# Analysis and Prediction of the Occurrence of an Earthquake Using ARIMA and Statistical Tests

**Rabbani Nur Kumoro[*1], Audrey Shafira Fattima[2], William Hilmy Susatyo[3], Dzikri Rahadian Fudholi[4]**

[1,2]Program Studi S1 Ilmu Komputer FMIPA UGM, Yogyakarta, Indonesia

[2]Departemen Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta, Indonesia

e-mail: [*1]rabbani.nur.kumoro@mail.ugm.ac.id, [2]audrey.shafira1503@mail.ugm.ac.id, [3]william.hilmy.susatyo@mail.ugm.ac.id, [4]dzikri.r.f@ugm.ac.id

***Abstrak***

*Salah satu tujuan utama dari seorang geoscientist adalah untuk memprediksi waktu dan besarnya gempa bumi. Pada penelitian ini, kami menunjukkan bahwa prediksi gempa bumi dapat dilakukan dengan Machine Learning menggunakan metode Time Series, dimana dilakukan analisis pada dataset untuk mendapatkan pola dari sinyal seismik berdasarkan runtun waktu. Berdasarkan hasil penelitian yang sudah dilakukan, dapat disimpulkan bahwa model dengan metriks evaluasi terbaik adalah ARIMA dengan Mean Absolute Error (MAE) sebesar 0.11.*

***Abstract***

*The primary objective of a geoscientist is to accurately forecast the timing and magnitude of earthquakes. In this study, we demonstrate the efficacy of employing Machine Learning techniques, specifically the Time Series method, to achieve earthquake predictions. This approach involves analyzing seismic signal patterns within the dataset, utilizing their chronological sequence. Through meticulous research and analysis, we have determined that the optimal evaluation metric for this purpose is ARIMA, which yielded a remarkable Mean Absolute Error (MAE) score of 0.11.*

***Keywords*** *— Earth Sciences, Forecasting, Machine Learning, Seismic Signals, Time Series.*

## 1. INTRODUCTION

The main focus of Earth Sciences research is to forecast the timing and severity of earthquakes. Earthquakes are natural phenomena that occur when there is movement or vibration within the Earth's layers. Essentially, earthquakes can transpire worldwide and are typically caused by tectonic activities such as plate movements or volcanic events. Predicting the timing of earthquakes is of utmost importance, given the high potential for devastating consequences resulting from such disasters. A tangible example of the adverse impact of such a disaster can be observed in the earthquake that occurred in Turkey and Syria on February 15, 2023, claiming a minimum of 41,232 lives, and causing estimated material losses of up to US$1 billion [1].

One of the factors utilized in predicting the timing of earthquakes is the utilization of information derived from seismic signals based on time series analysis. Seismic signals encompass vibrations or waves that propagate through the Earth's interior as a result of geological activities such as earthquakes, explosions, or magma movements within volcanoes. These signals are employed in the field of seismology to study the nature and structure of the Earth, as well as to

detect and monitor seismic activity that may pose risks to human safety and the environment. There are several types of seismic signals, including primary waves (P-waves), secondary waves (S-waves), and surface waves, that are commonly employed in seismological research [2]. In this study, we use a stochastic modeling approach known as the AutoRegressive Integrated Moving Average and compare it with Linear Regression, to forecast the timing of an earthquake.

## 2. METHODS

### *2.1 Dataset*

The dataset used in this study was obtained from the Los Alamos National Laboratory (LANL) via the Kaggle platform [3]. The dataset consists of a total of 577,060,473 rows. This data was artificially generated during an experiment conducted on rocks by LANL using a device called the "classic lab earthquake model." The model simulates the loading and failure cycles of a tectonic fault, reproducing the processes involved in natural earthquakes [4]. The laboratory model imitates tectonic faults in the earth's layers. Although this is a simplified version of an actual earthquake, it is claimed that it shares most of the physical characteristics of a real earthquake. The data is periodic with realistic behavior, which means that the data includes earthquakes that occur irregularly. Furthermore, the utilized dataset contains two features. As shown in Table 1, the two features are *acoustic_data* and *time_to_failure*. In this study, the target feature to be predicted is *time_to_failure*.

**Table 1.** Features in the Dataset.

| No. | Feature | Data Type | Explanation |
|---|---|---|---|
| 1. | *acoustic_data* | Integer | seismic signal |
| 2. | *time_to_failure* | Float | time until the occurrence of an earthquake |

### 2.3 *Data Pre-Processing*

To ensure that the dataset used in the machine learning model yields accurate results, data preprocessing is conducted as a preliminary step. The preprocessing stage involves checking for missing values and skewness. As a result, the missing value check revealed that there are no missing data in the dataset. In addition, the skewness check indicates that the *acoustic_data* feature displayed a slight right-skewed distribution with a skewness value of 0.82, but this skewness was deemed insignificant, suggesting a generally balanced data distribution.

### 2.3 *Exploratory Data Analysis*

In this stage, exploration of the utilized dataset is conducted. The first step is to visualize the correlation between the two features, namely *acoustic_data* and *time_to_failure*. As depicted in Figure 1, there is no observable correlation between the two features. This indicates that the two features possess distinct information and are effective for use within the same machine learning model.

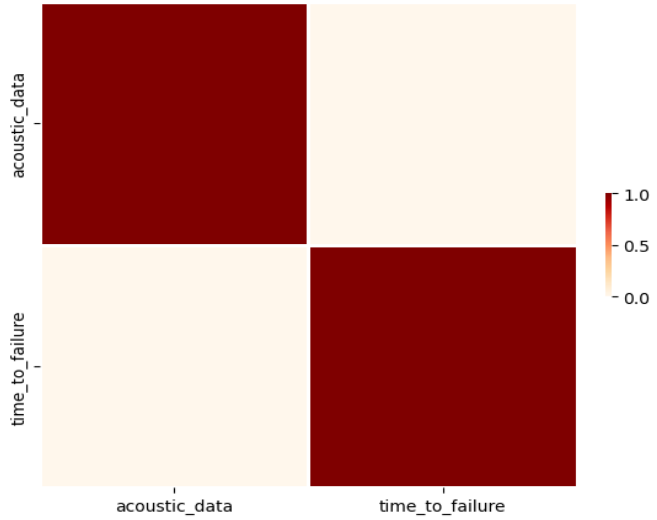


**Figure 1.** Correlation Matrix Visualization

Next, a scatter plot visualization is performed to examine whether there are any outliers in the distribution of *acoustic_data* with respect to *time_to_failure*. In Figure 2, it can be observed that there are outliers present in the *acoustic_data* feature. Essentially, the distribution of *acoustic_data* tends to exhibit greater variability as *time_to_failure* approaches 0.

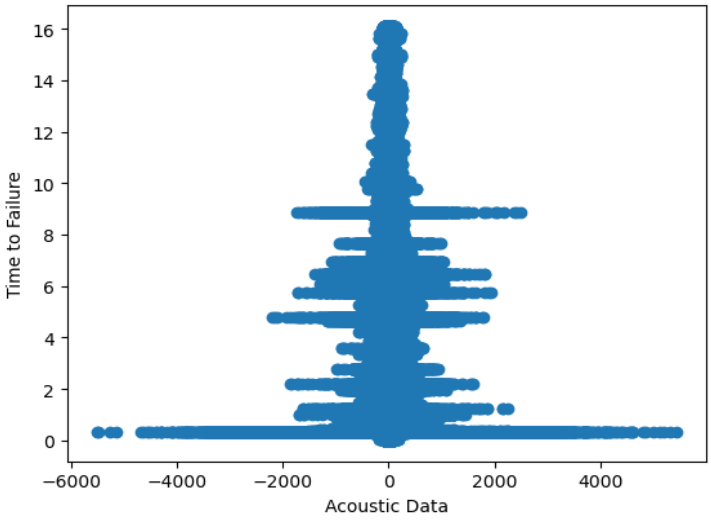


**Figure 2.** Scatter Plot Visualization

Subsequently, the Augmented Dickey-Fuller (ADF) statistical test was conducted. The ADF test is a method used to check if time series data is stationary, meaning its mean and variance remain constant over time [5, 6]. It uses linear regression with time series data as an independent variable. The null hypothesis of the test implies that the data contains a unit root, indicating non-stationarity. The test provides test statistics and p-values, and in this case, a very low p-value ($8.7221 \times 10^{-7}$) was obtained, leading to the rejection of the null hypothesis. This implies that the data is stationary, and there's no need to perform differencing for prediction.

### *2.4 Feature Engineering*

During this stage, the steps involved include dataset segmentation, feature manipulation within the dataset, and applying feature scaling using the Robust Scaler.

#### *2.4.1 Segmentation*

According to the findings of the exploratory data analysis, it was discovered that the number of rows in the dataset is extremely large. This will lead to a highly complex computational process. Therefore, at this stage, the dataset containing 577,060,473 rows is divided into 3,847 segments, with each segment consisting of 15,000 rows.

#### *2.4.2 Feature Manipulation*

From the previous stages, it can be observed that only one feature, namely *acoustic_data*, can be used as a predictor. This condition has the potential to result in suboptimal learning processes for the model, which in turn affects the model's inability to predict the target variable accurately. Therefore, in this stage, new features are added, including the average, standard deviation, maximum value, and minimum value of the *acoustic_value* for each row of the dataset within a specific segment. Additionally, the *rseismic* feature is also added, obtained from the difference between the *acoustic_mean* feature in two adjacent rows.

#### *2.4.3 Feature Scaling*

Based on the visualization of the scatter plot shown in Figure 2, it is evident that there is an outlier in the train acoustic feature. Therefore, the Robust Scaler is implemented to eliminate the outliers in this feature. The implementation of the Robust Scaler involves calculating the median and quartiles of the feature with outliers, followed by subtracting the median from each feature value and dividing the difference by the interquartile range (IQR). The description of features in the dataset after undergoing the feature engineering process can be seen in Table 2.

**Table 2.** Features in the Dataset after Feature Engineering.

| No. | Feature | Data Type | Explanation |
|---|---|---|---|
| 1. | *acoustic_mean* | Float | mean seismic signal on a specific segment |
| 2. | *acoustic_std* | Float | standard deviation of seismic signal on a specific segment |
| 3. | *acoustic_max* | Float | maximum value of seismic signal on a specific segment |
| 4. | *acoustic_min* | Float | minimum value of seismic signal on a specific segment |
| 5. | *rseismic* | Float | difference of the acoustic mean between two consecutive records. |
| 6. | *time_to_failure* | Float | time until the occurrence of an earthquake |

## *2.5 Modeling*

### *2.5.1 ARIMA*

AutoRegressive Integrated Moving Average (ARIMA) is a statistical method used to predict time series data. This model considers the relationship between the current value and the previous values in order to make accurate predictions. ARIMA is a combination of two statistical models, namely regression and moving averages, which are used to improve the accuracy of the model [7-9]. The model has three main parameters, which include the differencing level, autoregression, and moving average. Generally, the predictions made using the ARIMA model with orders (1, 1, 1), (4, 1, 1), or (2, 1, 2) can closely approximate actual data, as demonstrated in Figure 3.

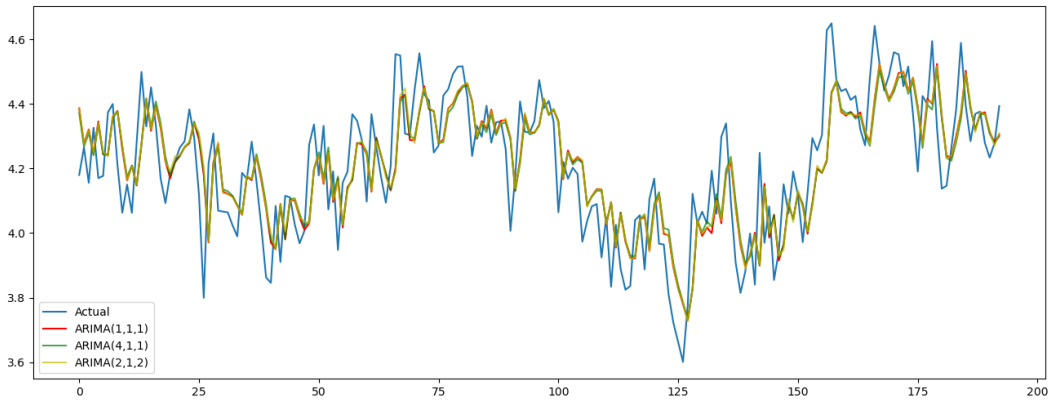


**Figure 3.** Comparison of 3 ARIMA Models Prediction Results

### *2.5.2 Linear Regression*

Linear Regression (LR) is a method used to predict the relationship between a dependent variable and one or more independent variables by assuming a linear relationship between the variables and creating the best-fitting line to represent that relationship [10, 11]. The result is an equation of the line that can be used to predict the value of the dependent variable based on the values of the independent variables. Based on the visualization in Figure 4, it can be observed that the predictions generated by the LR model tend to differ from the actual data. When the values in the actual data decrease, the predictions from the LR model do not decrease accordingly.

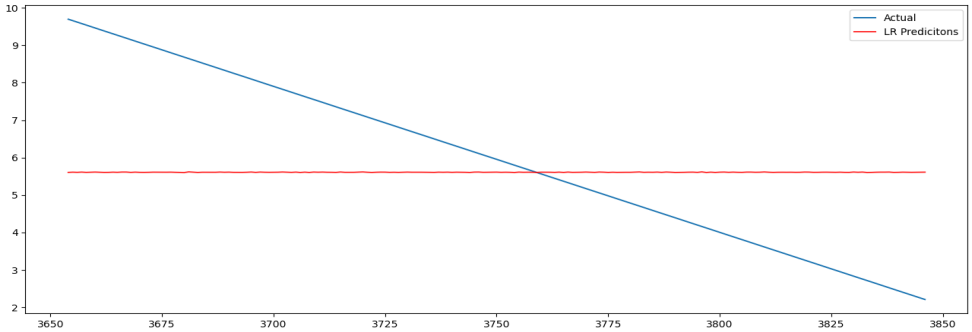


**Figure 4**. Linear Regression Model's Prediction Result

## 3. RESULTS AND DISCUSSIONS

### *3.1 Validation*

The custom train test split method is implemented on the dataset processed during the preprocessing and feature engineering stages. In the custom train test split method, the available dataset is divided into training data with a certain percentage, and the remaining data is used as validation data to measure prediction metrics. The difference between a custom train test split and a regular train test split lies in its flexibility. In the custom train test split, the data separation process is created in the form of a function, making it easier to modify the separation percentage if needed [12]. Furthermore, the use of custom train test split can also address issues of class imbalance and overfitting in the dataset.

### *3.2 Evaluation*

In this study, the metric used to evaluate the prediction results is the mean absolute error (MAE). Essentially, MAE measures the average of the absolute differences between the prediction results and the actual values or targets [13, 14]. MAE has a positive value and calculates the absolute error, thereby unaffected by the error's sign, whether the prediction is too high or too low. A smaller MAE value indicates better model performance. The formula used to calculate MAE is as follows:

$$MAE = \frac{1}{n}\sum |y - \hat{y}|$$

with $n$ representing the quantity of the dataset, $y$ denoting the actual values in the dataset, and $\hat{y}$ signifying the predicted outcome. The comparison of the MAE scores for each model used in the research is presented in Table 3. From the table, it can be concluded that the model ARIMA (1, 1, 1) has the lowest MAE score, compared with Linear Regression and ARIMA with different parameters, namely ARIMA (2, 1, 1) and ARIMA (4, 1, 1). This indicates that the ARIMA (1, 1, 1) model provides the most accurate predictions among the models tested, as it yields the smallest mean absolute error. These results highlight the potential of time series analysis using ARIMA models for forecasting the occurrence of the earthquake.

**Table 3.** Comparison of the ARIMA Model Performance.

| Model | Mean Absolute Error (MAE) |
|---|---|
| ARIMA (1,1,1) | **0.110628** |
| ARIMA (2,1,2) | 0.110654 |
| ARIMA (4,1,1) | 0.111187 |
| Linear Regression | 1.895926 |

Furthermore, to determine the presence of autocorrelation in the prediction results using the best model, the Ljung-Box Test was conducted [15]. Additionally, the Jarque-Bera Test was also implemented on the generated predictions to determine whether the prediction results follow a normal distribution or not [16]. Further explanations regarding each of these tests are as follows:

I. Ljung-Box Normalization Test
The Ljung-Box test is a statistical test used to determine whether a set of data exhibits autocorrelation by comparing the variance of the data with the expected variance if the data were independent [15]. A positive result from the Ljung-Box test indicates the presence of autocorrelation in the data. In this study, the Ljung-Box statistical test was performed on the prediction of the *time_to_failure* feature, resulting in a p-value of 0.44. This implies that the acceptance of the null hypothesis indicates that the generated predictions lack autocorrelation and are distributed independently.

II. Jarque-Bera Normalization Test
The Jarque-Bera Normality Test is utilized to examine whether a sample data distribution follows a normal distribution or not by calculating two statistics, namely skewness and kurtosis, and comparing them to the standard normal distribution [16]. If the test results are significant, it can be concluded that the data does not follow a normal distribution. Based on the Jarque-Bera normality test conducted on the estimation of the *time_to_failure* value, a p-value close to 0.24 was obtained, indicating that the generated predictions follow a normal distribution.

In order to implement the model on the dataset used, several stages need to be carried out as shown in Figure 5. The process involves acquiring and preprocessing data, then using machine learning models like ARIMA and Linear Regression. ARIMA performs best based on evaluation metrics. Statistical tests like Ljung-Box and Jarque-Bera check for autocorrelation and normality.

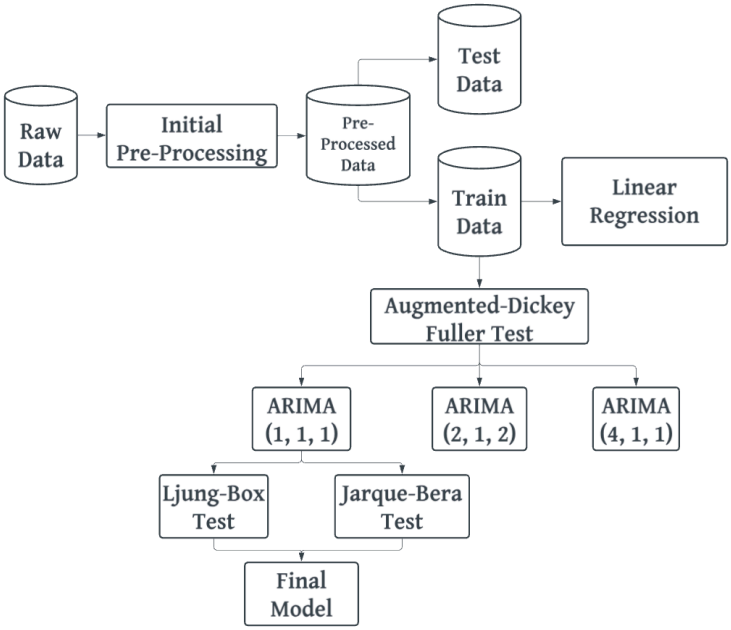


**Figure 5**. Flowchart Diagram of Research Methodology

## 4. CONCLUSION

Based on the conducted research, it can be concluded that the occurrence time of earthquakes can be predicted using the ARIMA (1, 1, 1) model with an MAE score of 0.11, which is superior to other machine learning models used in the study including ARIMA with different parameter and Linear Regression. Additionally, the Ljung-Box statistical test results indicated no autocorrelation in the prediction outcomes. Furthermore, the Jarque-Bera normality test results indicated that the distribution of the prediction outcomes follows a normal distribution. Moreover, this study aims to provide a valuable reference for institutions looking to utilize the ARIMA model for earthquake prediction. The goal is to use this model to plan and implement preventive measures to mitigate the adverse effects of earthquakes on both communities and the environment.

## 5. FUTURE WORKS

In this study, there are several suggestions for methods that can enhance research outcomes, both in terms of quality and performance metrics, these methods will include:

I. Creating new features from seismic signal information by utilizing the real and imaginary values of Fast Fourier Transform on the *acoustic_value* feature in each segment.
II. Implementing the Deep Neural Network (DNN) architecture as a predictive model, considering its characteristics in learning complex and large-sized feature representations with high-performance metrics.
III. Utilizing Vector Autoregressive (VAR) as a predictive model that can simultaneously utilize multiple features as predictors.

## REFERENCES


[1] N. Christiastuti, "Korban Jiwa GEMPA Turki-Suriah Bertambah jadi 41.000 orang," detiknews, https://news.detik.com/internasional/d-6569601/korban-jiwa-gempa-turki-suriah-bertambah-jadi-41-000-orang (accessed Feb. 15, 2023).

[2] K. Hirose, S. Labrosse, and J. Hernlund, "Composition and state of the Core," Annual Review of Earth and Planetary Sciences, vol. 41, no. 1, pp. 657–691, 2013. doi:10.1146/annurev-earth-050212-124007.

[3] A. Howard, B. Rouet-Leduc, and L. J. Pyrak-Nolte, "Lanl earthquake prediction," Kaggle, https://kaggle.com/competitions/LANL-Earthquake-Prediction (accessed Feb. 7, 2023).

[4] P. A. Johnson et al., "Laboratory earthquake forecasting: A machine learning competition," Proceedings of the National Academy of Sciences, vol. 118, no. 5, 2021. doi:10.1073/pnas.2011362118.

[5] E. Paparoditis and D. N. Politis, "The asymptotic size and power of the augmented dickey–fuller test for a unit root," Econometric Reviews, vol. 37, no. 9, pp. 955–973, 2016. doi:10.1080/00927872.2016.1178887.

[6] A. K.P, A. S. Oluwaseun, and V. G. Jemilohun, "Test for Stationarity on Inflation Rates in Nigeria using Augmented Dickey Fuller Test and Phillips-Persons Test," IOSR Journal of Mathematics, vol. 16, no. 3, pp. 11–14, 2020. doi:10.9790/5728-1603031114.

[7] X. Wang, Y. Kang, R. J. Hyndman, and F. Li, “Distributed Arima models for ultra-long Time Series,” International Journal of Forecasting, vol. 39, no. 3, pp. 1163–1184, 2023. doi:10.1016/j.ijforecast.2022.05.001.
[8] B. Dey, B. Roy, S. Datta, and T. S. Ustun, “Forecasting ethanol demand in India to meet future blending targets: A comparison of Arima and various regression models,” Energy Reports, vol. 9, pp. 411–418, 2023. doi:10.1016/j.egyr.2022.11.038.
[9] I. Unggara, A. Musdholifah, and A. K. Sari, “Optimization of Arima forecasting model using Firefly algorithm,” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 13, no. 2, p. 127, 2019. doi:10.22146/ijccs.37666.
[10] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “Linear regression,” Springer Texts in Statistics, pp. 69–134, 2023. doi:10.1007/978-3-031-38747-0_3.
[11] H. K. Prakosa and N. Rokhman, “Anomaly detection in hospital claims using K-means and linear regression,” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 4, p. 391, 2021. doi:10.22146/ijccs.68160.
[12] L. Wynants et al., “Prediction models for diagnosis and prognosis of covid-19: Systematic Review and Critical Appraisal,” BMJ, p. m1328, 2020. doi:10.1136/bmj.m1328.
[13] S. M. Robeson and C. J. Willmott, “Decomposition of the mean absolute error (mae) into systematic and unsystematic components,” PLOS ONE, vol. 18, no. 2, 2023. doi:10.1371/journal.pone.0279774.
[14] M. D. Fauzi, A. E. Putra, and W. Wahyono, “Estimation of average car speed using the haar-like feature and Correlation Tracker method,” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 14, no. 4, p. 353, 2020. doi:10.22146/ijccs.57262.
[15] H. Hassani and M. R. Yeganegi, “Selecting optimal lag order in ljung–box test,” Physica A: Statistical Mechanics and its Applications, vol. 541, p. 123700, 2020. doi:10.1016/j.physa.2019.123700.
[16] D. Abdellatif, K. El Moutaouakil, and K. Satori, “Clustering and Jarque-bera normality test to face recognition,” Procedia Computer Science, vol. 127, pp. 246–255, 2018. doi:10.1016/j.procs.2018.01.120.