Tilburg University

SPATIO-TEMPORAL ANALYSIS USING RNNS AND SEMIVARIOGRAM FOR FORECASTING GLOBAL MAJOR EARTHQUAKES

HARRY AVERKIADIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

917019

COMMITTEE

dr. Gonzalo Nápoles
dr. Chris Emmery

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8772

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Gonzalo Nápoles, for his unwavering support and invaluable guidance throughout this academic journey. His expertise, patience, and encouragement have been instrumental in the completion of this master's thesis.

I am also profoundly grateful to my family, friends, and beloved girlfriend. Their understanding and support have been a constant source of strength, especially during the most time-consuming moments of this thesis. Despite being 2,743 kilometers away, their love and encouragement have bridged the distance and made this achievement possible.

Thank you all for your unwavering belief in me.

SPATIO-TEMPORAL ANALYSIS USING RNNS AND SEMIVARIOGRAM FOR FORECASTING GLOBAL MAJOR EARTHQUAKES

HARRY AVERKIADIS

Abstract

This thesis addresses the problem of forecasting significant global earthquakes (5.5 Richter magnitude and above) using Recurrent Neural Networks (RNNs). The primary research question investigates the extent to which RNNs, enhanced with Semivariogram Analysis, can forecast major earthquakes using a global dataset from 1900 to 2023. Previous studies have not extensively explored this approach. Our method distinguishes itself by integrating Semivariogram Analysis to capture spatial correlations and employing attention mechanisms to enhance prediction accuracy. We utilized a comprehensive dataset from the United States Geological Survey (USGS), including earthquake characteristics such as magnitude, depth, time, and location. Our findings indicate that LSTM networks with Semivariogram Analysis had an error of 19.39 days, outperforming the other LSTM models in our research but only slightly outperforming the baseline ARIMA model. The ARIMA model had an error of 19.40 days and could also be a valid approach for forecasting earthquakes due to its simplicity and similar results. This similarity in performance also suggests that the added complexity of RNN models provides a limited improvement, further highlighting the challenging random patterns and unpredictability that major earthquakes exhibit. The feature importance analysis underscored the critical role of earthquake depth and spatial variability, suggesting that global spatial statistics in RNN models are important for better forecasting of earthquakes. This research aims to contribute towards the development of more robust earthquake forecasting models and to suggest future directions for enhancing prediction accuracy.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement*

The earthquake data and the data about the boundaries of countries and tectonic plates have been acquired from the United States Geological Survey (USGS) and Natural Earth (a public domain map dataset), respectively. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. However, the institution was informed about the use of this data for this thesis and potential research publications. All the figures belong to the author. The thesis code can be accessed through the GitHub repository following the link https://github.com/Harrys-Ave/ForecastingEarthquakes. In terms of writing, the author used assistance with the language of the paper. A generative language model (ChatGPT) and the software Grammarly were used to improve the author's original content, for paraphrasing, spell-checking, and grammar. No other typesetting tools or services were used.

The versions of the software and libraries used in our analysis are as follows:

- Python version: 3.11.4 | packaged by Anaconda, Inc. | (main, Jul 5 2023, 13:38:37) [MSC v.1916 64 bit (AMD64)]
- TensorFlow version: 2.16.1
- Keras version: 3.3.3
- Keras Tuner version: 1.4.7
- keras-self-attention version: 0.51.0
- TensorBoard version: 2.16.2
- numpy version: 1.26.4
- pandas version: 1.5.3
- matplotlib version: 3.7.1
- sklearn version: 1.3.0
- tensorflow version: 2.16.1
- keras version: 3.3.3
- shap version: 0.42.1

2 PROBLEM STATEMENT & RESEARCH GOAL

2.1 *Project Definition*

This study addresses earthquake prediction as a spatio-temporal, one-step-ahead time series forecasting task, using a global dataset of significant earthquakes (5.5 Richter Magnitudes and above) from 1900 to 2023. By employing Recurrent Neural Networks (RNNs), it aims to uncover spatio-temporal patterns and estimate days until the next major seismic event. The model's spatial analysis is enriched by geostatistical methods like Semivariogram Analysis introduced by Matheron (1963), alongside geographical variables such as "country" and "tectonic plate." Semivariogram Analysis explores the spatial relationship of data, measuring how similarity diminishes with distance.

In our research, a clear state-of-the-art model is not defined due to the absence of a benchmark for comparison (see Section 3). Additionally, forecasting the time of earthquakes remains largely unexplored. Nevertheless, since we approach the issue using time series analysis, we will utilize top-tier models and attention mechanisms, as highlighted in a comparative study of 84 research papers on AI techniques for earthquake prediction by Al Banna et al. (2020). Consequently, we will experiment with Long Short-Term Memory (LSTM) models introduced by Hochreiter and Schmidhuber (1997), which are proven in the literature (see Section 3) to be highly effective with time series data, and we will also explore attention mechanisms introduced by Bahdanau et al. (2014), on top of these models.

Predicting earthquakes has always been challenging due to their stochastic nature, and the task becomes even more difficult with significant and potentially unpredictable events. However, artificial neural networks (ANNs) show promise in this field by identifying patterns in complex time series data (see Section 3).

2.2 *Relevance*

Enhancing earthquake prediction can improve disaster preparedness, save lives, protect infrastructure, and reduce economic losses. Despite the general decline in fatalities from natural disasters, the death tolls from earthquakes remain high (Elliott, 2020), emphasizing the urgency of this research.

Accurate predictions are essential to protect lives and heritage, as underscored by devastating earthquakes in Haiti (2010) and Turkey (2023), and further examined in the paper by D'Amico (2015). Additionally, major earthquakes significantly impact societal and psychological well-being beyond physical damage. For instance, the 2012 Costa Rica earthquake caused widespread psychological distress, as highlighted by Fernandez et al. (2016). Thus, advancements in forecasting these events are vital not only for reducing physical and material harm but also for better-equipping communities to deal with the emotional aftermath.

The economic stakes of accurate earthquake forecasting are also high. D'Amico (2015) showed that earthquakes have historically impacted the economies of 154 out of 245 countries, with economic losses from seismic events showing no signs of abating. Significant implications also exist for the insurance industry refining its risk management strategies in the face of earthquake-related uncertainties. Additionally, as discussed by Kraemer et al. (2015), improved models can enhance other disaster forecasts, thereby mitigating wider risks.

Scientifically, this research introduces a unique approach by developing a universal RNN model that accounts for the global spatiotemporal dynamics of earthquakes. It integrates an attention-based LSTM network with an extensive dataset of significant global earthquakes, a novel application recommended by the literature. Additionally, this study aims to combine RNNs with Semivariogram Analysis to identify spatial correlations, representing also a novel approach. Furthermore, this work redefines global earthquake forecasting by treating it as a one-step-ahead time series forecasting challenge, an unexplored perspective in the existing literature that focuses more on classification, clustering, or sequence-to-sequence forecasting methods, as discussed by Berhich et al. (2023). Finally, our research, utilizing a general dataset encompassing global earthquakes, aims to establish a benchmark dataset for earthquake prediction comparisons, as the current literature indicates a lack of a clear benchmark.

The aforementioned innovative approaches promise to extend the frontiers of both data science and seismology, providing insights into earthquake predictability.

2.3 *Research Strategy*

Motivated by the literature, this research employs various LSTM networks, with an ARIMA model, introduced by Box and Jenkins (1970), serving as the baseline model. The ARIMA model was chosen due to its efficacy in capturing linear patterns within time series data. Using this baseline provides a robust benchmark against which the performance of more complex RNNs can be evaluated.

Given the LSTM network's proficiency with time series data, it will be used as one of the models and applied to a global dataset of earthquakes. The aforementioned model will be compared to an LSTM integrated with

Semivariogram Analysis to evaluate how well it enhances earthquake forecasting by capturing global spatial correlations within the context of RNNs. Additionally, we will examine an attention-based LSTM for its potential advantages.

To assess our approach and test model generalizability, we will train it with older global earthquake data and make predictions for the most recent years, serving as our unseen test set.

Our dataset maps each earthquake to a specific country and tectonic plate, also tracking the days until the next earthquake occurs in each country. These steps are also crucial for conducting Semivariogram Analysis.

Specifically, the Semivariogram Analysis will focus on earthquake metrics like magnitude and depth, using features like nugget, sill, and range to capture spatial structures. By incorporating these features, we aim to account for the influence of previous earthquakes' spatial distribution on the timing of future events. This approach aims to enhance the model's forecasting performance by leveraging spatio-temporal patterns in the data.

The foregoing prompts the framing of the central research question.

> *Main RQ: To what extent can Recurrent Neural Networks, enhanced with Semivariogram Analysis, forecast significant earthquakes within days using a global dataset?*

We will evaluate the ability of RNNs, enhanced with Semivariogram Analysis, to predict global earthquakes with the use of median absolute error and through the comparison with models without Semivariogram Analysis.

Additionally, motivated by the results of attention mechanisms in the literature and to gain deeper insights, two more sub-questions are explored:

> *RQ1: How does the performance of the ARIMA model and LSTM networks, including those enhanced with Semivariogram Analysis and attention mechanisms, compare in forecasting significant global earthquakes, in terms of Median Absolute Error?*

Starting with an ARIMA model as our baseline to evaluate the performance improvements offered by our more complex models, we will compare the performance of an LSTM network against an LSTM network enhanced with Semivariogram Analysis, an attention-based LSTM network, and an LSTM model that integrates both Semivariogram Analysis and attention mechanisms.

Next, we aim to assess which features are the most important for our best-performing model. This analysis will provide valuable insights into earthquake predictions and guide the selection of appropriate features for a benchmark dataset. Thus, we formulate our final sub-question:

RQ2: What is the relative importance of various features in the predictive performance of the best-performing model, as determined through the Permutation Importance method and SHAP (SHapley Additive exPlanations) method?

To assess the results from the two feature importance methods, we will use performance degradation curves.

These sub-questions aim to deepen the understanding of model performance and the factors influencing earthquake predictability.

3 RELATED WORK

Using the ARIMA model as a baseline, we will evaluate RNNs performance to determine the improvements offered by these more complex models. Traditional time series models like ARIMA have been crucial in forecasting methodologies, especially for linear and stationary data (Box & Jenkins, 1970). Despite the rise of sophisticated machine learning models, ARIMA remains an important benchmark due to its interpretability and effectiveness (Box & Jenkins, 1970). Additionally, Makridakis et al. (2018) also notes ARIMA's long-standing role in time series forecasting. Machine learning methods like neural networks have emerged as alternatives, but Makridakis et al. (2018) found statistical models often more accurate. Kobiela et al. (2022) also demonstrated that ARIMA could surpass LSTM models in forecasting the prices of NASDAQ-listed companies, reinforcing its value as a baseline for comparing our RNN models. Thus, our research needs comparative analyses between advanced machine learning techniques and traditional statistical methods to assess their predictive capabilities.

As previously mentioned, ANNs can identify patterns even in complex time series, and this capability has been leveraged for forecasting seismic activities using RNNs. Figure 1 also shows an increase in the number of publications in this field, highlighting the growing interest in applying RNNs to predict earthquakes.

The paper by Al Banna et al. (2020) highlights the limitations of earthquake timing forecasting due to erratic patterns, with forecasts potentially off by 20 days to 5 months. Their comparison of 84 research papers using AI methods for earthquake prediction suggests that top-tier models for time series analysis are required in combination with attention-based techniques, promising for enhancing prediction accuracy but not extensively explored. In their work, Al Banna et al. (2020) also note the lack of a benchmark dataset for model comparison, as studies use various parameters like seismic waves, precursory parameters, or animal behavior. At the same time, the limited data and recordings make deep learning approaches challenging to apply. Our study addresses these limitations by

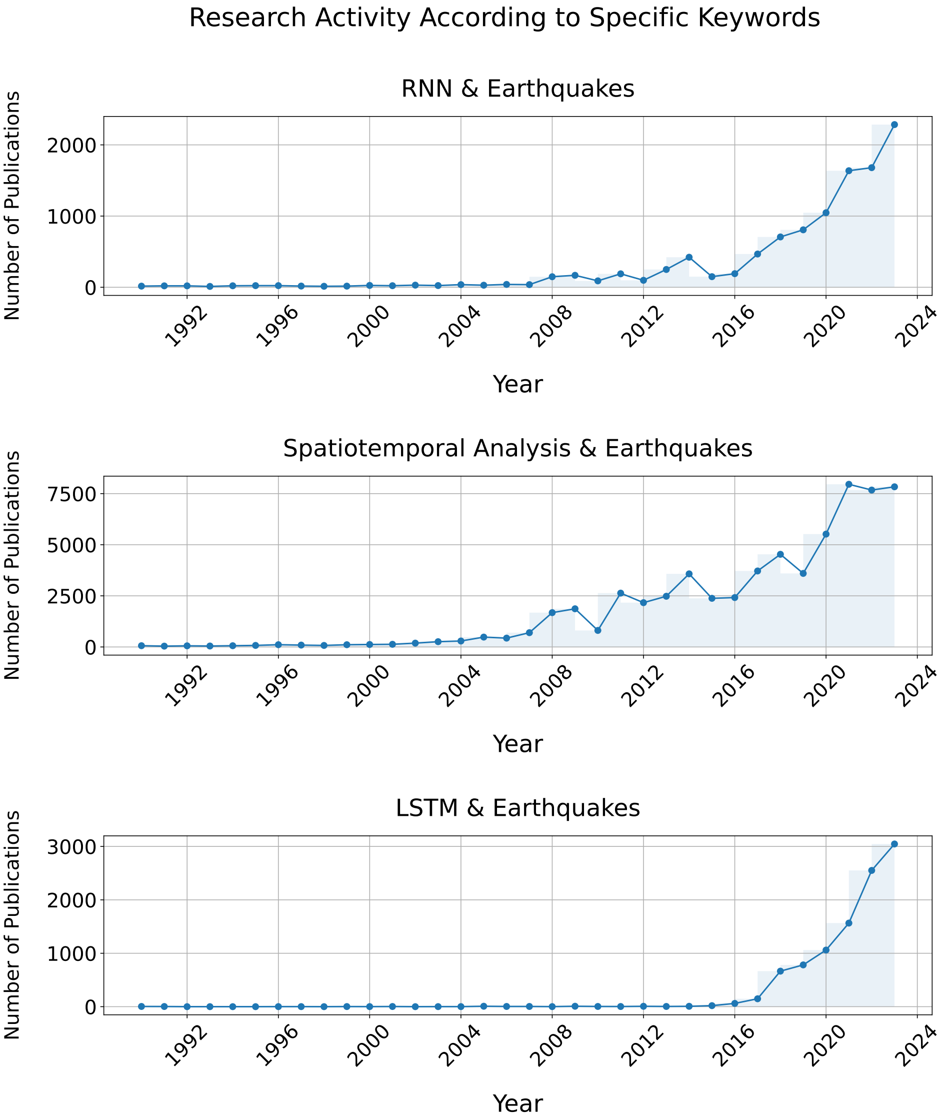


Figure 1: Research Activity

experimenting on a global dataset, providing a potential benchmark due to its generalizability and length, and containing key features like magnitude, depth, location, and time. In addition, as we are dealing with a time series analysis, we will compare RNNs for forecasting, specifically LSTM and attention-based LSTM networks, as recommended by the literature (Du Bois et al., 2023).

Additionally, the paper by Berhich et al. (2023) highlights the efficacy of attention-based LSTM networks for predicting significant earthquakes (magnitude > 5) in Japan, showing an improvement over traditional LSTM models. This finding aligns with our focus on high-magnitude earthquakes, underscoring the potential benefits of attention mechanisms. Additionally, the ability of LSTM networks to capture spatial correlations, as demonstrated by Wang et al. (2017) in the U.S. and China, supports our intent to apply these models to global spatio-temporal data.

As previously mentioned, our study addresses earthquake prediction as a one-step-ahead forecasting problem, a challenging approach that has not been fully explored in the literature due to the inherently unpredictable nature of earthquakes. The paper by Panakkat and Adeli (2009) significantly advanced this field by demonstrating the importance of RNNs in predicting earthquake location and timing. Their research focused on predicting the occurrence time in days and the specific locations of moderate and large earthquakes in southern California. The results indicated that for earthquakes with a threshold of 6.5 Richter magnitudes, the prediction error was 56 days and 17.5 miles. This research closely aligns with our methodology, as we will also predict significant earthquakes, calculate the error in days, and utilize RNNs, making it a valuable reference for comparing our results. To the best of our knowledge, there is no other existing literature that explores forecasting the timing of global earthquakes using RNNs.

Moreover, the role of spatial correlation in earthquake prediction is increasingly recognized. The paper by Zhang and Wang (2023) employed a Convolutional LSTM model for high-resolution, global earthquake prediction, focusing on binary classification. This method incorporates high-resolution global seismic maps, addresses spatial distortion through map rotation, and employs a specialized loss function to prioritize earthquake-prone areas. It demonstrates superior performance in predicting earthquakes with higher accuracy and resolution, showcasing its potential in understanding global seismic activity patterns. Additionally, the research of Puthran (2024) shows that a combined CNN-GRU model is highly effective in predicting earthquakes, leveraging spatial and temporal data analysis. Its predictive accuracy surpasses other models, emphasizing the value of merging spatial and temporal data for precise predictions. The importance of employing spatial correlation, as demonstrated by Zhang and Wang (2023) and Puthran (2024), underscores its relevance to our research. In contrast to these studies, our investigation will explore the integration of RNNs enhanced with Semivariogram Analysis to capture spatial correlations, a method not previously explored. We will also investigate the impact of the attention method on this approach, which has not yet been applied in conjunction with LSTM networks.

Regarding our goal of using semivariogram parameters to potentially enhance our model's ability to capture spatial correlation, to our knowledge, this approach has not been previously attempted in the literature in the manner we intend to leverage it. However, the semivariogram is a common geostatistical tool used to describe the spatial correlation between variables, which will be explained in detail later (see Section 4). Additionally, as seen in Figure 1, the approach of spatiotemporal analysis in earthquake research is widely utilized. This research aims to achieve this by combining RNNs and Semivariogram Analysis.

4 METHODOLOGY & EXPERIMENTAL SETUP

4.1 *Dataset Description*

This study uses a one-step-ahead forecasting approach, leveraging a United States Geological Survey (USGS) dataset of global earthquakes (5.5 Richter magnitude or higher) from 1900 to 2023. The dataset includes key variables such as magnitude, depth, latitude, longitude, and time. Additionally, we introduced the target variable "counter" to track days until the next earthquake in each country, making it the feature for predicting earthquake timing.

Each earthquake is mapped to a country and tectonic plate based on geographic coordinates. If the epicenter falls outside national boundaries, the nearest country is assigned. This mapping used GeoJSON files (admin zero level, national-level administrative boundaries) for the boundaries of countries and tectonic plates. Graphs in the Appendix depict the top five most seismically active countries and tectonic plates (see Figure 21 and Figure 20). The graphs show that a localized approach (by country) reveals fewer major earthquakes compared to a global approach. Additionally, using the GeoJSON file of tectonic plate boundaries (see Figure 2), we illustrate earthquake distribution across the world. The red lines delineating various tectonic plates demonstrate a fundamental seismological principle that earthquakes predominantly occur along tectonic plate boundaries which is also supported in the papers of McCann et al. (1978) and Stein and Sella (2002).

Furthermore, we calculated features such as Mean Magnitude, Magnitude Difference, and the number of earthquakes for each country. The "time" feature was split into "year," "month," and "day" for model comprehension, with additional features like "day of the week" and its sine and cosine transformations to capture cyclic patterns.

4.2 *Data Prepossessing*

Initial data pre-processing revealed fewer earthquakes in earlier records compared to recent years (see Figure 3). This reflects advancements in

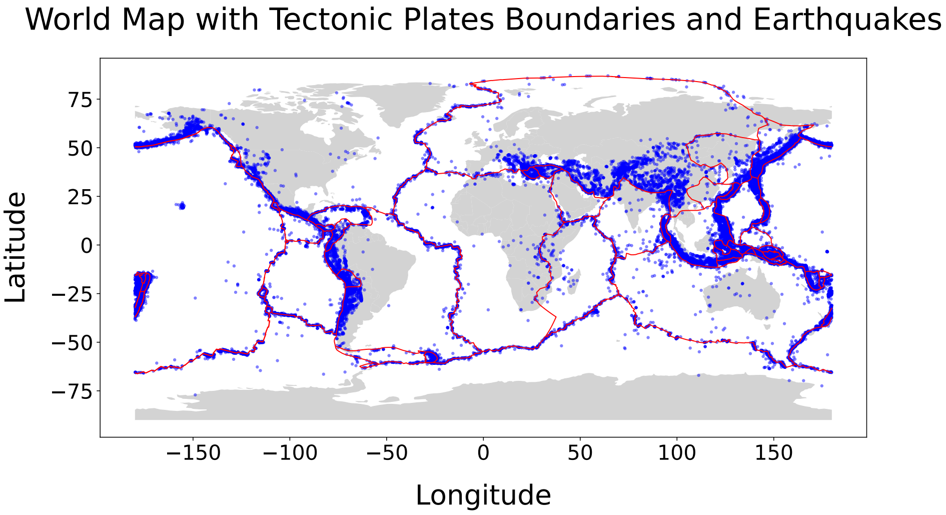


Figure 2: Earthquakes Occurrences in the World

seismic technology rather than an increasing trend in earthquake frequency. The introduction of the World Wide Standardized Seismographic Network (WWSSN) in 1961 improved earthquake detection capabilities, as noted by the USGS. To avoid the influence of historical technological limitations, data from 1961 and earlier were excluded, ensuring the model is trained on data from higher-quality instruments.

Next, outliers were addressed using the Interquartile Range (IQR) method, which is robust and less sensitive to extreme values (Huber & Ronchetti, 2011). Countries with infrequent earthquakes were removed. Using a median threshold, 70 countries with 11 or fewer earthquakes in 61 years were excluded, leaving 66 countries. This strict approach removed countries adding more noise than information, as the mean number of earthquakes per country was 705, with 75% having more than 1007 earthquakes.

Countries with very infrequent earthquakes were also removed. This involved calculating the IQR for the "counter" feature to establish outlier thresholds. The median "counter" for each country was computed, and countries with median "counter" values outside the established thresholds (three times the IQR below the first quartile and above the third quartile) were flagged as outliers, identifying countries with unusual earthquake activity patterns.

This refinement resulted in a dataset comprising 23 countries: the United States of America, Japan, France, Samoa, Greece, Tonga, New Zealand, Indonesia, the United Kingdom, Papua New Guinea, Chile, Colombia, India, Myanmar, the Solomon Islands, the Philippines, Panama, Brazil, Vanuatu, China, Iran, Peru, Russia, Mexico, Fiji, Portugal, Canada,

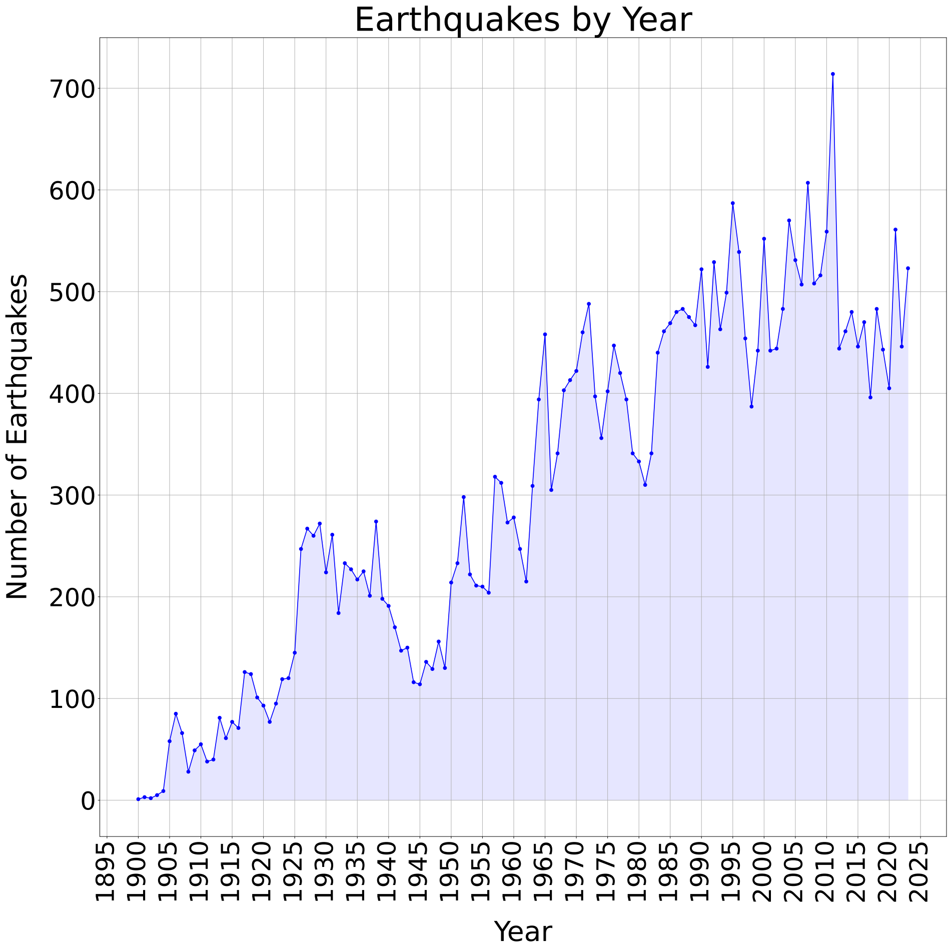


Figure 3: Earthquakes per Year

Antarctica, Afghanistan, Australia, Italy, Pakistan, Argentina, Taiwan, Ecuador, Turkey, Nicaragua, Norway, Guatemala, Nepal, Ethiopia, Kiribati, and Trinidad and Tobago.

The remaining outliers observed in the "counter" feature depicted in the boxplot (see Figure 4), were retained. Although these points appear distant from their country's median values, they represent genuine, rare scenarios that characterize each country's variability in earthquake occurrences. Including these data points could potentially enhance the model's predictive accuracy for such irregular events as they represent real but rare situations of each country separately (and not rare in the overall data).

To prepare our time series for use in an RNN, we ensured consistent data point intervals by inserting empty rows to fill gaps between recorded earthquakes, resulting in a uniformly spaced series where each row follows

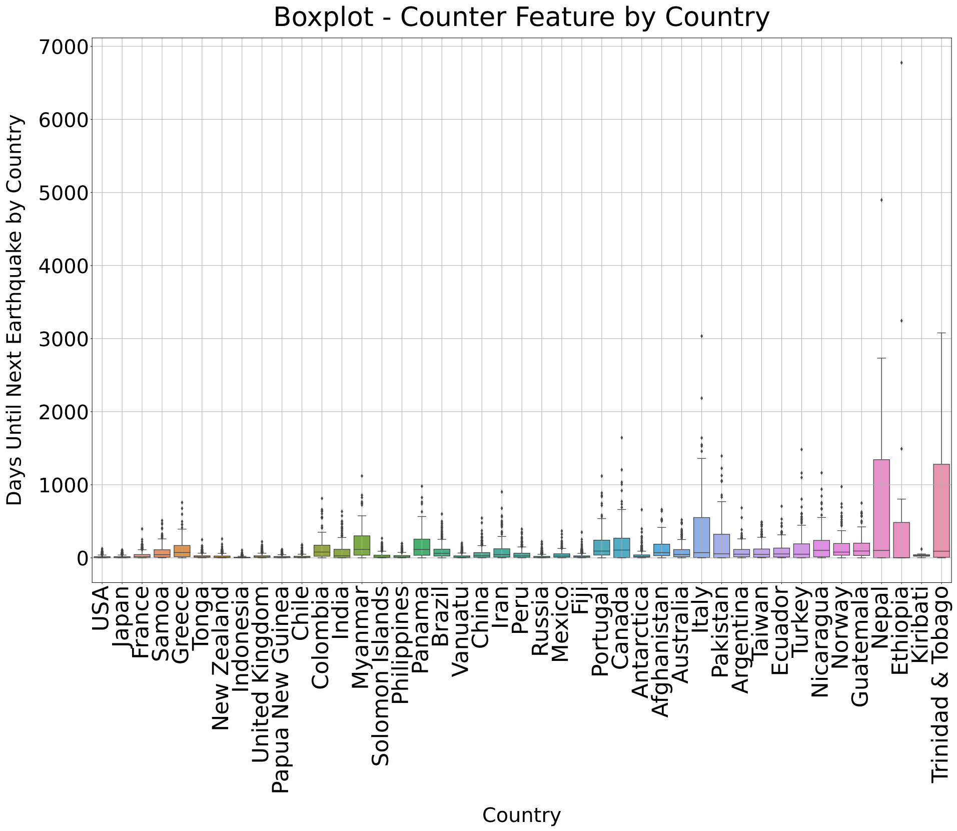


Figure 4: Counter of days for the next earthquake in each country

the previous day. This consistency is crucial as RNNs rely on stable temporal dynamics.

On certain days, multiple earthquakes were recorded either globally or within individual countries. To maintain data continuity and avoid loss of valuable information, these rows were not deleted. Instead, we applied a technique similar to data augmentation, akin to time shifting within a narrow window. Essentially, we redistributed earthquake records from days with multiple occurrences to adjacent days lacking events. Earthquake records from days with multiple occurrences were redistributed to adjacent days lacking events within a 10-day window, maintaining data integrity and reducing the missing data rate from approximately 7% to about 4%.

Major earthquake occurrences are scarce and unpredictable, leading to features that lack a normal distribution, as shown in Figure 5. Most features displayed right skewness and inconsistent variance, except for "Magnitude Difference." To address these issues, we applied log transformations on skewed features and upper clipping to depth, counter, and magnitude to mitigate the impact of outliers (Box & Cox, 1964). Additionally, all

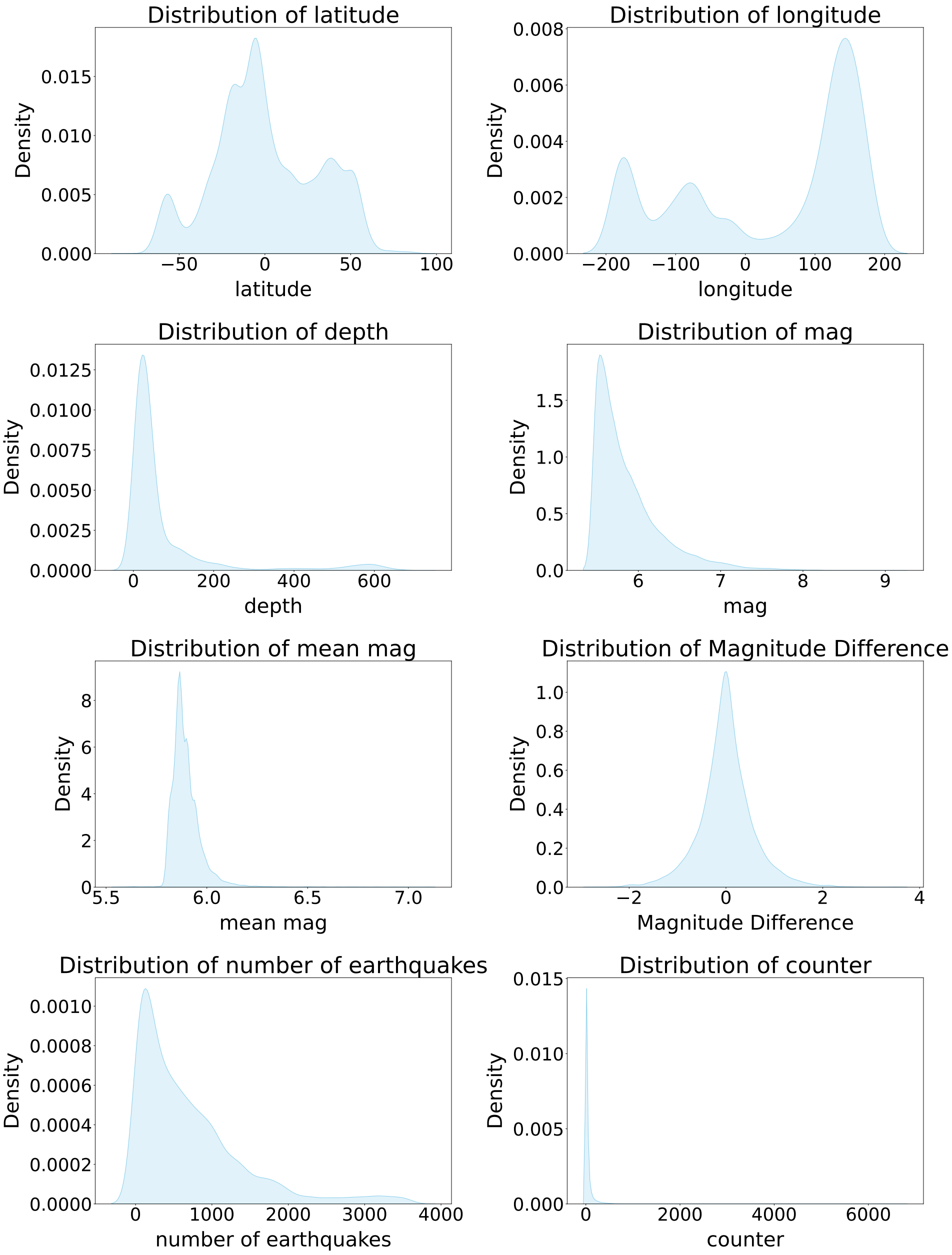
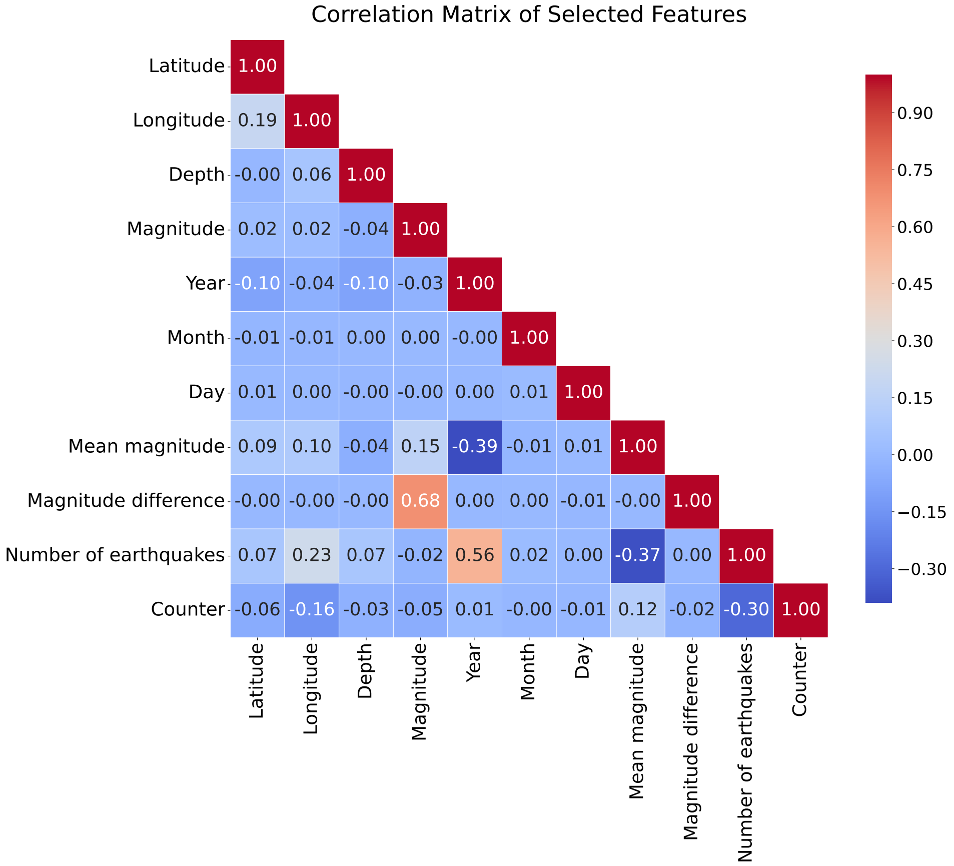


Figure 5: Distribution of Key Numerical Features about Earthquakes

features underwent standardization to ensure uniform contribution during the model training process. Without such scaling, features with larger variances or scales could disproportionately influence the optimization, leading to biased or suboptimal learning outcomes.

In our analysis, a correlation matrix (see Figure 6) was also utilized to understand feature relationships. A strong positive correlation of 0.68

between "Magnitude Difference" and "Magnitude" indicates that larger magnitudes coincide with greater differences in magnitude. Additionally, a positive correlation of 0.56 between "year" and "number of earthquakes" suggests an increasing trend in earthquake detection over the years, attributed to technological advancements rather than a rise in earthquake frequency. Weak correlations between "latitude" and "depth" with temporal features confirm the expected independence of geographic measurements from time variables. The negative correlation of -0.39 between "Mean Magnitude" and "Year" likely reflects the improved detection of smaller quakes in more recent records.


Correlation Matrix of Selected Features

	Latitude	Longitude	Depth	Magnitude	Year	Month	Day	Mean magnitude	Magnitude difference	Number of earthquakes	Counter
Latitude	1.00										
Longitude	0.19	1.00									
Depth	-0.00	0.06	1.00								
Magnitude	0.02	0.02	-0.04	1.00							
Year	-0.10	-0.04	-0.10	-0.03	1.00						
Month	-0.01	-0.01	0.00	0.00	-0.00	1.00					
Day	0.01	0.00	-0.00	-0.00	0.00	0.01	1.00				
Mean magnitude	0.09	0.10	-0.04	0.15	-0.39	-0.01	0.01	1.00			
Magnitude difference	-0.00	-0.00	-0.00	0.68	0.00	0.00	-0.01	-0.00	1.00		
Number of earthquakes	0.07	0.23	0.07	-0.02	0.56	0.02	0.00	-0.37	0.00	1.00	
Counter	-0.06	-0.16	-0.03	-0.05	0.01	-0.00	-0.01	0.12	-0.02	-0.30	1.00

Figure 6: Correlation Matrix

The matrix also reveals minimal correlations between the "counter" feature and geographical or direct temporal features, highlighting the challenges in predicting earthquake occurrences based solely on these factors. This supports the need for more complex models and machine learning techniques to capture nonlinear relationships, as linear correlations are weak.

Further insights from scatter plots (see Figure 7) reinforce that the relationships between the target feature and others are predominantly nonlinear, necessitating sophisticated modeling approaches, as employed in our research.

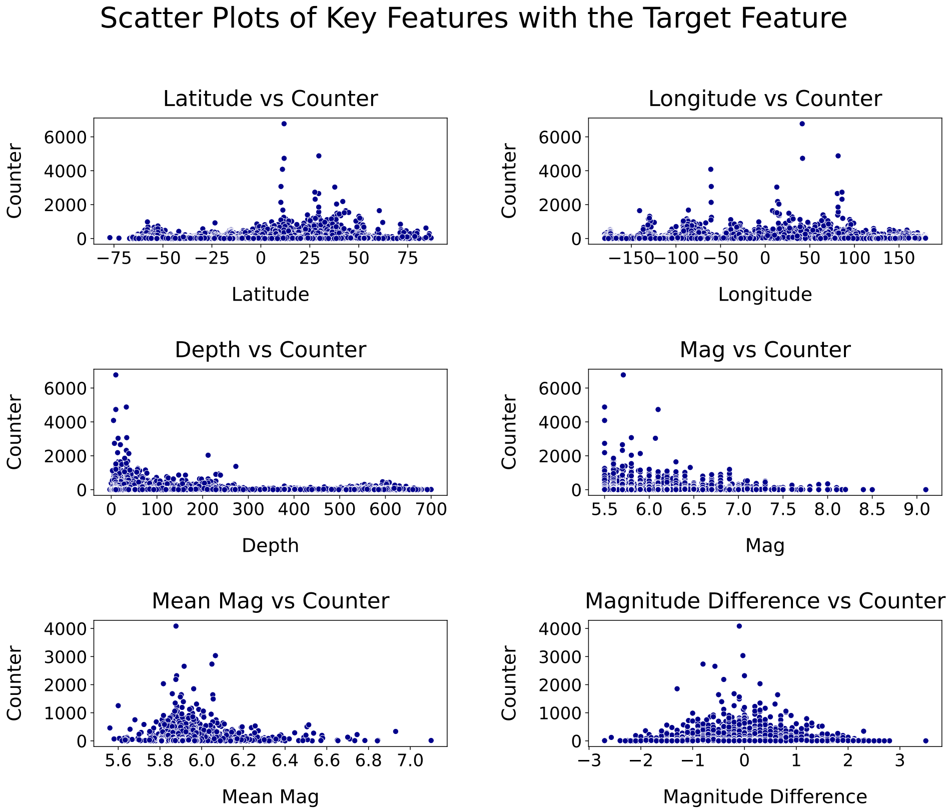


Figure 7: Scatter plots of target feature against other key features

The plots display a broad dispersion of points, indicating complex relationships that simple linear models cannot accurately represent. Notably, data clustering at lower "counter" values suggests that longer intervals between earthquakes are less common and influenced by different factors compared to more frequent occurrences.

Regarding potential outliers observed in the scatter plots, we chose not to remove them since they represent real-world data, including critical measures such as depth, latitude, and longitude, and were not detected by the IQR methods used earlier. These data points, although appearing as outliers, reflect valid seismic activities essential for comprehensive seismic analysis. Labeling these points as outliers based solely on their numerical value can be misleading, as they represent earthquakes that occurred in less frequent locations but are still valid for our study.

Additionally, we applied the Shapiro-Wilk test, introduced by Shapiro and Wilk (1965) for normality across the entire dataset with a significance level of 0.05 and we observed that we have a normal distribution. This test assesses whether the data distribution deviates from normality, producing a test statistic and a p-value.

For out-of-sample generalization and hyperparameter tuning, we structured our data into train, validation, and test sets. Given the time series nature of our data, we allocated the most recent 20% as the test set. From the remaining 80%, we segmented 25% as the validation set, leaving 60% for training. This temporal split ensures that the training, validation, and testing phases reflect the most current conditions and trends in seismic activity.

The training and validation sets had missing data rates of 8.7% and 1.6%, respectively, due to empty rows added for temporal continuity. We addressed this using kNN imputation, a popular method that is also explained in the paper of Wu et al. (2008). This method is versatile and efficient, as it does not assume any underlying distribution and considers the local context of data points, leading to more accurate imputations than simple mean or median methods.

For imputing geographical coordinates (latitude and longitude), we used the Haversine distance function (Robusto, 1957), crucial for accurately calculating distances between points on a sphere, such as Earth. The Haversine formula, outlined in Equation 1, ensures geographical accuracy by considering true spherical distances, unlike Euclidean distance, which assumes a flat surface and is unsuitable for latitude and longitude.

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (1)$$

where:

ϕ_1, ϕ_2 : are the latitudes of the two points in radians,

λ_1, λ_2 : are the longitudes of the two points in radians,

R : is the Earth's radius (approximately 6,371 km).

Following the imputation of latitude and longitude, these coordinates were then used to update the associated country and tectonic plate data for each point, aligning with the initial steps of our analysis.

Additionally, we implemented one-hot encoding for categorical features such as country and tectonic plate names, converting them into a binary matrix. This method is beneficial for neural network models, enhancing performance and interpretability by avoiding ordinality bias (Mikolov et al., 2013).

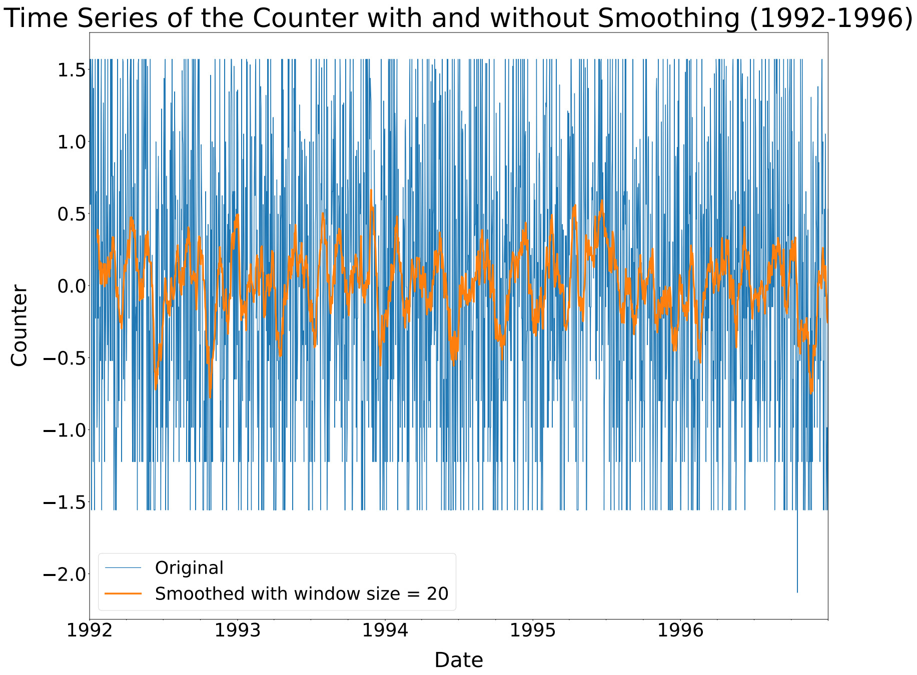


Figure 8: Time Series Visualization

To evaluate the stationarity of the target variable in our time series data, we applied the Dickey-Fuller test, as proposed by Dickey and Fuller (1979), a robust method for detecting unit roots and addressing non-stationarity. The test yielded a statistic substantially lower than the critical value at a 1% significance level, confirming that the time series is stationary.

For visualization, we concentrated on the earliest five years in our training dataset (see Figure 8). This decision was based on the expectation that more recent data would better correspond with the overall dataset due to progressive enhancements in seismic recording technology. We also visualized the most recent year available in our training dataset, 1996 (Figure 9), utilizing a smoothing technique with a rolling window size of 20 to mitigate the raw data's fluctuations and expose underlying trends. These visualizations underscore the significant volatility that our RNN models must accommodate. The smoothed data reveals seasonal cycles with periodic drops around December and mid-year (June and July). The absence of a clear directional trend supports the stationarity of the time series, as confirmed by the Dickey-Fuller test results. These observations are crucial for understanding the temporal dynamics our models need

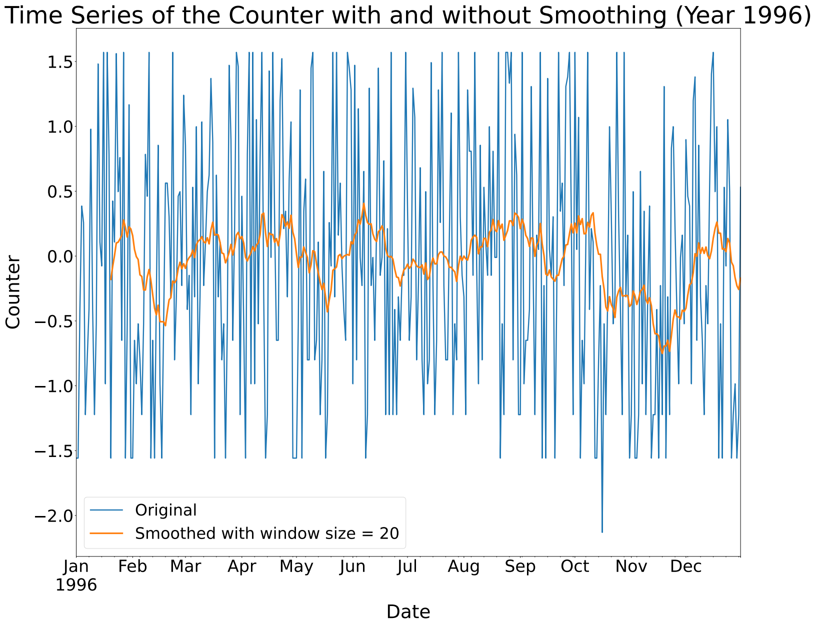


Figure 9: Time Series Visualization - Year 1996

to capture, indicating that while there is no long-term trend, seasonal fluctuations could be critical in forecasting future seismic activity.

We also explored seasonality through additive decomposition, assessing patterns over 7 and 91 days. No seasonality was found over a 7-day cycle, but a predictable pattern emerged over a 91-days period (approximately 3 months), as shown in Figures 22 and 23. This 91-day cyclical behavior suggests a potential natural periodicity, which could be integrated into predictive models.

To ensure that our target feature does not exhibit autocorrelation and partial autocorrelation, which could complicate model training and prediction accuracy, we conducted an analysis using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). This step was also crucial for determining the AR (autoregressive) and MA (moving average) components of the ARIMA model, which serves as our baseline. The results, depicted in Figure 10, show significant spikes only at Lags=0, with ACF and PACF values equal to 1, as expected since a time series is perfectly correlated with itself at Lags=0. For all other lags, the ACF and PACF values are near zero, indicating no substantial autocorrelation beyond the immediate past value.

Based on the analysis of these plots, the ARIMA model likely does not require high values for the AR and MA terms (Box & Jenkins, 1970).

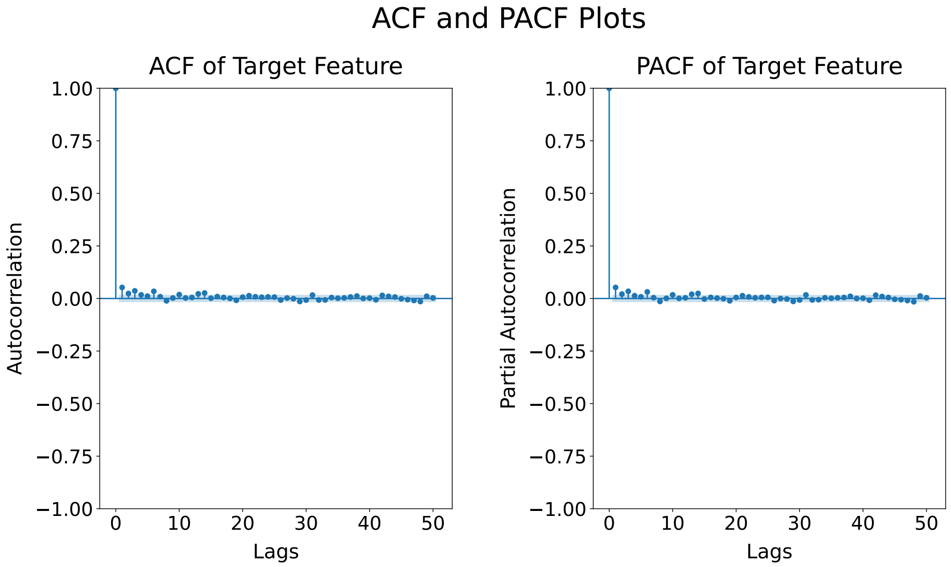


Figure 10: Autocorrelation and Partial Autocorrelation Function of Target Feature

Consequently, the values for the p and q parameters, representing the AR and MA terms respectively, were set to different combinations of zero and one, and a grid search was performed later to identify the optimal combination.

In our study, we also leveraged Semivariogram Analysis to explore and quantify the spatial correlations and dependencies within our earthquake data. Semivariograms measure the average dissimilarity between values of a spatial variable based on the distance separating them. Our selected variogram model was spherical, capturing how spatial correlation diminishes with increasing distance. The semivariance is computed as half the average squared difference between values at two locations over all pairs separated by a specific distance h (referenced in Equation 2).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(\mathbf{s}_i) - Z(\mathbf{s}_i + h)]^2 \tag{2}$$

where:

- $\gamma(h)$: Semivariance at distance h.
- h: Lag distance.
- $N(h)$: Number of pairs of observations separated by distance h.
- $z(x_i)$: Value of the variable at location x_i.
- $z(x_i + h)$: Value of the variable at location x_i shifted by distance h.

- $\sum_{i=1}^{N(h)}$: Summation over all pairs of observations separated by a distance h.

This analysis was executed across different levels—globally, by tectonic plate, and by country annually. Such a varied approach allows for dynamic semivariogram parameters, rather than static ones, enhancing the relevance and precision of our spatial analysis. This granularity ensures sufficient data is available to accurately calculate the semivariogram parameters.

The primary parameters derived from the Semivariogram Analysis were the nugget, sill, and range, applied specifically to the features of earthquake magnitude and depth. The nugget C_0 represents the semivariance at an infinitesimally small lag distance, accounting for measurement errors or micro-scale variations. Mathematically, it is defined as the semivariance when the distance h approaches zero:

$$C_0 = \lim_{h \to 0} \gamma(h)$$

The sill C is the value that the semivariogram approaches as the distance h increases, representing the total variance of the data:

$$C = \gamma(h \to \infty)$$

The range a is the distance at which the semivariogram reaches the sill, indicating the extent of spatial correlation. Beyond this distance, the spatial correlation is negligible:

$$\gamma(a) = C$$

We also included the interaction terms of these parameters with their corresponding feature for each earthquake in our dataset, allowing our model to learn more complex patterns and interactions.

4.3 *Experimental Setup of the Models*

Our baseline is the ARIMA model, a widely used statistical method for time series forecasting. ARIMA integrates Autoregression (AR), Integration (I), and Moving Average (MA), and is denoted by ARIMA(p, d, q), where:

- p represents the number of autoregressive terms (the order of the AR part).
- d denotes the number of times the raw observations are differenced (the degree of differencing).
- q indicates the number of moving average terms (the order of the MA part).

Further details on the mathematical formulation can be found in Appendix.

With our time series confirmed as stationary, *d* is set to zero. Moreover, ACF and PACF plots showed that both the autoregressive term (p) and the moving average term (q) do not require high values. The specific values tested during the hyperparameter tuning process are documented in Table 1 for clarity.

Table 1: Hyperparameter Tuning for ARIMA Model

Hyperparameter	**Tested Values**
p (Autoregressive term)	0, 1
d (Differencing term)	0
q (Moving average term)	0, 1

For our LSTM models, we formatted the time series data into 7-day sequence intervals, recognizing that RNNs excel in sequence-based processing. We experimented with different timestep configurations, but the weekly interval emerged as the most effective.

To enhance our LSTM models, we also experimented with an LSTM network integrated with a self-attention layer, proposed by Vaswani et al. (2017), which allows the model to focus on relevant parts of the input sequence, improving its prediction accuracy. Unlike conventional LSTMs that process information sequentially, self-attention allows the model to evaluate the significance of each token in the sequence relative to others, regardless of their position. This mechanism calculates attention weights for each token, indicating the importance of all other tokens in the sequence with respect to the current token. This is particularly useful for capturing long-range dependencies and intricate relationships within the data. We believe that these dependencies and relationships exist in our global earthquake data and the application of a mechanism that captures them will be beneficial. Mathematically, the self-attention mechanism is defined in Appendix.

The architecture of the simple LSTM model comprises a single LSTM layer and a Dense layer with a single unit output, as the objective is to predict the feature "counter." Through manual testing of various combinations of LSTM and Dense layers, this configuration was determined to be optimal. Similarly, for the LSTM model incorporating Semivariogram Analysis, the architecture consists of one LSTM layer, one Dense layer (with the number of neurons subject to hyperparameter tuning), and a final Dense layer with a single neuron.

For the two models integrating self-attention mechanisms, the architecture includes an LSTM layer, followed by a self-attention layer another

LSTM layer, and a final Dense layer with a single neuron. The selection of layers for each model was determined through experimentation with different complexities in terms of the number of layers. The activation function of the final Dense layer is linear, as the aim is to predict a continuous numerical value that can vary widely.

Lastly, in the model employing both LSTM and Semivariogram Analysis, the rectifier activation function (ReLU) was used for the Dense layer preceding the final Dense layer. This decision was made based on experimental results indicating that ReLU outperformed the hyperbolic tangent function (tanh).

Next, to counteract overfitting, which is a common challenge in training deep neural networks, we employed Ridge regularization (L2 regularization), proposed by Hoerl and Kennard (1970), across all models due to its ability to prevent overfitting by penalizing large weights. Specifically, this technique discourages learning overly complex models by penalizing the square values of the model weights. Additionally, for LSTM models that include an attention mechanism, we applied dropout regularization, introduced by Srivastava et al. (2014), after the self-attention layer. This method randomly omits neurons during training, compelling the network to learn more robust features.

An early stopping mechanism was implemented with a patience parameter set to five epochs. This approach terminates the training process if the model's performance does not improve for five consecutive epochs, thereby conserving computational resources and mitigating the risk of overfitting effectively and straightforwardly (Prechelt, 2002).

For the loss function, we chose the Huber loss, proposed by Huber (1992), which is less sensitive to outliers by combining attributes of both Mean Squared Error (MSE) and Mean Absolute Error (MAE). As an evaluation metric, we used MAE for its robustness to outliers and straightforward interpretability, essential for our analysis and facilitating comparisons in the field, a challenge discussed in the literature.

The optimization of our models was conducted using the Adam optimizer, as proposed by Kingma and Ba (2014), which is preferred for its adaptive learning rate capabilities, efficiency, and rapid convergence. Hyperparameter tuning for the LSTM models was executed through random search, as it is less computationally expensive than other methods like grid search while still providing valuable results (Bergstra & Bengio, 2012). The optimized variables included neuron count, L2 regularization strength, dropout rate, learning rate, adjustments in the Huber loss delta value, and batch size. Details of the parameters tested are documented in a corresponding table (see Table 2) for clarity and reproducibility.

Table 2: Hyperparameter Tuning Experimentation

Hyperparameter	Tested Values
Batch Size	16, 32, 64
Delta	5, 10, 20, 30
Dropout Rate	0.1, 0.2, 0.3, 0.4
L2 Regularization (l2_reg)	0.01, 0.001, 0.0001
Learning Rate	0.0001, 0.001, 0.01
Units	16, 32, 64

For the hyperparameter tuning process, we utilized Keras Tuner's RandomSearch. The specific self-attention mechanism employed was from Keras Self-Attention, and the TensorBoard was used for visualizing and monitoring the tuning process.

4.4 *Error Analysis & Feature Importance Analysis*

In our error analysis, we utilized Residual Distribution Analysis and Residuals vs Actual Values Scatter Plot. These techniques are preferred for their simplicity and clear, interpretable insights into model performance. Residual Distribution checks for normal distribution and constant variance, with deviations indicating issues like non-linearity or outliers. The Residuals vs Actual Values Scatter Plot examines the relationship between residuals and actual values to identify patterns indicating potential model weaknesses.

For feature importance, we implemented SHAP (SHapley Additive exPlanations), as proposed by Lundberg and Lee (2017), and permutation feature importance. These methods provide complementary insights into feature significance. SHAP values offer a unified measure of feature importance by calculating the contribution of each feature to the model's predictions based on cooperative game theory, making it robust for understanding feature interactions. Permutation Importance assesses impact by randomly shuffling each feature, directly measuring its effect on accuracy.

To assess feature importance results, we used performance degradation curves. These curves involve sequentially neutralizing the top features (top 9 in our case) identified by each method and measuring the cumulative change in prediction error. Neutralization is achieved by using the Totally Random Time Series (TRTS) approach, where each feature is replaced with random values, effectively breaking any potential structure in the data. We preferred this method because, unlike mean or mode replacement, TRTS does not preserve any statistical properties of the original data. This approach is particularly useful for understanding the extent to which the model relies on specific features' internal structure

Moreover, before deploying our models on the test set for final performance evaluation, we combined the training and validation sets. This integration allowed us to utilize the complete range of data for final adjustments, ensuring the models are well-calibrated across the entire dataset. This comprehensive approach enhances the robustness and predictive accuracy of the models, leveraging all available data for optimal performance before generalization testing.

Lastly, for clarity and simplicity, a flowchart is included to illustrate all the aforementioned steps of the Methodology and Experimental Setup in a more straightforward manner (see Figure 11). The methodology flowchart involves obtaining an earthquake dataset from the USGS, followed by data cleaning and exploratory data analysis. The data is split into training (60%), validation (20%), and test (20%) sets. Semivariogram Analysis is used to capture spatial dependencies and enhance RNNs performance. The ARIMA model serves as the baseline for comparison. All models undergo hyperparameter tuning and are evaluated using MAE. For the best-performing model, feature importance is assessed with SHAP values and Permutation Importance, using degradation curves to identify the most impactful method.

5 RESULTS

In this section, we report the outcomes of the data preprocessing steps, model performance comparisons, and feature importance analysis on our earthquake prediction models. We discuss the effects of feature transformations, the performance of the ARIMA model and the various LSTM-based models, hyperparameter tuning results, error analysis, and insights from the feature importance analysis.

During data preprocessing, we applied log transformations to right-skewed features and upper clipping to the depth, counter, and magnitude features, followed by standardization. In Figure 12, we can see that while the magnitude feature's skewness was slightly reduced, the depth and counter features saw notable distribution improvements. In addition, both the magnitude and counter features achieved more symmetric and uniform distributions, enhancing model performance.

Scatter plots (see Figure 13) of key features vs. the target feature illustrate also these adjustments' impact. Initially, features like latitude, longitude, and depth showed heavy clustering and skewness. Post-transformation, the data points are more uniformly distributed, reducing central clustering and better-representing variance. These transformations improve model

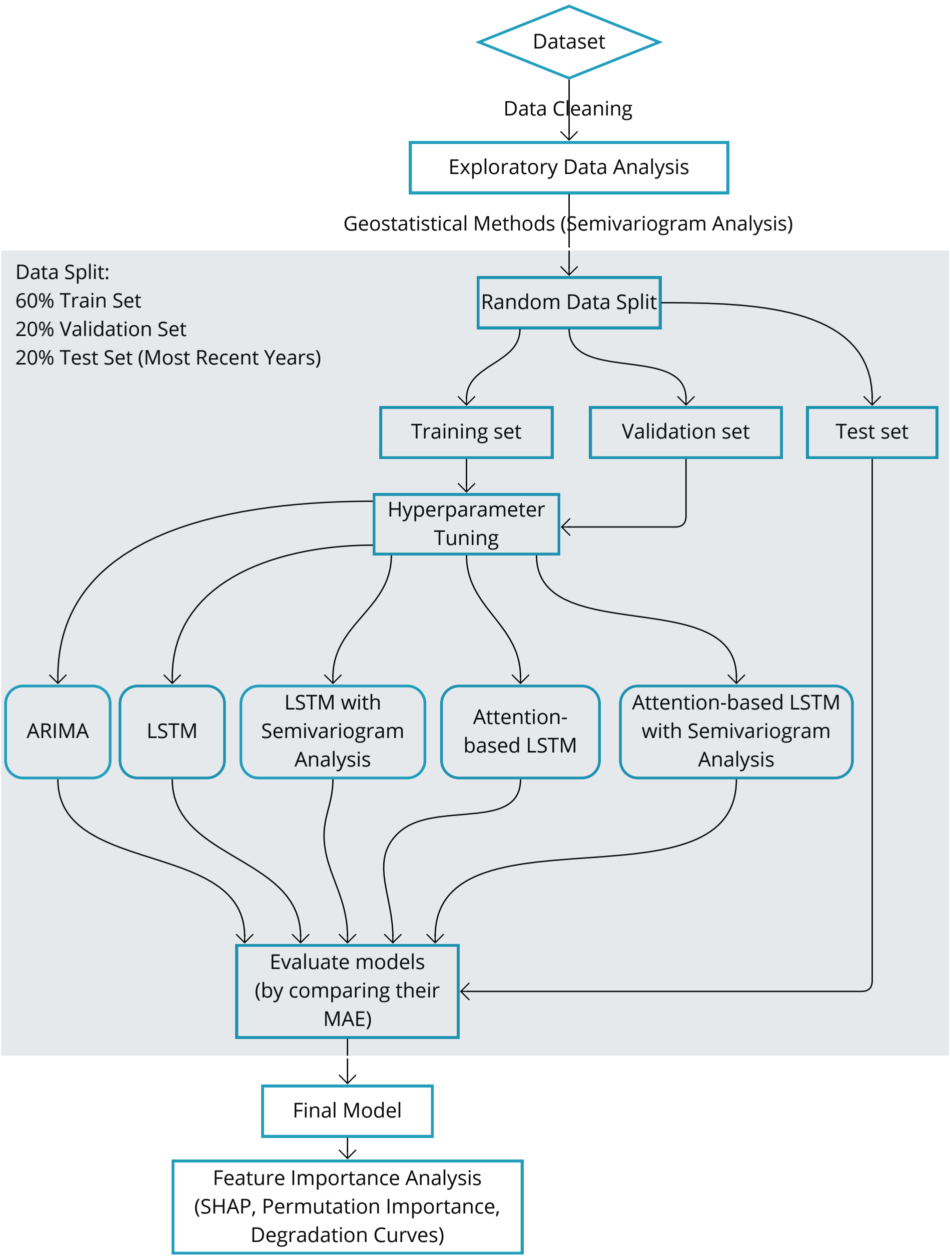

Dataset
Data Cleaning
Exploratory Data Analysis
Geostatistical Methods (Semivariogram Analysis)
Data Split:
60% Train Set
20% Validation Set
20% Test Set (Most Recent Years)
Random Data Split
Training set
Validation set
Test set
Hyperparameter Tuning
ARIMA
LSTM
LSTM with Semivariogram Analysis
Attention-based LSTM
Attention-based LSTM with Semivariogram Analysis
Evaluate models (by comparing their MAE)
Final Model
Feature Importance Analysis (SHAP, Permutation Importance, Degradation Curves)

Figure 11: Methodology Flowchart

performance by ensuring balanced features suitable for capturing underlying data patterns.

Our analysis showed that the best-performing model was the LSTM network with Semivariogram Analysis, outperforming the baseline model by a minimal error difference of 0.01 days (see Table 3). Additionally, incorporating attention mechanisms in the top-performing model did

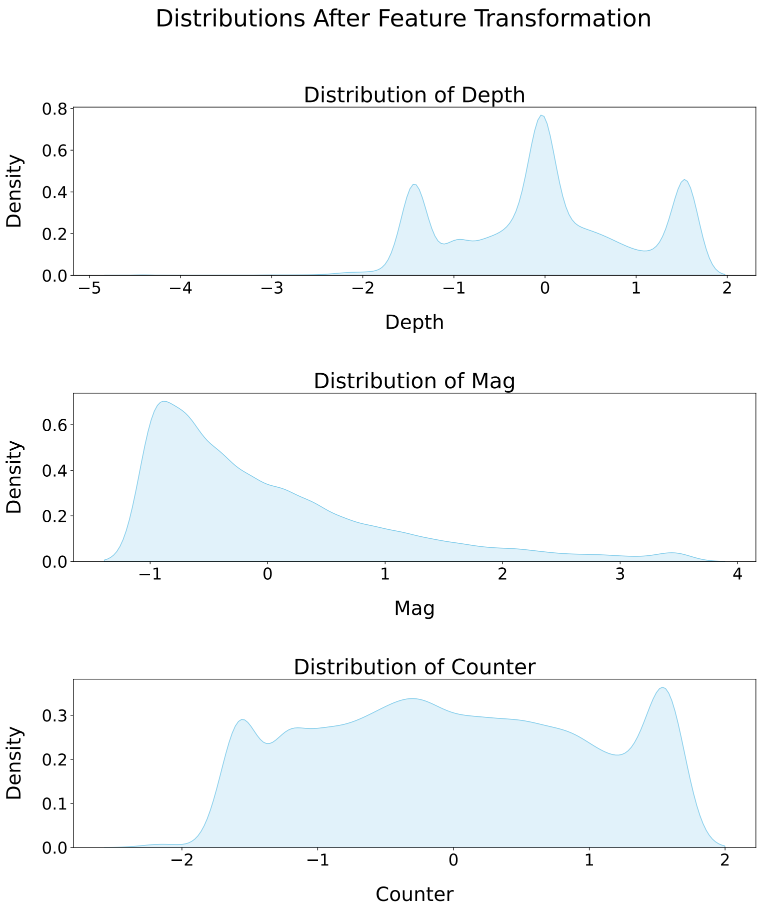


Figure 12: Distributions After Feature Transformation

Table 3: Models predicting the timing of the next earthquake and their performance in terms of MAE which represents the error in days

Models	Validation (MAE)	Test (MAE)
ARIMA (baseline)	19.53	19.40
LSTM	19.59	19.61
LSTM with Semivariogram Analysis	**19.57**	**19.39**
LSTM with Attention Mechanism	19.61	19.40
LSTM with Attention and Semivariogram	19.53	19.41

not enhance performance. That minimal MAE difference between the baseline ARIMA(0,0,0) and the best-performing RNN model suggests limited predictive advantage from the added complexity. This could be due to the nature of earthquake data or potential overfitting in the complex

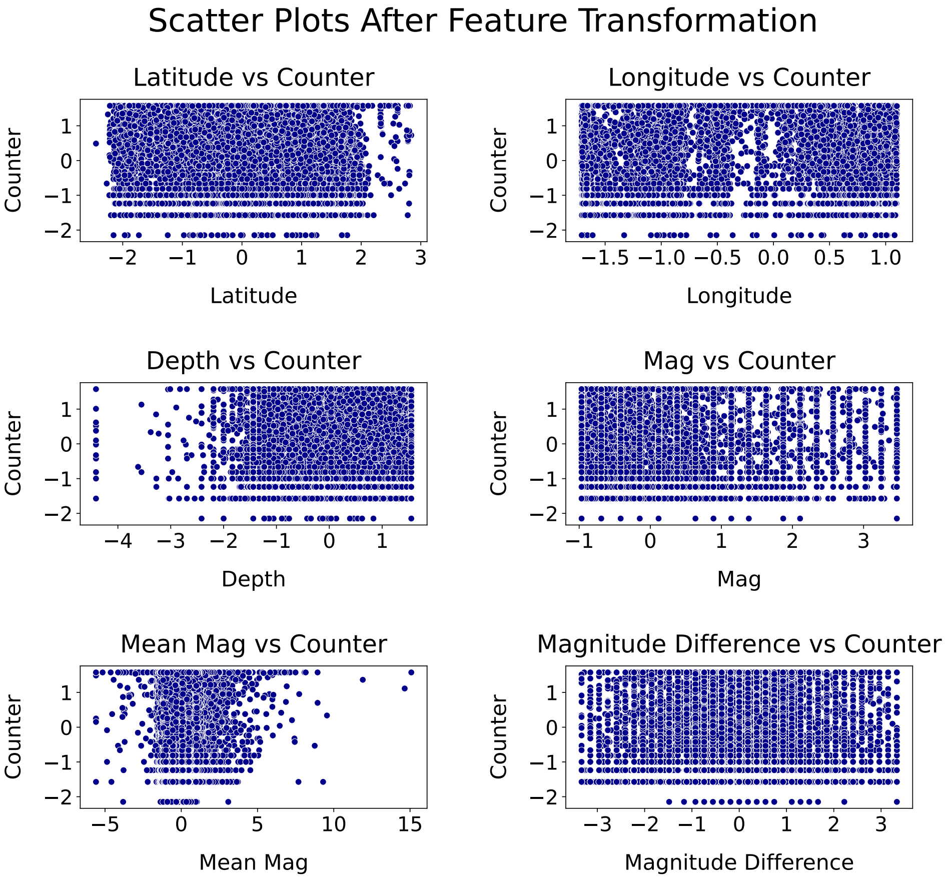


Figure 13: Scatter plots of target feature vs. key earthquake features, after feature transformation

model. Moreover, that minimal MAE difference might suggest that in some cases the simple and more interpretable ARIMA model can be preferred.

Our model comparison shows no clear evidence of overfitting. The models generalize well to unseen data, with validation errors generally larger than test errors, except for the simple LSTM model where the difference is minimal.

Hyperparameter tuning optimized the RNN models' performance, and a grid search identified ARIMA(0,0,0) as the best combination of parameters, representing the most basic model in time series analysis. The optimal parameters for each model are summarized in Table 4. From the table, we can also observe that LSTM models with Semivariogram Analysis benefit from more LSTM units (64 units), indicating a need for greater capacity to handle additional complexity. A low learning rate (0.0001) was consistently chosen, emphasizing the importance of a slower and more gradual optimization process. While the simple LSTM model

Table 4: Hyperparameter Tuning Results for Different Models

Model	Parameter	Best Value
ARIMA (baseline)	p (Autoregressive Term)	0
	d (Differencing Term)	0
	q (Moving Average Term)	0
LSTM	LSTM Units	16
	L2 Regularization	0.01
	Learning Rate	0.0001
	Huber Delta	20
	Batch Size	32
LSTM with Semivariogram Analysis	LSTM Units	64
	Dense Units	32
	L2 Regularization	0.0001
	Learning Rate	0.0001
	Huber Delta	20
	Batch Size	16
LSTM with Attention Mechanism	LSTM Units (first layer)	16
	L2 Regularization	0.0001
	Dropout Rate	0.2
	Learning Rate	0.0001
	Huber Delta	20
	Batch Size	16
LSTM with Attention and Semivariogram	LSTM Units (first layer)	64
	L2 Regularization	0.01
	Dropout Rate	0.1
	Learning Rate	0.0001
	Huber Delta	5
	Batch Size	16

prefers a larger batch size (32), more complex models favor a smaller batch size (16), likely due to the need for more frequent updates and improved generalization during training.

TensorBoard was used to visualize the tuning of the best-performing model. The scatter plot matrix (see Figure 14) illustrates the relationships between different hyperparameters and the MAE. Key insights indicate that certain hyperparameters significantly influence performance. Larger learning rates increase MAE, emphasizing the need for slow convergence. Smaller batch sizes improve performance, supporting frequent updates. An increased number of dense layer units correlates with lower MAE. Variability in the parameters delta (for the Huber loss function) and LSTM units highlights the importance of tuning these hyperparameters. Summarizing for the tuning of the best model, more LSTM units, lower L2 regularization,

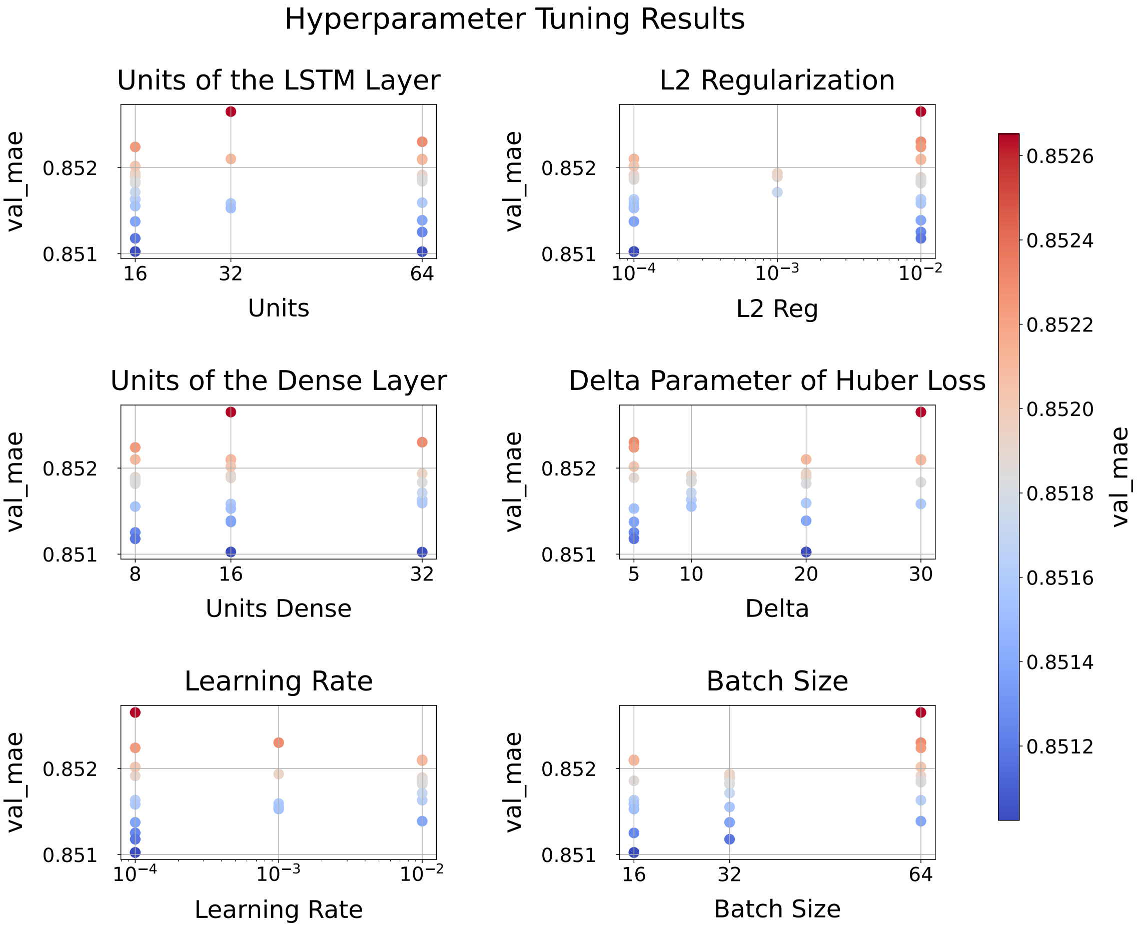


Figure 14: Scatter Plot Matrix - Parameter Tuning

more dense units, and a lower learning rate contribute to lower MAE and better performance.

Error analysis was conducted on both validation and test sets to assess generalization. Due to similar results, only test set graphs are presented here, with validation set graphs in the Appendix. The results were also similar across all our models, consistent with the close performance results observed for each model.

The errors representing the difference in days between predictions and actual earthquake occurrences, show a right-skewed residual distribution for the best model (see Figure 15). Additionally, it is important to note that the negative values of residuals observed in our error analysis represent instances where our model underpredicts the number of days until the next earthquake (i.e., the feature counter). Most residuals range between 0 and 20 days, with a peak around 80, indicating a tendency to underestimate actual values and suggest specific outliers or systematic errors.

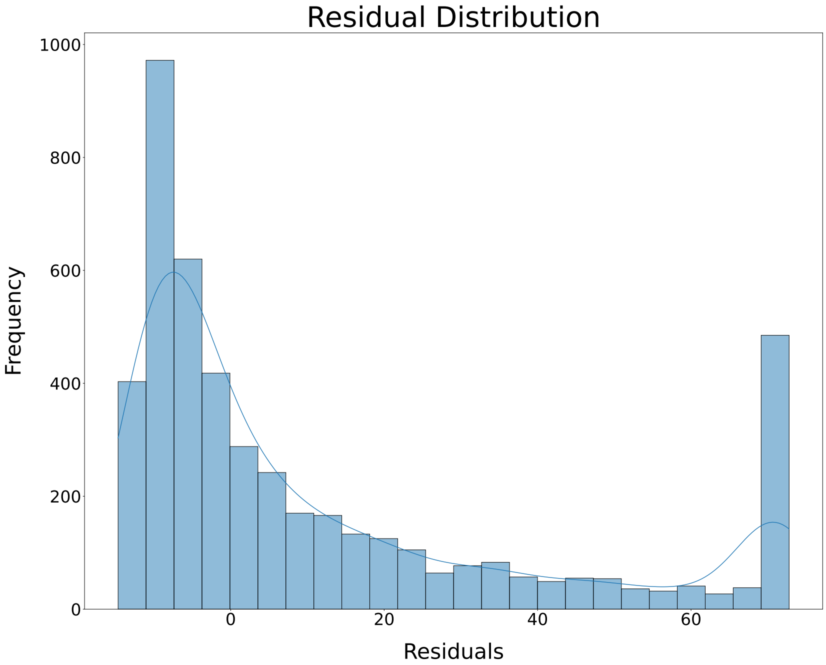


Figure 15: Residuals Distribution - Test Set

The concentration of residuals around zero indicates a reasonable level of accuracy in predicting seismic activity timing.

The scatter plot of residuals against actual values (for both validation and test sets) reveals a pattern, indicating heteroscedasticity (see Figure 16). Residuals increase with actual values, suggesting that error variance grows with the magnitude of actual values. This violates the assumption of constant variance (homoscedasticity), critical for many predictive models. This indicates that the model's predictions are less reliable for certain ranges of data, such as predicting earthquakes with long time intervals.

However, this linear increase in residuals might also indicate that earthquakes occurring after long periods are challenging to predict and happen randomly, without any spatial or temporal correlations, unlike those closer in time. Earlier scatter plots of the target feature also support this, showing that longer intervals between earthquakes are less common.

Continuing with our feature importance results, the Permutation Importance graph shows feature importance based on their absolute change in MAE (see Figure 17). The SHAP values graph orders features by cumulative importance, representing each feature's contribution to predictions (see Figure 18).

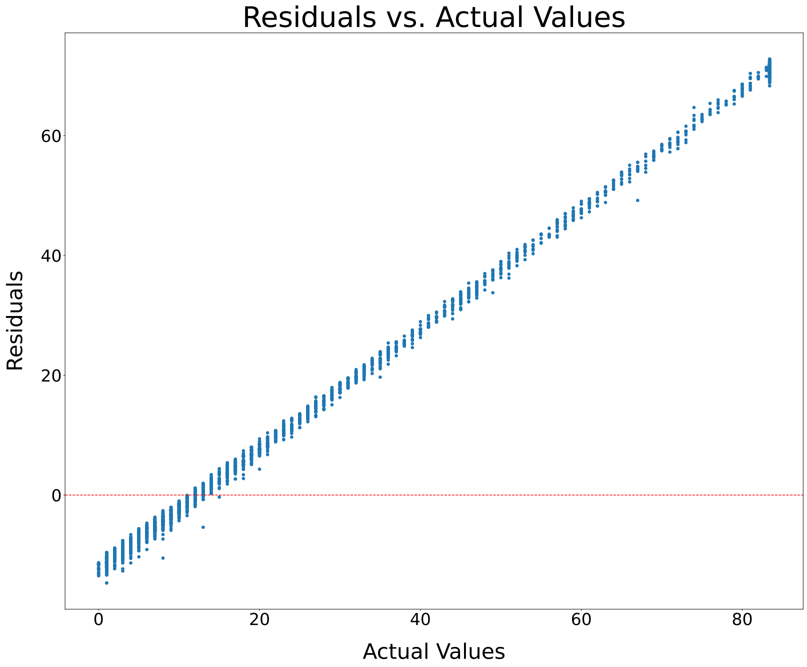


Figure 16: Residuals vs. Actual Values - Test Set

Both methods identified the same top two features: the interaction terms of each earthquake's depth with the global "sill" for depth and each earthquake's depth with the global "range" for depth that year. The global "sill" reflects the overall variability in earthquake characteristics (e.g., magnitude or depth) worldwide annually, quantifying how these characteristics vary across different locations. An interesting difference in the methods is that SHAP found categorical features of countries and tectonic plates important (3rd and 4th place respectively), while Permutation Importance did not include them.

The most important feature was the interaction term of each earthquake's depth with the global "sill" of depth, suggesting that considering earthquake depths with the overall variance in global depth data is crucial for predicting future seismic activity. This may highlight the importance of depth variability in understanding and forecasting earthquake patterns.

Analysis of global, tectonic plates, and country levels revealed that global features consistently rank higher in importance in both methods, emphasizing the significance of global seismic patterns and correlations. The depth and global scale features, whether interactions or standalone parameters are critical for predicting seismic events. These insights may

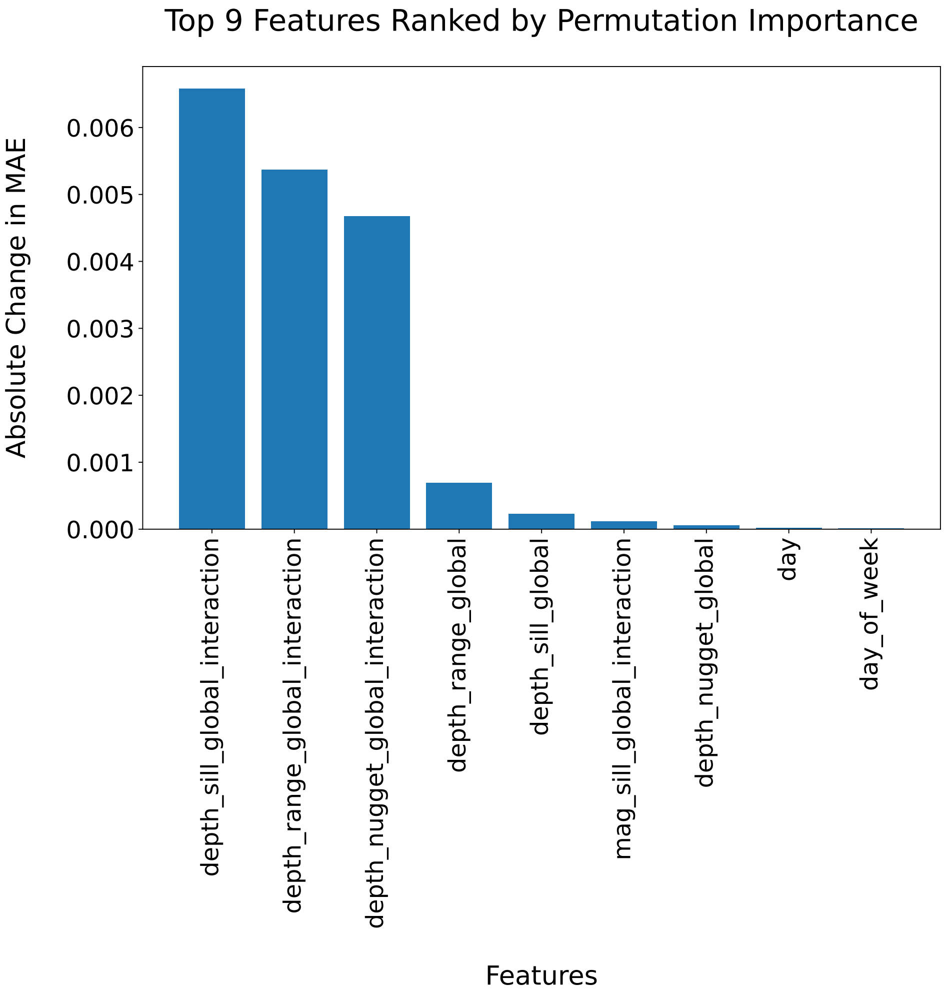


Figure 17: Top 9 Features According to Permutation Feature Importance

highlight the necessity of considering global spatial statistics and earthquake depth to enhance model accuracy and predictive capabilities.

Performance degradation curves indicate that the Permutation Importance method has the most significant impact on model performance (see Figure 19). This conclusion is based on comparing the areas under the error curves (the grey area is the smallest error area).

Initially, we observe an increase or, in some instances, a saturation in error when the most important features are neutralized. Surprisingly, neutralizing less important features (e.g., the 4th most important in Permutation Importance or the 7th in SHAP) led to a temporary drop in MAE, although still higher than our baseline. We initially expected a consistent increase in MAE with each neutralization, followed by saturation. This unexpected drop may be connected to the model's marginally better

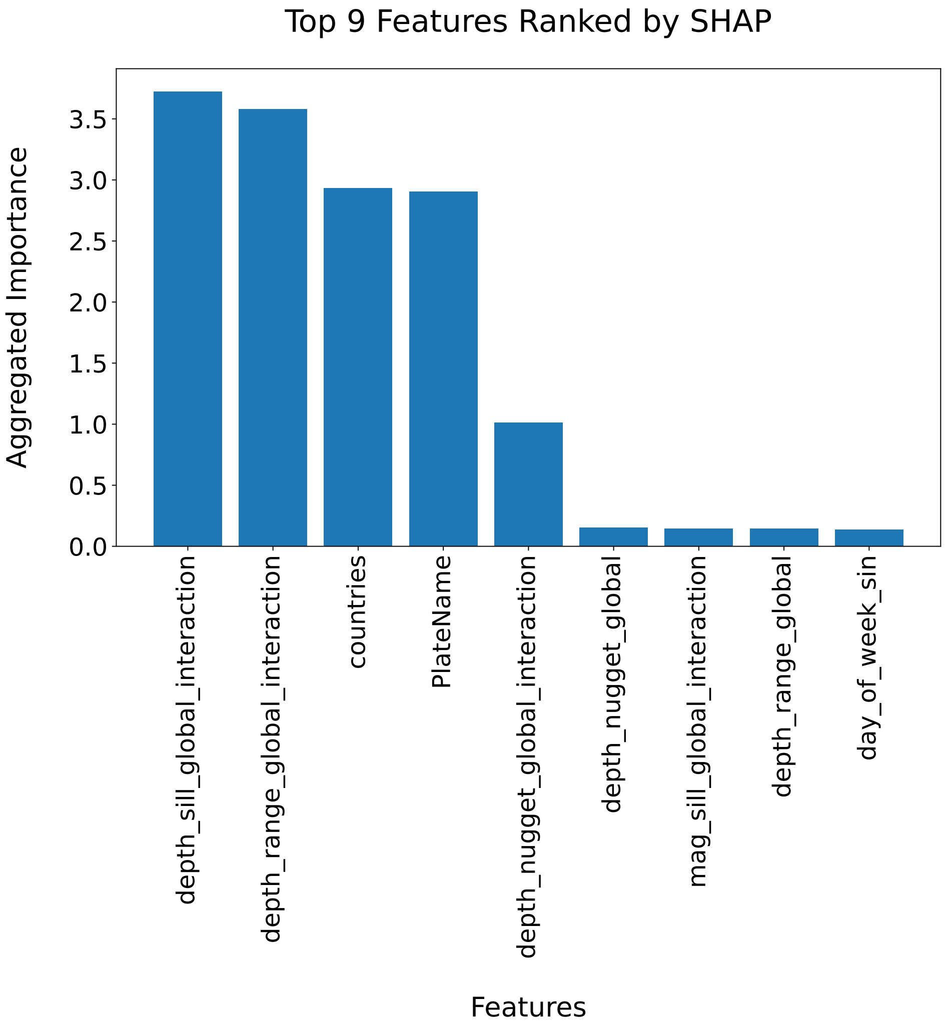


Figure 18: Top 9 features according to SHAP

performance than the ARIMA model, which indicates that the features may not considerably enhance pattern capture. Additionally, the drop occurs in features of low importance (e.g., the 7th feature in SHAP with negligible importance), which may introduce more noise than value to the predictions.

6 DISCUSSION

6.1 *Summary & discussion of the results*

This study evaluates the effectiveness of Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, in fore-

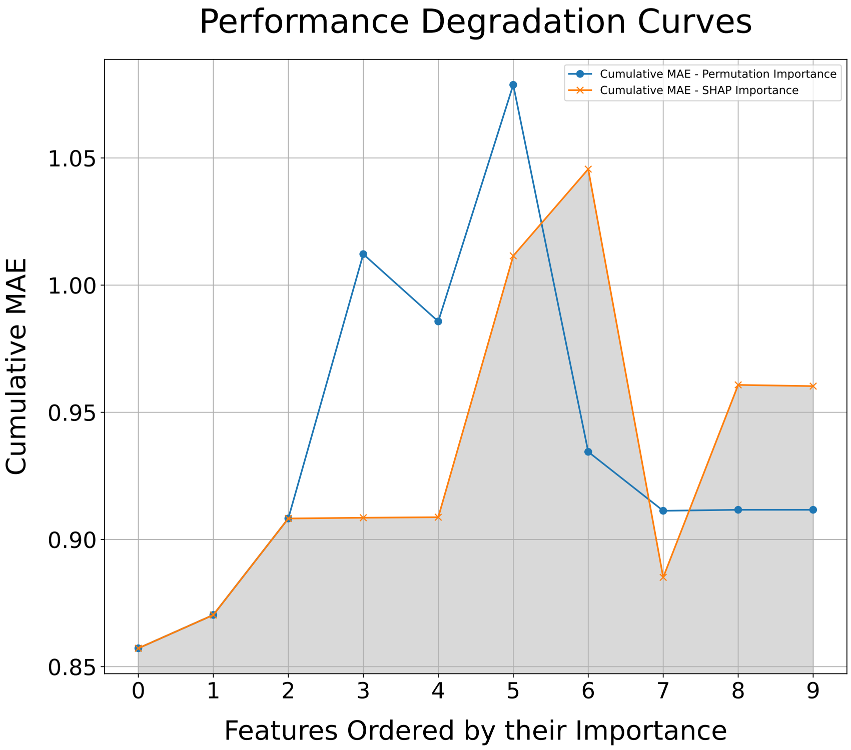


Figure 19: Performance Degradation Curves for the feature importance methods of Permutation Importance (blue curve) and SHAP (orange curve)

casting significant global earthquakes (magnitude 5.5 and above) using a dataset spanning from 1900 to 2023. It investigates how integrating Semivariogram Analysis enhances predictive accuracy, examines the impact of attention mechanisms, and analyzes feature importance. Additionally, an ARIMA model serves as a baseline for comparison due to its simplicity and established efficacy in capturing linear patterns in time series data.

The research revisits the main question and two sub-questions, linking them to the findings and relevant literature. The main research question is:

> *Main RQ: To what extent can recurrent neural networks, enhanced with Semivariogram Analysis, forecast significant earthquakes within days using a global dataset?*

The LSTM network with Semivariogram Analysis achieved the best performance, with a Median Absolute Error (MAE) of 19.39 days. This slight improvement over the baseline ARIMA model (MAE of 19.40 days) and the larger improvement over the simple LSTM model without enhancements (MAE of 19.61 days) indicates that Semivariogram Analysis in RNNs can help capture spatial correlations in global earthquake data, thereby slightly enhancing predictive accuracy of RNNs. While the improvements of the LSTM with Semivariogram Analysis are marginal over the LSTM without Semivariogram Analysis, specifically a 0.22-day MAE difference, they suggest that spatial dependencies contribute to earthquake forecasting and that integrating geostatistical methods can provide enhancements, albeit modest ones.

Furthermore, while Semivariogram Analysis is a straightforward geostatistical tool for capturing spatial correlations, the marginal improvements observed in our study may be due to its simplicity. More sophisticated methods for capturing spatial correlations could potentially yield better results (Cressie, 2015), especially in our case where we use global data with potentially very complex spatial dependencies.

Our LSTM with Semivariogram Analysis reduced the prediction error to 19.39 days, outperforming the 56-day error in Panakkat and Adeli (2009). Our baseline model, a simple ARIMA model with an error of 19.40 days, also outperformed the study's results by Panakkat and Adeli (2009). This might be attributed to the dataset, which in our case utilized a global dataset of significant earthquakes, while the study of Panakkat and Adeli (2009) focused on regional earthquake data. Therefore, our results may suggest that leveraging a global dataset for forecasting significant earthquakes can enhance the timing predictions, probably because we have more data than if we had focused on one region.

Continuing, the next sub-question is:

> *RQ1: How does the performance of the ARIMA model and LSTM networks, including those enhanced with Semivariogram Analysis and attention mechanisms, compare in forecasting significant global earthquakes, in terms of Median Absolute Error?*

The results from comparing all the models show that the LSTM network with Semivariogram Analysis is the best-performing model. The ARIMA model, serving as a baseline, had an MAE of 19.40 days. The simple LSTM model had a higher MAE of 19.61 days. The LSTM model with Semivariogram Analysis improved the MAE to 19.39 days, demonstrating the benefit of capturing spatial correlations.

Additionally, attention mechanisms improved the MAE of the simple LSTM from 19.61 days to 19.40 days, but combining attention with Semivariogram did not yield better results (MAE 19.41 days) over the LSTM with Semivariogram Analysis. It should be noted, though, that occasionally simpler models sometimes outperform complex ones, as noted by Bishop (1995).

Through our analysis and experimentation, it is suggested that incorporating Semivariogram in an LSTM network enhances the spatial understanding of earthquakes, yielding slightly better predictions than an LSTM without a Semivariogram. Additionally, consistent with the literature (Al Banna et al., 2020; Berhich et al., 2023; Du Bois et al., 2023), we observed that applying attention mechanisms in an LSTM network improves predictions, achieving an MAE of 19.40 days, while the simple LSTM without attention mechanism had an error of 19.61 days. Given the small difference of 0.01 days from the best-performing model, which used a Semivariogram, it can be argued that applying attention mechanisms yields similar results.

Interestingly, the ARIMA model and the LSTM with attention both achieved an MAE of 19.40 days, while the best model's MAE which was 19.39 days shows minimal improvement. This suggests that an ARIMA model might be preferred in some cases, as it provides similar results while being computationally less expensive and more interpretable than an RNN. The marginal MAE differences indicate that the added RNNs complexity does not enhance the predictive performance to the extent we expected. This is likely because the earthquake data lacks substantial patterns or dependencies that the RNN could exploit. This inherent unpredictability of significant earthquakes is also discussed by Sykes et al. (1999). It should also be noted that it is not uncommon for ARIMA models to be preferred over more complex models such as RNNs. Papers by Makridakis et al. (2018) and Kobiela et al. (2022) have shown similar findings.

Our final sub-question is:

> *RQ2: What is the relative importance of various features in the predictive performance of the best-performing model, as determined through the Permutation Importance method and SHAP (SHapley Additive exPlanations) method?*

Having the LSTM with Semivariogram Analysis as the best-performing model according to MAE, we performed a feature importance analysis. The analysis revealed that the interaction between earthquake depth and the global sill (total variance) was the most important feature, with an absolute change in MAE of 0.006 using the Permutation Importance method and an aggregated importance value of 3.5 using the SHAP method. Other important features include the interaction between depth and global range (spatial correlation distance) and depth and global nugget (micro-scale variations), all ranking in the top 5 for both methods.

SHAP method also highlighted the importance of categorical features like the country and tectonic plate (3rd and 4th most important features), each with an aggregated importance value of 3, suggesting that regional and geological characteristics also play an important role in the model's predictions. In addition, the performance degradation curves showed that the feature importance method with the most impact was the Permutation Importance method.

The dominance of features from Semivariogram Analysis highlights the importance of spatial statistics for RNNs in earthquake forecasting. In addition, the high importance of the semivariogram features of depth also showcases the importance of that feature for the spatial understanding of earthquakes and consequentially for enhancing the predictions of earthquake forecasting.

These findings emphasize that a comprehensive, multi-scale spatial analysis, including the interactions of semivariogram parameters with earthquake characteristics and the consideration of regional and geological factors, is essential for enhancing LSTM models in earthquake predictions. This insight is also demonstrated by Zhang and Wang (2023) and Puthran (2024) papers, which emphasize the importance of spatial analysis in earthquake forecasting.

While our study overall provides important insights into forecasting significant global earthquakes, it is essential to consider the potential disparate impacts of our models. One major concern is data representation; regions with fewer recorded earthquakes or less advanced recording technology may have less accurate predictions due to the model's reliance on historical data. This can lead to biased results favoring areas with richer and more frequent datasets. Additionally, socioeconomic disparities significantly affect the implementation and benefits of earthquake prediction models. Wealthier regions can better leverage predictions for disaster preparedness and response by investing in advanced infrastructure, early warning systems, and comprehensive emergency plans. In contrast, under-resourced areas may lack the funding, technology, and organizational frameworks to act effectively on these predictions, limiting their benefits. This disparity can worsen existing inequalities in disaster mitigation and recovery, leaving vulnerable populations in less affluent regions at higher risk of damage and loss from seismic events.

6.2 *Scientific and Societal Impact*

Inspired by the literature, we used LSTM networks due to their superior performance in previous studies (see Section 3). Our LSTM model outperformed vanilla RNNs that were used for earthquake forecasting in Panakkat and Adeli (2009). Incorporating attention mechanisms further enhanced these predictions, demonstrating their efficacy even on a global level, which had not been applied before.

Additionally, by integrating Semivariogram Analysis, we contributed to the scientific field by exploring a novel approach to capturing spatial correlations in earthquakes using RNNs. Our findings suggest that Semivariogram Analysis captures the spatial correlation of global earthquakes to some extent, indicating its potential as a valuable tool for improving earthquake forecasts and warranting further examination in other contexts, such as local earthquake predictions.

Another significant contribution is our use of a global dataset to predict major earthquakes. Previous research often focused on specific locations, limiting data and making predictions challenging. By leveraging global data, we obtained information about major earthquakes and accurate predictions about their timing. This approach could indicate a new direction for earthquake predictions, suggesting the use of global data either exclusively or to supplement local data.

Our feature importance analysis highlighted the importance of earthquake depth and its global variance, suggesting the need for further improving measurement and identification of earthquake depths.

Moreover, as discussed in the literature review, a key issue in this field is the lack of a benchmark dataset for comparing earthquake prediction results. By using a general global dataset and achieving promising results, we contribute to establishing a possible benchmark for future research.

The societal impact of improved earthquake predictions is significant. Knowing the timing of the next earthquake would be invaluable for disaster preparedness and risk mitigation. Governments, emergency services, and urban planners could use such a model to implement timely evacuation plans, reinforce critical infrastructure, and allocate resources effectively to minimize damage and loss of life. Additionally, insurance companies and businesses could leverage these predictions to better manage financial risks and prepare for potential economic impacts.

6.3 *Limitations and future directions*

A major limitation in studying significant earthquakes is data scarcity, which affected our research and may have limited the contributions of Semivariogram Analysis to our predictions. This issue is inherent to major earthquakes but can be mitigated by advancements in recording technology.

The inherent unpredictability and stochastic nature of earthquakes present another significant limitation. Despite using advanced models like RNNs with geostatistical tools, capturing the complex, random patterns of seismic activities remains challenging. This unpredictability might limit performance improvements with current modeling techniques.

While our study used semivariogram parameters to capture spatial correlations, other more sophisticated geostatistical methods and spatial-temporal analysis techniques might offer additional improvements. Future research could explore a broader range of geostatistical tools, such as kriging or spatial point processes, to enhance model performance.

Given the promising results of Semivariogram Analysis, further experimentation with geostatistical tools to enhance RNN models is recommended. Future research could apply these techniques to less significant, more frequent earthquakes, addressing data scarcity and improving earthquake forecasting robustness. Expanding the scope to include other seismic-related phenomena, such as aftershocks and foreshocks, could also provide a more comprehensive understanding of earthquake dynamics.

7 CONCLUSION

This study tackled the challenge of forecasting significant earthquakes using RNNs enhanced with Semivariogram Analysis. Incorporating semivariogram parameters into LSTM models improved their ability to capture spatial correlations in earthquake data, resulting in more accurate predictions. Our best-performing model, an LSTM with Semivariogram Analysis, achieved an error of 19.39 days, slightly better than the ARIMA model's 19.40 days. This shows that while the LSTM with Semivariogram Analysis provided a slight improvement, the simplicity and effectiveness of the ARIMA model could render it also a valid approach for forecasting earthquakes. Feature importance analysis in the LSTM network with Semivariogram Analysis revealed the critical role of earthquake depth and its global spatial variability, with interaction terms like depth with global sill and depth with global range being highly important. These findings underscore the importance of integrating geostatistical tools in RNN models for forecasting earthquakes. Our work suggests that future research should continue exploring this integration and further refining predictive models to mitigate the societal impacts of major seismic events.

REFERENCES

Al Banna, M. H., Taher, K. A., Kaiser, M. S., Mahmud, M., Rahman, M. S., Hosen, A. S., & Cho, G. H. (2020). Application of artificial intelligence in predicting earthquakes: State-of-the-art and future challenges. *IEEE Access*, *8*, 192880–192923.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Berhich, A., Belouadha, F.-Z., & Kabbaj, M. I. (2023). An attention-based lstm network for large earthquake prediction. *Soil Dynamics and Earthquake Engineering*, *165*, 107663.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211–243.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

D'Amico, S. (2015). *Earthquakes and their impact on society*. Springer.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Du Bois, N., Hollywood, L., Coyle, D., et al. (2023). State-of-the-art deep learning models are superior for time series forecasting and are applied optimally with iterative prediction methods.

Elliott, J. (2020). Earth observation for the assessment of earthquake hazard, risk and disaster management. *Surveys in geophysics*, *41*(6), 1323–1354.

Fernandez, M., Saenz, L., Carranza, M., Matamoros, C., Duran, O., Brenes, M., Alfaro, A., Solis, C., Macluf, S., Zarate, A., et al. (2016). Psychosocial support to people affected by the september 5, 2012, costa rica earthquake. *Earthquakes and Their Impact on Society*, 693–701.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution* (pp. 492–518). Springer.

Huber, P. J., & Ronchetti, E. M. (2011). *Robust statistics*. John Wiley & Sons.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kobiela, D., Krefta, D., Król, W., & Weichbroth, P. (2022). Arima vs lstm on nasdaq stock exchange data. *Procedia Computer Science*, *207*, 3836–3845.

Kraemer, M., Mrsnik, M., Petrov, A., & Glass, B. S. (2015). Storm alert: Natural disasters can damage sovereign creditworthiness. *S&P Global Ratings*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

McCann, W. R., Nishenko, S. P., Sykes, L. R., & Krause, J. (1978). Seismic gaps and plate tectonics: Seismic potential for major plate boundaries. *Proceedings of Conference VII—Methodology for Identifying Seismic Gaps and Soon T Break Gaps J. Evernden*, 441–584.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Panakkat, A., & Adeli, H. (2009). Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators. *Computer-Aided Civil and Infrastructure Engineering*, *24*(4), 280–292.

Prechelt, L. (2002). Early stopping-but when? In *Neural networks: Tricks of the trade* (pp. 55–69). Springer.

Puthran, R. (2024). Spatio-temporal analysis of hybrid cnn-gru model for prediction of earthquake for disaster management. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 270–281.

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly*, *64*(1), 38–40.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stein, S., & Sella, G. F. (2002). Plate boundary zones: Concept and approaches. *Plate Boundary Zones*, *30*, 1–26.

Sykes, L. R., Shaw, B. E., & Scholz, C. H. (1999). Rethinking earthquake prediction. *Pure and Applied Geophysics*, *155*, 207–232.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, Q., Guo, Y., Yu, L., & Li, P. (2017). Earthquake prediction based on spatio-temporal data mining: An lstm network approach. *IEEE Transactions on Emerging Topics in Computing*, *8*(1), 148–158.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1–37.

Zhang, Z., & Wang, Y. (2023). A spatiotemporal model for global earthquake prediction based on convolutional lstm. *IEEE Transactions on Geoscience and Remote Sensing*.

APPENDIX A

The ARIMA model for a stationary time series is mathematically defined as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{3}$$

This equation can be broken down into its components:

- **Autoregressive (AR) part:**
$$\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} \tag{4}$$
- **Moving Average (MA) part:**
$$-\theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{5}$$
- **Error term:**
$$e_t \tag{6}$$

Mathematical explanation of the self-attention mechanism:

Computing Attention Scores: The relevance of each pair of tokens is computed with an attention score, using the dot product of their representations:

$$\text{Score}(Q, K) = QK^T$$

where Q (queries) and K (keys) represent the token embeddings.

Attention Weights: A softmax function is applied to the scores to obtain attention weights, which are normalized probabilities:

$$\text{AttentionWeights}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

where d_k is the dimension of the key vectors.

Calculating Weighted Sum: The attention weights are then applied to the value vectors$V$ to compute a weighted average, producing the final output for each token:

$$\text{SelfAttention}(Q, K, V) = \text{AttentionWeights}(Q, K)V$$

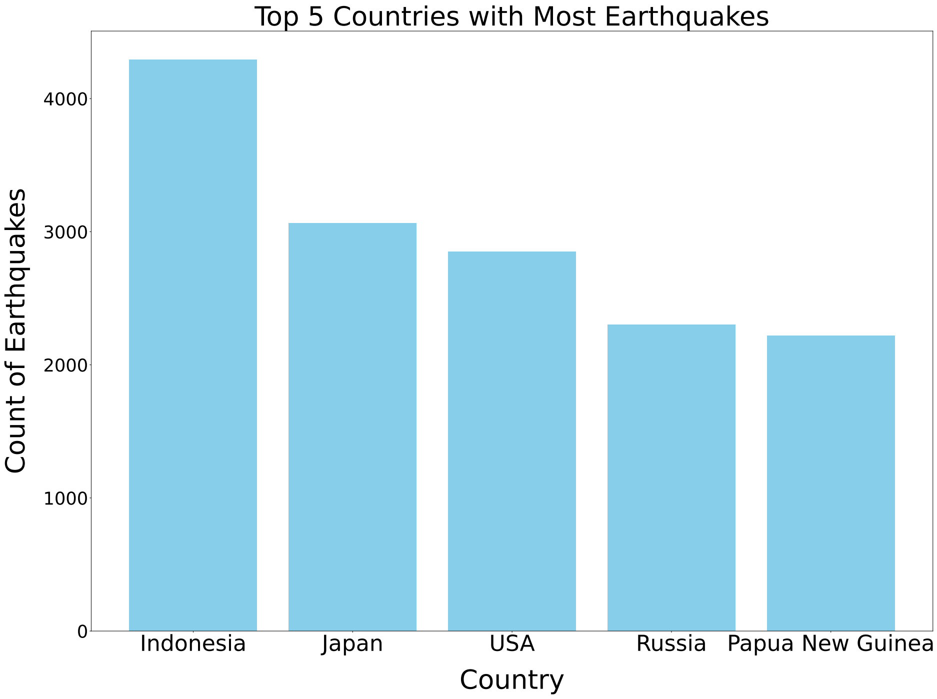


Figure 20: Top 5 Seismic Countries

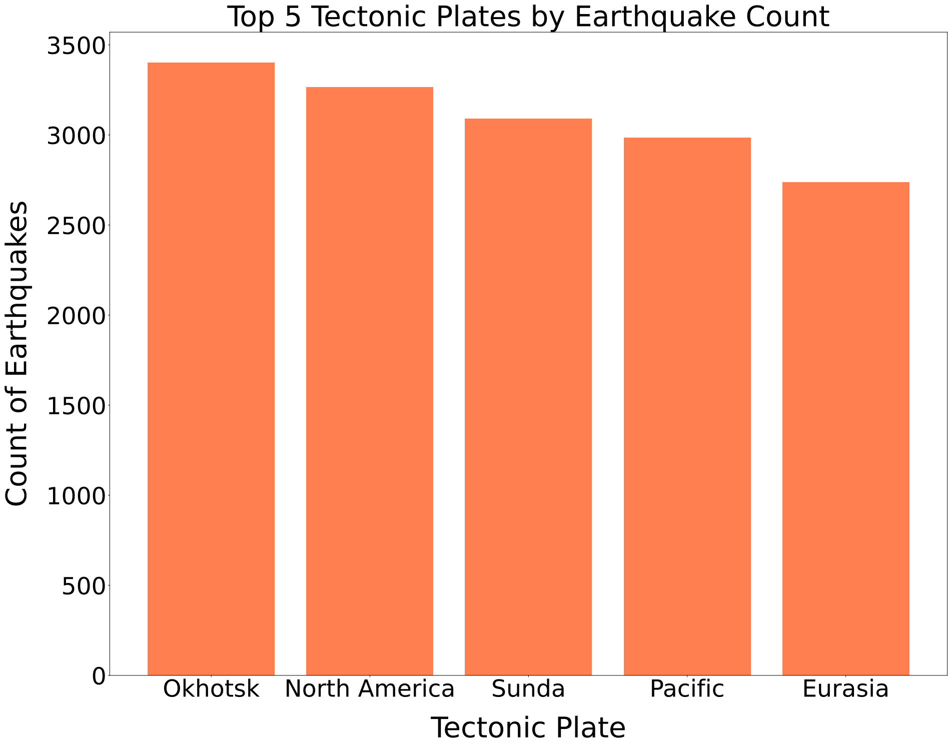


Figure 21: Top 5 Seismic Plates

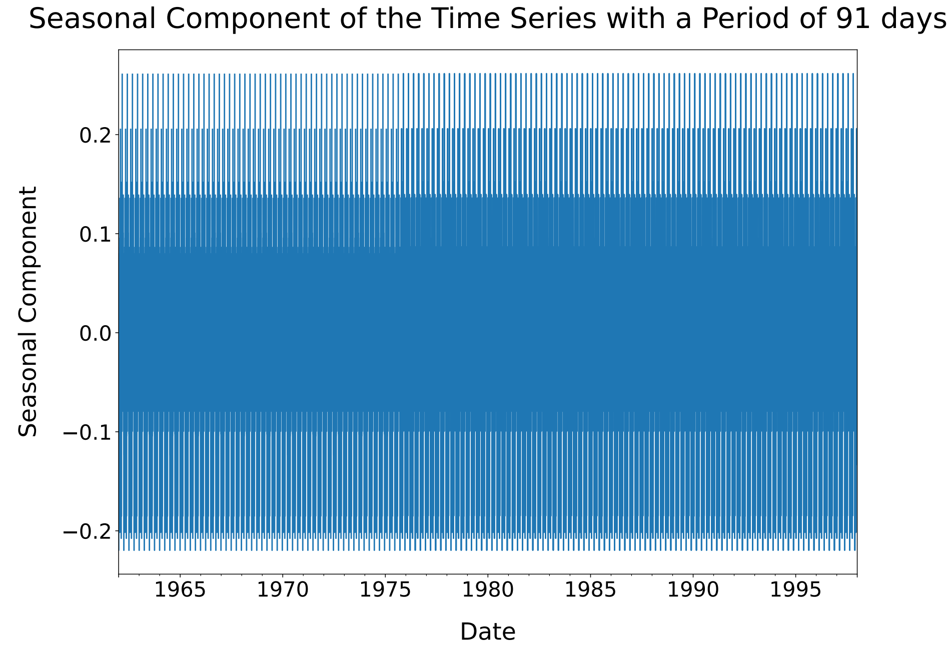


Figure 22: Seasonal Component of the Time Series with a Period of 91 days

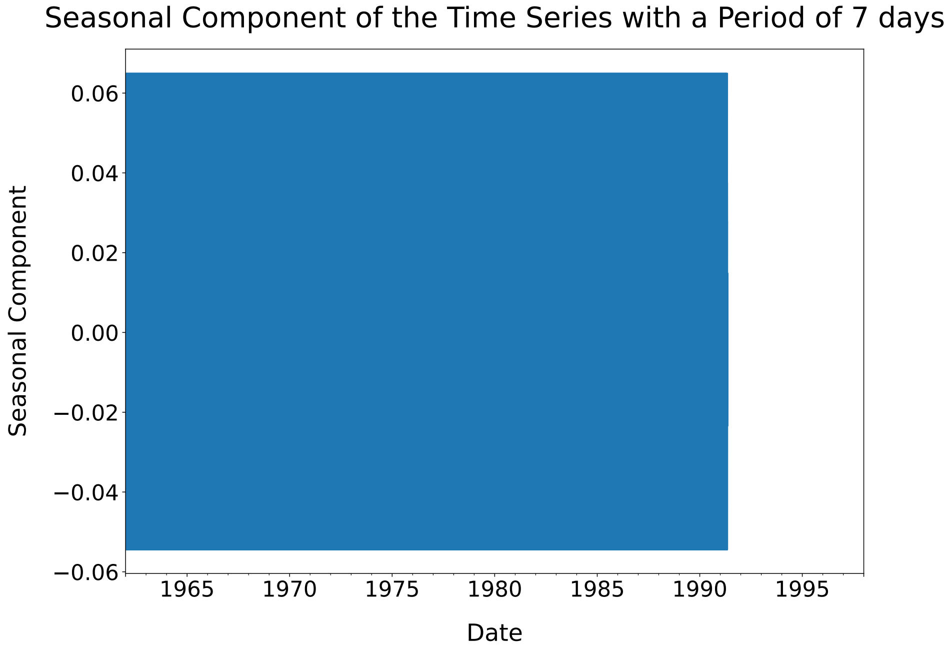


Figure 23: Seasonal Component of the Time Series with a Period of 7 days

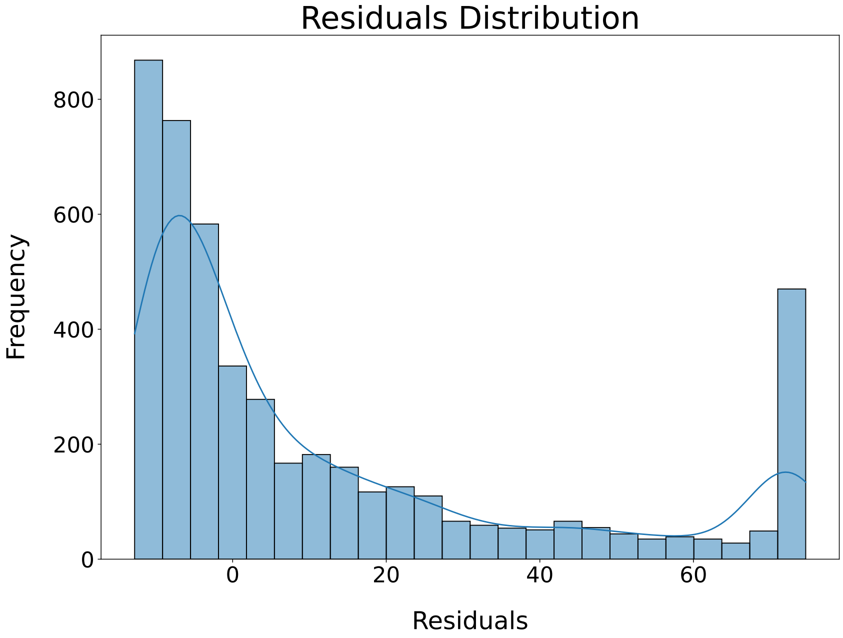


Figure 24: Residual Distribution - Validation Set)

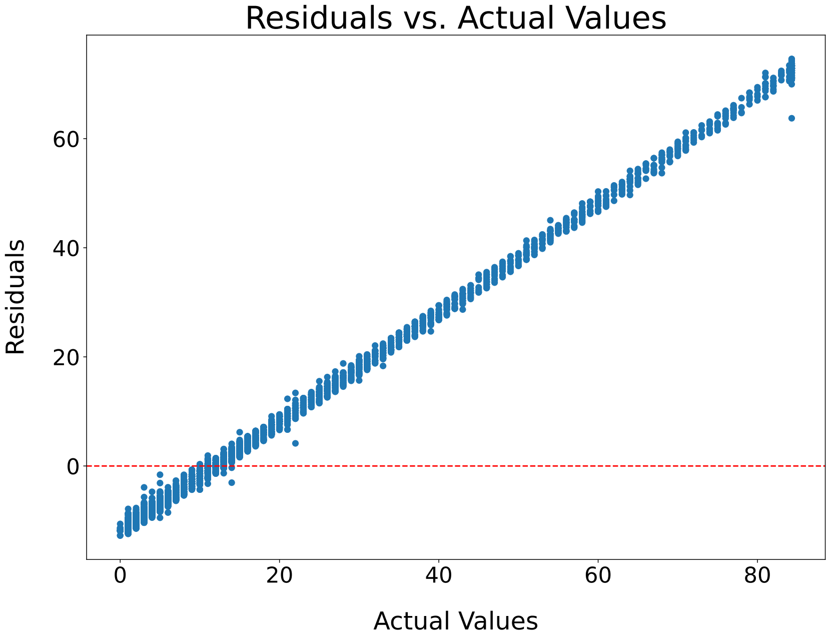


Figure 25: Residuals vs. Actual Values - Validation Set)