as guidelines regarding which residential areas are better/less prepared for earthquakes (should they occur). Other efforts in predicting earthquakes also existed, which were

[1]http://content.time.com/time/specials/packages/completelist/0,29569,1953425,00.html

[2]https://www.cs.waikato.ac.nz/ml/weka/

978-1-7281-9279-6/20/$31.00 ©

based on electromagnetic waves, seismic waves, foreshocks, seismicity & chemistry changes, or plate tectonic settings, as surveyed in [6]. Given these research results, predicting earthquakes in general is, however, still deemed to be a difficult (if not impossible) task [7].

Time series analysis is performed over observations collected in time sequences [8]. Earthquakes are inherently time series data, as their occurrences can be grouped by time (e.g., year). In this paper, we propose *an approach to time series analysis over earthquake data, consisting of two steps: exploration and prediction.* Our exploration step makes use of descriptive statistics as well as data visualization, while our prediction step applies different ML techniques to forecast the number of earthquakes in future years. We perform our time series analysis over a global earthquake dataset in the period of 1965-2016, provided by the US Geological Survey [9]. We hope that the exploration and prediction results of this work may shed some light on how to approach the problem of analyzing earthquakes wrt. specificities in terms of tectonic plates and time zones.

The rest of the paper is structured as follows. Section II provides preliminaries, while Section III describes the research methodology. Section IV reports and visualizes our exploration. Section V discusses how we predict the number of earthquakes in the future based on past trends. We conclude our work in Section VI.

II. Preliminaries

Earthquakes. Earthquakes are mainly studied in geology, which is the science that deals with the composition and dynamics of the Earth. The Earth's lithosphere (i.e., outer layer) is composed of tectonic plates, such as the Pacific Plate, the North American Plate, and the African Plate. An earthquake is an intense shaking of Earth's surface, caused by the shifting of the Earth's tectonic plates. The Ring of Fire is a seismically and volcanically active area where the Pacific Plate meets its neighboring tectonic plates, accounting for around 90% of the world's earthquakes.[3]

[3]https://www.nationalgeographic.org/article/plate-tectonics-ring-fire

Exploratory Data Analysis. Exploratory Data Analysis (EDA) is a (graphical) approach for analyzing data in order to learn data characteristics, detect outliers/anomalies, and test underlying assumptions [10]. EDA provides guidelines as to how to look and interpret data, and is usually a precursor to more advanced data analysis techniques (e.g., statistical modeling and machine learning). Visualization techniques in EDA often leverage the use of raw data plots, such as histograms and barcharts, as well as simple statistical plots, such as boxplots and meanplots. With respect to earthquake data, which is temporal and spatial by nature, EDA is also equipped with line charts (suitable for showing data changes over *time*) and map plotting (suitable for visualizing the *spatial* aspect of data). Details regarding how we perform EDA over earthquakes are given in Section III.

Machine Learning. Machine Learning (ML) concerns how to build a model from data/experience. Supervised learning is an ML approach where a predictive model is learned from labeled (training) data. In this paper, we specifically rely on two supervised learning methods: linear regression and Recurrent Neural Networks (RNNs). Linear regression assumes a linear relationship between the input variables and a numerical output. A linear regression task involves the finding of the best-fitting straight line through data points. RNNs are a neural-network family with cyclic connections, suitable to model sequence data. Long Short-Term Memory (LSTM) is an improved RNN developed to overcome RNN's modeling weaknesses (e.g., vanishing gradients) [11]. LSTM has been applied to various domains [12], such as sentiment classification, handwriting recognition, and time series prediction, and is often regarded as a forefront in deep (machine) learning. In addition to linear regression and RNNs, we also use Prophet [13], a forecasting modeling approach by Facebook. Prophet is based on an additive model that is able to fit to non-linear trends. Prophet is claimed to be robust to missing data, trend shifts, and outliers. As for model evaluation, we will use R^2, a common metric for evaluating time series. The metric R^2 measures the proportion of variation over a dependent variable that can be attributed to the independent variables.

III. Methodology

In this section, we describe the methodology of our research. We first introduce the earthquake dataset used in this research, followed by explaining approaches to explore and make predictions based on the dataset.

Dataset. For our work, we use the Significant Earthquakes dataset, provided by the U.S. Geological Survey [9]. The dataset consists of 23,412 recorded global earthquake occurrences with magnitude of 5.5 or higher from the period between 1965 and 2016. The dataset consists of 21 columns, which can be categorized into the main ones (i.e., Date, Latitude, and Longitude) and the geology-specific ones (like Azimuthal Gap, Depth Seismic Stations, and so on). As we investigate mainly the temporal and spatial aspect of earthquakes, we only take the main columns. Additionally, we also rely on a dataset about plate boundaries by Peter Bird [14], which has been parsed and cleaned.[4] This dataset consists of earth plate coordinates as well as boundaries, and can be used to group earthquakes by plates. We will later refer to this dataset as the GeoJSON dataset.

Exploration. First of all, we need to explore and see how our dataset looks like. The exploration step relies on data visualization, so that we may learn the characteristics of our earthquake data. The visualization is done by drawing tables, graphs, and plot earthquake locations on the map.

In this paper we propose two categories of visualization based on the spatial nature of earthquakes. The first categorization is based on *earth plates*. The GeoJSON dataset, as mentioned before, contains boundary coordinates of earth plates from all over the world. To determine on which plate an earthquake occurs, we check whether the earthquake coordinate is contained within the polygon shape of some plate. The second category is based on *time zones*. We rely on the Greenwich Mean Time (GMT) standard to divide the earthquake locations into 24 time zones.

Prediction. Here, we propose an approach to predict the number of earthquakes in a certain year based on the number of earthquakes in the previous years. The workflow of our prediction step is shown in Fig. 1. As previously mentioned, earthquakes can be grouped based on two categories, resulting in: 52 groups based on earth plates, and 24 groups based on time zones. Moreover, for each group there are two time series variants: the normal and stationarized time series. A stationary time series would have constant statistical properties (e.g., mean, variance) over time, which might improve its predictability. We rely on three different prediction modeling techniques: linear regression, LSTM, and Prophet. In total, there are 152 model settings: 52×2 models from earth plate categorization and 24×2 models from time zone categorization.

All the steps mentioned above are implemented using Python with external libraries. We use pandas[5] for feature engineering, matplotlib[6] and folium[7] for data visualization, and keras[8] and scikit-learn[9] for machine learning.

IV. Exploration

In this section, we report on the results of our time series exploration over the global earthquake dataset. Fig. 2 plots the earthquake locations (based on the earthquake coordinates) on the world map along with plate boundaries. At first glance, we observe that most earthquakes occurred near the plate boundaries. This might be due to the existence of subduction zones at plate boundaries, which may generate many earthquakes [15]. Moreover, we notice that different plate boundaries have different earthquake frequencies. For example, there are much more earthquakes happening on the Pacific Plate and Australian Plate than, e.g, the African Plate. This phenomenon could be due to the fast slip-rate characteristics of those two plates [15].

Fig. 3 visualizes the number of earthquakes per year from 1965 to 2016. In general, we observe an increasing trend of earthquake occurrences, reaching the peak at 712 occurrences in 2011. Note that the year 2011 is when the Great Tohoku earthquake happened, considered to be the strongest earthquake

[4] https://github.com/fraxen/tectonicplates

[5] https://pandas.pydata.org/
[6] https://matplotlib.org/
[7] https://python-visualization.github.io/folium/
[8] https://keras.io/
[9] https://scikit-learn.org/stable/

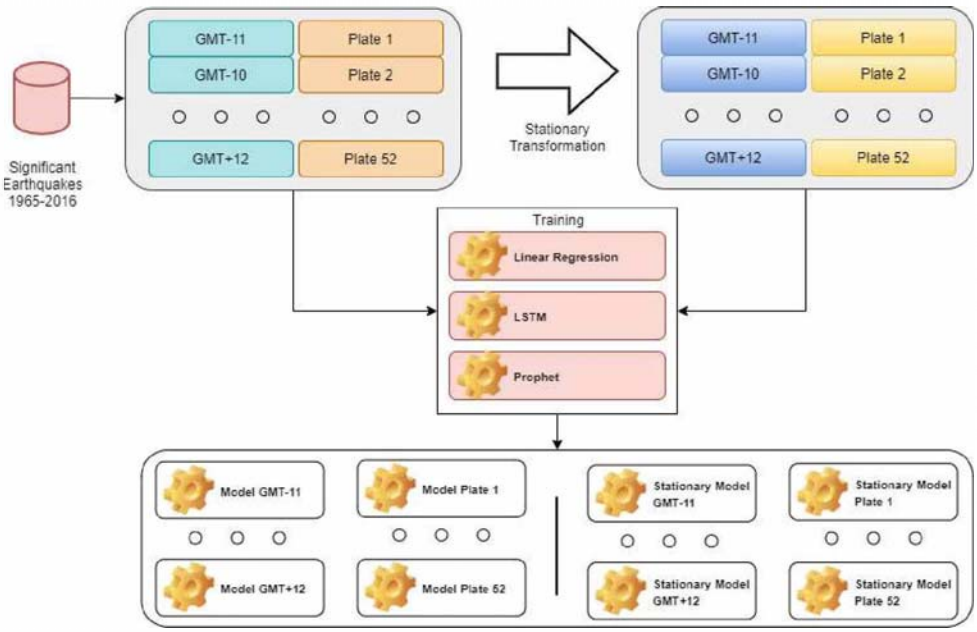

Fig. 1. Flowchart of Earthquake Prediction

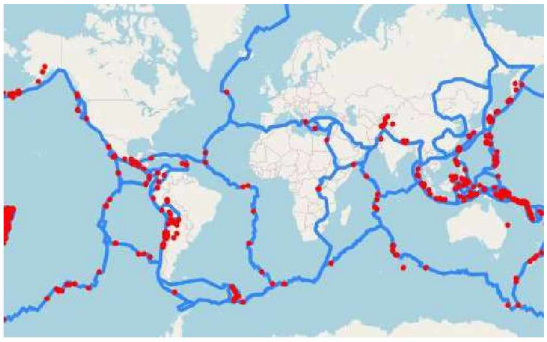

Fig. 2. Earthquakes and Plate Boundaries Plotted on the World Map

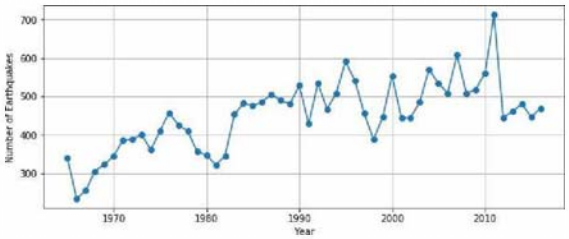

Fig. 3. Number of Earthquakes per Year

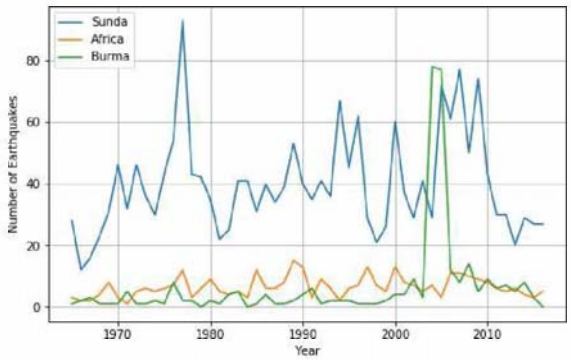

Fig. 4. Comparison of the Number of Earthquakes per Year in the Plates of Sunda, Africa, and Burma

recorded in the Japan history.[10] Also, we notice that the earthquake occurrences from 1990 onwards tend to fluctuate more. This might be due to the more frequent occurrences of large earthquakes (with a magnitude of ≥ 8.3 M_W) in the following years after 1990 [16] along with the theory of energy release in aftershocks [17].

Now, we would like to examine the interplay between the spatial and temporal aspect of earth-

[10]https://www.nationalgeographic.org/thisday/mar11/tohoku-earthquake-and-tsunami/

quakes. As previously observed, the earthquake occurrences are unevenly distributed across tectonic plates. Fig. 4 gives a closer look at the number of earthquakes per year in three different plates:[11] the Sunda Plate, which is tectonically active; the African Plate, which is tectonically passive; and the Burma Plate, which exhibits irregularity. As seen in the figure, the number of earthquakes in the Sunda Plate is consistently higher than that of the African Plate during the whole period. Indeed, on average there are about 40 earthquakes per year in the Sunda Plate compared to just 6 in the African Plate. Furthermore, the average number of annual earthquakes per plate is around 9, so the Sunda Plate significantly deviates from the majority. As for the Burma Plate, we observe an anomaly: though it generally has a low number of earthquakes per year, in the year of 2004 and 2005, there are spikes in the number of earthquakes, that is, 78 and 77, respectively. We suspect that this irregularity is associated with the magnitude 9.1 Sumatra-Andaman earthquake on December 26, 2004, which took place on the interface between the India and Burma Plates.[12]

As an alternative to the tectonic plate categorization, we also explore the grouping of earthquakes based on time zones. Fig. 5 shows the number of earthquakes per year on the time zones of GMT+10, GMT-10, and GMT+6. The figure exhibits a similar pattern as that in Fig. 4:[13] the GMT+10 time zone has a significantly higher number of earthquakes (averaging at 67 of annual earthquakes) than that of the GMT-10 time zone (averaging at just 5 of annual earthquakes) for the whole period, and the GMT+6 time zone (where the Burma Plate is located) has an anomaly in year 2004 and 2005.

V. Prediction

In this step, we aim to predict the number of earthquakes in the future based on that of past years. The implementation of the prediction follows the methodology (see Section III). In particular, we vary

[11]We take these three as representatives out of all the 52 plates.

[12]https://www.usgs.gov/news/indian-ocean-tsunami-remembered-scientists-reflect-2004-indian-ocean-killed-thousands

[13]By similar, we mean that there are groups with a high number of earthquake occurrences, groups with a low number, and groups with an anomaly.

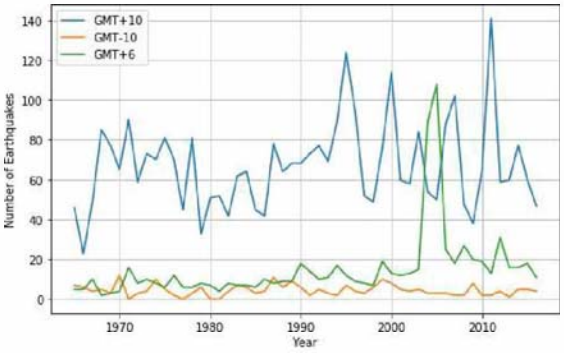

Fig. 5. Comparison of the Number of Earthquakes per Year in the Time Zones of GMT+10, GMT-10, and GMT+6

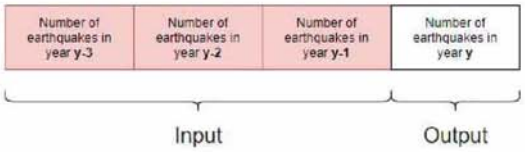

Fig. 6. Input-Output Pair for Window Size 3

the window size for training (and testing) from 3 to 5. Fig. 6 illustrates the input-output pair for the window size of 3. In the figure, the number of earthquakes in year y is predicted based on the number of earthquakes in the past 3 years. The input-output pairs for the other window sizes follow the same scheme. Note that the window size settings would only apply to linear regression and LSTM, while Prophet does not have such a window size setting (as per the Prophet documentation).[14] When building our models, we set the earthquakes in year 1965–2006 as the training data, and the earthquakes in year 2007–2016 as the testing data.

Let us now describe the evaluation results. Note that there are mainly two categorizations, based on tectonic plates and time zones. Table 1 and Table 2 show the average of our model evaluation for representative plates and time zones, respectively.[15] We take the window size of 3 to be shown in Table 1 since the overall prediction performance for that size wrt. plate categorization is better compared to the other sizes. On the other hand, wrt. time zones the

[14]https://facebook.github.io/prophet/

[15]We take representatives as the raw data is too large to be shown here.

Table 1
EVALUATION OF PREDICTION RESULTS FOR PLATE CATEGORIZATION

Model	Plate	Normal Data	Stationarized Data
LSTM, Window Size 3	Niuafo'ou	0.19	0.65
	Easter	-0.04	0.63
	Pacific	-0.71	0.21
	Burma	-3168.72	-238.56
Linear Regression, Window Size 3	Niuafo'ou	-0.22	0.50
	Easter	-0.08	0.34
	Pacific	-0.87	0.47
	Burma	-138.70	-17.31
Prophet	Niuafo'ou	-0.01	0.01
	Easter	-0.20	-0.10
	Pacific	-0.47	-0.04
	Burma	-15.54	-0.21

Table 2
EVALUATION OF PREDICTION RESULTS FOR TIME ZONE CATEGORIZATION

Model	Plate	Normal Data	Stationarized Data
LSTM, Window Size 4	GMT+1	-0.04	0.68
	GMT-7	-0.03	0.66
	GMT+10	0.15	0.62
	GMT+6	-135.32	-1591.69
Linear Regression, Window Size 4	GMT+1	0.08	0.21
	GMT-7	0.16	0.71
	GMT+10	0.02	0.45
	GMT+6	-148.81	-35.18
Prophet	GMT+1	-1.30	-0.91
	GMT-7	-0.98	-0.23
	GMT+10	-1.04	-0.94
	GMT+6	-2.77	-0.09

window size of 4 has the best performance, hence shown in Table 2. Note we regard the Burma Plate and the time zone GMT+6 (where the Burma Plate resides) as an outlier, and that we do not consider such an outlier in computing the overall prediction performance. We nevertheless show the Burma Plate results as a reference.

In both categorizations, we observe from the results that the stationary transformation does improve the prediction performance in terms of R^2. For instance, the R^2 scores for the Niuafo'ou Plate and the Pacific Plate in Table 1 increase from 0.19 to 0.65 and -0.71 to 0.21, respectively.

Now, let us compare the prediction performance in terms of modeling techniques. Both tables show a mixed observation: that in some plates/time zones, LSTM performs better that linear regression, and the other way around in other plates/time zones. Nevertheless, the average R^2 among *all* the plates in our stationarized data for LSTM is lower than that of linear regression (i.e., -0.35 vs. 0.18), and the average R^2 wrt. *all* the time zones for LSTM is also lower compared to linear regression (i.e., -0.14 vs. 0.19). The Prophet-based modeling, however, does not exhibit a satisfactory result in our experiments.

A closer look at the comparison between the stationary-model prediction for the window size of 3 of LSTM and linear regression, and the ground truth is exemplified in Fig. 7. The x-axis of the figure represents the year, whereas the y-axis represents the difference of the number of earthquakes from the previous year. Both the models are able to capture the earthquake fluctuations, though the exact predicted number of earthquakes still deviates from the ground truth.

Let us sum up the results in the prediction step. First, the stationary transformation may improve the prediction performance. Second, plates and time zones may vary in their prediction results. Finally, though in some cases LSTM could perform better

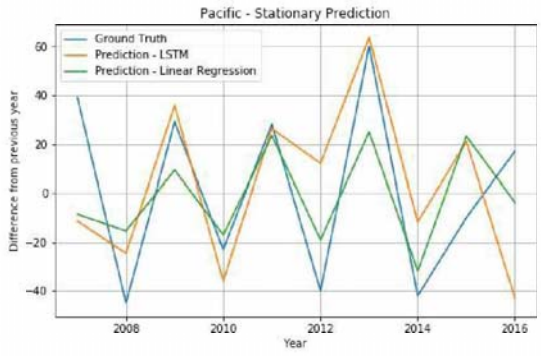

Fig. 7. Comparison of Stationary Prediction and Ground Truth for the Pacific Plate

than linear regression, in general linear regression still has a better performance than LSTM, possibly due to the small size of data.[16]

VI. Conclusions

Analyzing earthquakes is indeed a challenging task. Nevertheless, we have proposed an approach to time series analysis over earthquake data, consisting of two steps: exploration and prediction. As for the exploration step, we have made visualizations using map plotting and line charts. From these, we have gained insights, for example, the earthquake distribution on interfaces between plate boundaries (particularly wrt. convergent boundaries, such as the Pacific Plate and Australian Plate, and divergent boundaries, such as the African Plate and Arabian Plate)[17], and the temporal characteristics of earthquakes. As for the prediction step, we have compared the performance of prediction models in terms of normal vs. stationarized data, plate vs. time zone categorizations, and modeling techniques (i.e., LSTM, linear regression, and Prophet). The prediction models basically have various performance results depending on the plate and time zone.

Future directions include the inclusion of other time series modeling techniques (e.g., ARIMA), more fine-grained time units for prediction, and the incorporation of geology-specific aspects in the exploration and prediction.

[16]LSTM tends to require large data than that of linear regression.

[17]By convergent and divergent, we mean colliding and spreading, respectively.

Acknowledgments

This work is supported by the 2020 PUTI research grant "Knowledge Graph-based AI – Analysis and Applications" from Universitas Indonesia. We thank the anonymous reviewers for their valuable feedback.

References

[1] S. Earle, *Physical Geology – 2nd Edition*. Victoria, B.C.: BCcampus, 2019. [Online]. Available: https://opentextbc.ca/physicalgeology2ed/

[2] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, "Machine learning predicts laboratory earthquakes," *Geophysical Research Letters*, vol. 44, no. 18, pp. 9276–9282, 2017.

[3] F. Martínez-Álvarez, J. Reyes, A. Morales-Esteban, and C. Rubio-Escudero, "Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula," *Knowl. Based Syst.*, vol. 50, pp. 198–210, 2013.

[4] K. M. Asim, A. Idris, T. Iqbal, and F. Martínez-Álvarez, "Earthquake prediction model using support vector regressor and hybrid neural networks," *PLOS ONE*, vol. 13, no. 7, pp. 1–22, 2018.

[5] S. Anagnostopoulos and M. Moretti, "Post-earthquake emergency assessment of building damage, safety and usability—part 1: Technical issues," *Soil Dynamics and Earthquake Engineering*, vol. 28, no. 3, pp. 223–232, 2008.

[6] L. R. Sykes, B. E. Shaw, and C. H. Scholz, "Rethinking earthquake prediction," *Pure and Applied Geophysics*, vol. 155, no. 2, pp. 207–232, 1999.

[7] H. Joffe, T. Rossetto, C. Bradley, and C. O'Connor, "Stigma in science: the case of earthquake prediction," *Disasters*, vol. 42, no. 1, pp. 81–100, 2018.

[8] J. Cryer and K. Chan, *Time Series Analysis: With Applications in R*, ser. Springer Texts in Statistics. Springer, 2008.

[9] USGS, *Significant Earthquakes, 1965-2016*, 2017 (accessed June 1, 2020). [Online]. Available: https://www.kaggle.com/usgs/earthquake-database

[10] NIST/SEMATECH, *e-Handbook of Statistical Methods*. NIST, 2012. [Online]. Available: https://doi.org/10.18434/M32189

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[12] J. Wang, T. Liu, X. Luo, and L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," in *ROCLING*, 2018.

[13] S. J. Taylor and B. Letham, "Forecasting at scale," *PeerJ Prepr.*, vol. 5, p. e3190, 2017.

[14] P. Bird, "An updated digital model of plate boundaries," *Geochemistry Geophysics Geosystems*, vol. 4, p. 1027, 03 2003.

[15] P. Kearey, K. A. Klepeis, and F. J. Vine, *Global Tectonics*. NJ, USA: Wiley-Blackwell, 2009, vol. 3rd Edition.

[16] G. Hayes, G. Smoczyk, V. A.H., K. Furlong, and H. Benz, "Seismicity of the Earth 1900–2018," in *U.S. Geological Survey Scientific Investigations Map 3446*, 2020.

[17] T. Utsu, "Aftershocks and earthquake statistics: Further investigation of aftershocks and other earthquake sequences based on a new classification of earthquake sequences," *Journal of the Faculty of Science, Hokkaido University*, vol. 3, no. 4, pp. 197–266, 1970.

978-1-7281-9279-6/20/$31.00 © 2020 IEEE