МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2010
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2010
Proceedings of the Thirty Ninth Spring Conference of
the Union of Bulgarian Mathematicians
Albena, April 6–10, 2010

APPLICATION OF MARS FOR THE CONSTRUCTION OF NONPARAMETRIC MODELS*

Snezhana Gocheva-Ilieva

This paper presents the main features of the relatively new statistical technique called Multivariate Adaptive Regression Splines (MARS) and the corresponding software product. The MARS method is designed for statistical analysis of data, when standard parametric modeling by multiple regression or logistic regression methods is not applicable. A case study from the area of laser technology, especially for modeling of UV Cu+ Ne-CuBr laser is performed. The obtained results are in a good agreement with practical issues. It is shown that the constructed nonparametric MARS models can be applied in estimation and prediction of current and future experiments in order to improve the output laser power.

1. Introduction. The method of multivariate adaptive regression splines (MARS) is a kind of nonparametric regression method for studying relationships. It was initially developed for data mining from large multidimensional datasets, but it displayed excellent qualities in a number of other fields. Today it is successfully used as a prediction and description technique in economics, sociology, ecology, geographical information systems, meteorology, engineering, etc.

Being a nonparametric regression method, MARS does not impose the limitations on the normal distribution of data, characteristic of classic parametric methods such as multidimensional regression or logistic regression. It is comparable to other nonparametric methods used for describing complex relationships between variables, namely Classification and Regression Tree – CART, Artificial Neural Networks – ANN, additive models, in particular the generalized additive model (GAM) etc. MARS has a number of advantages over these methods, including simple interpretation of obtained models, efficiency with large and small sample size, higher speed of realization of algorithms (for example 100 to 1000 higher than ANN) etc. [1, 2]

This paper introduces MARS and its capabilities relating to the construction of nonparametric regression models. Its main elements are described – nodes and algorithms for adaptive selection, basic functions, testing techniques, validation and avoidance of overestimation of the model, graphic representation of the results etc. The features of evaluation of local nonlinearity of data through determination of first, second or higher order interactions and construction of nonlinear models are demonstrated.

*2000 **Mathematics Subject Classification:** 62G08, 62P30.

Key words: multivariate regression, nonparametric model, ultraviolet laser.

This paper is partially supported by projects VU-MI-205, NSF of the Bulgarian Ministry of Education, Youth and Science and RS09–FMI–013, ISM-4 of NPD, Plovdiv University "Paisii Hilendarski".

In this paper, MARS method is applied in a case study regarding experimental data for a deep ultraviolet copper ion excited neon copper bromide (DUV Cu+ Ne-CuBr) lasers.

Because of its unique capabilities (high power, various wavelengths) DUV Cu+ Ne-CuBr laser is widely used in a number of fields [3–5]. The opportunity of extending its application justifies its theoretical and experimental study. Using the methodology of MARS, the following problems are solved: investigation on the influence of input parameters on the output laser power; establishment of the best MARS models of the 0^{th}, 1^{st} and 2^{nd} order for this dependence; comparison between the models constructed and interpretation of the results.

Experimental results obtained in the Metal Vapor Lasers Laboratory, at the Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences are used for basis of the statistical study. This laboratory is leader in the DUV Cu+ laser development [5–9].

2. Brief description of the method. The mathematical basis for MARS was developed by the american statistician J. H. Friedman in 1990–1991 [1]. The algorithms created by Friedman and their first program implementation have been integrated in the currently existing MARS software product. The product has gained popularity and has been applied with increasing success in the last few years [2].

In essence MARS generates adaptive models through partially linear regressions, i.e. data nonlinearities are approximated using separate sloped intercepts in different subintervals of the set defined for each predictor variable. A broken line is used, instead of looking for one common regression curve approximating the data. The slope of the regression line varies from one interval to another at the so called nodes.

The node is a basic element of the model. It shows where the behavior of the function changes. In the classic spline, nodes are given in advance, while with MARS, the most suitable place for them is determined using a fast algorithm when certain suitable optimization conditions are met (for example a SSE minimum – the sum of the squares of the errors). The initial node of the searching procedure is always at the lowest value of the predictor (the minimum). In addition, it is possible to find new relationships between variables and to determine new variables to be included in the model. In practice, determining the best distribution of the nodes for a large number of variables, especially as they are usually an unknown number, is a very complex task which requires intensive calculations. This problem is solved by using the so called "forward-backward stepwise procedure" and intermediate (fictitious) points between data.

The other basic element of MARS is the basic function for transformation of predictors. The basis function is called a "hockey stick" and has the following form:

$$\max(0, X - c) \quad \text{or} \quad \max(0, c - X),$$

where c is a constant, with which the function X is mapped in$X*$. An example with four basic functions is given in Fig. 1. In fact, such a function is generated for each value of X.

In the case of the function $BF20$ the transformation for X is

$$BF20 = \max(0, X - 20).$$

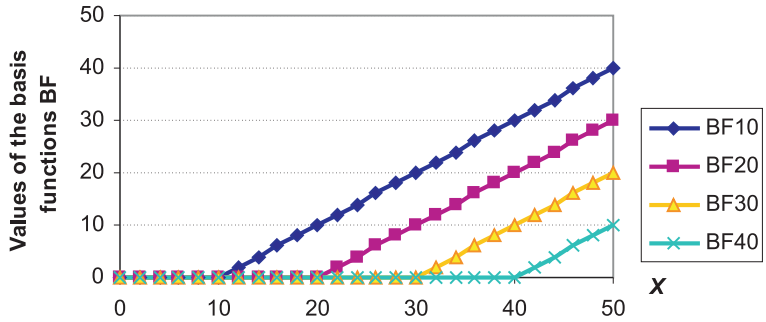


Fig. 1 Graphics of four basic "hockey stick" functions for the predicator variable X, defined in the interval [0,50].

If in the regression equation X is replaced with the basis function $BF20$, we get

$$y = \text{const} + b_1.BF20 + \text{error}.$$

The spline is written in the following form:

$$y = \begin{cases} \text{const} & X \in [0, 20] \\ \text{const} + b_1(X - 20) & X \in [20, 50] \end{cases}.$$

More complex linear splines can be constructed analogically. The coefficients of the spline are determined by additional optimization conditions, minimizing the total regression error when using different algorithms, adaptively in relation to the data.

Another important feature of MARS is the application of different smoothing methods. To name a few: floating means method, floating median method, weighted least squares method, LOWESS – locally-weighted regression smoothness curve etc.

We are also going to give abriefly look at some techniques for testing, validation and avoidance of overestimation of the model. First, the number of candidate basic functions, for which it is assumed to be part of the "best model", is estimated. After that, their number is at least doubled. The strategy is to exclude those functions which contribute the least to the accuracy of the model. The exclusion procedure is iterative and can be described as follows: (1) in the model with the biggest number of basic functions, MARS determines the one which contributes the least to the sum of the squares of the residuals (the least squares criterion), after which that function is removed; (2) the new model is calculated and in the same way the next least influential basic function is removed and so on; (3) the process is repeated until all basic functions are eliminated.

We have to specify that the "naïve model" suggested by MARS which has the biggest determination coefficient R^2 corresponds to the model with the maximum number of functions. In order to avoid the overestimation the specified R^2 and other criteria are used. The best MARS model satisfies a special general criterion [1, 2].

Finally, we have to note that usual regression splines, similarly to a normal linear one-dimensional regression, "mediate" the data grouped around the regression line. For this reason, when predicting results in the multidimensional dataset, predicted values are respectively lower or higher for a positive or negative spline slope.

3. Case study for experimental data for UV Cu+ Ne-CuBr laser.

3.1. Data description. Further we investigate the data of UV Cu+ Ne-CuBr laser by using MARS method. Development of copper halide lasers continues to be actul [5]. That is due to the fact that in the visible range ($\lambda_1 = 510.6\,\text{nm}$, $\lambda_2 = 578.2\,\text{nm}$), as well as in the ultraviolet range ($\lambda_1 = 248.6\,\text{nm}$, $\lambda_2 = 252.9\,\text{nm}$, $\lambda_3 = 260.0\,\text{nm}$ and $\lambda_4 = 270.3\,\text{nm}$), these lasers operate at highest output power.

We examine data of eight input basic variables which determine the UV Cu+ Ne-CuBr laser operation. They are: D (mm) – inside diameter of the laser tube, L (cm) – electrode separation (length of the active area), PIN (W) – input electrical power, PRF (KHz) – pulse repetition frequency, PNE (Torr) – neon gas pressure, $PH2$ (Torr) – hydrogen gas pressure, Dr (mm) – inside diameter of the rings, PL (W/cm/4) - specific electrical power per unit length. The response variable is $Pout$ (mW) – output laser power.

The data is of historical type. The sample size is $n = 176$. Here we have to mention the complexity, long duration and high cost of each conducted experiment.

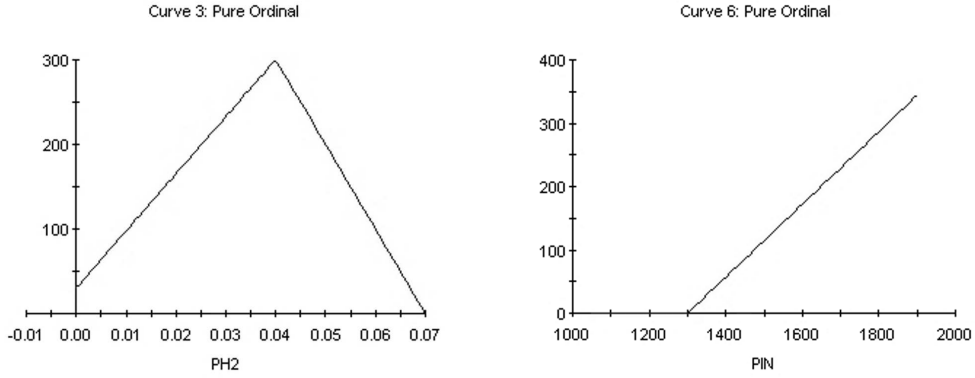
The data of consideration are not of Gaussian type, which must be checked by applying the nonparametric Kolmogorov-Smirnov test. This is why the well known parametric regression methods as multiple linear regression would not give satisfactory results for our data.

3.2. Zero order MARS model. Within this study only the best MARS models are presented. They are calculated by using the original eight independent laser variables D, L, PIN, PRF, PNE, $PH2$, DR, PL as predictors and the response variable $Pout$. The models are selected so as to allow no over fitting of the model, as well as by using the algorithm for applying the least squares method [1, 2]. The obtained basic statistic indexes are given in Table 1. All models and statistics are significant at level 0.000.

Firstly we construct the MARS model of the 0^{th} order, without interaction between predictors. It includes the following eleven piecewise linear basic functions, as described in Section 1:

$$
\begin{aligned}
&BF1 = \max(0, D - 4.0);\\
&BF4 = \max(0, 4.5 - PL);\\
&BF7 = \max(0, PH2 - 0.04);\\
&BF8 = \max(0, 0.04 - PH2);\\
&BF9 = \max(0, PNE - 16.5);\\
&BF11 = \max(0, PNE - 19.37);\\
&BF13 = \max(0, DR - 11.8);\\
&BF14 = \max(0, 11.8 - DR);\\
&BF15 = \max(0, PNE - 18.75);\\
&BF17 = \max(0, PNE - 21.88);\\
&BF19 = \max(0, PIN - 1300).
\end{aligned}
\tag{1}
$$

The basic functions include six predictors D, PIN, PNE, $PH2$, DR, PL. As an example, some graphs are shown in Fig. 2 – a, b. The estimated values of laser output power are calculated using the expression



(a) (b)

Fig. 2. The piecewise linear dependences between: a) hydrogen gas pressure PH2 and Pout; b) input electrical power PIN and Pout

$$
\begin{aligned}
Pout = {} & 294.7174 + 11.9923 * BF1 + 71.5160 * BF4 - 10013.2227 * BF7 \\
& - 6750.2637 * BF8 + 83.5532 * BF9 - 417.2946 * BF11 \\
& + 333.0301 * BF13 - 12.3079 * BF14 + 188.3967 * BF15 \\
& + 149.6411 * BF17 + 0.5734 * BF19.
\end{aligned} \tag{2}
$$

With the help of MARS model (1)–(2) it is easy to calculate the estimate of $Pout$ when predictor values are known. The same is valid for predicting a future response. For example, a maximum laser output power $Pout = 1300$ mW has been measured at $D = 26$ mm, PIN=1900 W, $PNE = 19.37$ Torr, $PH2 = 0$ Torr, $DR = 12$ mm, $PL = 5.52$ W/cm/4, $L = 80$ cm and $PRF = 25$ KHz. After substituting the latter in (1)–(2) we find the approximate estimate $Pout = 1055.8$ mW. More results are presented in the next section 4.

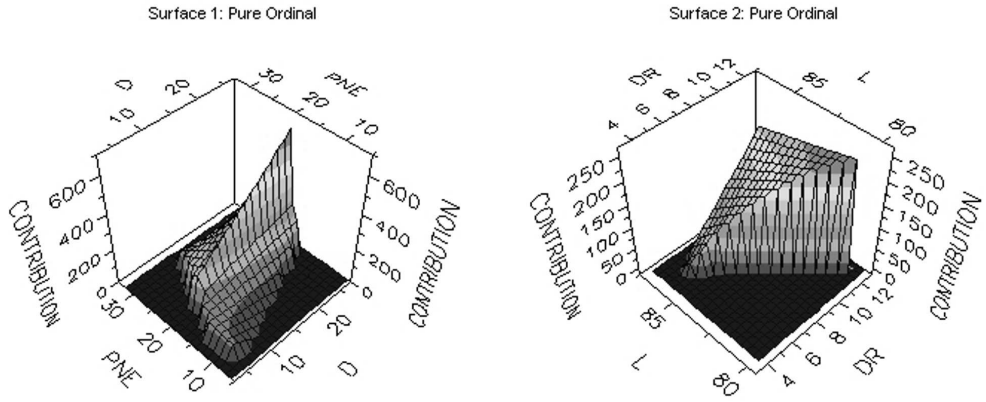
3.3. First order MARS model. The second MARS model which we consider is the one which accounts for possible first order interactions between predictors. The resulting best model includes the following twelve basic functions:

$$
\begin{aligned}
& BF1 = \max(0, D - 4.0); \\
& BF2 = \max(0, L - 80.0); \\
& BF3 = \max(0, PNE - 19.75) * BF1; \\
& BF4 = \max(0, 19.75 - PNE) * BF1; \\
& BF5 = \max(0, PNE - 18.75) * BF1; \\
& BF7 = \max(0, DR - 4.0) * BF2; \\
& BF11 = \max(0, DR - 11.8); \\
& BF13 = \max(0, PH2 - 0.04) * BF11; \\
& BF14 = \max(0, 0.04 - PH2) * BF11; \\
& BF17 = \max(0, PNE - 19.25) * BF1; \\
& BF19 = \max(0, PIN - 1400.0) * BF1; \\
& BF20 = \max(0, 1400.0 - PIN) * BF1;
\end{aligned} \tag{3}
$$

We can see that basic functions in the best model also include six predictors: D, L, PIN, PNE, $PH2$ and DR, almost the same as in (1)–(2). The respective equation that can be used to calculate the estimates of $Pout$ is:

$$\begin{aligned} Pout = {} & 284.3228 + 11.6263 * BF1 - 46.4685 * BF2 + 17.1403 * BF3 \\ & - 2.5005 * BF4 + 38.0011 * BF5 + 4.9766 * BF7 \\ & + 1259.7822 * BF11 - 56044.7969 * BF13 - 27103.3535 \\ & * BF14 - 57.6090 * BF17 + 0.0240 * BF19 - 0.0495 * BF20. \end{aligned} \tag{4}$$

The contributions of some of the interactions to $Pout$ are shown in Fig. 3.



(a) (b)

Fig. 3. The influence of the first order interactions between predictors on $Pout$: a) contribution of PNE and D to $Pout$; b) contribution of L and D to $Pout$

For the same maximum value of $Pout = 1300$ mW, the model (3)–(4) predicts approximately $Pout = 1184.9$ mW which is much better than the corresponding prediction by the model (1)–(2).

3.4. Second order MARS model. The 2nd order best MARS model consists of fifteen basic functions

$$\begin{aligned} BF1 &= \max(0, D - 4.0); \\ BF2 &= \max(0, L - 80.0); \\ BF3 &= \max(0, PNE - 19.75) * BF1; \\ BF4 &= \max(0, 19.75 - PNE) * BF1; \\ BF5 &= \max(0, L - 80.0) * BF4; \\ BF6 &= \max(0, L - 80.0) * BF3; \\ BF7 &= \max(0, PH2 - 0.04) * BF4; \\ BF8 &= \max(0, 0.04 - PH2) * BF4; \\ BF9 &= \max(0, PNE - 17.5) * BF1; \\ BF11 &= \max(0, PIN - 1600.0) * BF9; \end{aligned}$$

$$
\begin{aligned}
(5)\qquad & BF13 = \max(0, DR - 11.8);\\
& BF14 = \max(0, 11.8 - DR);\\
& BF15 = \max(0, PNE - 19.37);\\
& BF16 = \max(0, 19.37 - PNE);\\
& BF17 = \max(0, PIN - 1600.0) * BF13;\\
& BF18 = \max(0, 1600.0 - PIN) * BF13;\\
& BF19 = \max(0, D - 4.0) * BF16;\\
& BF20 = \max(0, PL - 3.28) * BF15.
\end{aligned}
$$

The respective equation to calculate the estimates of $Pout$ is:

$$
\begin{aligned}
Pout = {} & 453.9386 - 10.7268 * BF1 - 42.7187 * BF2 + 36.3037 * BF4\\
& + 0.8727 * BF5 + 0.3679 * BF6 - 381.6795 * BF7\\
(6)\qquad & - 177.4236 * BF8 + 5.2942 * BF9 + 0.0530 * BF11(6)\\
& + 991.4677 * BF13 - 25.4270 * BF14 - 4.7253 * BF17\\
& - 3.5428 * BF18 - 35.0331 * BF19 - 111.1146 * BF20;
\end{aligned}
$$

In particular, the predicted value for $Pout = 1300$ mW by the model (5)–(6) is approximately $Pout = 1248.5$ mW.

4. Discussion. We briefly discuss the obtained results from nonparametric MARS models. From Table 1 it can be seen that the 2^{nd} order model (5)–(6) gives the best results.

Table 1. The basic statistics of constructed best MARS models. GSV is the generalized cross validation measure criterion [10].

MARS model	0^{th} order	1^{st} order	2^{nd} order
R^2	0.9366	0.9533	0.9700
R^2adjusted	0.9318	0.9498	0.9672
GCV	0.9117	0.9244	0.9451
Direct predictors	6	6	7
Terms in model	12	12	15

The quality of estimation of the models is seen in Figs 4–6, respectively. It is given by drawing the correlation between measured and predicted values of the response variable $Pout$. The best results are also observed for the 2^{nd} order interaction. We have to mention that the models of higher order of interaction between predictors do not show further improvement of the obtained results.

Our results can be easily interpreted. For instance, the increase of the inside diameter of the laser tube D and the input electric power PIN leads to increase in the output power $Pout$ (see also Fig. 2b). Also the influence of the other laser parameters is observed and estimated. These results are fully agreed with particular experiment studies [5–9].

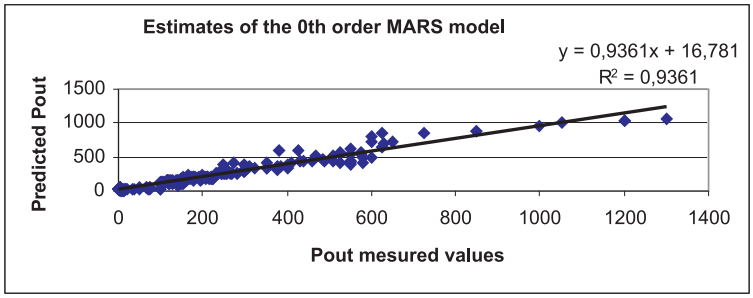


Fig. 4. Correlation of predicted versus measured values of *Pout*, obtained by the first order best MARS model (without interactions)

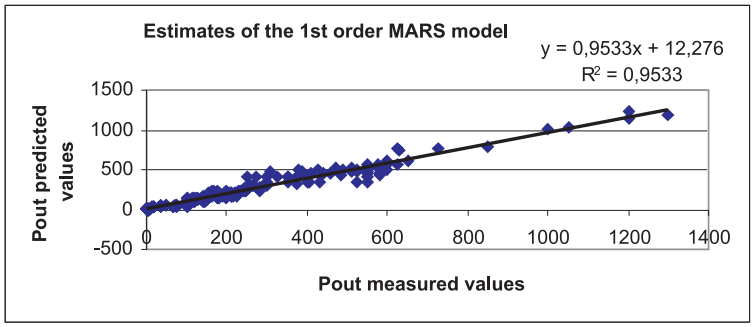


Fig. 5 Correlation of predicted versus measured values of *Pout*, obtained by the first order best MARS model

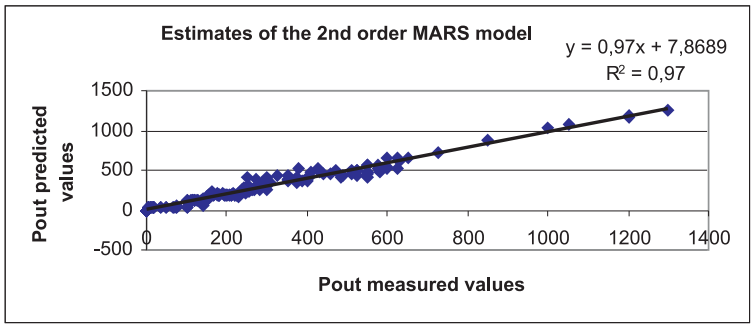


Fig. 6 Correlation of predicted versus measured values of *Pout*, obtained by the second order best MARS model

In our investigation we combine all influences of the input variables and what is more, we obtain the expressions for estimating the values of *Pout* with a necessary statistical inference.

5. Conclusion. Often in a statistical study of real data the standard parametric techniques of regression analysis as multiple regression analysis and logistic regression analysis are not applicable or give unsatisfactory results. The nonparametric method of MARS presented here can be successfully applied for solving many difficult problems in this type of investigation, especially when data is characterized by a multicolinearity, has no normal distribution and the mutual dependences between predictors can not be easily transformed.

In the performed case study of the data arising in laser technology we constructed and compared three MARS models. The obtained results show that the second order model is the best. In fact, this model is strongly nonlinear and contains the piecewise second degree terms of predictor variables.

The obtained models can be used in estimation and prediction of current and future experiments in order to improve the output laser characteristics, in our case the very important one – the output laser power.

REFERENCES

[1] J. H. Friedman. Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, **19**, (1991), No 1, 1–141.

[2] `http://salford-systems.com`

[3] K. Beev, K. Temelkov, N. Vuchkov, Tz. Petrova, V. Dragostinova, R. Stoycheva-Topalova, S. Sainov, N. Sabotinov. Optical properties of polymer films for near UV recording. *J. Optoelectr. and Advanced Materials*, **7**, (2005), 1315–1318.

[4] M. Ilieva, V. Tsakova, N. K. Vuchkov, K. A. Temelkov, N. V. Sabotinov. UV copper ion laser treatment of poly-3,4-ethylenedioxythiophene. *J. Optoelectr. and Advanced Materials*, **9**, (2007), 303–306.

[5] N. K. Vuchkov. High Discharge Tube Resource of the UV Cu+ Ne-CuBr Laser and Some Applications. In: New Development in Lasers and Electric-Optics Research (Ed. W. T. Arkin), New York: Nova Science Publishers Inc, 2007, 41–74.

[6] N. K. Vuchkov, K. A. Temelkov, N. V. Sabotinov. UV Lasing on Cu+ in a Ne-CuBr pulsed longitudinal discharge. *IEEE J. Quantum Electron.*, **35**, (1999), 1799–1804.

[7] N. K. Vuchkov, K. A. Temelkov, P. V. Zahariev, N. V. Sabotinov. Optimization of a UV Cu+ laser excited by pulse-longitudinal Ne-CuBr Discharge. *IEEE J. Quantum Electron.*, **37**, (2001), 511–517.

[8] N. K. Vuchkov, K. A. Temelkov, P. V. Zahariev, N. V. Sabotinov. Influence of the active zone diameter on the UV-Ion Ne-CuBr laser performance. *IEEE J. Quantum Electron.*, **35**, (2001), 1538–1546.

[9] N. K. Vuchkov, K. A. Temelkov, N. V. Sabotinov. Effect of hydrogen on the average output of the UV Cu+ Ne-CuBr laser. *IEEE J. Quantum Electron.*, **41**, (2005), 62–65.

[10] P. Craven, G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, (1979) 377–403.

Snezhana Gocheva-Ilieva
Faculty of Mathematics and Informatics
Plovdiv University "Paisii Hilendarski"
24, Tsar Assen Str.
4000 Plovdiv, Bulgaria

ПРИЛОЖЕНИЕ НА МНОГОМЕРНИТЕ АДАПТИВНИ РЕГРЕСИОННИ СПЛАЙНИ ЗА ПОСТРОЯВАНЕ НА НЕПАРАМЕТРИЧНИ МОДЕЛИ

Снежана Гочева-Илиева

В тази статия са представени основните възможности на сранително новата статистическа техника – Многомерни Адаптивни Регресионни Сплайни (MARS) и съпътстващия софтуерен продукт. Методът MARS е предназначен за статистически анализ на данни, когато стандартното параметрично моделиране с методите на многомерна регресия или логистична регресия не са приложими. Проведено е конкретно изследване на експериментални данни от областта на лазерните технологии, по-специално за моделиране на ултравиолетов йонен лазер с пари на меден бромид. Получените резултати имат добро съвпадение с реално изследваните случаи. Показано е, че построените непараметрични MARS модели могат да се използват за оценка и предсказване на настоящи и бъдещи експерименти, с цел подобряване на изходната лазерна мощност.