Randwick International of Social Sciences (RISS) Journal
Vol. 4, No.1, January 2023 | Page: 80-90
ISSN Online: 2722-5674 - ISSN Print: 2722-5666
http://www.randwickresearch.com/index.php/rissj

Predictive Analysis of Spatial Data and Time Series to Predict Earthquake Magnitudes by Using Data Mining Approach

DOI: https://doi.org/10.47175/rissj.v4i1.607

| Ignazio Ahmad Pasadana[1,*] | Herman Mawengkang[2] | Syahril Efendi[3] |

[1,2,3] Deparment of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

[*]ignazioahmadpasadana@gmail.com

ABSTRACT

The suggested methodology presented in this work uses data mining to identify seismic zones and time series to forecast earthquake magnitudes. Utilize historical earthquake data gathered from the United States Geological Survey (USGS) and obtained by utilizing hierarchical and fartherst first clustering to predict seismic activity. Latitude and longitude cluster data were used to create a prediction model to forecast the size of upcoming earthquakes in the Nanggroe Aceh Darussalam region and its nearby areas.

KEYWORDS

Future Pattern; Hierarchical Clustering; Fatherst First Clustering; Forecasting; Magnitudes; Earthquakes; Data Mining

INTRODUCTION

On the island of Sumatra, Nanggroe Aceh Darussalam (NAD), one of Indonesia's provinces, has one of the highest seismic areas. The likelihood of earthquakes that move actively along Sumatra is increased by the presence of subduction zones, as well as by the association of the Sunda Arc on the island's west coast, and a sizable shear fault known as the Great Sumatran Fault. This fault zone hosts the majority of the strike-slip movement (H. Harjono, 2017). The northern terminus of this active fault lies beneath Banda Aceh, which was shaken by the Indian Ocean earthquake in 2004. Since the earthquake ceased, the strain on the Great Sumatran Fault has dramatically grown, particularly in the northern part of the island of Sumatra (Deutsche Welle Indonesia, 2014).

In this work, two clustering algorithms are utilized to pinpoint specific NAD regions that are likely to experience intense seismic activity.. To find appropriate groupings, we contrasted hierarchical clustering and fartherst first clustering. After clustering, we identified two seismic zones that are most likely to experience an earthquake.

Three well-known classification techniques, including Decision Tree (DT), k-Nearest Neighbor (KNN), and Support Vector Machine (SVM), were used to assess the effectiveness of the chosen clusters. These findings should describe the most precise evaluation model discovered. Then, a time series analysis is carried out to forecast upcoming seismic activity patterns. Predictive models that forecast the possibility of a seismic event are created using predictive algorithms that include Simple Linear Regression, Gaussian Process, Multilayer Perceptron, and Support Vector Machine for Regression (SMOreg). With the help of the proposed model, NAD can estimate the likelihood of earthquakes and take precautions that should lessen their damaging effects in the future.

One of the research that has been released can calculate the likelihood of an earthquake (Absar et al., 2017). These elements include the distance from the epicenter, the connectivity, the population density, the level of development, and the severity of the earthquake. The risk of a particular location can be calculated more precisely thanks to this approach. However,

because so much of this work depends on user input, it is ineffective for assessing the overall risk of earthquakes.

According to Hashemi et al. (2016), there is a suggested way to take advantage of hierarchical clustering methods. Three clusters of five traits each were chosen. At first, only non-spatial attributes are taken into account, then spatial (location) attributes, and lastly, all of them are used. The clusters were trained using the DT, KNN, and SVM. They claimed that after examining the RMSE, KNN performs terribly and that DT performs just marginally better than SVM.

Another piece of literature (Kulkarni et al., 2015) compares various techniques for earthquake prediction. The k-nearest neighbor graph takes into account all distributions of training data. The research suggests that geographical, temporal, and magnitude forecasts are crucial for earthquake prediction.

While neural networks are better at handling temporal and magnitude predictions, data mining and clustering are better at handling spatial predictions. To anticipate earthquakes, several computational intelligence techniques are applied (Martínez-Álvarez, 2011).

Several pre-processing methods were employed to enter data for 33 seismic occurrences over a 28-year period (Last, 2016). Seven different classification methods, including KNN, SVM, ANN, J48, and others, were tested on this data. The Multi-Objective Info-Fuzzy Network (M-IFN) produces the best accuracy out of all the methods. By designating spatial features as independent variables, multi-linear regression is additionally employed to determine earthquake magnitudes (Dutta et al., 2017).

A higher coefficient of determination was obtained when comparing the anticipated magnitudes. Quantitative association rules and regression are used to analyze the time pattern and date of occurrence in order to predict earthquakes (Nivedhita et al., 2016).

On seismic activity scales, *k*-means clustering is utilized to group together homogeneous data (Hoque et al., 2017). This rule has a 90% accuracy rate in predicting the number of seismic vents given a year and a quadrant of the planet. Based on actual data sets, spatial data mining is used to identify seismic zones. In order to uncover intriguing patterns, spatial data mining differs from standard data mining in that it takes spatial data into account. After the spatial datasets have been normalized and discretized, a clustering algorithm is used to group related spatial data together because they actually correspond to the same seismic zone.

There is also discussion of long-term forecasting methods to determine the likelihood of strong earthquakes ($M \geq 4.95$). (M.J. Werner et al., 2011). By assuming that future earthquakes are more likely to occur in regions where previous earthquakes, particularly small ones ($M \geq 2$), have happened, the suggested model expands the general recurrence method. According to this presumption, the number of previous earthquakes in nearby regions can be used to forecast the number of earthquakes that will occur in a certain area in the future.

The study made use of data from 300,278 earthquakes with magnitudes of $M \geq 1.7$ that occurred in California between January 1, 1981, and April 1, 2010. The predictive feature is the degree of seismicity (the number of $M \geq 2$ events in a given cell smoothed by the kernel function) over the 24 year training period, as opposed to the anticipated characteristic, which is the number of $M \geq 3.5$ or $M \geq 4.95$ events over the upcoming five-year target period.

The smoothing parameter (number of nearby places) in this experiment is set to 10. Effects of farther sites are totally disregarded. The Gutenberg-Richter (GR) ratio is used to evaluate the magnitude distribution at a specific place (cell) over the target period. The average probability per earthquake in relation to the reference model is used to assess each prediction model's performance. Around 6.0 is the largest documented probability gain.

RESEARCH METHODS

Proposed Workflow

The major goal of this research is to examine past seismic occurrences in order to identify earthquake hotspots using spatial data from the USGS. Two pertinent clustering techniques are used for this. The cluster accuracy is then determined by comparing the clusters using three distinct algorithms. With the help of time series analysis, we can forecast upcoming seismic events. The proposed model's process is shown in Figure 1.

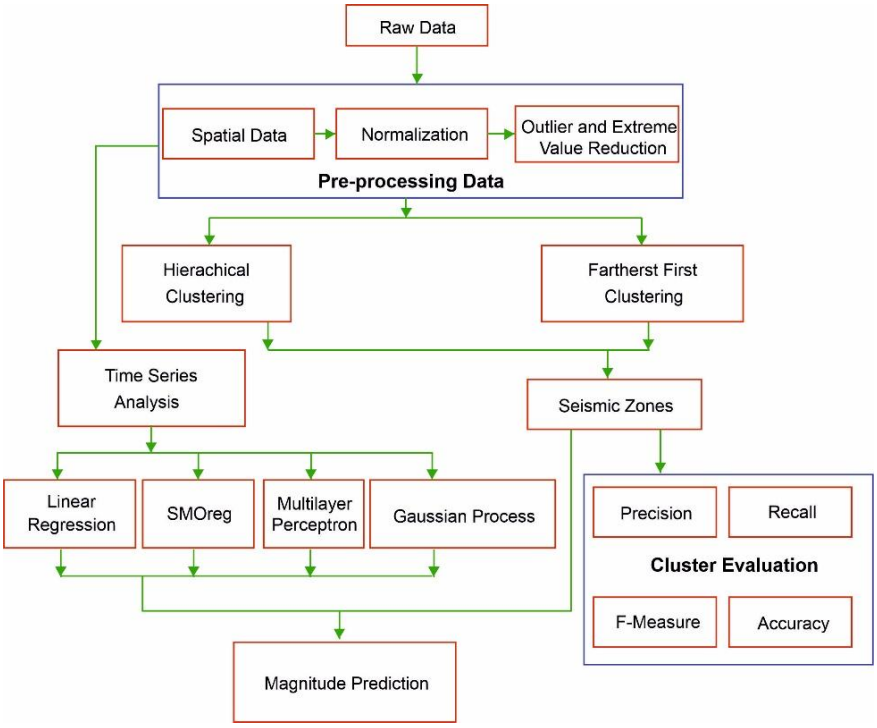

Figure 1. Block Diagram of Workflow

Data Collection

The study considers earthquakes that occurred within a 60-year time frame between 7.362° North Latitude and 2.153° North Latitude and 92.505° East Longitude and 99.492° East Longitude (01 January 1907 to 10 February 2021). NAD falls under our preferred field. The specifics of the earthquake and its processes were discovered via a website managed by the United States Geological Survey (USGS), which provided timely, important, and beneficial information. The table contains information on the earthquake's depth, magnitude, time of occurrence, and other pertinent details. We found that 7,461 (seven thousand four hundred and sixty one) earthquakes with a magnitude of 4.5 or greater have occurred in this area. An explanation

Table 1. Description of the dataset's attributes

Attribute	Description
DateTime	Month, Year, Day, Hour, Minute, and Second
Long	Decimal longitude degrees
Lat	Decimal latitude degrees
Depth	Kilometers between the epicenter of an earthquake and the surface
Mag	Earthquake's magnitude

Data Pre-processing

Normalizing variables is a crucial step before combining data points or creating a prediction model in order to eliminate outliers and extreme values. The zero mean and standard deviation approaches are utilized for that. The dataset's normalization is displayed in Equation 1. Here, the terms "mean," "standard deviation," and "normalized data" are used.

$$\text{Gap Days} = \text{NDays- IP} \quad (1)$$

Normalizing variables is a crucial step before combining data points or creating a prediction model in order to eliminate outliers and extreme values. The zero mean and standard deviation approaches are utilized for that. The dataset's normalization is displayed in Equation 2. Here, the terms "mean," "standard deviation," and "normalized data" are used.

$$N = (m - m')/s \quad (2)$$

It handle problems including erroneous or missing location data, the choice of pertinent features, and other concerns. The writers use a variety of pre-processing methods on the dataset when it comes to the spatial data. Here, spatial position, together with latitude, longitude, and depth, are used as spatial properties.

As utilized by Hashemi, et al., we added two additional properties to the dataset called "N-Days" (elapsed earthquake occurrence days since 1907) and "Gap Days" (days since the previous earthquake occurred) (M. Hashemi et al., 2016). Equation 1 is used to calculate "Gap Days" by ordering the dataset in ascending order based on "N-Days," where IP is the immediate predecessor of "Passed Days."

Clustering

To identify earthquake-prone areas in NAD, we group data based on longitude and latitude. In order to establish how frequently significant earthquakes occur in these regions, the data is then divided by magnitude, depth, and time. We execute hierarchical clustering using the average link approach since it is less sensitive to outliers, and we use two alternative clustering methods for this purpose. The following formula is used in this method to determine how far apart clusters are from one another (Equation 3).

$$\frac{1}{|A|.|B|}\sum_{x\varepsilon A}\sum_{y\varepsilon B} d(x,y) \quad (3)$$

In addition to being essentially deterministic, hierarchical clustering algorithms do not demand a specific number of clusters.

The farthest first traversal of a compact metric space in computational geometry is a series of points in space, where the first point is randomly selected and each following point is as far away from the set of previously chosen points as is possible. By limiting the selected points to the set or, alternatively, by taking into account the finite metric space produced by these points, the same idea can also be used to a finite collection of geometric points (Dasgupta et al., 2005). The resulting sequence creates a point permutation, also referred to as a greedy permutation, for a finite metric space or finite geometric point set (Har-Peled et al., 2006).

Cluster Evaluation

In order to evaluate clusters, the author separates the data set based on the number of clusters. Next, SVM, KNN, and DT were used to assess the cluster accuracy. A magnitude model based on longitude and latitude attributes will be produced after applying two clusters to the mapper. a map that will show seismic zones based on low, medium, and high densities. Latitude is represented by the x axis, and longitude is represented by the y axis. A seismic event is thus represented by each point. From this, we can infer that a particular place is more seismically sensitive than other areas.

Forecasting Future Pattern

Data forecasting can be done based on past occurrences to forecast future values. A set of data points in a time-dependent data set are modeled using time series analysis. The four methods utilized in this case include Simple Linear Regression, Gaussian Process, Sequence Minimum Optimization (SMOreg), and Multilayer Perceptron, among many more regularly used techniques. Using each of these four methods, our seismic event data is used to predict the magnitude of upcoming earthquakes. The optimal prediction method for the seismic data is chosen after comparing various algorithms based on the results attained.

Time is designated as the independent variable in each approach, while the magnitude of the earthquake is designated as the dependent output variable. The magnitude prediction was examined in phases (1, 5, and 13 of the validation process). Lagging data values are used to get rid of the mean's temporal variation. Because the variable being taken into account is the shift in time, this will result in the finding of a pattern that is marginally different from the pattern generated from the real data.

Time Series Analysis

We use a variety of techniques, including Linear Regression, Gaussian Process, Support Vector Machine of Regression (SMOreg), and Multilayer Perceptron, to forecast time-dependent occurrences, such as time series analysis for earthquakes. Mean Absolute Error (MAE) and Root Mean Squared Error have been compared (RMSE). We then try to identify which algorithm is more accurate and efficient.

RESULTS AND DISCUSSION

Clustering

We ultimately obtained a magnitude model based on the longitude and latitude attributes after applying the two clusters to the mapper. This map gives us a clear visual of the low, medium, and high density seismic zones in NAD and its surrounding regions, which are depicted in Figures 2 and 3. Latitude is represented by the x axis, and longitude is represented by the y axis. Therefore, each dot stands for a seismic event. As we can see, the spots at longitude 99-100 and latitude 5.88-1, which correspond to the northern and northwest portions of NAD, are particularly dense. This leads us to believe that these two areas are more seismically sensitive than the others.

Cluster Evaluation

Recall, precision, and F-measure calculations are made across a pair of clustering points in order to assess the clustering outcomes. To analyze clusters, the author has employed a DT, SVM, and KNN. Out of a total of 3322 examples, the author discovered 3218 instances that were correctly classified, with a DT accuracy of 99.8796%, while SVM and KNN both displayed the same accuracy of 99.127%.

The several measures employed for cluster evaluation are displayed in Table 2. Table 3 compares the two clustering approaches with regard to cluster evaluation.

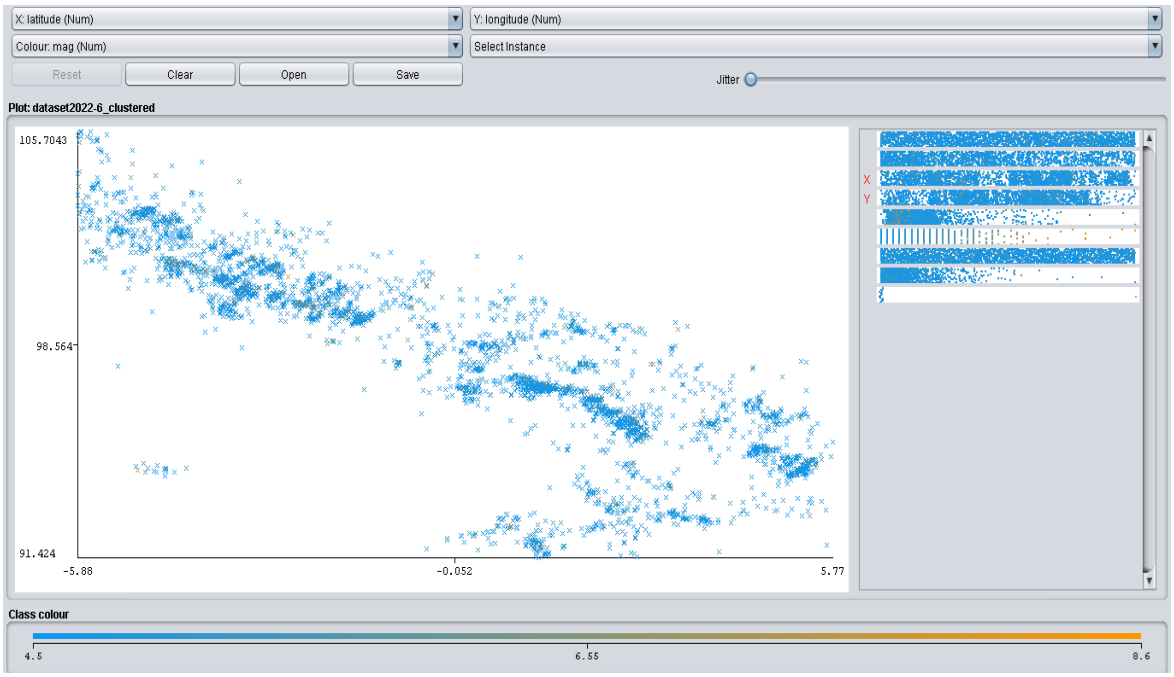

Figure 2. Clustering Using Hierarchical Clustering Algorithm

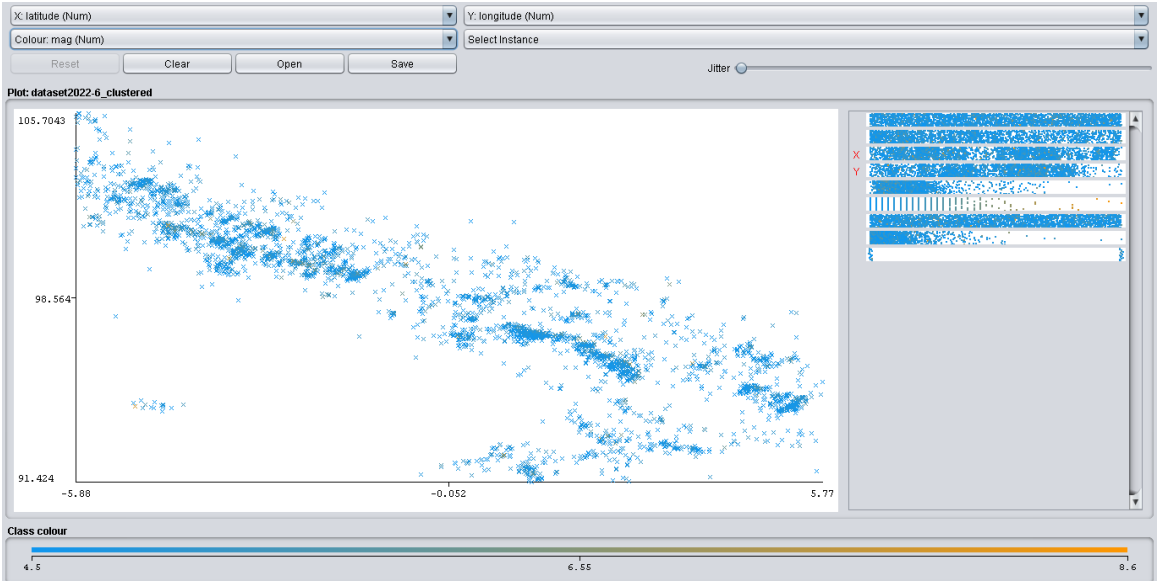

Figure 3. Clustering Using Fartherst First Algorithm

Table 2. Matrix of Classifiers for Confusion

Output Class	**Target Class**	
	Negative	**Positive**
Classifiers for negative	tn	fn
Classfiers for positive	fp	tp

$$Recall = tp/(tp + fn) \quad (4)$$

$$Precision = tp/(tp + fp) \quad (5)$$

$$Accuracy = (\sum mf - \sum af \,/ \sum af) * 100 \quad (6)$$

Where tn, fn, fp, tp, mf, and af stand for true negatives, false negatives, false positives, and true positives, respectively.

Table 3. Accuracy of Classifiers

Algorithm	Precision	Recall	F-Measure	RMSE	Accuracy
Decision Tree (DT)	0.999	0.999	0.999	0.0347	99.8796 %
Support Vector Machine (SVM)	0.991	0.991	0.996	0.0934	99.127 %
k-Nearest Neighbor (KNN)	0.999	0.999	0.999	0.0245	99.9398 %

Forecasting Future Patterns

We are able to display the actual value using all four time series approaches and visualize the anticipated value 1, 5, or 13 steps in advance. Additionally, we can say that SMOreg is the best method for predicting earthquake magnitude values with the least amount of inaccuracy.

Model Evaluation with MAE and RMSE: The magnitude of the error is measured using the Mean Absolute Error (MAE) (error). Each unique difference in MAE is given equal weight by calculating a linear score and disregarding direction. It can be stated as follows:

$$MAE = \frac{1}{n}\sum_{i}^{n}|e_i| \quad (7)$$

The Root Mean Squared Error is the primary indicator of the size of the average error (RMSE). The projected value and the actual value are squared to perform calculations on the dataset. The sample dataset is then averaged after that. These two procedures are used to identify and prevent larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad (8)$$

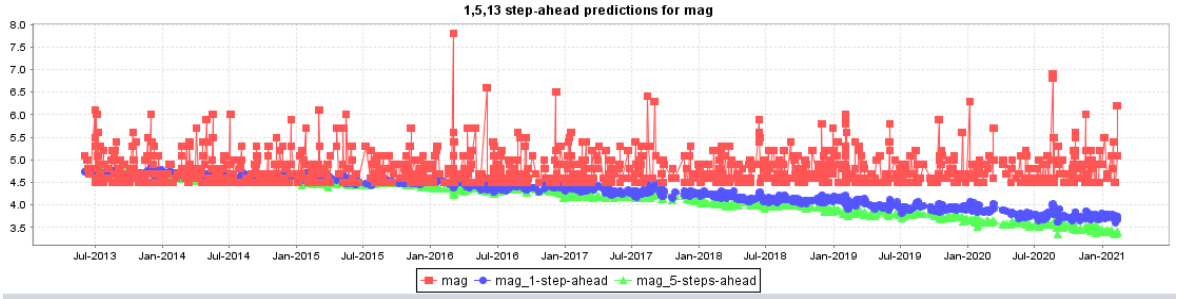

Figure 4. Prediction Test Using the Gaussian Process

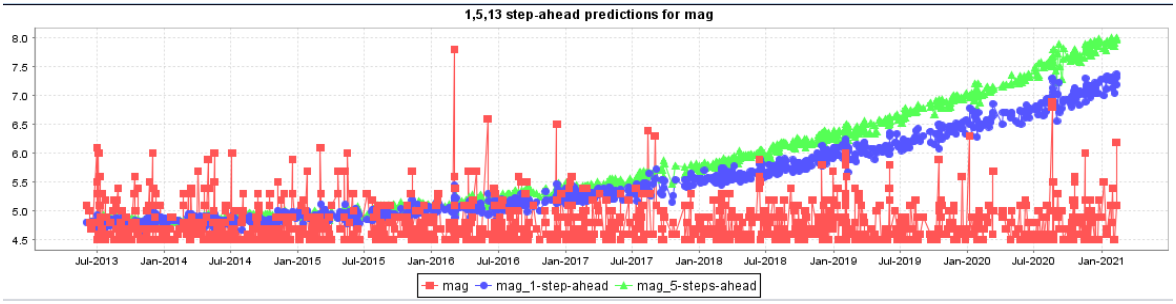

Figure 5. Prediction Test Using Linear Regression

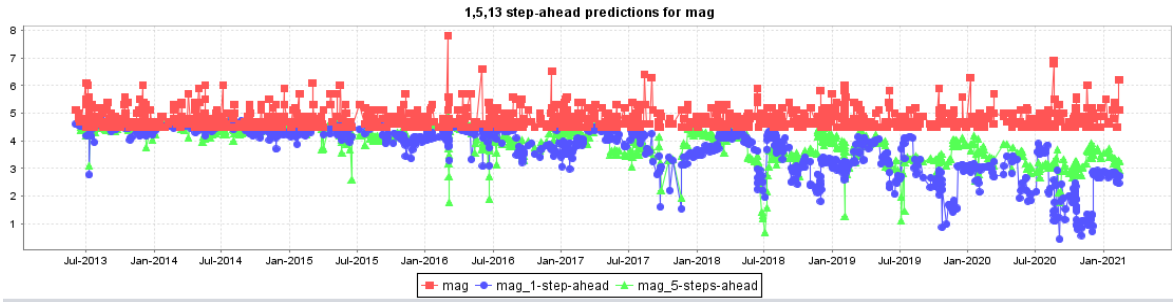

Figure 6. Prediction Test Using Multilayer Perceptron

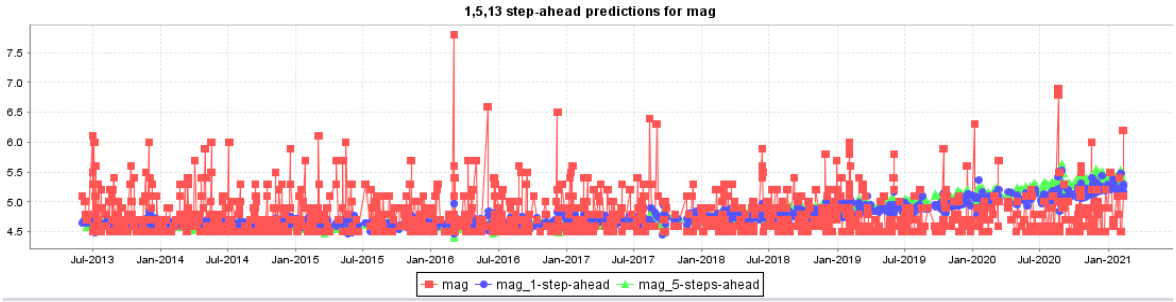

Figure 7. Prediction Test Using SMOreg

By assessing the MAE and RMSE in the four time series methods, the training and testing data were examined at 1, 5, and 13 steps forward to determine the performance of the final model. Test results acquired using the 5-fold method. Table 4 (training data phase) and Table 5 compare the MSE and RMSE values for the four time series algorithm types (testing phase).

Table 4. Error in Training Data

Algorithm	Training Data								
	1-step ahead			5-step ahead			13-step ahead		
	MAE	RMSE	Inst.	MAE	RMSE	Inst.	MAE	RMSE	Inst.
Linear Regression	0.3314	0.5242	2313	0.32	0.5321	2309	0.3014	0.5276	2302
Gaussian Process	0.2841	0.4198		0.2858	0.4208		0.2848	0.4223	
SMOreg	0.2803	0.4024		0.301	0.4275		0.3189	0.439	
Multilayer Perceptron	1.1086	1.4839		0.8771	1.1262		0.7411	1.0032	

Table 5. Error in Testing Data

Algorithm	Training Data								
	1-step ahead			5-step ahead			13-step ahead		
	MAE	RMSE	Inst.	MAE	RMSE	Inst.	MAE	RMSE	Inst.
Linear Regression	0.7591	1.0399	997	1.0165	1.3428	993	1.2917	1.6951	986
Gaussian Process	0.5424	0.705		0.6603	0.8369		0.7419	0.9217	
SMOreg	0.2803	0.4024		0.301	0.4275		0.3189	0.439	
Multilayer Perceptron	1.1086	1.4839		0.8771	1.1262		0.7411	1.0032	

Where, Inst. = instances.

The minimal value in SMOreg is represented by the root mean squared error (in Tables 4 and 5). As a result, the SMOreg algorithm has superior seismic event forecasting accuracy than others.

CONCLUSION

For a country like Indonesia, which frequently experiences earthquakes, it can be extremely challenging to precisely predict natural events like rapid changes of the earth's crust. However, there are still a lot of significant seismic occurrences in the vicinity of NAD that can have an impact here as well. Therefore, the authors took these occurrences into account and employed various data mining approaches. We can locate earthquake-prone regions in NAD by evaluating the data using the suggested model.

Future seismic occurrences that can have an impact on the NAD region can also be predicted using time series analysis. If these forecasts are more accurate, it will encourage people to take preventative measures.

REFERENCES

Absar, N., Shoma, S.N., Chowdhury, A.A.(2017). Estimating the occurrence probability of earthquake in Bangladesh. Int. J. Sci. Eng. Res. 8(2). ISSN 2229-5518. https://www.researchgate.net/publication/316437975_Estimating_the_Occurrence_Probability_of_Earthquake_In_Bangladesh

Adetiloye, T., Awasthi, A.(2017). in Handbook of Neural Computation, https://doi.org/10.1016/C2016-0-01217-2

Beitzel., Steven M. (2006). On Understanding and Classifying Web Queries (Ph.D. thesis). CiteSeerX 10.1.1.127.634. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.127.634

Dasgupta, S.; Long, P. M. (2005), "Performance guarantees for hierarchical clustering", *Journal of Computer and System Sciences*, 70 (4): 555–569, https://doi:10.1016/j.jcss.2004.10.006

Deutsche Welle Indonesia: (2014) Apa Yang Sebenarnya Terjadi Dalam Tsunami 2004?. https://p.dw.com/p/1E7Wk

Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. Proceedings of the International Conference on Language Resources and Evaluation. https://aclanthology.org/L16-1040.pdf

Dutta, P., Naskar, M., Mishra, O.P. (2011). South Asia earthquake catalog magnitude data regression analysis. Int. J. Stat. Anal. 1(2), 161–170. ISSN 2248-9959.

https://www.researchgate.net/publication/260481178_South_Asia_Earthquake_Catalog_Magnitude_Data_Regression_Analysis
Goodchild, M. (2008). Data Analysis, Spatial. In: Shekhar S., Xiong H. (eds) Encyclopedia of GIS. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-35973-1_236
Harjono, H.: (2017). Seismotektonik Busur Sunda. Jakarta: LIPI Press. http://penerbit.lipi.go.id/data/naskah1502855463.pdf
Har-Peled, S.; Mendel, M. (2006). "Fast construction of nets in low-dimensional metrics, and their applications", *SIAM Journal on Computing*, 35 (5): 1148–1184, arXiv:cs/0409057, https://doi:10.1137/S0097539704446281
Hashemi, M., Karimi, H. (2016). Seismic source modeling by clustering earthquakes and predicting earthquake magnitudes. *In: Smart City 360°*, pp. 468–478. https://doi.org/10.1007/978-3-319-33681-7_39
Hoque, S., Istyaq, S., Riaz, M.M. (2013). A clustering method for seismic zone identification and spatial data mining. Int. J. Adv. Res. Comput. Sci. Eng. Inf. Technol. 1(2). ISSN 2321-3337. https://www.researchgate.net/publication/299470942_A_Clustering_Method_For_Seismic_Zone_Identification_And_Spatial_Data_Mining
Kulkarni, A.D., More, A.: Analysis of the effect of cell phone radiation on the human brain using electroencephalogram. Orient. *J. Comput. Sci. Technol*. 9(3) (2015). http://dx.doi.org/10.13005/ojcst/09.03.07
Last, M., Rabinowitz, N., Leonard, G. (2016). Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. PLOS ONE 11(1), e0146101. https://doi.org/10.1371/journal.pone.0146101
Leonard, L.C. (2017). Advances in Computers. Cambridge: Elsevier.
Martínez-Álvarez, F., Troncoso, A., Morales-Esteban, A., Riquelme, J.C. (2011). Computational intelligence techniques for predicting earthquakes. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) Hybrid Artificial Intelligent Systems, HAIS 2011. Lecture Notes in Computer Science, vol. 6679. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-21222-2_35
Neill, S.P., Hashemi, M.R. (2018) Fundamentals of Ocean Renewable Energy, Cambridge: Elsevier. https://doi.org/10.1016/C2016-0-00230-9
Nivedhitha, U.S., Krishna, A. (2016). Development of a predictive system for anticipating earthquakes using data mining techniques. Indian J. Sci. Technol. 9(48). https://doi.org/10.17485/ijst/2016/v9i48/107976
Powers, David M W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" . *Journal of Machine Learning Technologies*. 2 (1): 37–63. https://doi.org/10.48550/arXiv.2010.16061
Sammut C., Webb G.I. (2011) Encyclopedia of Machine Learning. Mean Absolute Error. In: Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_525
Sasaki, Y. (2007). "The truth of the F-measure". https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure
Sethi I. K. (2001). Data Mining: An Introduction. In: Braha D. (eds) Data Mining for Design and Manufacturing. Massive Computing, vol 3. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-4911-3_1
Tableau (2021). http://tableau.com
Technopedia (2021). http://technopedia.com
United States Geological Survey (USGS) (2021). http://earthquake.usgs.gov/earthquakes/search/

Werner, M. J., Helmstetter, A., Jackson, D.D., Kagan, Y.Y. (2011). High-Resolution Long-Term and Short-Term Earthquake Forecasts for California. Bull. Seism. Soc. Am. August; 101: 1630–1648. https://doi.org/10.1785/0120090340

X. Li; Y.Y. Wang; A. Acero (2008, July). Learning query intent from regularized click graphs. Proceedings of the 31st SIGIR Conference. p. 339. https://doi.org/10.1145/1390334.1390393. ISBN 9781605581644. S2CID 8482989

Zaidi, F., Archambault, D., Melançon, G. (2010). Evaluating the quality of clustering algorithms using cluster path lengths. In: Perner, P. (ed.) Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2010. Lecture Notes in Computer Science, vol. 6171. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-14400-4_4