A Spatio-Temporal Exploration of Significant Earthquakes

Tasdeed Haq

Abstract— In this paper, we provide an effective spatio-temporal analysis of significant earthquakes spanning from 1965 to 2016. We highlight how visual displays and human judgement are used heavily in conjunction via a visual analytics approach to aide our analysis helping to result in comprehensive conclusions.

The report utilizes cluster maps and time series to identify temporal patterns, possible seasonality and peculiar time periods. Density-based clustering methods such as DBSCAN and OPTICS are implemented to aide in identifying oceanic hotspots, used in combination with tree maps to identify the most susceptible countries. Time-series analysis is implemented to identify countries expected to receive increased earthquakes in the future through assessment of plate boundaries with risk analysis implemented to locate which countries will become further unsafe. ARIMA models are developed to help predict the total number of earthquakes in following years with close examinations of ACF and PACF plots as well as model diagnostics by consideration of residuals.

We expect the outcomes of this paper to benefit planners in identifying the most at-risk regions whereby construction plans should incorporate significant earthquake protection. We further hope this aides residents in susceptible regions deciding whether additional earthquakes protection measures would be required in the future.

1 Problem Statement

Earthquakes are one of the most dangerous phenomena which strike unexpectedly. It is a result of releasing tension at faults where slowly moving tectonic plates get stuck. As Earth's lithosphere is made up of multiple plates, earthquakes occur frequently worldwide which can cause severe disruptions to services and more destructive impacts such as tsunamis wreaking havoc amongst cities like the Boxing Day Tsunami [1], ultimately leading to loss of life.

Hence, much literature is available focusing on this for example proposals on constructing 'earthquake-proof' buildings [2] and developments of deep-learning models for earthquake detection [3]. We aim to answer the following:

1. Is there a detectable seasonality in earthquakes or peculiar years in terms of magnitude?
2. Can we identify the most susceptible countries, and identify countries expected to receive increased impact if earthquakes exhibit temporally increasing trends?
3. Can we develop a reliable time series model to predict the expected number of earthquakes in following years?

This analysis will prove useful in possibly identifying earthquake patterns and suggesting to planners which countries require more protection to minimize future damages.

The dataset available [4] is suitable as it contained spatio-temporal data on earthquakes spanning from 1965 to 2016 where temporal patterns can be investigated under varying granularities. Extensive feature engineering will be implemented such as plate boundary proximity, allowing for spatial assessments of earthquakes along these. Whilst factors such as yearly tectonic plate speed naturally unavailable, our data should nonetheless suffice and with over 20,000 observations, produce a detailed analysis.

2 State of the Art

Earthquakes have always been a major concern due to its impacts and therefore has various subtopics of interest, for example, developments of earthquake monitoring software [5] and resistant structural engineering [6]. Combined with the rise in visual analytics approaches [7][8], it is now easier to efficiently investigate earthquakes and their consequences.

Azis et al. [9] worked with identical data and used time series visualizations assessing how numbers of earthquakes varied annually, aggregated by plate boundaries and by time zone. To predict the total number of earthquakes in following years, 152 models were developed; for each of the 24 time zones, there were two time series, a regular time series and its stationary counterpart, similarly for the 52 plate boundaries. A combination of Linear Regression and LSTM Models were deployed, where both models showed varying performance depending on the time zone and plate. Predictive results were primarily considered by assessing R^2 values. It was found that stationary models performed better, and hence suggested that further research should look at implementations of stationary time series modelling techniques.

Hence, we plan to implement ARIMA modelling to predict the number of earthquakes in following years and furthermore, use diagnostic visualizations to evaluate model effectiveness which the author did not carry out, and consider time granularity.

Yang et al. [10] explored spatio-temporal properties of earthquakes between 1960 and 2014 using a spatio-temporal scanning technique to identify two kinds of clusters, burst and persistent. This was carried out by modelling the whole space as a spatio-temporal cube, detecting the type of clusters using cylinders, expanding their base and height until a threshold was met. The relative risk of each cluster was calculated. Space-Time Plots were used to visualize clusters as well as two-dimensional spatial plots examining how clusters varied along plate boundaries.

Our spatial methodology will differ here slightly, as we will be using instead OPTICS and DBSCAN Clustering due to their abilities to adapt with geo-spatial data [11][12]. This will be applied on spatial data to eliminate noise and locate clusters of dangerous at-sea earthquake hotspots since most occur off-shore. This report will aid our research as we will similarly be utilizing two-dimensional spatial plots to help visualize clusters.

Battul et al. [13] explored spatio-temporal techniques using a similar dataset, for earthquakes between 1912 to 2009 in India. Feature engineering was carried out by binning features such as magnitude, and visualizations such as bar charts and histograms were used to assess distributions of features, as well as spatial visualizations assessing the distribution of earthquakes, colored by binned groups. A linear regression model was fitted examining how numerical features related to earthquake size, visualized by plotting comparisons of true values against predicted.

Whilst some histograms had yielded some conclusions, the author did not consider temporal changes, this would be more meaningful to assess which is what we will implement in combination with binning to improve insight. We will also build on this by examining how earthquakes vary given plate boundary locations and examine seasonality via heatmaps.

3 Properties of the Data

The dataset, extracted from Kaggle [4] contained 23,230 earthquakes between 1965 and 2016, giving a sufficient timescale to identify patterns, consisting of earthquakes with magnitudes over 5.5 making up 21 features; the ‘main’ features included Date, Time, Latitude, Longitude, Depth and Magnitude. Since magnitudes below 5.5 were not included, we implicitly assumed each earthquake was distinct rather than related by foreshocks, mainshocks and aftershocks which would be unrepresentative. Spatial features spanned worldwide; precision-wise, Latitude and Longitude were given to 3 decimal places, sufficient for earthquakes accounting for random error, with Time given down to the second. Remaining features were geology specific for example Earthquake ID and Magnitude Source which were irrelevant, as well as Azimuthal Gap etc. which contained 30% to 99% missing values. Given these proportions and that imputing these would yield unfair analysis, these were dropped. No duplicates appeared.

Missing value analysis was carried out; where only 3 missing values were found for Magnitude Type (how magnitude was measured), these were imputed using the mode since it made up miniscule amounts. Earthquake magnitudes were measured via different metrics for example Moment Magnitude, Richter Magnitude etc., these measurements are valid for certain frequency and distance ranges; the range of validity for these all lie on the same scale. Where possible, magnitudes were converted to the uniformly applicable Moment Magnitude [14], suitable for large magnitudes. Where conversions were not possible, we implicitly assumed their magnitudes were measured using the Moment Magnitude since previously converted magnitudes did not deviate much from their original values.

Outlier analysis was carried out using histograms and boxplots with the latter acclaimed at its ability to detect outliers [15]. Only Depth exhibited these across the 50 years. Negative depths had their depths set to zero, which were effectively 0 anyways. 300 earthquakes were identified with larger than usual depths (Fig. 1) - we decided to not drop these – we interpreted these as rare earthquakes in which we cannot know when or how often they occur, attributed to the Black Swan Principle [16].

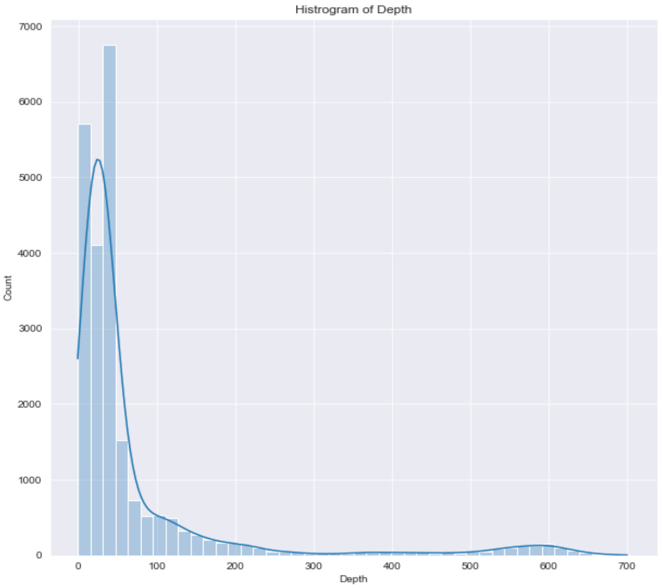


Fig. 1 - Histogram of Depth highlighting a small cluster of earthquakes with very deep depths

Extensive feature engineering was implemented; geocoding was employed retrieving country names based on latitude and longitude, and subsequently continent names. Over 50% of earthquakes occurred at sea, hence their Country and Continent label was ‘Undefined’. Magnitude and Depth were grouped creating two new features using MTU’s Magnitude Groupings [17] and the USGS Depth Groupings [18] respectively. The Time feature was kept but broken down into Year and Month allowing for analysis over different granularities.

K-Means Clustering was applied using a large number of clusters (40) not for the purpose of clustering, but to create many small groupings of points. By assessing the spatial distributions of each cluster, two new features were created – the Plate Boundary these earthquakes lie on and the Type of Plate Boundary (Convergent, Divergent and Transform). K-Means surprisingly performed well by producing many tight clusters lying precisely on plate boundaries, given that latitude and longitude were used with the algorithm implementing Euclidean distances, yielding the final dataset shown in Fig. 2.

Feature:	Description:
Date	YYYY-MM-DD
Time	HH:MM:SS
Latitude	Spatial Coordinates of Earthquakes
Longitude	Spatial Coordinates of Earthquakes
Depth	How Deep Underground an Earthquake Occurred in Kilometres (km), (0 km +)
Magnitude	Relative Size of Earthquakes; Measured Using Moment Magnitude (M_w)
Year	1965, 1966, …, 2016
Month	1, 2, …, 12
Country	Undefined (At Sea), Indonesia, Japan…
Continent	Undefined (At Sea), Asia, Europe…
'MagGroup'	Grouped Magnitude; Band 0: [5.5 – 6], Band 1: (6 – 6.9], Band 2: (6.9 – 7.9] and Band 3 (7.9 - 9.1]
'DepthGroup'	Grouped Depth: Shallow: [0 - 70), Intermediate: [70 - 300) and Deep: [300 - 700]
'Tecto-Setting'	Type of Plate Boundary: Convergent, Divergent and Transform
Region	(Closest) Plate Boundary, Australian-Pacific, Philippine-Eurasian…

Fig. 2 – Description of the final feature set for analysis

4 ANALYSIS

4.1 APPROACH

We now discuss the analysis approach taken to answer our questions, highlighting how human reasoning with visual displays are repeatedly utilized together to aide our analysis. Fig. 3 summarizes our approach.

Our approach to pre-processing was explained earlier, human reasoning was essential in understanding visualizations like histograms and determining how to analyze outliers in an earthquake context. It was used to engineer a further meaningful dataset through determining suitable numbers of clusters for K-Means and assignment of plate boundaries. It was required to rectify errors during geocoding ensuring correct country assignments, identify suitable magnitude transformations, to make assumptions and ensure a suitable final dataset. We now discuss our analysis in further detail.

To expose possible seasonality and peculiarities, heatmaps will be implemented, human reasoning will be used by examining how magnitudes differ across varying combinations of years and months aiding in seasonality identification. Given the high number of years, time-series plots by both year and month will also be used to help reasoning. Tweaking color schemes of the heatmap also will assist reasoning in identifying peculiar periods by highlighting these in dark colors. Where peculiar periods occur, deeper investigation will be performed providing insight on responsible plate boundaries for increased magnitudes via bar charts aggregated by magnitude group. Spatial visualizations will be utilized, examining most hit

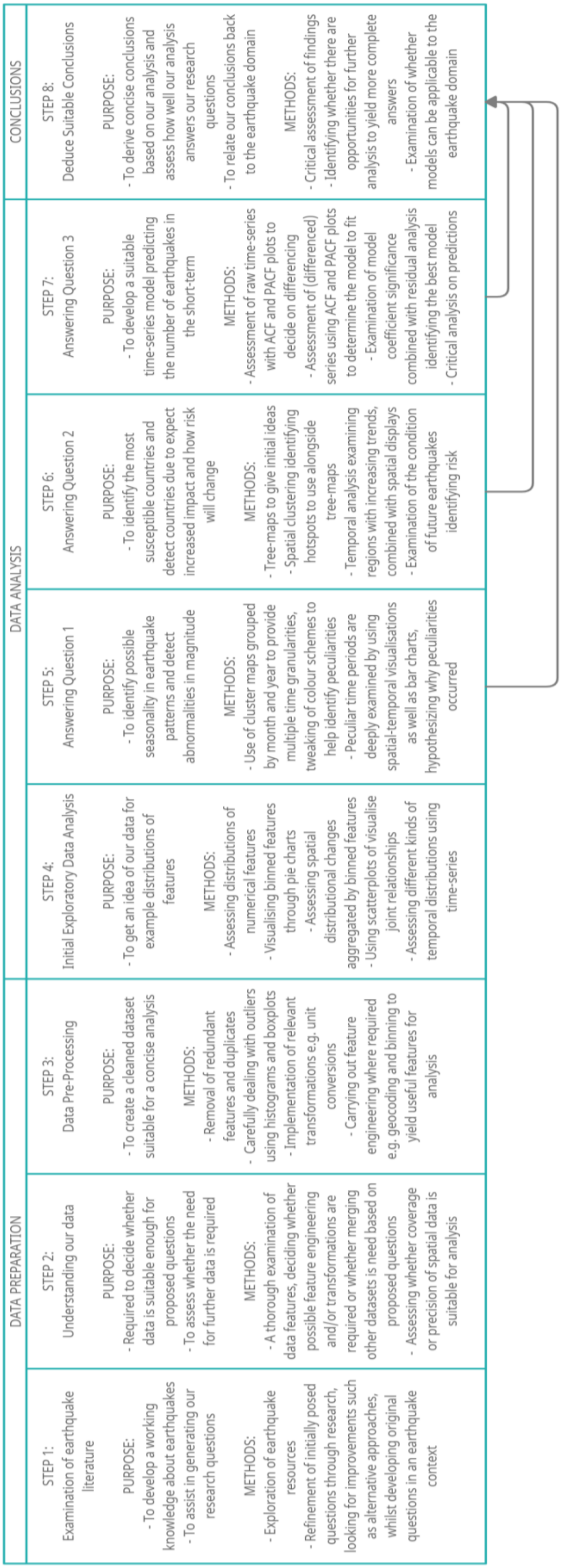

DATA PREPARATION
DATA ANALYSIS
CONCLUSIONS
STEP 1:
Examination of earthquake literature
PURPOSE:
- To develop a working knowledge about earthquakes
- To assist in generating our research questions
METHODS:
- Exploration of earthquake resources
- Refinement of initially posed questions through research, looking for improvements such as alternative approaches, whilst developing original questions in an earthquake context
STEP 2:
Understanding our data
PURPOSE:
- Required to decide whether data is suitable enough for proposed questions
- To assess whether the need for further data is required
METHODS:
- A thorough examination of data features, deciding whether possible feature engineering and/or transformations are required or whether merging other datasets is need based on proposed questions
- Assessing whether coverage or precision of spatial data is suitable for analysis
STEP 3:
Data Pre-Processing
PURPOSE:
- To create a cleaned dataset suitable for a concise analysis
METHODS:
- Removal of redundant features and duplicates
- Carefully dealing with outliers using histograms and boxplots
- Implementation of relevant transformations e.g. unit conversions
- Carrying out feature engineering where required e.g. geocoding and binning to yield useful features for analysis
STEP 4:
Initial Exploratory Data Analysis
PURPOSE:
- To get an idea of our data for example distributions of features
METHODS:
- Assessing distributions of numerical features
- Visualising binned features through pie charts
- Assessing spatial distributional changes aggregated by binned features
- Using scatterplots of visualise joint relationships
- Assessing different kinds of temporal distributions using time-series
STEP 5:
Answering Question 1
PURPOSE:
- To identify possible seasonality in earthquake patterns and detect abnormalities in magnitude
METHODS:
- Use of cluster maps grouped by month and year to provide multiple time granularities, tweaking of colour schemes to help identify peculiarities
- Peculiar time periods are deeply examined by using spatial-temporal visualisations as well as bar charts, hypothesizing why peculiarities occurred
STEP 6:
Answering Question 2
PURPOSE:
- To identify the most susceptible countries and detect countries due to expect increased impact and how risk will change
METHODS:
- Tree-maps to give initial ideas
- Spatial clustering identifying hotspots to use alongside tree-maps
- Temporal analysis examining regions with increasing trends, combined with spatial displays
- Examination of the condition of future earthquakes identifying risk
STEP 7:
Answering Question 3
PURPOSE:
- To develop a suitable time-series model predicting the number of earthquakes in the short-term
METHODS:
- Assessment of raw time-series with ACF and PACF plots to decide on differencing
- Assessment of (differenced) series using ACF and PACF plots to determine the model to fit
- Examination of model coefficient significance combined with residual analysis identifying the best model
- Critical analysis on predictions
STEP 8:
Deduce Suitable Conclusions
PURPOSE:
- To derive concise conclusions based on our analysis and assess how well our analysis answers our research questions
- To relate our conclusions back to the earthquake domain
METHODS:
- Critical assessment of findings
- Identifying whether there are opportunities for further analysis to yield more complete answers
- Examination of whether models can be applicable to the earthquake domain

Fig. 3 – Analysis Approach Workflow

countries during these periods, attempting to hypothesize the reasoning behind these using domain knowledge.

Tree-maps will be utilized providing brief ideas of most susceptible countries. Human judgement will be used identifying whether fair conclusions from this can be derived. Given over 50% of earthquakes occur offshore, density-based clustering methods will be implemented on spatial data using default and tuned parameters (more detail later) to identify spatial clusters at sea visualized using a map. Human judgement will be used identifying oceanic hotspots, further considering how depth and magnitude vary in these regions and considering proximity to countries, in combination with the tree-map helping identify overall vulnerable countries.

In examining countries expected to receive increased impact, temporal visualizations will be used assessing whether increasing trends appear, aggregated by continent. Where trends appear, analysis showing changes in earthquakes numbers on surrounding plate boundaries will be explored, including magnitude and depth changes assessing risk. Spatial visualization of earthquakes surrounding worsening plate boundaries will be examined using judgement to identify further at-risk countries in the future.

Human judgement will be utilized deciding the time granularity to predict the future number of earthquakes on; an initial time series plot complimented with its ACF and PACF plots will be made where judgement is required identifying whether differencing is needed. Once differenced, if required, human judgement is needed to determine ARIMA model parameters based on (new) ACF and PACF visualizations whereby a model is fitted. Assessment of the model coefficient significances, results of Ljung-Box and Jarque-Bera Tests and examination of residual diagnostics will be done via human judgement. Using the final model, predictions will be done, to be compared to true values whereby conclusions will be made on quality and predictive power.

4.2 PROCESS

4.2.1 QUESTION 1

Heatmap analysis in Fig. 4 showed average magnitude did not yield trends by month, year or a combination, supported also by time series examination by year and month even when accounting for magnitude and depth, suggesting absent seasonality supporting expectations that earthquakes unexpectedly strike.

However, the heatmap suggested peculiarities during August 1965, December 1966 and February 1969 exhibiting unusually high average magnitudes.

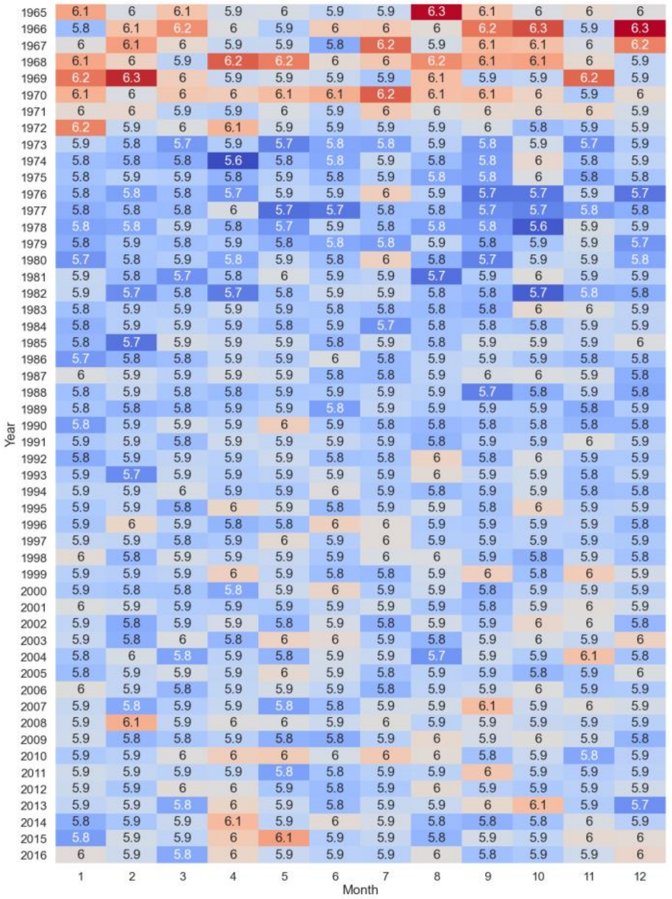


Fig. 4 – Heatmap of average magnitudes by Month and Year

August 1965 found the Australian-Pacific boundary primarily responsible here. Analysis showed annually, roughly 65-70% of earthquakes magnitudes are Band 0. With 53% of earthquakes in Bands 1 and 2 with remaining in Band 0 during this period, this explained increases in average magnitude. Highest magnitude earthquakes were concentrated on the Australian-Pacific boundary, with depth mostly shallow. Combinations of high magnitude and shallow earthquakes imply high chances of land damages in close proximity. Spatial analysis showed much of these earthquakes surrounding Vanuatu which borders the boundary – correlating with literature mentioning Vanuatu being hit with a series of destructive earthquakes then [19], visualized in Fig. 5.

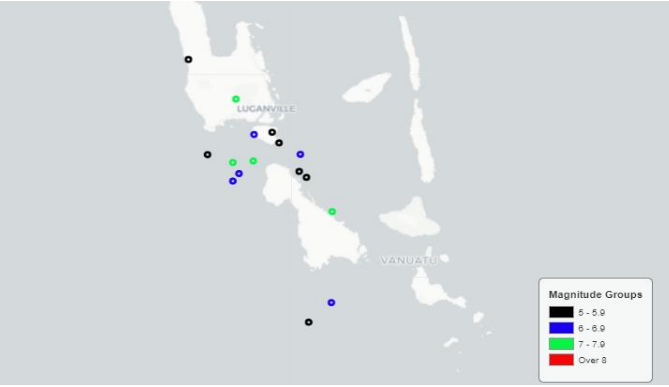


Fig. 5 – Spatial Distribution of the barrage of earthquakes hitting Vanuatu with varying magnitudes in August 1965

Similar conclusions yielded compared to August 1965 with most earthquakes focused on the Australian-Pacific plate, including the highest magnitude ones, although a shallow Band 2 earthquake occurred on the South American – Nazca boundary. Most earthquakes were shallow; spatial analysis showed these surrounded Papua New Guinea and Vanuatu, as well as a destructive earthquake on the Chilean coast causing $400,000 dollars’ worth of damages (1966 rate) [20].

February 1969 was interesting, whilst there was dominance of higher band magnitude earthquakes like before, with much of the earthquakes focused on the Philippine-Eurasian plate boundary, what was interesting was that strongest earthquakes still occurred on Australian plate boundaries, e.g., Australian-Pacific, with Band 0 earthquakes occurring along Philippine-Eurasian plate and elsewhere. In contrast, earthquakes that occurred along Australian plates were deep depth-wise hence not deemed as destructive due to seismic waves losing energy travelling far towards the surface [21]. Spatial visualization showed the Philippines and parts of Indonesia primarily affected here.

In conclusion we observe the source of peculiarity narrowed down to high magnitude earthquakes surrounding the Australian plate in the 60s, where boundaries are convergent, primarily affecting Vanuatu and Papua New Guinea. Since the 60s, earthquake magnitudes surrounding Australian plates have dropped and stabilized to reasonable levels via time series visualizations explaining peculiarities in this period, although no domain literature found explaining why. With insufficient data, we cannot conclude why these boundaries had such high numbers however literature suggested the Australian and Pacific plates are the fastest moving [22], hence we could hypothesize that the Australian plate was moving at its fastest in the 60s; given its speed, huge build-ups of stress were created where the plate got stuck. Once weaker crust slipped, releases of tension created a series of dangerous earthquakes affecting surrounding countries compared to slower moving plates where stress build-up is naturally weaker. We may further hypothesize that since the 60s, this plate speed has slowed down.

4.2.2 Question 2

Fig. 6 gives a tree-map showing the most susceptible countries; however, this was deemed unrepresentative. Over 50% of earthquakes occurred offshore, with spatial visualizations showing many earthquakes occurring in close proximity to countries where geocoding failed at assigning countries. Spatial visualizations were highly congested, so DBSCAN and OPTICS clustering were employed on spatial data to identify oceanic hotspots due to their flexibilities working with varying distance metrics [11][12]. We will be working with offshore earthquakes here.

Spatial coordinates were given via latitude and longitude, hence the Haversine metric was used in calculating great-circle distances between earthquakes assuming Earth was spherical to develop a distance matrix, first converting spatial data into radians.

An initial default DBSCAN algorithm was implemented – yielding poor results, with essentially all points allocated to a single cluster, likely attributed to the epsilon parameter in scikit-learn being unsuitable; DBSCAN results are highly sensitive to epsilon [11] with large values yielding large clusters with less noise and vice versa. Attempts at tuning both the epsilon and minimum number of samples parameters were implemented using heuristic approaches [23][24]. This was our best DBSCAN outcome - clusters resembled plate boundaries, shown in Fig 7a. Whilst helpful in showing most earthquakes occurred on boundaries, this was still too noisy.

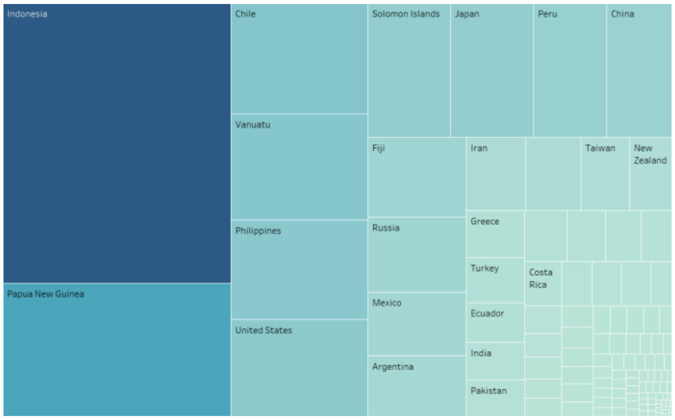
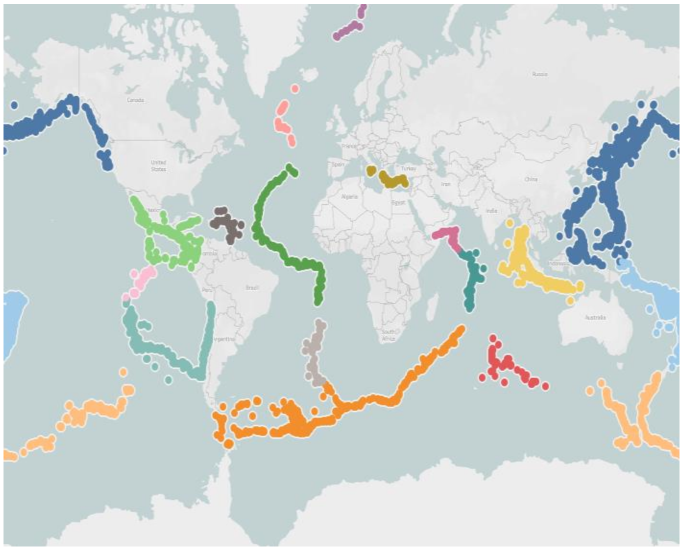
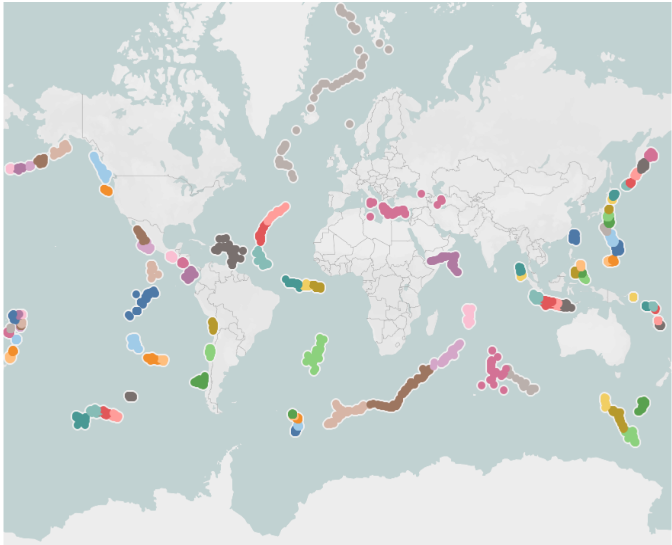


Fig. 6 – An initial tree-map highlighting the most susceptible countries, excluding the country 'Undefined'

We expected OPTICS would improve results given independence of epsilon and consideration of clusters having local densities [12]. Using the same minimum number of samples as the 'tuned' DBSCAN model, results are shown in Fig. 7b, showing improved results with noisier observations such as those far from countries removed, hence easier to identify susceptible countries. Many countries in (south) eastern Asia such as Japan, Indonesia, Papua New Guinea and Vanuatu, and most lying on the west of South America such as Chile were found to be highly susceptible alongside New Zealand. Most countries within the Greater Antilles are susceptible and also those around the Mediterranean and Arabian Sea, all which noticeably lie close to plate boundaries; highlighting why utilizing tree-maps alone were unrepresentative; for example, Japan, known as one of the most susceptible countries, was not reflected in Fig. 6 as nearly all earthquakes happened offshore. By combining the tree-map with above clusters, it was easily identifiable what the most susceptible countries are, namely Japan, Indonesia, Papua New Guinea, Chile, Vanuatu and Philippines in which the latter 5 dominated the tree-map, also exhibiting oceanic hotspots.





Figs. 7a and 7b – Clustering results obtained from our ‘tuned’ DBSCAN algorithm (top) and OPTICS algorithm (bottom), with observations considered noisy removed

Time series plots were used assessing changes in numbers of earthquakes annually by continent (Fig. 8a). All continents showed stability except ‘Undefined’, exhibiting significant increases i.e., rises in offshore earthquakes, prompting temporal examinations of changes in earthquakes aggregated by plate boundary only applied to ‘Undefined’, identifying responsible boundaries. Most boundaries remained stable, except five (Australian-Pacific, Caribbean-Cocos, Nazca-Pacific, Antarctic-Pacific and South America-Nazca) exhibiting slow increases which collectively gave significant increases overall (Fig. 8b). An investigation into risk was done. In all mentioned plates, increases in earthquakes were attributed to rises in Band 0 earthquakes, which were all shallow hence low risk, consequences for example light ground shaking. Higher bands of earthquakes remained stable on most plates hence risk levels only marginally increasing for surrounding countries, although the Australian-Pacific and Caribbean-Cocos boundaries are experiencing gradual increases in Band 1 earthquakes (Fig. 8c) alongside with shallow depths (Fig. 8d), so risk levels are further increased, with consequences such as heavier ground shaking.

Whilst risk levels to countries surrounding most boundaries seem unchanged, many countries should expect further serious earthquakes in the upcoming years primarily those in the Greater Antilles, and those bordering the Australian-Pacific plate such as New Zealand, Vanuatu and Fiji.

4.2.3 Question 3

The granularity of time was selected to be months, this struck a balance between years and days. Predictions based on year leads to an insufficient 50 data points, with days being too granular leading to a noisy series being too random.

An increasing trend was visible showing the series was not stationary combined with slow decay occurring and points not staying within bounds on the ACF plot at high lags. The trend looked roughly linear or slightly quadratic suggesting testing of first and second differences and comparing results. (Fig. 9a)

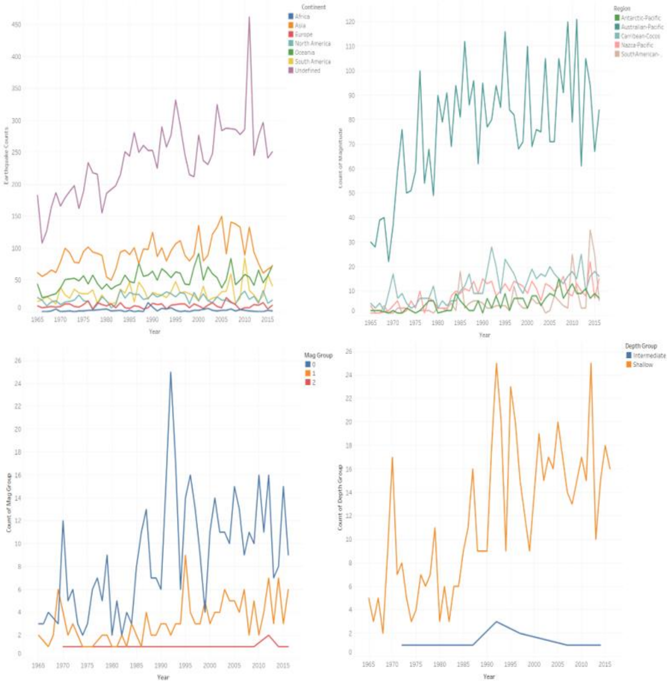


Fig. 8 – Top Left, 8a: Change in earthquake numbers by continent, Top Right, 8b: Plate boundaries exhibiting increases in offshore earthquakes, Bottom Left, 8c: Changes in magnitude groups of offshore earthquakes for the Caribbean-Cocos boundary, Bottom Right, 8d: Changes in offshore depth groups for the Caribbean-Cocos boundary

Both first and second differenced series looked stationary by visualization of the series albeit some large spikes in the latter (likely attributed to introduction of more noise with higher-order differencing), and in ACF plots although this is clearly not perfect with multiple points reaching out of bounds at higher lags, expected with real word data (Fig. 9b). No pattern was determined with out of bounds spikes, indicating no seasonality.

Considering the first differenced series, one significant spike was shown at lag 1 on the ACF plot, indicating an MA(1) model, supported by exponential decay in the PACF plot. Fitting this model returned statistically significant coefficients through p-value assessment. The model passed the Ljung-Box Test [25] for the first 20 sufficient lags [26] except the first using 5% significance level, indicating residual autocorrelations being statistically zero. We hope model residuals resemble white noise which has this property; however, the model failed the Jarque-Bera Test [27] indicating residuals not following a Gaussian Distribution. Examination of the Kernel Density estimate showed a slight skew and kurtosis supporting this.

The second differenced model yielded an MA(2) model through similar examinations. We observed significant coefficients and improved results observed for the Ljung-Box Tests; whilst the Jarque-Bera Test still failed, visualizations of the residual KDE more closely resembled a Gaussian Distribution as well as the raw series visualization albeit some large spikes, roughly resembling white noise, with improved skew and kurtosis values, this was our model of choice although not perfect (Fig. 9c).

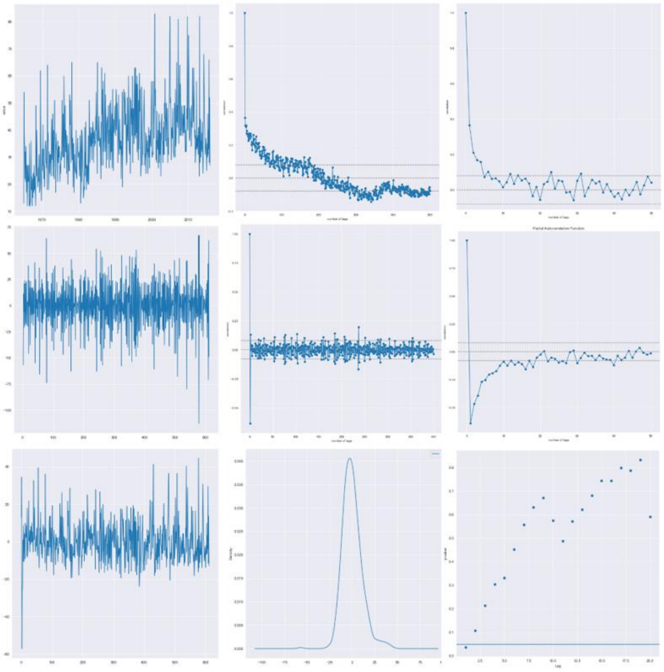


Fig. 9 – First Row, 9a: Original time-series showing changes in earthquakes by month complimented by its ACF and PACF plot respectively. Second Row, 9b: Identical to first row with the series differenced twice. Third Row, 9c: Visualization of twice-differenced model's residuals, Residual Kernel Density Estimate and p-values of the Ljung-Box Test for the first 20 lags with a p-value bar at 0.05 respectively

Using this model, predictions were made forecasting the monthly number of earthquakes in 2015-2016 as long-term predictions are unreasonable, fitting a model on 1965-2015 data (Fig. 10). Our model exhibited effectively constant predictions over the 12 months, which is expected for MA(2) models (mean prediction after first two predictions). Whilst predictions were particularly unhelpful, it supports the randomness of earthquakes highlighting the difficulties in earthquake prediction. In terms of predictive power, empirical evidence shows average forecasts outperforming ARIMA models [28], therefore providing a benchmark for earthquakes in following months.

4.3 RESULTS

We concluded that there was an absence of seasonality, both in terms of months and years, successfully supporting expectations that earthquakes strike unexpectedly. Peculiarities during the 60s were identified with increased magnitudes, which was narrowed down to high magnitude earthquakes on Australian plate boundaries. Whilst literature on why this occurred during the 60s was unavailable, we developed a suitable hypothesis explaining this phenomenon.

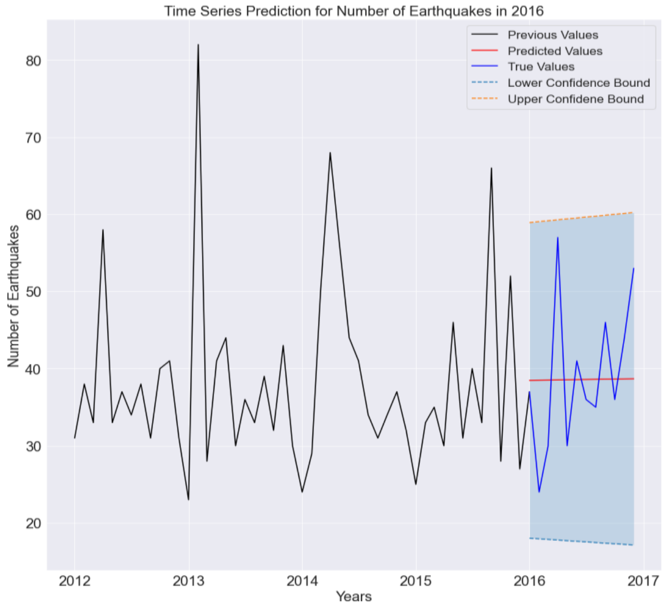


Fig. 10 – Comparison of predictions for the next 12 months using the second differenced model (red) compared to true values (blue), shading represents the 95% confidence interval for predictions

We successfully identified oceanic hotspots via OPTICS clustering, which in combination with tree-maps, developed an insight into the most susceptible countries, particularly located in East and Southern Asia and South America. Repetitive use of time series visualizations identified countries expecting increased earthquakes, particularly those by the South American-Nazca, Australian-Pacific and Caribbean-Cocos boundaries. A complimentary risk analysis via investigation of magnitude band and depth changes highlighted countries that should be expecting increased risk, useful for respective residents and planners to implement suitable earthquake-proofing measures.

Through diagnostics and visualizations, an ARIMA model was developed to predict the number of earthquakes in the short-term. We found predictions were not as useful as hoped with the mean of the process constantly being predicted, although it highlighted the randomness of earthquakes and provided a baseline for future earthquakes useful for earthquake geologists.

5 CRITICAL REFLECTION

Heatmap visualizations were a valuable tool allowing magnitude analysis against differing time granularities, aiding reasoning in peculiarity identification and temporal patterns, particularly seasonality. Given high number of years however, the heatmap was crowded, thus alternative visualizations were utilized alongside such as time series visualizations by months and years leading to our conclusion of seasonality. We considered average magnitudes overall, hence disregarded possible temporal trends on country and continent levels or via plate (made possible through feature engineering using K-Means) providing future analysis steps. For further insight, alongside, changes in magnitude bands could be considered rather than average magnitude, answering our initial research question better, further possibly proving useful to planners for example, choosing when to begin construction, maximizing chances of avoiding major earthquakes.

Feature engineering identifying countries and continents using spatial coordinates proved useful; geocoding filtered out offshore earthquakes due to inability to assign countries, allowing for examinations at more granular levels such as continent and allowing us to observe that over 50% of earthquakes occurred offshore. Dense spatial visualizations of offshore earthquakes prompted judgement in deducing methods to improve visualizations. DBSCAN and OPTICS were hence implemented to identify oceanic hotspots. Whilst OPTICS clustering produced impressive results, DBSCAN proved troublesome given high dependencies on epsilon and minimum number of samples. Whilst heuristic approaches improved results, it could not yield expected results, further investigation should be implemented tweaking parameters. Spatial visualizations were useful in deciding meaningfulness of clusters. Other clustering approaches like DENCLUE-IM [29], a DENCLUE variant, could be attempted, improving efficiency, since OPTICS took large periods of time, a lesson to remember for future projects. Time series plots proved invaluable in determining risk levels to countries also.

Through repetitive visualizations with judgement via considerations of time series, ACF and PACF plots, we managed to devise a time-series model, deciding how many times to difference and model type to fit, though residual visualizations and hypothesis test results indicated a slightly imperfect model, expected with real-life data. Whilst predictions were somewhat unhelpful given the nature of the model, it provided baselines for future earthquake numbers. Further steps in improving prediction results aside from improving Regression and LSTM approaches [9] include fitting Artificial Neural Networks (ANN) given the non-linear nature of the series; LÖK et al. [30] showed ANN's performed well predicting smaller earthquakes but struggled on larger magnitude earthquakes attributed to insufficient data. With a dataset three-times larger, applying ANN's may yield more promising results.

Our analysis suffered from limitations however, only containing earthquakes with magnitudes over 5.5, hence unreasonable to group earthquakes by foreshocks, mainshock and aftershocks potentially having allowed examination for how earthquakes spread via foreshocks and aftershocks from the epicenter, interactively via space-time visualizations. Hence, we assumed each earthquake was distinct, which is not entirely accurate. Geology-specific features were removed given high proportions of missing values; imputation may have benefitted analysis. Missing data on volcanic eruptions meant we could not examine whether eruptions possibly triggered certain earthquakes or vice versa as these are related, yielding another avenue for analysis.

6 TABLE OF WORD COUNTS

Excluding Student Details, Headings, Subheadings, Figure Captions, Text Contained in Tables and References.

Abstract	200/200
Problem Statement	250/250
State of the Art	500/500
Properties of the Data	500/500
Analysis: Approach	500/500
Analysis: Process	1500/1500
Analysis: Results	200/200
Critical Reflection	500/500
Total Word Count	**4150/4150**

7 REFERENCES

[1] Jackson, L.E., Jr., Barrie, J.V., Forbes, D.L., Shaw, J., Manson, and G.K., Schmidt, 'Effects of the 26 December 2004 Indian Ocean tsunami in the Republic of Seychelles. Report of the Canada UNESCO Indian Ocean Tsunami Expedition', *Geological Survey of Canada, Open File 4539*, 2005 p. 73.

[2] M. Ozaki and S. Hayashi, 'Earthquake resistant design of offshore building structures', *IEEE Journal of Oceanic Engineering*, vol. 3, no. 4, pp. 152–162, Oct. 1978

[3] M. Maya and W. Yu, 'Short-term prediction of the earthquake through Neural Networks and Meta-Learning', in *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Sep. 2019, pp. 1–6.

[4] Significant Earthquakes, 1965-2016, published by USGS. https://www.kaggle.com/usgs/earthquake-database

[5] R. Hoque, S. Hassan, MD. A. Sadaf, A. Galib, and T. F. Karim, 'Earthquake monitoring and warning system', in *2015 International Conference on Advances in Electrical Engineering (ICAEE)*, Dec. 2015, pp. 109–112.

[6] H. Sun and Q. Li, 'Research and Development of Seismic Base Isolation Technique for Civil Engineering Structures', in *2010 International Conference on E-Product E-Service and E-Entertainment*, Nov. 2010, pp. 1–5.

[7] S. Chen, S. Li, L. Xie, Y. Zhong, Y. Han, and X. Yuan, 'EarthquakeAware: Visual Analytics for Understanding Human Impacts of Earthquakes from Social Media Data', in *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2019, pp. 122–123.

[8] N. Adulyanukosol, 'Earthquake Damage Report Interactive Dashboard Using Bayesian Structural Time Series and Value-Suppressing Uncertainty Palettes', in *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2019, pp. 106–107.

[9] M. F. A. Azis, F. Darari, and M. R. Septyandy, 'Time Series Analysis on Earthquakes Using EDA and Machine Learning', in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2020, pp. 405–412.

[10] J. Yang, C. Cheng, C. Song, S. Shen, T. Zhang, and L. Ning, 'Spatial-temporal Distribution Characteristics of Global Seismic Clusters and Associated Spatial Factors', *Chin. Geogr. Sci.*, vol. 29, no. 4, pp. 614–625, Aug. 2019.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 'A density-based algorithm for discovering clusters in large spatial databases with noise', in *Proceedings of the Second*

International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, Aug. 1996, pp. 226–231.

[12] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, 'OPTICS: ordering points to identify the clustering structure', *SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999.

[13] V. Battul, M. Helen Santhi, and G. Malathi, 'Earthquake data analysis and data visualization of Maharashtra state, India from 1912 to 2009 using R programming', *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 989, no. 1, p. 012029, Nov. 2020.

[14] S. Baruah, 'Moment magnitude – local magnitude relationship for the earthquakes of the Shillong-Mikir plateau, Northeastern India Region: a new perspective', *Geomatics, Natural Hazards and Risk*, vol. 3, no. 4, pp. 365–375, Nov. 2012.

[15] Y. Sun and M. G. Genton, 'Adjusted functional boxplots for spatio-temporal data visualization and outlier detection', *Environmetrics*, vol. 23, no. 1, pp. 54–64, 2012.

[16] S. Hajikazemi, A. Ekambaram, B. Andersen, and Y. J.-T. Zidane, 'The Black Swan – Knowing the Unknown in Projects', *Procedia - Social and Behavioral Sciences*, vol. 226, pp. 184–192, Jul. 2016.

[17] Earthquake Magnitude Scale, published by the Michigan Technological University Engineering Department at: https://www.mtu.edu/geo/community/seismology/learn/earthquake-measure/magnitude/

[18] US Geological Survey Earthquake Depth Scale: https://www.usgs.gov/programs/earthquake-hazards/determining-depth-earthquake

[19] M. Benoit, D. Jacques, 'The earthquake swarm in the New Hebrides archipelago, August 1965', *Royal Society of New Zealand Bulletin* , 9, pp. 141-148, 1971.

[20] National Centre for Environmental Information (NCEI), Hazard Earthquake Information: December 1966 Chilean Earthquake: https://www.ngdc.noaa.gov/hazel/view/hazards/earthquake/event-more-info/4383

[21] H. Houston, '4.13 - Deep Earthquakes', in *Treatise on Geophysics (Second Edition)*, G. Schubert, Ed. Oxford: Elsevier, 2015, pp. 329–354.

[22] S. Zahirovic, R. D. Müller, M. Seton, and N. Flament, 'Tectonic speed limits from plate kinematic reconstructions', *Earth and Planetary Science Letters*, vol. 418, pp. 40–52, May 2015.

[23] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, 'Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications', *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, Jun. 1998.

[24] N. Rahmah and I. S. Sitanggang, 'Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra', *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 31, p. 012012, Jan. 2016.

[25] G. M. Ljung and G. E. P. Box, 'On a measure of lack of fit in time series models', *Biometrika*, vol. 65, no. 2, pp. 297–303, Aug. 1978.

[26] Ruey S. Tsay, *Analysis of Financial Time Series*, 2nd Ed. Hoboken, NJ: John Wiley & Sons, Inc., 2005, p. 33

[27] C. M. Jarque and A. K. Bera, 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals', *Economics Letters*, vol. 6, no. 3, pp. 255–259, Jan. 1980.

[28] K. C. Green and J. S. Armstrong, 'Simple versus complex forecasting: The evidence', *Journal of Business Research*, vol. 68, no. 8, pp. 1678–1685, Aug. 2015.

[29] H. Rehioui, A. Idrissi, M. Abourezq, and F. Zegrari, 'DENCLUE-IM: A New Approach for Big Data Clustering', *Procedia Computer Science*, vol. 83, pp. 560–567, Jan. 2016.

[30] S. LÖK and M. Karabatak, 'Earthquake Prediction by Using Time Series Analysis', in *2021 9th International Symposium on Digital Forensics and Security (ISDFS)*, Jun. 2021, pp. 1–6.