Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Using multivariate adaptive regression splines to estimate pollution in soil

Betul Kan Kilinc [1,*], Semra Malkoc [2], A. Savas Koparal [2], Berna Yazici [1]

[1]Department of Statistics, Science Faculty, Anadolu University, 26470, Eskişehir, Turkey
[2]Applied Research Centre for Environmental Problems, Anadolu University, 26555 Eskişehir, Turkey

ARTICLE INFO

Article history:
Received 16 September 2016
Received in revised form
17 November 2016
Accepted 10 December 2016

Keywords:
Response surface
Piecewise regression
Regression spline
Heavy metal

ABSTRACT

Heavy metal pollution is one of the main factors of the traffic pollution. The public authorities have been monitoring the concentration of heavy metal by means of sampling stations. This paper describes the response surface models and an intelligent regression algorithm, multivariate adaptive regression splines (MARS) models to data collected from soil at the stations where there were high density of buildings, roads, traffic and tramways. The model variables included the number of car and tramways and the concentration levels of Cadmium (Cd), Zinc (Zn) and Lead (Pb), at depth of 0-100mm. The objective of this study was to apply MARS to analyze the model output when there are a few numbers of design points. Several MARS models developed to simulate the concentration of each heavy metal. The performance of MARS was compared to that of response surface methodology (RSM) using 1st and 2nd order response surface models with respect to the accuracy metrics; root mean square error and adjusted R^2. The results showed that MARS models were successful in goodness of fit, suitable and also reliable as compared to the RSM models. Additionally, use of MARS in selection of the variables indicating great contribution on the response was effective.

1. Introduction

With the continuous increase in industrial development, heavy metal contamination has posed a serious threat for the environment. Elevated concentrations in air, water, and soil may occur close to industrial emission sources, particularly nonferrous mining and metal refining industries. Human activities, such as the atmospheric deposition of industrial soot, dust and aerosols, and coal burning exhausts, the application of fertilizers, livestock manures, and agrochemicals, and the disposal of anthropic wastes were the main sources of Cd, Pb and Zn. Increasing deposition of heavy metals on land and air has given a considerable concern about its impact on human health by the society to provide a sustainable environment. In recent years, various ways of approaching the distribution of pollution were analyzed by the scientists. Silva et al. (2001) studied on the main factors of air pollution in Santiago, Chile. They modelled and predicted the atmospheric pollution using the meteorological variables by means of MARS and non-parametric discriminant analysis. Gruszczynski (2005) examined the soil spatial distribution pollution of Chromium (Cr) where pollution by this element was high using interpolation algorithms and artificial neural networks. Covelo et al. (2008) examined the tree fitted regression models on the data containing six heavy metal. The sorption and retention of mixtures of heavy metals was reproduced by binary decision-tree regression models using classification and regression trees (CART) algorithm by an accompanying paper of Vega et al. (2009). Cheng et al. (2009) presented a case study in assessment of the distribution of soil Zinc (Zn) in an area severely polluted. Nieto et al. (2012) improved his work about cyanotoxins, a kind of poisonous substances produced by cyanobacteria, prediction from some experimental cyanobacteria concentrations in the Trasona reservoir (Asturias, Northern Spain) using (MARS). Piedade et al. (2014) applied a new approach of visualization based on tridimensional images of lead (Pb) concentrations in soil of a mining and metallurgy area to determine the spatial distribution of this pollutant and to estimate the most contaminated volumes. Lee and Toscas (2015) estimated the spatial distribution of the lead

* Corresponding Author.
Email Address: bkan@anadolu.edu.tr (B. K. Kilinc)
https://doi.org/10.21833/ijaas.2017.02.002

concentration levels that may affect exposed humans by using penalized regression and tensor product smooths.

In keeping with the above, the contaminated soil was also investigated by different statistical methods as well. Long et al. (2013) used response surface methodology (RSM) based on Box-Behnken experimental design for the analysis of variables of surfactant flushing treatment to optimize toluene removal efficiency from contaminated soil. Martínez-Fernández et al. (2014) studied response surface methodology to develop predictive models from central composite designs. Zhu et al. (2015) studied confirmed the use of Box-Behnken experimental design for analyzing the variables of ultrasound-assisted surfactant extraction treatment.

Different from conventional models, Govaerts and Noel (2005) discussed the analysis of a designed experiment when the response was a curve using three different approaches: two-step nonlinear modeling, pointwise functional regression and smoothed functional regression.

Developed in 1990 by Friedman, MARS is an intelligent, flexible, fast and accurate in prediction for various types of variables. Many applications have shown the successful prediction of MARS; Chun et al. (2003) showed the performance of MARS for simulating the pesticide transport in soils and confirmed that it can simulate complex phenomena in a simple and straightforward way rather than artificial neural networks. Woods and Lewis (2006) gave a method for constructing all-biased designs for polynomial spline regression models. Crino and Brown (2007) combined MARS with a response surface methodology. In his study, MARS showed low computational cost and better interpretability when compared to neural networks and generalized additive models.

Since 1990's, the successful applications of MARS were appeared in several field of studies (Crino and Brown, 2007). The primary objective of this study was to apply MARS to simulate the concentrations of three heavy metal at soil depth of 0-100mm. Several MARS models with interaction terms were developed and compared to the results obtained by 1st and 2nd order response surface models. Importance sequence of input variables to each heavy metal was determined respectively. The performance of MARS models was compared to the response surface models with respect to root mean squared error (RMSE), number of variables, number of observations, and the adjusted determination coefficient (R^2_{adj}).

2. Materials and methodology

2.1. Model description

RSM comprises a group of statistical techniques for empirical model building and model exploitation (Box and Draper, 2007). In this study, it is assumed that some true physical relationships between the expectation of the response y and two factors (ξ_1 and ξ_2) and via physical constants λ exist as follows (Eq. 1):

$$E(y) = f(\xi_1, \xi_2, \lambda) \quad (1)$$

The nature of the expectation function in E(y) is unknown and it is replaced by an approximating function as either of the following (Eqs. 1 and 2):

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \quad (2)$$

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i=1}^{k-1} \sum_{\substack{j=1 \\ 1<j}}^{k} \beta_{ij} X_i X_j + \varepsilon \quad (3)$$

where, $X_i = f(\xi_1, \xi_2)$ is a linear coding of a factor ξ_i, i=1, 2 (Box and Draper, 2007). The response is y and $X_1, , X_2, \ldots, X_k$ are k known explanatory variables. X_i^2 is a higher-order term and $X_i X_j$ is the interaction term for i=1, ..., k-1, j=1, ..., k. $\beta_0, \beta_1, \ldots, \beta_k, \beta_{11}, \ldots, \beta_{kk}$ are unknown parameters and ε is a random error (Khuri and Cornell, 1987). The response is modeled by a linear function or the model is upgraded by adding higher-order terms if there is a curvature in the system.

In order to build a MARS model, a response and a set of exploratory variables are required. MARS splits the data into several splines and approximates the regression model using basis functions (BFs) as follows (Eq. 4):

$$y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{X}) \quad (4)$$

where, $h_m(\boldsymbol{X})$is BF that represents the data in each subgroup. $\beta_0, \ldots, \beta_M$ are unknown parameters, i=1, ..., M, that are estimated by ordinary least squares method once BFs are investigated. BFs can be represented by (Eq. 5):

$$h_m(\boldsymbol{X}) = \prod_{k=1}^{K_m} [S_{k,m}(X_{V(k,m)} - t_{k,m})]_+^q \quad (5)$$

where, "+" means the argument is a truncated power function, $K_m$is the number of variables (interaction order) in the mth basis expansion. $X_{V(k,m)}$ is the vth variable, $1 \leq V(k,m) \leq n$. $t_{k,m}$ is a knot on each of the corresponding variable where the two BFs in two adjacent domains of data intersect? Therefore, MARS creates knots which can be located among different exploratory variables. The BF represents the relationship between the knots using the reflected pairs of hockey stick function as follows (Eq. 6):

$$f(X_i) = \max(0, X - t)$$
$$f(X_i) = \max(0, t - X) \quad (6)$$

Where, $f(X_i)$ is a new variable with values 0 for all values X up to some threshold t whereas $f(X_i)$ is equal to x for all values of x larger than the threshold value. The second pair of hockey stick function generates a reflected effect of the first one and illustrates the variation in BFs for changes of t values for variable X. Thus, t denotes the knot where the

behavior of the function changes. Each BF is unique between any two knots and replaced by another BF at each knot (Abraham and Steinberg, 2001; Friedman, 1991). Thus, a knot is located at the beginning of a region and the end of another. Finding the best knot is a search process in MARS. In each spline, the data is splitted in regions. The MARS model fits a regression line from region to region using splines whereas model's response is continuous.

MARS procedure adaptively selects the BF set by two iterative approaches: forward and backward selection. It uses the residual squared error in iterations to compare the partition points. Determination of knot locations is adaptive to data characteristics (Abraham and Steinberg, 2001; Friedman, 1991). The criterion used to set the final model is a modified generalized cross validation (GCV) of the first proposed one by Craven and Wahba (1978) (Eq. 7).

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(\boldsymbol{X})\right]^2 \Big/ \left[1 - \frac{C(M)}{N}\right]^2 \quad (7)$$

MARS generates and compares the models in terms of the importance of the inputs in the models using ANOVA. Analysis of variance decomposition of the MARS model is given by the following expression (Friedman, 1991) (Eq. 8):

$$y = \beta_0 + \sum_{B=1} f_i(X_i) + \sum_{B=2} f_{ij}(X_i, X_j) + \sum_{B=3} f_{ijk}(X_i, X_j, X_k) + \cdots \quad (8)$$

where, $\sum_{B=1} f_i(X_i)$ is overall BFs involve only a single variable, $\sum_{B=2} f_{ij}(X_i, X_j)$ is overall BFs represent the contribution from two variables, $\sum_{B=3} f_{ijk}(X_i, X_j, X_k)$ is overall BFs represent the contribution from three variables, and so on. The best (or final) MARS model is the one with the smallest GCV value and the largest R^2_{adj} value (Friedman, 1991; Hastie et al. 2001).

2.2. Data description

The data used to develop 1st and 2nd order response models and MARS included the number of cars, the number of tramways, and the concentration measurements of three heavy metals that cause traffic pollution: Cd, Zn, Pb. Urban soils collected at locations where there were high density of buildings, roads and tramways in Eskisehir, Turkey. A systematic sampling was adopted to prove a sampling strategy over the entire workspace. Sample points within topsoil layers 0-10 cm were located roads alongside. Portions of the soil samples which were hold approximately 25gr were grounded in a mechanical agate grinder until fine particles (<200μm) were obtained. The coordinates of the sample locations were recorded with a GPS. All soil samples were dried for 3 h at 105°C (to a constant weight), milled and passed through a nylon sieve (0.5 mm). 0.5g samples were weighed and transferred into reaction vessels. The descriptive statistics of the heavy metals were presented in Table 1.

Table 1: Heavy metal contents in soil samples of Eskişehir (μg/g)

Response	Minimum	Maximum	Mean	Median	Geometric Mean	Standard Deviation	Skewness	Guideline
Cd	0.60	3.39	1.28	1.19	1.23	0.45	2.86	3 [a]
Zn	34.10	136.68	64.93	55.84	60.16	28.42	1.49	300 [a]
Pb	7.92	101.12	31.50	21.93	24.24	23.57	1.15	300 [a]

[a] Values recommended by Turkey Ministry of Environment and Forestry (2005)

2.3. Method description

In this study, the software SAS (version 9.0 for Windows) was applied to build the response surface models. A response surface statistical experimental design was used to optimize the concentration of Cd, Zn, and Pb separately. This design is based on a 3^2 factorial design, five replicates of the experiment, leading to 45 observations at nine different sample stations investigated. To properly represent the heavy metal concentration, 1st and 2nd order response models were investigated.

Thus six different models were developed in this part of the study. There were two inputs to the response models: the number of cars and the number of tramways. The levels of each input were chosen on the basis of the minimum and the maximum number of vehicles passing at chosen stations. Each level of the two factors was run in all combinations for Cd, Zn, and Pb. The levels of two factors coded as (-1), (0), and (+1) due to a computational ease are listed in Table 2.

Table 2. Levels of factors for 3^2 factorial design

Factors	Symbol	Levels		
		Low	Medium	High
Tramway	X_1	12 (-1)	18 (0)	24 (+1)
Car	X_2	174 (-1)	852 (0)	1530 (+1)

In the second part of the study, the same input variables in Table 2 were used to develop MARS models. As one MARS model has only one response variable, the total number of MARS models in this study is three for the three outputs: Cd, Zn, and Pb. The MARS software, version 2.0 was used in the analysis (Salford Systems, 2010).

2.4. Motivation

ANOVA decomposition of the MARS model captures the main idea of this study. The input variables X_1 and X_2 are considered as the main factors (variables) whereas X_1^2 and X_2^2 represent the squared effects in the model. Thus, a second order MARS model has been used so that the BFs of the final MARS model consist both linear and second-order splines. Besides the interaction terms were also included in the model. The maximum number of interactions was set to 2. Higher order of interaction was not allowed as the number of experimental points was not enough in this study. Hence the interactions were controlled before and only interactions of (X_1X_1), $(X_1\,X_2^2)$, $(X_2\,X_1^2)$ and $(X_1^2\,X_2^2)$ were taken into account. The same data collected from road were used to develop the MARS models. As there are three outputs of interest, three MARS models were simulated. The coded values of factors were used to provide orthogonality. Each model was assumed to describe the effect of the factors over the interest region (linear coding: -1, 0,+1, quadratic coding: +1,-2,+1).

3. Results

3.1. The analysis of variance

The quality of the fitted models was checked by F-test at 0.01 significance level. To make detailed information regarding to the structure of the variation in main and interaction effects, those were divided into linear and quadratic terms. The statistical insignificance of linear, quadratic and interaction terms were determined by using p-value>0.01. Thus a p-value with '*, **, ***' codes indicates the terms significant at the corresponding level. The results in Table 3 are satisfactory for a good prediction for the experiments carried out.

In Table 3, the results suggested that the 2nd order model was considered to approximate the surface curvature's nature better than 1st order response model with respect to RMSE for three outputs.

The results from the canonical analysis of the 2nd order response model for each output were listed in Table 4.

Table 3: Analysis of variance of calculated models for Cd, Zn, and Pb

	Cd			Zn			Pb		
	MS	F	p-value	MS	F	p-value	MS	F	p-value
X_1	0.024	3.879	0.029*	21277	3024.42	2.2e-16***	4949.6	295.727	<2.2 e-16***
X_1	0.037	6.035	0.019*	33285	4731.31	2.2e-16***	681.6	40.727	2.15e-07***
X_1^2	0.010	1.724	0.197	9269	1317.53	2.2e-16***	9217.6	550.728	<2.2e-16***
X_2	0.244	39.586	8.1e-10***	9184	1305.43	2.2e-16***	4440.3	265.297	<2.2e-16***
X_2	0.485	78.837	1.4e-10***	1879	267.13	2.2e-16***	6841.3	408.747	<2.2e-16***
X_2^2	0.002	0.336	0.566	16488	2343.73	2.2e-16***	2039.4	121.847	4.15e-13***
X_1X_2	0.398	64.653	6.2e-16***	2939	417.82	2.2e-16***	1447.3	86.471	<2.2e-16***
X_1X_2	0.120	19.508	8.8e-05***	2926	415.86	2.2e-16***	477.9	28.552	5.24e-06***
$X_1^2X_2$	1.181	191.968	5.4e-16***	1635	232.37	2.2e-16***	4166.7	248.949	<2.2e-16***
$X_1X_2^2$	0.002	0.2438	0.625	5965	847.84	2.2e-16***	1144.6	68.385	7.66e-10***
$X_1^2X_2^2$	0.289	46.889	5.2e-08***	1233	175.23	2.12e-15***	0.0	5.85e-07	0.999
Residuals	0.006			7			16.7		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*'; MS: Mean Squares

Table 4: Canonical analysis of response surface based on Cd, Zn, and Pb

Response	RMSE	R^2_{adj}	Stationary point	Mean	Predicted value
Cd	0.19	0.26	minimum	1.07	0.89
Zn	16.62	0.83	maximum	65.09	123.24
Pb	12.80	0.71	maximum	27.37	60.27

The obtained results suggested that the response surface predictions were in good agreement with the experimental results. The exception in here was when Cd was the output, the inputs were not satisfying to explain the response surface, but the predictive value was still in the interval of guideline. Thus, the experimental designs were reliable and effective in determining the optimum conditions.

3D plots of response surfaces

Fig. 1 indicates the three dimensional surfaces of the response for Cd, Zn, and Pb contamination. In Fig. 1, the maximum predicted Cd contamination was located at the corresponding levels of $(X_1,\ X_2) = (\pm1, \pm1)$. The minimum of the response was located at the levels of $(X_1,\ X_2) = (0, 0)$. The minimum predicted value of Zn contamination was located at the corresponding levels of $(X_1,\ X_2) = (-1, +1)$ and the maximum value of the response was located at the levels of $(X_1,\ X_2) = (0, 0)$. The response surface showed to be a mount shaped. The maximum predicted Pb is located at the point $(X_1,\ X_2) = (+1, 0)$. The plot shows that there is a saddle point in the surface.

3.2. MARS modeling results

The evaluated final MARS model and its estimated coefficients were given for each heavy metal below. The models were the linear combinations of five, six, and three BFs, respectively which used two original variables (number of tramways, number of cars). In Cd concentration MARS model, five BFs were found to be statistically important. The effects $X_1$and X_2 appeared both individually and interactively (i.e. their product with

each other). For instance, the BFs $max\ (0, X_2\text{-}0)$ and $max\ (0, X_1\ +1)$ represented the single effects on Cd concentration.

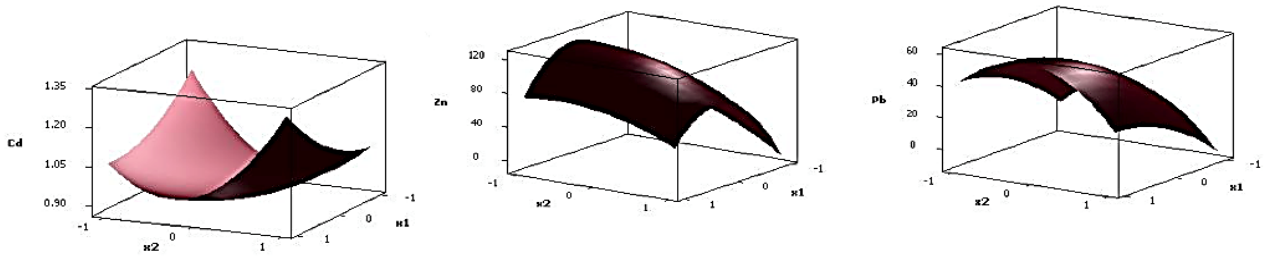

Fig. 1: Response surface plots for Cd, Zn, and Pb contamination

The single effects whose impact was positive (with coefficients of 0.529, 0.753, and 0.299) on the response were sensitive to the knot values 0 and (-1). The two BFs, $max\ (0, X_1\ +1)\times max\ (0,0\text{-}X_2)$ and $max(0, X_1\ +1)\times max(0, X_2\ -0)$ denoted that the impact of $X_1$was emerged through the interaction with X_2. It was also worth to mention that the impact of the contribution from the two effects was negative (with coefficients of -0.541 and -0.301). Hence it can be said that the effect X_1 played a role in the Cd concentration when it was larger than the knot (-1) while X_2<0 and X_2>0.

$$y_{\text{Cd}} = 0.629 + 0.529 \times max(0,\ X_2 - 0) + 0.753 \times max(0,0 -\ X_2) - 0.541 \times max(0,\ X_1\ + 1) \times max(0,0 -\ X_2) +\ 0.299 \times max(0,\ X_1\ + 1) - 0.301 \times max(0,\ X_1\ + 1) \times max(0,\ X_2\ - 0)$$

Having the similar interpretation for Zn and Pb concentration MARS models, it appeared the single effects and the contributions from two effects were significant. Hence the models were composed of six and three BFs, respectively.

$$y_{\text{Zn}} = 131.560 - 77.488 \times max(0,\ X_1\ - 0) - 89.738 \times max(0,0 -\ X_1) - 84.026 \times max(0,\ X_2\ - 0) +\ 92.449 \times max(0,\ X_1\ - 0) \times max(0,\ X_2\ - 0) + 43.403 \times max(0,0 -\ X_1) \times max(0,\ X_2\ - 0) +\ 19.668 \times max(0,\ X_1\ - 0) \times max(0,0 -\ X_2)$$

$$y_{\text{Ph}} = 9.844 + 34.658 \times max(0,\ X_1\ + 1) - 30.640 \times max(0,\ X_2\ - 0) \times max(0,\ X_1\ + 1) -\ 20.747 \times max(0,0 -\ X_2) \times max(0,\ X_1\ + 1)$$

To further access the capability of MARS and RSM, some of the important statistics were given in Table 5.

Table 5: Comparison of the MARS and RSM models

Method	MARS			RSM		
Heavy Metal	Cd	Zn	Pb	Cd	Zn	Pb
RSS	0.3801	369.5410	3727.8580	1.5240	10778	6391.6818
MSE	0.0097	9.7247	90.92337	0.0390	276.359	163.8892
F	40.3656	1243.6083	78.6146	4.2100	498.6900	115.3000
GCV	0.0506	41.5733	211.8050	-	-	-
RMSE (SE)	0.0987	3.1184	9.5353	0.1976	16.6242	12.8019
R^2_{adj}	82%	99%	84%	26%	83%	71%

RSS: Residual Sum of Squares; MSE: Mean Square Error; SE: Standard Error of Regression

MARS uses R^2_{adj} and GCV criteria to assess the goodness of a fit. For the three concentration models, the corresponding R^2_{adj} obtained by MARS are 82%, 99%, 84%, respectively. By considering 45 observations, overall R^2_{adj} scores indicate good fit than the scores R^2_{adj} of 26%, 83%, and 71% obtained by RSM models, respectively. MARS also computed GCV scores of the corresponding models as 0.051, 41.573, and 211.805, respectively. Table 5 compares also the accuracy of MARS and RMS models in estimating the concentration of Cd, Zn, and Pb. The MARS models indicate better than RSM models in terms of RMS accuracy. It was apparent that RMSE scores were decreased by MARS models.

MARS allows displaying the predicted response as a function of the others (Salford Systems, 2010).

In Fig. 2, the interactions of the input variables were presented for each concentration model. It is clear that given the X_1 and X_2 levels of (-1), the largest contribution to Cd concentration have been obtained. That is, the number of tramways and the number of cars both affected the contamination in soil in terms of Cd concentration although there were caused less traffic at the time. Secondly, it can be concluded that the X_1 of level (0) and the X_2 levels of (-1, 0) increased the contamination of the Zn concentration in soil. In other words, being moderate of the number of tramways and being increased from medium to high of the number of cars caused the largest contribution to Zn concentration. Finally, the largest contribution to Pb contamination was obtained when the number of tramways at highest level (+1) whereas the number of cars was at medium level (0). This result might be considered that the reason of the contamination in soil in terms of Pb was mostly from the reason of the number of tramways.

3.3. Relative importance

It can be seen from Table 6 that MARS possesses more information regarding to the important variables which RSM could not easily produce.

MARS explicitly indicates the important variables and ignores the unneeded ones. The most important variables are the ones that have the largest impact on the goodness-of-fit or GCV score (Steinberg and Colla, 1999). The relative importance of the variables for each concentration model was summarized in the Table 6. As can be seen, both variables car and tramway have great contribution to Cd, Zn, and Pb MARS models.

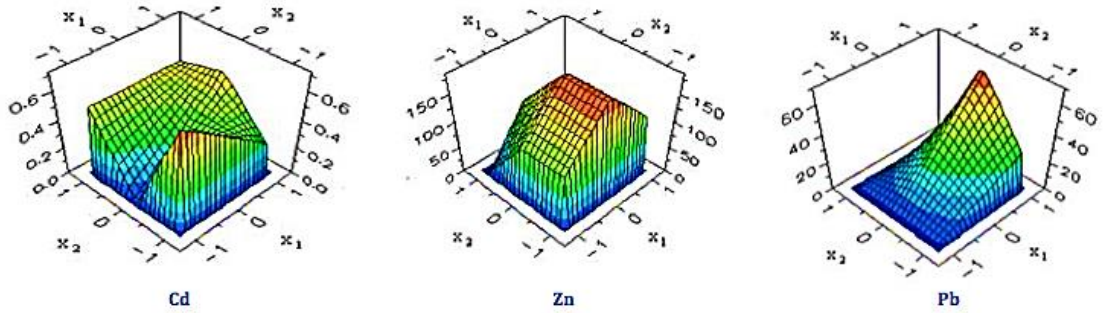

Fig. 2: Cd, Zn, and Pb concentration MARS models

Table 6: Relative variable importance for Cd, Zn, and Pb (%)

	Cd	Zn	Pb
Tramway	100.00	100.00	100.00
Car	91.03	81.83	84.61

4. Conclusions

This paper used a nonparametric method MARS and RSM based on a polynomial experimental design to approximate the concentration of some heavy metals in soil. Our objective was to develop response surface models to characterize the concentration of Cd, Zn, and Pb in soil and compare them with MARS. This method was used to explore the relationship by splitting the variables over its region and transform the original ones to new variables. The MARS algorithm was run to illustrate the concentration models combined with a group of spline base inputs.

The RSM based on a 3^2 factorial design was used first and then MARS was applied to the same datasets. The results of the analysis of variance on metal contaminations were given in tables for Cd, Zn, and Pb, respectively. The results of MARS showed that it extracted more information as RSM stood improper when there were a few numbers of runs in the experiment. MARS proposed extra information using the distribution of the design points. The MARS contamination models showed a moderate improvement in goodness of fit than the second order response surface models in terms of RMSE and the adjusted R^2_{adj}.

It was also noticeable that although the MARS models evaluated seemed more complex than the second order response surface models, the number of the covariates was still reasonable as compared to the number of observations. The successful applications of MARS in these three problems indicate that MARS is computationally efficient and easy to interpret. It can also estimate the contributions of the input variables and enable the scientists have an insight and understanding of the significant variables occur in the data.

5. Discussion

Our conclusions provide a better understanding of the response surfaces that are obtained by MARS. Based on the early studies of Kan and Yazici (2009a, 2009b) on modeling the response surfaces by MARS, our approximation is still adaptable to the first order and second order response surface models so that the conclusion has not been only referenced to one very specialized data set given in the frame of this paper but others as well. Since in low dimensions, the set of the design points were not proper for modeling, the weakness of RSM can be removed by using MARS.

References

Abraham A and Steinberg D (2001). MARS: Still an alien planet in soft computing?. In International Conference on Computational Science, Springer Berlin Heidelberg: 235-244.

Box GEP and Draper NR (2007). Response surfaces, mixtures, and ridge analyses. John Wiley and Sons, New Jersey, USA.

Cheng W, Zhang X, Wang K, and Dai X (2009). Integrating classification and regression tree (CART) with GIS for assessment of heavy metals pollution. Environmental Monitoring and Assessment, 158(1-4): 419-431.

Chung-Chieh Y, Prasher SO, Lacroix R, and Kim SH (2003). A multivariate adaptive regression splines model for simulation of pesticide transport in soils. Biosystems Engineering, 86(1): 9-15.

Covelo EF, Matías JM, Vega FA, Reigosa MJ, and Andrade ML (2008). A tree regression analysis of factors determining the sorption and retention of heavy metals by soil. Geoderma, 147(1): 75-85.

Craven P and Wahba G (1978). Smoothing noisy data with spline functions. Numerische Mathematik, 31(4): 377-403.

Crino S and Brown DE (2007). Global optimization with multivariate adaptive regression splines. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37(2): 333-340.

Friedman JH (1991). Multivariate adaptive regression splines (with discussion). The Annals of Statistics, 19(1): 79-141.

Govaerts B and Noel J (2005). Analysing the results of a designed experiment when the response is a curve: Methodology and application in metal injection moulding. Quality and Reliability Engineering International, 21(5): 509-520.

Gruszczynski S (2005). The assessment of variability of the concentration of chromium in soils with the application of neural networks. Polish Journal of Environmental Studies, 14(6): 743-751.

Hastie T, Tibshirani R, and Friedman J (2001). The elements of statistical learning-data mining, inference and prediction. Springer, New York, USA.

Kan B and Yazici B (2009a). Assessment of fuel consumption using factorial experiments, regression trees and MARS. In the 14th WSEAS International Conference on Applied mathematics. World Scientific and Engineering Academy and Society (WSEAS): 196-201.

Kan B and Yazıcı B (2009b). Determining the coordinates of an experimental data set based on multivariate adaptive regression splines. Proceedings of Joint Statistical Meetings Program Committee, Washington, USA: 3098-3104.

Khuri I and Cornell JA (1987). Response Surfaces. Dekker, New York, USA.

Lee DJ and Toscas P (2015). Flexible geostatistical modeling and risk assessment analysis of lead concentration levels of residential soil in the Coeur D'Alene River Basin. Environmental and Ecological Statistics, 22(3): 551-570.

Long A, Zhang H, and Lei Y (2013). Surfactant flushing remediation of toluene contaminated soil: Optimization with response surface methodology and surfactant recovery by selective oxidation with sulfate radicals. Separation and Purification Technology, 118: 612-619.

Martínez-Fernández, D, Bingöl D, and Komárek M (2014). Trace elements and nutrients adsorption onto nano-maghemite in a contaminated-soil solution: a geochemical/statistical approach. Journal of Hazardous Materials, 276: 271-277.

Nieto PG, Fernández JA, Lasheras FS, de Cos Juez FJ, and Muñiz CD (2012). A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. Science of the Total Environment, 430: 88-92.

Piedade TC, Souza LCP, and Dieckow J (2014). Three-dimensional data interpolation for environmental purpose: lead in contaminated soils in southern Brazil. Environmental Monitoring and Assessment, 186(9): 5625-5638.

Salford Systems (2010). Overview of MARS methodology. Available online at: https://www.salford-systems.com/resources/ whitepapers/113-an-overview-of-mars

Silva C, Pérez P, and Trier A (2001). Statistical modelling and prediction of atmospheric pollution by particulate material: two nonparametric approaches. Environmetrics, 12(2): 147-159.

Steinberg D and Colla PL (1999). MARS™ user guide. Salford Systems, San Diego, California, USA.

Vega FA, Matías JM, Andrade ML, Reigosa MJ, and Covelo EF (2009). Classification and regression trees (CARTs) for modelling the sorption and retention of heavy metals by soil. Journal of Hazardous Materials, 167(1): 615-624.

Woods D and Lewis S (2006). All-bias designs for polynomial spline regression models. Australian and New Zealand Journal of Statistics, 48(1): 49-58.

Zhu M, Yao J, Masakorala K, Chandankere R, Chen H, and Ceccanti B (2015). Ultrasound-assisted extraction of pah-contaminated clay soil in the middle Yangtze River basin, China: Optimisation with response surface methodology. Fresenius Environmental Bulletin, 24(10B): 3426-3435.