AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

Estimating Earthquake Strength Using Multivariate Adaptive Regression (MARS)

Domoaa Mohamed Radi

Master's student, College of Administration and Economics, University of Basra, Iraq

Email : domoaamohamed@gmail.com

PROF.Dr .Sahera Hussein Zain AL-Thalabi

Administration and Economics College, Basra University, Iraq

Email: sahera.zain@uobasrah.edu.iq

Abstract

In this study, the Multivariate Adaptive Regression Spline (MARS) model was used to predict earthquakes using a database of 18,727 seismic records Collected from earthquake tracking systems. The study examined the influence of explanatory variables (earthquake depth, distance, longitude, and latitude) on earthquake intensity (the dependent variable). The results demonstrated that the MARS model has high predictive efficiency and accuracy, thanks to its ability to detect complex, nonlinear relationships in large datasets. Forecasting using MARS is a measure to save lives and enable communities to plan, respond, and prepare for future disasters, in addition to providing early warnings of earthquakes as they occur.

Keywords: Earthquakes, adaptive multivariate regression, generalized cross-validation, nonlinear relationships.

Introduction: One of the most prominent modern technologies for information systems is artificial intelligence (AI). AI focuses on understanding, studying, and simulating the nature of human intelligence, such as adaptation, learning, and logical reasoning. It plays a fundamental role in solving difficult problems and making important decisions in various fields of research and application, using advanced models and algorithms. The study aimed to utilize one of the branches of AI that has received widespread attention in scientific research: machine learning. Machine learning is one of the most important methods for building systems capable of learning from data and improving their performance over time without the need for explicit programming or instruction. This branch is an effective tool for analyzing complex and nonlinear data, such as data related to natural phenomena like earthquakes, due to its dynamic and nonlinear nature, which is difficult to handle using traditional methods.[9]

Previous studies

- The study (Friedman, 1991) This study introduced a new technique, the multivariate fit partitioned regression function, for data analysis, to overcome the traditional limitations of conventional statistical analysis methods. The aim of the study was to develop a flexible and adaptive approach that can automatically and efficiently capture nonlinear relationships and interactions between variables. It was concluded that the MARS model is a powerful and useful tool for data analysis, and can contribute significantly to understanding and predicting complex and nonlinear phenomena in several different research and application fields. [5]
- The study (Friedman and Roosen, 1995) aimed to explore and explain the MARS method and how to apply it to the analysis of multidimensional data with traditional methods. This study also included a comparison with statistical analysis and highlighted its benefits and ability to deal with high-dimensional data. It was concluded that the ability of MARS to accommodate a wide range of types of predictors naturally, including continuous and categorical variables, as well as the ability of MARS to deal with overlapping variables and adjust for missing values without discarding data, all of this reflects the main benefits and capabilities of the MARS model in analyzing data and providing understandable and accurate predictive models. [6]
- The study (York and Eaves, 2001) used MARS to analyze genetic and non-genetic factors associated with common diseases, using simulated data from a genetic analysis workshop, The study also identified variables that influence disease status through a comprehensive analysis of a set of independent variables. The results showed that the MARS model could identify important variables such as age and specific genetic variants as the best predictors of disease

status. Specific candidate gene loci were identified as factors influencing disease status, Interactions between quantitative traits were also identified, suggesting that genetic factors may play a complex role in determining disease status. MARS also uses cross-validation techniques to reduce the risk of overfitting, thereby increasing model reliability.[12]

Multivariate Adaptive Regression Spline : [3][8][11]

MARS is a form of regression analysis introduced by Friedman in 1991. It is a nonparametric regression model characterized by speed, accuracy, and flexibility. MARS automatically detects nonlinear relationships between independent and dependent variables in statistical models and also identifies interactions between them. MARS generates basis functions for each independent variable, and by incorporating the extracted basis functions, it creates flexible regression models to predict the output parameters. MARS is particularly useful in predictive modeling because of its ability to handle complex, nonlinear patterns in data, which traditional linear models find difficult to capture. MARS model is built as follows:

$$\hat{y} = B_0 + \sum_{m=1}^{M} B_m h_m(x) \quad \ldots (1)$$

:Whereas

$\hat{y}$: represents the dependent variable predicted by the model MAR.

B_0: Represents the constant

$h_m\ (x)$:Represents the basis function .

B_m.: represents the coefficient of the basis function of order m.

M: represents the number of basis functions (BFs) in the model.

Building the MARS Model :[2][5][10]

1- Forward Stage: This stage begins with a simple model containing only one parameter, the constant term. Pairs of basis functions are then added to the model, one by one, until the maximum error, the sum of squared residuals, is reached.

2- Backward Stage: The model is shortened by repeating the previous steps. In this stage, the least contributing basis functions are removed to achieve a balance between goodness of fit and simplicity of the model. The performance of the extracted submodels is then compared using generalized cross-validation (GCV) until the best submodel is found.

Generalized Cross-Validation (GCV) : [1][4][10]

It is a criterion used to select the best model by balancing model complexity with its fit to the data. This criterion helps avoid overfitting , i.e., reduce prediction error by penalizing models with underlying functions. The GCV is calculated for different MARS models by adding or removing underlying functions, which helps determine the optimal number of underlying functions. MARS attempts to process data one step forward in a two-dimensional space using piecewise linear regression lines. After this step, a backscattering process is performed and the model is rerun. In the pruning step, factors that contribute less to the model are gradually removed using GCV until the best submodel is found. It is defined by the following formula.

$$\mathbf{GCV(\ M)} = \frac{\frac{1}{n}\sum_{i}^{n}[y_i-\hat{f}_M(X_i)]^2}{[1-P(M)^*/\ n\ ]^2} \quad \ldots(2)$$

Whereas:

n: represents the number of observations.

$\boldsymbol{y_i}$: represents the response variable.

$\boldsymbol{\hat{f}_M}$: represents the MARS estimation model based on the basis function.

$\boldsymbol{X_i}$: represents the vector of the independent variable, $X_i = \left(X_{i1}، \ldots \ldots \ldots X_{ip}\right)^T$.

$\boldsymbol{P(M)^*}$: represents the complexity function.

Statistical analysis:

This total data (ALL) was divided into two data sets: training data and testing data, with 75% for the training set (14,045 observations) and 25% for the testing set (4,682 observations), as shown in the following table:

Table (1): Data division into training and testing

Data Group	Sample Size
Training	14045
Testing	4682
ALL	18727

Estimation Using the MARS Model:

When using the multivariate adaptive regression (MARS) method with the generalized crossover (GCV) method which gave a smoothing parameter equal to (0.42713) based on the training data, and when using that and depending on the trained model, it was applied to the test data, we obtained the results of estimating the dependent variable as follows:

Table (2): Estimated data using the MARS model

1	2.36951	11	1.79163	21	2.62066	31	2.64900	41	2.84290
2	2.45611	12	1.78379	22	2.84669	32	2.72419	42	2.84153
3	4.81644	13	1.85665	23	2.62984	33	2.69125	43	2.75269
4	2.25735	14	1.78389	24	2.48088	34	2.73200	44	2.76035
5	1.75718	15	1.86860	25	2.35622	35	2.68581	45	2.69768
6	2.09649	16	1.89976	26	2.52489	36	2.99454	46	2.49819
7	1.65372	17	1.29228	27	1.83636	37	2.76818	47	2.75607
8	1.69202	18	3.03257	28	1.56727	38	3.15113	48	2.62884
9	1.80035	19	2.53538	29	1.85858	39	2.69864	49	3.02717

10	1.80410	20	2.52909	30	1.75528	40	2.81043	50	2.51065
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4633	2.02316	4643	2.93041	4653	2.67420	4663	2.61235	4673	2.70640
4634	1.70361	4644	2.89876	4654	2.60272	4664	2.65889	4674	2.64652
4635	4.98669	4645	2.76908	4655	2.74630	4665	5.06722	4675	3.74722
4636	1.85065	4646	2.77833	4656	2.64633	4666	2.53621	4676	2.72579
4637	2.57086	4647	2.67023	4657	2.82321	4667	3.15669	4677	2.61726
4638	2.85182	4648	3.08183	4658	2.59245	4668	2.15676	4678	2.34514
4639	2.67147	4649	3.04814	4659	2.68085	4669	1.91840	4679	2.43073
4640	2.66683	4650	2.80939	4660	2.60621	4670	1.72083	4680	2.35035
4641	2.60754	4651	2.68722	4661	2.59560	4671	1.96603	4681	1.48680
4642	2.64870	4652	3.14318	4662	2.48195	4672	1.54819	4682	1.74847

Partial estimated values representing the effect of each basis function on predicting the values of the dependent variable for each observation were extracted. These values express the components entering into the composition of the final prediction of the dependent variable, as follows:

Table (3): Effects of basic functions

No.	Depth	Latitude	Longitude
1	15.74567	-0.17891	0.46887

2	15.74567	-0.25566	0.46510
3	15.69936	-0.22658	0.50393
4	15.83503	-0.40736	-0.70208
5	16.07149	-0.30147	0.52606
6	15.76111	-0.25450	0.44800
7	15.76111	-0.53847	0.46637
8	15.71480	-0.52636	0.57784
9	15.71480	-0.49279	0.64296
10	16.21124	-0.29124	0.51187
⋮	⋮	⋮	⋮
14040	15.93173	-0.21148	-0.50578
14041	15.73024	-0.27611	0.53734
14042	15.83586	-0.16872	0.47796
14043	16.04311	-0.26617	0.51814
14044	15.76111	-0.19868	0.48265
14045	15.71480	-0.04740	0.43171

It's worth noting that the distance variable did not appear in the output of the basis functions in the MARS model. This is because the model does not automatically retain all original variables. We note that distance did not significantly improve the model's performance according to the mean sum of squares (SSE) criterion, so it was automatically excluded during the model pruning process. This reflects MARS's flexibility in selecting only important variables.

The following figure represents a graph of the actual test data with what was estimated using the MARS model.

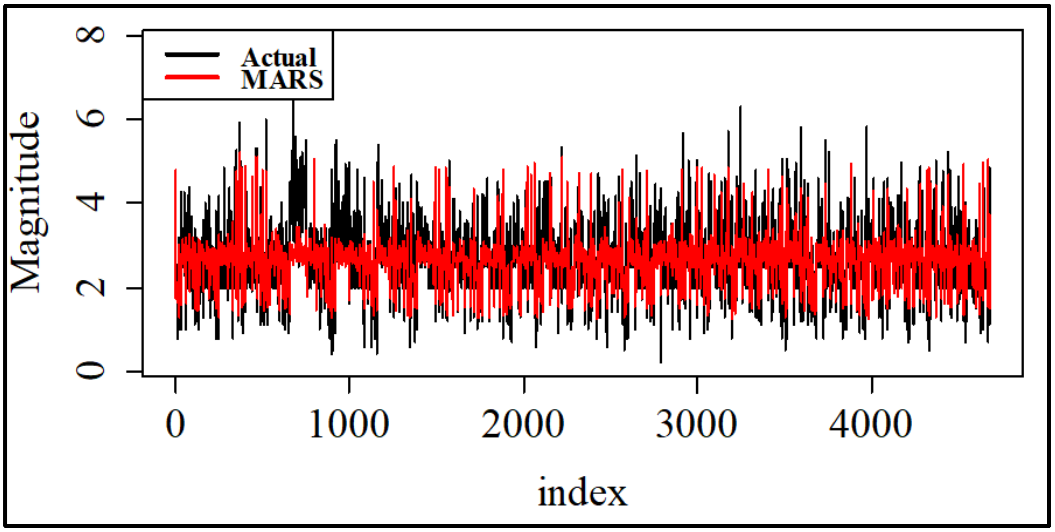

Figure (1): Data Estimated Using the MARS Model

The figure above shows a trade-off between the actual values of the response variable (in black) and the values estimated using MARS (in red) across all test observations. The red lines intersect and approach the actual values at most locations, demonstrating the effectiveness of MARS in capturing the underlying patterns in the data despite its variability and the presence of some outliers.

Conclusions:

The results showed that MARS is an effective tool for discovering complex and nonlinear relationships in data. It also proved to be the most effective model for handling large amounts of data , with a high ability to explain variance , as well as its efficiency and accuracy in estimation. This makes it a powerful model for analyzing earthquake data . MARS can contribute to providing more accurate predictions, improving earthquake and tremor forecasting, and increasing the effectiveness of early warning systems.

Sources:

1. Abd El-Sadah, H., Al-Thalabi, S. H. Z. (2023). Modeling and Forecasting Using Support Vector Regression and Chaotic Algorithms/Applied Study, University of Basra, Faculty of Administration and Economics, Department of Statistics, Iraq. 20 Journal of Theory, Mathematics and Physics Vol. 3, No. 12, 2024 ISSN: 2181-4376.
2. Al-Abadi, F., O., E., Al-Thalabi, S. H. Z., (2024). "using simulation to estimate a fuzzy regression model", Administration and Economics College, Basrah University, Iraq, NO.12, VOL.01, ISSN (E): 2347-6915.
3. Al-Lami, N. S. H. (2022). Adaptive estimators in nonparametric regression with an applied study (Master’s thesis, University of Basrah, College of Administration and Economics, Department of Statistics).
4. Çanga, D. (2022). Use of MARS Data Mining algorithm based on training and test sets in determining carcass weight of cattle in different breeds. Journal of Agricultural Sciences, 28(2), 259-268
5. Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19(1).
6. Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. Statistical methods in medical research, 4(3), 197-217.
7. Haghiabi, A.H. (2016). "Prediction of energy dissipation on the stepped spillway using the multivariate adaptive regression splines." Scientia Iranica, Transactions A: Civil Engineering, 24(6), 1789-1801.
8. Hassan, N., S., Al-Thalabi, S. H. Z. (2022). " nonparametric adaptive smoothing with practical application", Department of Statistics, College of Administration and Economics, Basra University, Basra, Iraq, NO.3, VOL.9.
9. Jarrah, N. B. (2019). Artificial intelligence techniques for developing statistical machine learning. Iraqi Journal of Information Technology, 9(3), 41-57.
10. López, F. A., & Kholodilin, K. A. (2020). The MARS algorithm in the spatial framework: non-Linearities and spatial effects in hedonic models..
11. Samadi, M., Jabbari, E., Azamathulla, H. M., & Mojallal, M. (2015). Estimation of scour depth below free overfall spillways using multivariate adaptive regression splines and artificial neural networks. Engineering Applications of Computational Fluid Mechanics, 9(1), 291-300.

12. York, T. P., & Eaves, L. J. (2001). Common disease analysis using multivariate adaptive regression splines (MARS), Genetic Epidemiology, 21(S1), S649-S654.