Analysis of Seismic Hazard Prediction Using Non Parametric Conic Multivariate Adaptive Regression Splines (C-Mars) Methods

Dadang Priyanto\textsuperscript{1*}, Muhammad Zarlis\textsuperscript{2}, Herman Mawengkang\textsuperscript{3}, and Syahril Efendi\textsuperscript{4}

\textsuperscript{1}Graduate Program Of Computer Science, Department Of Computer Science, Faculty Of Computer Science And Information Technology, Universitas Sumatera Utara, Medan, Indonesia.
\textsuperscript{2,3,4}Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia.

E-mail: dadangpriyanto@students.usu.ac.id

Abstract. The data mining process requires a data set that can be used in determining a number of specific patterns to gain new knowledge. Large data sets (Big data) require special methods to get effective results. Included in this study are using big data related to the earthquake in Lombok. Earthquake research, especially in Lombok, is needed because Lombok is on three active plates in Indonesia, so that the danger of earthquake damage can be minimized. Earthquake data obtained from the Geophysical Station (BMKG) of Mataram has different characteristics and is complex, an appropriate method is needed, namely by detecting a non-parametric method with the Multivariate Adaptive Regression Spline (MARS). The use of the backward stepwise algorithm with the Conical Quadratic Programming (CQP) framework of MARS, referred to as CMARS (Conic Multivariate Adaptive Regression Splines), is used for optimizing the results. The conclusions of this study are 1. A mathematical model with a total of 12 basis functions (BF) has contributed to the prediction analysis of the PGA dependent variable. 2. Contributions of the influence of independent variables on the PGA value are the epicenter distance ($R_{epi}$) of 100\% and the Magnitude (Mw) of 31.08608\% , while the temperature of the incident location (SUHU) of 5.48525\% and depth (Depth) of 3.52988\% . 3. Acquired areas that have earthquake hazard levels in the order of the most vulnerable are Malaka, Genggelang, Tegal Maja, Senggigi and Mangsit.

1. Introduction

Data which is one of the parts of information technology resources nowadays has a very important role because it cannot be separated from other technological re-sources. In the current digitalization era, the data available is very abundant or quite a lot (big data) which can be processed further to get more useful information. This data processing is included in the data mining process section. Some experts define data mining as the process of finding certain patterns and knowledge from a large amount of data (Big Data). Other experts according to Turban are data processing that uses statistics, mathematics, artificial intelligence and machine learning techniques to extract and add useful information and knowledge related to large databases. Big data can come from data warehouses, repository information, the web,
and other data that are dynamic via the internet. Some things that can be done in data mining, such as Description, Estimation, Prediction, Classification, Clustering, and Association. [3] There are several mathematical functions that can be used in the data mining process such as the functions of Association and Correlation, Classification and Regression, Classification and outlier analysis. [2, 3] This mathematical function can be used to find certain patterns as desired in the data mining process. In general, the data mining process can be grouped into two categories, namely descriptive data mining and predicted data mining. [2] Descriptive Data Extracting is the process of characterizing data properties into a target data set, while the Data Mining Prediction process is an induction on current data to make future predictions.

Many Data Mining Prediction methods can be used in forming mathematical or statistical relationships between several desired factors. One way is to approach the Multivariate Adaptive Regression Spline (MARS) method. [4] Prediction Analysis is also called Regression analysis, which is a statistical method that is widely used to investigate and model relationships between variables. [5] In Regression Analysis to estimate the regression curve can be done in two ways, namely Parametric Regression and Nonparametric Regression. Nonparametric regression is one of the approaches used to determine the pattern of relationships between predictor variables and responses that are not known for their regression curves or no complete past information about the shape of the data pattern. [6] Multivariate Adaptive Regression Spline (MARS) is one of the nonparametric regression models, which assumes the shape of the functional relationship between response variables and predictors is unknown. MARS is a complex combination of the spline method and recursive partitioning to produce estimates of continuous regression functions, and is used for prediction and classification. [7] For the management of high dimensional data, MARS has been developed into C-MARS (Conic Multivariate Adaptive Regression Splines Method) and is an effective non-parametric regression approach. [8] The flexible nature of the CMARS model can be implemented in various fields of data mining applications including in seismic hazard analysis.

Research on seismic hazards is very important to be carried out especially in areas that are in the area where the Earth’s plates meet. One of the areas in the Lombok Plateau region of Lombok that is active in Indonesia is making active areas affected by seismic activities. Earthquake data that occur in Lombok are available at the Mataram city Geophysics (BMKG) station and can also be accessed on the BMKG website. Research on earthquakes with a parametric approach has been done, but earthquake data that has parameters with uncertain nature then more precisely using the non parametric method approach. Several studies have been conducted such as, (Panakkat and Adeli, 2007), with the Neural Network approach, [16] (Kannan, 2015) with a mathematical innovation model, [17] (Yerlikaya, et al, 2014, 2016) with a robust computational approach, [10], [13] (Asim et al, 2018) with the Support of Vector Regressors and Hybrid Neural Networks approach, [18] (Zubaidah et al, 2014) by detecting changes in the earth’s electromagnetic fields. (Yazici et al, 2015) using a non parametric regression approach to bootstrapping CMARS, [19] and (Yerlikaya et al 2014) with a non-parametric adaptive regression approach. [4]

2. Method

Data sets are mostly discrete, and regression modeling usually uses a discrete approach. The problem of discrete data sets can be solved by a parametric approach with a linear regression model, but the data set in this study related to earthquake data has different and complex character traits, more precisely using a more effective method that is the non parametric approach. Non-parametric approach with Multivariate Adaptive Regression Spline (MARS) and to optimize the model using the Conic Multivariate Adaptive Regression Spline (C-MARS) method. In general, the data set in this study is used to build the relationship between the dependent variable and the independent variable. Non-parametric regression approaches can
Generally be written as follows: \[ \text{yi} = (\text{xi}) + E_i \] (1)

With \( \text{yi} = \) the response variable on observation \( i \),

\( (\text{xi}) = \) Vector predictor function

\( E_i = \) Is a free error \( i \)

For functions \( f(\text{xi}) \) with each component input \( \text{xi} \), where \( \text{p} \) is the knots dimension \( \tau = (\tau_{i,1}, \tau_{i,2}, ..., \tau_{i,p})^r \), \( (i = 1,2, ..., N) \) each input vector \( \text{xi} = (\text{xi}, 1, \text{xi}, 2, ..., \text{xi}, p)^r \)

Furthermore, the nonparametric regression function commonly used is splines regression which is a linear piecewise function whose form is as follows:

\[ c^+(x, \tau) = \left[ + (x - \tau) \right]_+, \quad c^-(x, \tau) = \left[ - (x - \tau) \right] \] (2)

Where \([q]_+ := \max(0; q) \) and \( \tau \) are univariate knot, so the base function becomes:

\( (x) = f(x_j - \tau) +: (\tau - x_j) + | \tau \epsilon \{ X_1, j, X_2, j, ..., xN, j, j \epsilon \{ 1, 2, ..., p \} \} \) (3)

The function \((x)\) is a representation of equation (1) which involves intercepts and linear combinations so that the model generally becomes: \[ y = \theta_0 + \sum_{m=1}^{M} \theta_m \beta_m(\text{xi}) + \varepsilon \] (4)

Where \( \beta_m \) with \( (m = 1,2, ..., M) \) is the basis of the function or product of equation (3). For \( \theta_m \) is the basis function parameter to \( m \) \( (m = 1,2, ..., M) \) or a constant value \( (m = 0) \). The interaction of basis functions is obtained by multiplying the basis functions by the truncated linear functions by involving new predictors. For the data provided \((\tilde{x}_i, \tilde{y}_i) \) \((i = 1,2, ..., N)\) with the base function \( m \) being: \[ \beta_m(\text{xi}) = K_m \prod_{j=1}^{r} \left[ S_{kj}^m (X_{kj} - \tau^m_{kj}) \right]_+ , \] (5)

Where \( K_m \) is the number of truncated linear functions multiplied in basis functions to \( m \). For \( X_{kj}^m \) is an input variable that corresponds to \( j \), and is related to a truncated function on the \( m \) to function base. \( \tau^m_{kj} \) is the value of knots in \( X_{kj}^m \) and \( S_{kj}^m \) is is +/-.

According to Friedman, MARS has two algorithms that must be passed for computational calculations, namely the first is the Forward Stepwise Algorithm in which the base function (BF) is chosen to minimize the base function that lacks according to criteria until the maximum number of base functions (BF) specified by the user, \( M_{\text{max}} \), is reached. The second algorithm, Backward Stepwise Algorithm, which is to overcome weaknesses in the basic stepwise functions that contribute the smallest amount of residual squared error is eliminated, thus making the model simpler. Of the two algorithms, MARS in the selection of variables uses Generalized Cross-Validation (GCV). [12] in [10] The C-MARS model as a form of backward stepwise development of the algorithm of the MARS model with Penalized Residual Sum of Squares (PRSS) which is described as follows: \[ \text{PRSS} = \sum_{1=1}^{N} (\tilde{y}_i - f(\text{xi}))^2 + \sum_{m} \lambda m \left| \sum_{|\alpha| = 1}^{2} \int_{r < s} \theta^2 \left[ D_{\alpha}^m (\tau^m_{\alpha}) \right] \right| d\tau^m \] (6)
In this equation, the set element $V(m) = (K_j^m) | j = 1, 2, ..., K_m$ calculate variables related to Function Base to m. $\beta_m, and \ t_m = (t_1, ..., t_mK_m)^T$ represents the variable predictor vector which gives the basis function to m. Next:

$$D^\alpha_{r,s} \beta_m(m) \frac{\partial \mid \alpha \beta_m \mid}{\partial \alpha l_{m}^r \partial \alpha 2_s^r}(m), \quad (7)$$

For $\alpha = (\alpha_1, \alpha_2)^T$, $|\alpha| := \alpha_1 + \alpha_2$ where $\alpha_1, \alpha_2 \in [0, 1)$

In equation (6), there are two parts of PRSS that are related to accuracy and complexity using the $\lambda$ penalty parameter. The integral symbol " $\int$ " indicates the dummy, which is $\int Q_m$ where $Q_m$ is a number of parallel-pipe dimensions of integrated $K_m$. Because the integrals in equation (6) are multi-dimensional and difficult to evaluate, the integrals in the PRSS equation are set to:

$$PRSS \approx \sum_{i=1}^{N} (y_i - \theta^T \beta(d_i))^2 + \sum_{m=1}^{M_{max}} \lambda_m \theta_m^2 \sum_{i=1}^{(N+1)K_m} \left( \sum_{\alpha = (\alpha_1, \alpha_2)^T}^{2} \sum_{r<s} |D^\alpha_{r,s} \beta_m(X_i^m)|^2 \right) \Delta X_i^m, \quad (8)$$

Where $\beta(d_i)(1, \beta_1(x_i^M)_1, \beta M + 1(x_i^{M+1})_1, ..., \beta M_{max}(x_i^{M_{max}})_1)^T, \theta := (\theta_0, \theta_1, ..., \theta_{M_{max}})^T$ with point $\tilde{d}_i := (\tilde{x}_i^1, \tilde{x}_i^2, ..., \tilde{x}_i^M, \tilde{x}_i^{M+1}, ..., \tilde{x}_i^{M_{max}})^T$ with argument $(\sigma^K_j)_{j \in \{1,2,...,p\} \in \{0,1,2,...,N+1\} K_m}$ and

$$\tilde{x}_i^m = \begin{pmatrix} \tilde{x}_i^1 \\ l^{K_m^i}_1 \\ \sigma^{K_m^i}_1 \cdot \tilde{x}_i^{K_m^i} \end{pmatrix}_{K_m^i}, \Delta \tilde{x}_i^m := \prod_{j=1}^{K_m^i} \begin{pmatrix} \tilde{x}_i^m \\ l^{K_m^i}_j \\ \sigma^{K_m^i}_j \cdot \tilde{x}_i^{K_m^i} \end{pmatrix}_{K_m^i} \quad (9)$$

(Equation (8) can be simplified by using $\lambda$ penalization on each derivative. Furthermore, the PRSS equation with Tikhonov regularization problem is as follows:

$$PRSS \approx \|(y - \beta(d))\theta\|^2 + \lambda \|(L)^2, \quad (10)$$

Can be reformulated into Conic Quadratic Problem (CQP) as follows [14, 15] in [10]

$$\min_{t,\theta} t, \quad (11)$$

subject to $\|\beta(d)\theta - y\|_2 \leq t,$

$$\|L\theta\|_2 \leq \sqrt{M}$$

with $t \geq 0$

In equation (11) optimization can be solved by the interior point method (IPMs). There are several solutions to the problem of different M values, one of which is closest to the angle of the efficient curve or (L) with $\|L\|_2$ plotted versus $\|y - \beta(d)\theta\|_2$. [5, 13]
3. Use of Data Sets
This study uses earthquake data sets taken from the Mataram city Geophysics station, in the time span from 2010 to 2019. Data were taken as many as 8,053 records with magnitudes of 1 Mw to 7 Mw. Location coordinates are coordinates (-4.0636o) LS - (-13.0636o) LS and (111.5798o) BT - (120.5798o) BT. The data set is in the form of an earthquake catalog so empirical calculations need to be done to get the value of the epicenter distance, and Peak Ground Acceleration (PGA). The prediction analysis process requires processing of the data set by filtering and classification based on the provisions of magnitude of more than 4.5 Mw, depth of less than 300 Km, and also the distance to the location of the less than 500 km. the data outside the provision is less impactful due to earthquake, or even not felt at all. The magnitude magnitude data processing results can be classified based on the number of events as in table 1 below:

| No | Magnitude (Mw) | Frequency |
|----|----------------|-----------|
| 1  | 4.5 – 5        | 283       |
| 2  | 5 – 6          | 121       |
| 3  | 6 – 7          | 15        |

The pattern of earthquake spread in Lombok based on magnitude and epicenter distance gives a wide spread at various depths as shown in Figure 1 below:

**Figure 1.** Distribution of Earthquakes in Lombok with Magnitude 4.5 - 7 from 2010 to 2019.

Peak Ground Acceleration (PGA) values that will be used in the prediction analysis process need to be done empirically using the Joyner and Boore attenuation function equations. The Joyner and Boore Attenuation function as shown in equation (12) follows:

\[
PGA(gal) = 10(0.71 + 0.23(M - 6) - \log(r) - 0.0027r)
\] (12)
Furthermore, the results of pre-processing data with classification and calculation of the attenuation function. The next step is sorting and determining the data variables in two categories, namely dependent and independent variables. This research has determined the dependent variable data is Peak Ground Acceleration (PGA), and the independent variables are Magnitude ('MW'), Depth('Depth'), Epicenter Distance ('Re,pi'), and earth-quake location temperature ('SUHU').

4. Results and Discussion
Regession analysis with a non-parametric model approach using C-MARS needs to be done solving problems with the approach of MARS through two stages. The first stage is the completion of the Forward Stepwise algorithm and the second stage is the Backward Stepwise algorithm. The first step is to do a combination of inputs by determining the number of base functions (BF) and input of the maximum interaction (MI) and the minimum number of observations (MO) to get the best MARS model. The second step is to choose the basis function (BF) which does not contribute to the best MARS model, which is deleted or removed to simplify the model formed. From the trial results obtained the results of a combination of input BF, MI, and MO with a minimum GCV value of 0.00000, minimum MSE of 0.00000, and the largest R2 (Square) of 0.99723, so based on a combination of the results of the BF, MI, and MO can be determined the best model of MARS. After going through the backward stepwise algorithm stage a mathematical model is obtained based on the relationship between the dependent variable and independent variable relationships. A number of 12 basis functions (BF) are formed, namely BF 1, 2,3,5,7,9,10,11,13,14,15 and 16. Base functions that have no contribution are removed or deleted, namely the base function (BF) 4, 6, 8, and 12 so as to obtain a mathematical model for the Peak Ground Acceleration (PGA) variable as in equation (13) follows:

\[
Y(PGA) = -0.0175733 - 0.00211487 \times BF1 + 0.0029936 \times BF2 + 0.000556472 \times BF3 + 0.00172513 \times BF5 + 0.000369563 \times BF9 - 0.000160793 \times BF10 - 0.000689482 \times BF11 + 0.0003626173 \times F13 + 0.00329239 \times BF14 - 0.00125948 \times BF15 + 6.46282e - 05 \times BF16
\]

Based on the level of interrelation of the independent variables can be explained as in table 2 below:

| Variable | Importance | GCV  |
|----------|------------|------|
| R_EPI    | 100.00000  | 0.00067|
| MW       | 31.08608   | 0.00007|
| SUHU     | 5.48525    | 0.00000|
| DEPTH    | 3.52988    | 0.00000|

It can be seen in Table 2 that the variables that are very influential in the PGA value are the epicenter distance (Re,pi) by 100% and the Magnitude (Mw) by 31.08608%, while the location temperature (SUHU) is 5.48525% and depth (Depth) amounted to 3.52988%.
Based on empirical calculations and classification of data sets, an overview of the regions that have the highest potential for earthquake hazard in Lombok can be obtained as shown in table 3 below.

| No | Time     | Lat  | Long | Depth | Mw | R-epi | PGA (g) | PGV (cm/s) | SUHU (°) | Area Location |
|----|----------|------|------|-------|----|-------|---------|------------|----------|---------------|
| 1  | 22-06-2013 | -8.44| 116.04| 16    | 5.2| 14.42381995 | 0.183715166 | 1.832770384 | 26.7     | Malaka        |
| 2  | 09-08-2018 | -8.36| 116.22| 12    | 6.2| 27.39175594  | 0.16732396  | 1.814869245 | 24.9     | Genggelang    |
| 3  | 05-08-2018 | -8.41| 116.16| 17    | 5.5| 19.22254717  | 0.16068049  | 1.836118384 | 24.9     | Tegal Maja    |
| 4  | 31/03/2016 | -8.52| 115.99| 12    | 4.5| 10.60668207  | 0.163065868 | 1.830562422 | 27.5     | Senggigi      |
| 5  | 06-08-2018 | -8.42| 116.03| 23    | 5  | 16.8807379  | 0.143937973 | 1.829356187 | 24.9     | Senggigi      |
| 6  | 04-05-15  | -8.43| 116.03| 13    | 4.6| 15.83302429  | 0.123357614 | 1.830810815 | 26.3     | Mangsit       |

5. Conclusion
This research was conducted an empirical calculation of data sets and training data has been carried out using non-parametric MARS and CMARS methods to obtain the following conclusions:

1. A mathematical model with a total of 12 basis functions (BF) contributes to the prediction analysis of the PGA dependent variable.
2. Contributions of the infl of independent variables on the PGA value are the epicenter distance \((R_{epi})\) by 100% and the Magnitude (Mw) by 31.08608%, while the location temperature (SUHU) is 5.48525% and the depth (Depth) is 3.52988%.
3. Obtained areas that have earthquake hazard levels in the order of the most vulnerable are Malacca, Genggelang, Tegal Maja, Senggigi and Mangsit. The area is located in three regencies namely North Lombok, West Lombok and part of Mataram City.

References

[1] Han, J., Kamber, M., Pei, J. 2012, Data mining : concepts and techniques, Morgan Kaufmann, 225Wyman Street, Waltham, MA 02451, USA
[2] Turban, Efram, Aronson, Jay E, dan Peng-Liang, Ting, 2005, Decision Sup- port Systems and Intelligent Systems, pearson
[3] Larose, D.T, 2005, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley-Interscience, John Wiley & Sons, Inc., Hoboken
[4] Yerlikaya, F., Askan, A., Weber, G.W., 2014, An alternative approach to the ground motion prediction problem by a non-parametric adaptive regression method, Engineering Optimization, Vol. 46, No. 12, 1651–1668.
[5] Weber, G.W., I. Batmaz, G. Köksal, P. Taylan, and F.Yerlikaya-Özkurt. 2012. “CMARS: A New Contribution to Nonparametric Regression with Multivari- ate Adaptive Regression Splines Supported by Continuous Optimization.” Inverse Problems in Science and Engineering 20 (3): 371–400.
[6] Eubank, R.L., 1999, Nonparametric Regression and Spline Smoothing, Sec- ond Edition, Marcel Dekker, New York.
[7] Friedman, J.H., 1991, Multivariate Adaptive Regression Spline (With Dis- cussion), The Annals of Statistics, Vol. 19, hal. 1-141.
[8] Weber, G.W., Cevik, A., 2018, Voxel-MARS and CMARS: Methods for Early De- tection of Alzheimer’s Disease by Classification of Structural Brain MRI, https://www.researchgate.net/pu [9] Putra, J,W,G, 2018, Pengenalan Konsep Pembelajaran Mesin Dan Deep Learning, https://wiragotama.github.io/ [10] Yerlikaya, F., Batmaz, I., Weber, G.W., 2014, A Review and New Contribu- tion on Conic Multivariate Adaptive Regression Splines (CMARS): A Powerful Tool for Predictive Data Mining, In book: Modeling, Dynamics, Optimization and Bioeconomics I Edition: Springer Proceedings in Mathematics & Statistics Volume 73, 2014 Chapter: 38 Publisher: Springer Verlag.
[11] Weber, G.W., I. Batmaz, G. Köksal, P. Taylan, and F.Yerlikaya-Özkurt. 2012. CMARS: A New Contribution to Nonparametric Regression with Multivariate Adaptive Regression Splines Supported by Continuous Optimization, Inverse Problems in Science and Engineering 20 (3): 371–400.
[11] P. Craven and G. Wahba, Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, Numerische Mathematik 31, 1979, pp. 377–403

[12] F. Yerlikaya, A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Application to Data Mining for Quality Control in Manufacturing, MSc., Middle East Technical University, 2008

[13] A. Nemirovski, 2020. A lectures on modern convex optimisation, Israel Institute of Technology. Available at http://iew3.technion.ac.il/Labs/Opt/opt/LN/Final.pdf.

[14] P. Taylan, G.W. Weber and A. Beck, 2007. New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology, Journal Optimisation 56, pp. 675–698.

[15] Panakkat, A. and Adeli, H. 2007 Neural Network Models For Earthquake Magnitude Prediction Using Multiple Seismicity Indicators, vol. 17, no. 1, pp. 13–33

[16] Kannan, A. 2015. Innovative Mathematical Model For Earthquake Prediction. https://www.researchgate.net/publication/280764638

[17] K. M. Asim, A. Idris, and T. Iqbal, 2018, Earthquake prediction model using support vector regressor and hybrid neural networks, pp. 1–22.

[18] Yazici C, Yerlikaya, F.O, Batmaz I, 2015. A computational approach to nonparametric regression: bootstrapping CMARS method. https://www.researchgate.net/publication/277907298