Study on Optimization of Mars Model for Prediction of Pile Drivability Based on Cross-Validation

C.Z. Wu[1], Anthony T.C. Goh[2], and W.G. Zhang[1*]

[1]School of Civil Engineering, Chongqing University, Chongqing, China 400045.
E-mail: cheungwg@126.com
[2]School of Civil & Environmental Engrg., Nanyang Technological University, Nanyang Avenue, Singapore 639798.
E-mail: ctcgoh@ntu.edu.sg

Abstract: Piles are long, slender structural elements used for transfering the loads from the superstructure through weak strata onto stiffer soils or rocks. For driven piles, the impact of the piling hammer induces compression and tension stresses in the piles. Hence, an important design consideration is to check that the strength of the pile is sufficient to resist the stresses caused by the impact of the pile hammer. Due to its complexity, pile drivability lacks a precise analytical solution with regard to the phenomena involved. This paper investigates the use of a fairly simple nonparametric regression algorithm known as multivariate adaptive regression splines (MARS), to approximate the relationship between a series of inputs and dependent maximum compressive stress (MCS) response, and to mathematically interpret the relationship between the various parameters. For obtaining the model of superior generalization ability and better persuasiveness results, the 10-fold cross-validation method and Bayesian information criterion are adopted to select the optimal model. This paper demonstrates that the MARS algorithm is capable of producing simple, accurate and easy-to-interpret models and estimating the contributions of the input variables.

Keywords: Multivariate adaptive regression splines; pile drivability; cross-validation; Bayesian information criterion; optimal model.

1 Introduction

For driven piles, the impact of the piling hammer induces compression and tension stresses in the piles. Hence, an important design consideration is to check that the strength of the pile is sufficient to resist the stresses caused by the impact of the pile hammer. One common method of calculating driving stresses is based on the stress-wave theory (Smith 1962) which involves the discrete idealization of the hammer-pile-soil system. As the conditions at each site is different, generally a wave equation based computer program is required to generate the pile driving criteria for each individual project. However, this process can be rather time consuming and requires very specialized knowledge of the wave equation program. Actually, for nonlinear and multidimensional geotechnical problems like this, soft computing techniques based on a large database are generally adopted to capture the intrinsic relationship.

This paper explores the use of promising procedure known as multivariate adaptive regression splines (MARS) (Friedman 1991) to develop models for multivariate geotechnical problems. No prior knowledge of the form of the function is required in MARS. The main advantages of MARS are its capacity to extract the complex data mapping in high-dimensional data and produce simple, easy-to-interpret models, and its ability to estimate the contributions of the input variables. Previous applications of MARS algorithm in civil engineering include modeling doweled pavement performance (Attoh-Okine et al. 2009), predicting shaft resistance of piles in sand (Lashkari 2013), estimating deformation of asphalt mixtures (Mirzahosseini et al. 2011), determining the undrained shear strength of clay (Samui and Karup 2011), predicting surface settlement associated with tunneling operation (Adoko et al. 2013), estimating building energy performance and lateral spreading induced by earthquakes (Cheng and Cao 2014), predicting liquefaction-induced lateral spread (Goh and Zhang 2014), analysis of geotechnical engineering systems (Zhang and Goh 2013; 2014; Zhang et al. 2015b; 2017; Goh et al. 2018).

For driven piles, the piling hammer induces compression stresses in the piles. Therefore, the strength of the pile needs to be considered whether it is sufficient to resist the stresses induced by the impact of the pile hammer. The analysis relates to numerous statistically dependent inputs, thus, a fairly large database is utilized to develop the pile drivability in relation to maximum compressive stresses (MCS). By comparing the cross-validation modeling results of MARS by Bayesian information criterion, a superior generalization ability model and better persuasiveness results are obtained.

2 MARS Methodology

Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR)
Editors: Jianye Ching, Dian-Qing Li and Jie Zhang

Published by Research Publishing, Singapore.
ISBN: 978-981-11-2725-0; doi:10.3850/978-981-11-2725-0_MS2-7-cd

MARS is non-parametric modeling and the entire modeling process is driven by data. The resulting piecewise curves, known as basis functions (BFs), give greater flexibility to the model, allowing for bends, thresholds, and other departures from linear functions. It generates BFs by searching in a stepwise manner, and it searches over all possible univariate knot locations and across interactions among all variables. An open source code ARESLab from Jekabsons (2016) is used in carrying out the analyses presented in this paper.

Let y be the target output and X = (X_1,…,X_P) be a matrix of *P* input variables. Then it is assumed that the data are generated from an unknown"true" model. In case of a continuous response this would be

$$y = f(X_1, \dots, X_p) + \ e = f(X) + \ e \tag{1}$$

where *e* is the distribution of the error. MARS approximates the function f by applying basis functions (BFs). BFs are splines (smooth polynomials), including piecewise linear and piecewise cubic functions. For simplicity, only the piecewise linear function is expressed. Piecewise linear functions are of the form max (*0, x−t*) with a knot occurring at value t. The equation max (·) means that only the positive part of (·) is used otherwise it is given a zero value. Formally,

$$max(0, x - t) = \begin{cases} x - t, if \ \ x \geq t \\ 0 \quad otherwise \end{cases} \tag{2}$$

The MARS model *f(X)*, is constructed as a linear combination of BFs and their interactions, and is expressed as

$$f(X) = \beta_0 + \ \sum_{n=1}^{n=N} \beta_n BF(X) \tag{3}$$

where each *BF(X)* is a basis function. It can be a spline function, or the product of two or more spline functions already contained in the model (higher orders can be used only when the data warrants it; for simplicity, at most second-order is assumed in this paper and the predictive accuracy based on it is proved to be satisfactory). The coefficient $\beta_0$is a constant, and $\beta_n$is the coefficient of the mth basis function, estimated using the least-squares method.

At the end of the backward phase, from those "best" models of each size, the one with the lowest Generalized Cross-Validation (GCV) is selected and outputted as the final one. GCV, as an estimator for prediction Mean Squared Error, for an ARES model is calculated as follows (Friedman 1991):

$$GCV = \frac{MSE_{train}}{(1 - \frac{enp}{N})^2} \tag{4}$$

where *enp* is effective number of parameters, enp=k + c×(k-1)/2, *k* basis functions in MARS model (including the intercept term), *c* penalty in the range of about 2 to 4 (Friedman 1991). (k - 1)/2 is the number of hinge-function knots, so the formula penalizes the addition of knots.

3 Feature Variable Selection

In the condition of excessive input features, while a small subset of features is sufficient to approximate the label well, additional features of **xi** can lead to smaller training errors, but in the testing set, they will interfere with the prediction of **y**. Under this circumstance. In statistics and machine learning, least absolute shrinkage and selection operator (Lasso) is a regression analysis method that can accomplish both feature selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. It was independently rediscovered and popularized (Tibshirani 1996), who coined the term and provided further insights into the observed performance. Lasso was introduced to enhance the prediction accuracy and interpretability of regression models by changing the model fitting process to pick only a subset of the provided covariates for using in the ultimate model rather than using all of them. Lasso can achieve both of these targets by compelling the sum of the absolute value of the regression coefficients to under a specified value, which causes the coefficients to be set to zero, effectually choosing a simpler model that does not include those coefficients.

In this case study, the database comprised of 4072 piles that were installed for bridges in the State of North Carolina (Jeon and Rahman 2008). The seventeen input variables included the hammer, hammer cushion material, pile, soil parameters, ultimate pile capacities, and stroke. The target output is MCS. There are 17 variables here, and we can delete 10 feature variables by Lasso regularization. A summary of the input variables and output is listed in Table 1.

4 Cross-Validation and Optimal Model

Model selection and evaluation play a crucial role in machine learning because the quality of the model directly affects the accuracy of the prediction. In the selection and evaluation of models, many methods have been

proposed and applied to practice. Cross-validation is considered to be an effective method because of its simplicity and universality. K-fold cross-validation is a commonly used sample reuse method. By using a large number of data sets, statisticians perform a large number of experiments using different methods, indicating that 10-fold is the right choice for obtaining the best error estimate. In this paper, 10-fold cross validation is adopted.

Figure 1 is a flowchart for the entire calculation process. For obtaining superior generalization ability model and better persuasiveness results, ten sets of 90% training data and 10% testing data is produced by 10-fold cross-validation, and then they are put in the MARS algorithm separately, and we can obtain these 10 cross-validation models. Choosing the optimal model is considered from two aspects: one is the maximization of the likelihood function, and the other is the minimization of the number of unknown parameters in the model. The larger the value of the likelihood function, the better the effect of the model fitting, but we can't simply measure the pros and cons of the model with the accuracy of the fitting. This leads to more and more unknown parameters in the model, and the model becomes more and more complicated and causes overfitting. So a good model should be a comprehensive optimization of the fitting accuracy and the number of unknown parameters. Bayesian information criterion (BIC) considers the number of samples and the number of samples is too large, it can effectively prevent the model complexity caused by excessively high model accuracy (Chen and Gopalakrishnan 1998; Weakliem 1999). Therefore BIC is used to select the optimal model and BIC is defined as

$$BIC = n \cdot ln\left(\frac{RSS}{n}\right) + \quad k \cdot ln(n) \tag{5}$$

where n is the sample size; k is the number of basis functions in here; RSS is the residual sum of squares. We score each candidate model, and the lowest score is the optimal model.

Table 1. Summary of input variables and outputs.

Inputs and outputs		
Input variables	Hammer weight (kN)	Variable 1 (x1)
	Energy (kN m)	Variable 2 (x2)
	Helmet weight (kN)	Variable 3 (x3)
	Length (m)	Variable 4 (x4)
	Section area (m^2)	Variable 5 (x5)
	Ultimate pile capacity (kN)	Variable 6 (x6)
	Stroke (m)	Variable 7 (x7)
Outputs	Maximum compressive stress MCS (MPa)	

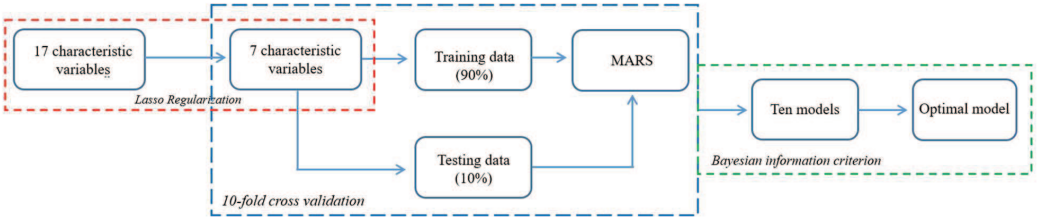

Figure 1. Flowchart of entire calculation process.

5 MARS Modeling Results

The adjusted R-square, root mean square error (RMSE), basis function (BF), and Bayesian information criterion (BIC) of the testing set are calculated, and the point dashed lines of these values with 10-fold cross-validations are plotted. Figure 2(a) shows the result of an adjusted R-square with 10-fold models. From Fig 2(a), it can be seen that the R2 value ranges from 0.923 to 0.945, with the mean value is 0.939, and the coefficient of variation (COV) is 0.004, which indicates that the model fits well with the presented data and the result is fairly stable, where K = 8, the maximum R2 value of 0.946 is observed. It is easy to interpret that when K = 8, the BIC score is the lowest, indicating the comprehensive model's complexity and desirable performance, and the model corresponding to K = 8 is the optimal model, thus, it is selected as the final model.

The predicted results are shown in Fig. 3 and 4 along with the training and testing modes (K = 8). It is obvious that the MARS model has been able to learn the complicated relationship between the maximum compressive stresses (MCS) and other basic information. The majority of the estimations of MCS were within ±20 % of the target values. The MARS model predictive capacities are satisfactory considering that the

geotechnical capacities analyzed in this case study are highly multivariate with 7 design variables with a large data set of 4072 observations.

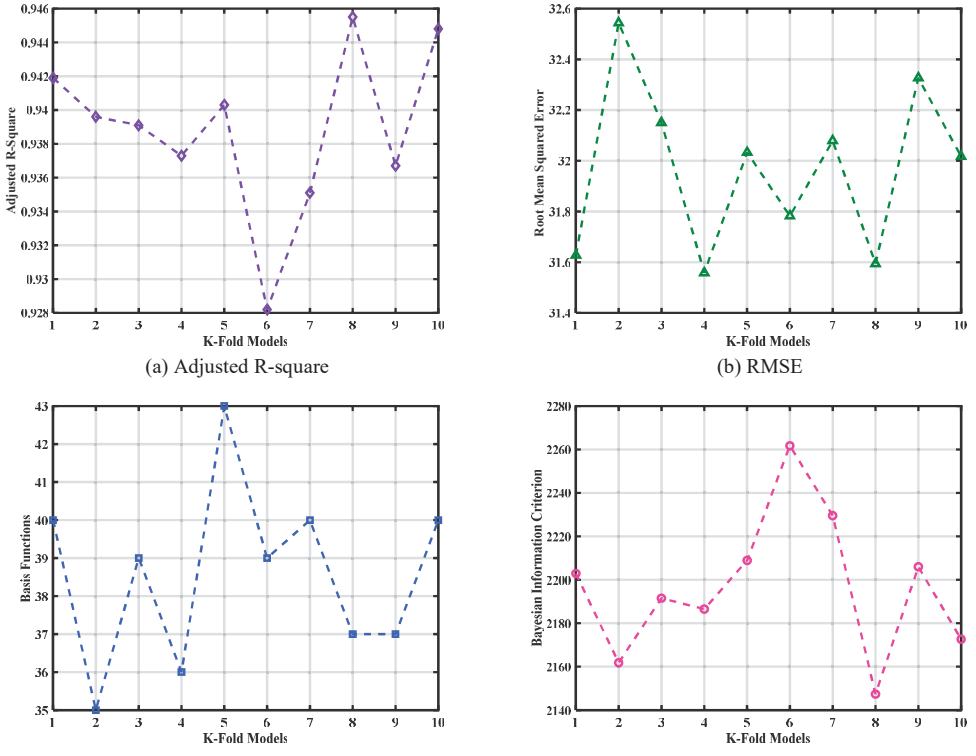

Figure 2. Tendency chart of parameter under 10-fold cross validation: (a) Adjusted R-square, (b) RMSE, (c) Basis function and (d) BIC.

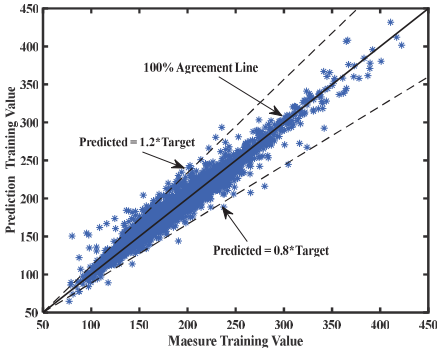

Figure 3. Performance of training dataset (K=8).

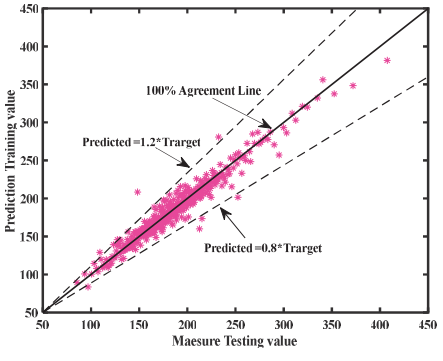

Figure 4. Performance of testing dataset (K=8).

6 Parameter Relative Importance

The relative importance of a variable is defined as the square root of the GCV of the model with all basis functions involving that variable removed, minus square root of the GCV score of the corresponding full model scaled so that the relative importance of the most important variable (using this definition) has a value of 100. Figure 5 gives the plot of the relative importance of the input variables for the HP drivability models developed

by MARS. It can be observed that both MCS is mostly influenced by x17 (Stroke)，followed by x16 (Ultimate pile capacity), x7 (Length) are significantly important in determining MTS.

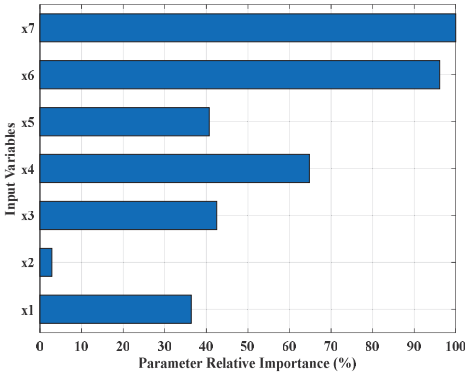

Figure 5. Relative importance of the input variables in MARS for MCS.

Table 2 lists the BFs of the MARS model for MCS and their corresponding equations. The interpretable MARS model to predict MCS is given by Eq. (3).

7 Summary and Conclusions

A database containing 4072 pile data sets with a total of seventeen variables is adopted to develop MARS models for drivability predictions in relation to the MCS prediction. The main conclusions arrived at include:

1. It is demonstrated that the MARS algorithm is capable of producing simple, accurate and easy-to-interpret models and estimating the contributions of the input variables.
2. Via lasso regularization, a small subset of the feature variables is sufficient to approximate the target response well, instead of using the full 17 variables.
3. Considering the complexity and predictive power of the models by cross-validation, we can obtain superior generalization ability model and better persuasiveness results.

Table 2. Basis functions and corresponding equations of MARS model for MCS overall datasets.

Coefficient	Basis Function	Coefficient	Basis Function
Intercept	176.49	0.99554	BF19=BF2 * max(0, x4 -3.05)
0.043769	BF1=max(0, x6 -1601.3)	0.15581	BF20=BF5 * max(0,18.29 -x4)
6.8992	BF2=max(0,1601.3 -x6)	0.51164	BF21=BF4 * max(0, x3 -6.67)
22.588	BF3=max(0, x7 -2.271)	9.7228	BF22=BF4 * max(0,6.67 -x3)
56.77	BF4=max(0,2.271 -x7)	118.96	BF23=max(0,0.014 -x5) * max(0, x3 -12.43)
2.7502	BF5=max(0, x3 -7.38)	629.27	BF24=max(0,0.014 -x5) * max(0,12.43 -x3)
5.4654	BF6=max(0,7.38 -x3)	989.2	BF25=max(0,0.014 -x5) * max(0, x4 -10.67)
3.5265	BF7=max(0, x1 -12.2)	1807.5	BF26=max(0,0.014 -x5) * max(0,10.67 -x4)
8.556	BF8=max(0,12.2 -x1)	0.99749	BF27=BF2 * max(0, x4 -10)
0.81755	BF9=max(0, x4 -3.05)	1.0005	BF28=BF2 * max(0,10 -x4)
1757.2	BF10=max(0,3.05 -x4)	16.224	BF29=BF5 * max(0, x7 -2.895)
0.0044212	BF11=BF9 * max(0, x6 -1668)	0.24888	BF30=BF9 * max(0, x3 -19.88)
0.0012404	BF12=BF2 * max(0, x3 -9.21)	0.050592	BF31=BF6 * max(0, x2 -58.1)
0.0062182	BF13=BF2 * max(0,9.21 -x3)	0.22353	BF32=BF6 * max(0,58.1 -x2)
63.235	BF14=BF4 * max(0, x1 -29.4)	63.61	BF33=BF3 * max(0, x2 -57)
765.91	BF15=max(0,0.014 -x5) * max(0, x4 -3.05)	65.476	BF34=BF3 * max(0, x2 -57.5)
3.4414e+05	BF16=max(0,0.014 -x5) * max(0,3.05 -x4)	0.78917	BF35=BF3 * max(0, x4 -9.14)
6.3097	BF17=max(0,0.014 -x5) * max(0, x6 -1067.5)	3.8881	BF36=BF3 * max(0,9.14 -x4)
2.2775	BF18=max(0,0.014 -x5) * max(0,1067.5 -x6)	8.6628	BF37=BF3 * max(0, x1 -17.8)
Resulting Expression			$f(X) = \beta_0 + \sum_{n=1}^{n=N} \beta_n BF(X)$

References

Adoko, A.C., Jiao, Y.Y., Wu, L., Wang, H., and Wang, Z.H. (2013). Predicting tunnel convergence using multivariate adaptive regression spline and artificial neural network. *Tunnelling and Underground Space Technology*, 38, 368-376. doi:10.1016/j.tust.2013.07.023.

Attoh-Okine, N. O., Cooger, K., and Mensah, S. (2009). Multivariate adaptive regression (mars) and hinged hyperplanes (hhp) for doweled pavement performance modeling. *Construction & Building Materials*, 23(9), 3020-3023. doi:10.1016/j.conbuildmat.2009.04.010.

Chen, S.S. and Gopalakrishnan, P.S. (1998). Clustering via the Bayesian information criterion with applications in speech recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE. doi: 10.1109/ICASSP.1998.675347.

Cheng, M.Y. and Cao, M.T. (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing*, 22(5), 178–188. doi:10.1016/j.asoc.2014.05.015.

Friedman, J.H. (1991). *Multivariate Adaptive Regression Spline. Annals of Statistics*, 19(1), 1-67. doi:10.1214/aos/1176347963.

Goh, A.T.C. and Zhang, W.G. (2014). An improvement to MLR model for predicting liquefaction-induced lateral spread using multivariate adaptive regression splines. *Engineering Geology*, 170, 1-10. doi: 10.1016/j.enggeo.2013.12.003.

Goh, A.T.C., Zhang, W., Zhang, Y., Xiao, Y., and Xiang, Y. (2018). Determination of earth pressure balance tunnel-related maximum surface settlement: a multivariate adaptive regression splines approach. *Bulletin of Engineering Geology and the Environment*, 77(2), 489-500. doi:10.1007/s10064-016-0937-8.

Jekabsons, G. (2016). *ARESLab: Adaptive Regression Splines Toolbox for Matlab/Octave Ver. 1.13.0*, Riga Technical University. http://www.cs.rtu.lv/jekabsons/.

Jeon, J. and Rahman, M.S. (2008). *Fuzzy Neural Network Models for Geotechnical Problems* (No. FHWA/NC/2006-52).

Lashkari, A. (2013). Prediction of the shaft resistance of nondisplacement piles in sand. *International Journal for Numerical & Analytical Methods in Geomechanics*, 37(8), 904-931. doi: 10.1002/nag.1129.

Mirzahosseini, M.R., Aghaeifar, A., Alavi, A.H., Gandomi, A.H., and Seyednour, R. (2011). Permanent deformation analysis of asphalt mixtures using soft computing techniques. *Expert Systems with Applications*, 38(5), 6081-6100. doi:10.1016/j.eswa.2010.11.002.

Samui, P. and Kurup, P. (2012). Multivariate adaptive regression spline and least square support vector machine for prediction of undrained shear strength of clay. *International Journal of Applied Metaheuristic Computing (IJAMC)*, 3(2), 33-42. doi:10.4018/jamc.2012040103.

Smith, E.A. (1962). Pile-driving analysis by the wave equation. *American Society of Civil Engineers Transactions*, 127, 1145-1170.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B: Methodological, 73(1), 273-282. doi:10.1111/j.1467-9868.2011.00771.x.

Weakliem, D.L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3), 359-397. doi: 10.1177/0049124199027003002.

Zhang, W.G. and Goh, A.T.C. (2013). Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics*, 48, 82-95. doi:10.1016/j.compgeo.2012.09.016.

Zhang, W. and Goh, A.T.C. (2014). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, S1674987114001364. doi:10.1016/j.gsf.2014.10.003.

Zhang, W., Goh, A., Zhang, Y.M., Chen, Y.M., and Xiao, Y. (2015). Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. *Engineering Geology*, 188, 29-37. doi:10.1016/j.enggeo.2015.01.009.

Zhang, W., Zhang, R., and Goh, A.T.C. (2017). Multivariate adaptive regression splines approach to estimate lateral wall deflection profiles caused by braced excavations in clays. *Geotechnical and Geological Engineering.* doi:10.1007/s10706-017-0397-3.