The Journal of Animal & Plant Sciences, 30(2): 2020, Page: 431-441
ISSN: 1018-7081

DESCRIPTION OF THE RELATIONSHIPS BETWEEN DIFFERENT PLANT CHARACTERISTICS IN SOYBEAN USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS) ALGORITHM

S. Celik[1] and E. Boydak[2]

[1]Department of Animal Science, Biometry Genetics Unit, Agricultural Faculty, Bingol University, Bingol, Turkey
[2]Department of Field Crops, Faculty of Agriculture, University of Bingol, Turkey.
Corresponding Author E-mail: senolcelik@bingol.edu.tr

ABSTRACT

The aim of this study was to reveal the relationships between several morphological characteristics of the soybean (*Glycine max (L.*) Merr.) plants in the year 2014. For this aim, plant height (PH), first pod height (FPH), branch number (BN), number of nodes (NN), pod number per plant (PNP), seed number per pod (SNP), 1000-seed weight (1000SW), yield per decare (YD) and harvest index (HI) were measured. Five different MARS models were developed for the plant height, first pod height, pod number, harvest index and yield per decare characteristics. The constructed models were evaluated based on the criteria of minimum generalized cross-validation (GCV), SDratio, RMSE, AIC, AICc and maximum coefficient of determination (R^2) in predictive performance. The R^2 values of the MARS models were determined to be 0.902, 0.924, 0.949, 0.987 and 0.998, respectively. For the prediction of PH, FPH and HI, the second degree interaction model was determined to be the most suitable model. For predicting PNP and YD, the third degree interaction MARS model was determined to be the best model. The dependent variables considered here was predicted with a high accuracy by all models established with the MARS algorithm.

As a result, application of the MARS algorithm may allow plant breeders to obtain influential clues in selecting promising soybean varieties.

Keywords: MARS algorithm, soybean, plant characteristics.

https://doi.org/10.36899/JAPS.2020.2.0037 *Published online March 02, 2020*

INTRODUCTION

Soybean agriculture is now carried out across an area of 90.000.000 hectare around the world and annual soybean production is 200.000.000 tons in the year 2014. The largest soybean-growing countries in descending order are USA, Brazil, Argentina, China and India (FAO, 2014). In Turkey, until the 1980s, the largest growers of soybean as a main crop were Central Anatolia Region and Eastern Black Sea Region. However, since 1981, soybean has found a widespread growing area in Çukurova Region as an aftercrop project that was undertaken after the efforts to steer soybean agriculture towards new areas. Ninety percent of the soybean-growing areas in Turkey are located in Adana City (Yosmaoğlu, 2002).

Soybean is the most valuable food source of Asian countries for centuries. It is an agricultural product of great importance because of its nutritional value, benefits for human health and use in hundreds of industrial products.

In animal nutrition, soybean is the most commonly preferred feed raw material in rations for the production of cattle, poultry and aquaculture products due to its high fat and protein content and easy digestibility (Unal and Onder, 2008).

Soybean is a legume that is native to East Asia and grown worldwide for its oil and protein content (FAO, 2015). Soybeans were originally used as nitrogen fixers in former crop rotation systems (Sall and Sinclair, 1991). The development of certain technologies such as fermentation and processing for oil has led to various new enforcements for this beneficial plant.

As a plant from the legume family, its contribution to the nitrogen fixation in soils leads to increases in the yield of following products and allows savings in fertilizer use. Thus, it is one of the most suitable plants for plant rotation (Nazlıcan, 2018). In literature, agronomic performances of various soybean varieties grown under different ecological conditions were evaluated by several authors (Karasu *et al*., 2002; Arslan and Arıoğlu, 2003; Gür *et al*., 2004; Çalıskan *et al*., 2007; Kleinschmidt 2009; Worku and Astatkie, 2015).

In literature, there is great number of the previous studies in the prediction of yield or other properties by using plant characteristics in numerous plants all over the world. In general, basic statistical methods i.e. simple and multiple linear regressions have been adopted for the plant characteristics estimation, while the reliability of the basic methods can be adversely affected by the violation of distributional assumptions (Eyduran *et al*., 2017). The accuracy of this technique depends upon selecting the improved statistical methods such as data mining algorithm CHAID (Akın *et al*., 2016a, b; Akın *et al*., 2017a, b), CART (Kovalchuk *et al*., 2017), ANNs (Karadas *et al*., 2017) and MARS (Eyduran, 2016) thus the precision satisfies to make the right decision on the ideal management conditions. As in other scientific fields, powerful statistical methods i.e. data mining algorithms are required for achieving more effective results on selection of promising soybean varieties within the scope of plant breeding. Among those, MARS is a statistically valuable tool that can capture relationship between sets of dependent and independent variables. There are other studies on MARS algorithm in agricultural sciences (Celik *et al*., 2017; Eyduran *et al*., 2017; Aytekin *et al*., 2018; Aksoy *et al*., 2018).

MARS is an algorithm that can produce a powerful prediction equation in the response variable. However, the MARS algorithm has not yet been investigated in the prediction of some important characteristics from argonomic measurements taken from economically significant plants such as soybean. Therefore, the present study aimed to ascertain the plant characteristics that affect the yield (YD), plant height (PH), first pod height (FPH), pod number (PNP) and harvest index (HI) of soybeans using the Multivariate Adaptive Regression Splines (MARS) method, as a powerful data mining tool.

MATERIALS AND METHODS

MATERIALS: This research was conducted in the year 2014 at the experimental area of Department of Field Crops, Faculty of Agriculture, Bingol University, Bingol province, located in Eastern Anatolia Region of Turkey. Bingol province is located in the High Euphrates District over eastern longitudes from 38° to 40° and over northern latitudes from 38.5° to 39.5° (Bingöl Municipality, 2015). Usually, two types of soil are observed in the province. Brown and brown-red soil is mainly observed in the sloping-rough lowlands, while alluvial soil is observed in the valleys; these soils can vary in terms of organic materials (Bingöl Province Environmental Status Report, 2011).

In the present study, 12 soybean varieties (Yemsoy, Ataem-7, Nazlıcan, Cinsoy, Türksoy, Adasoy, Erensoy, Nova, Yeşilsoy, May 5312, Umut 2002 and Blaze) provided from diverse sources were evaluated to define the circumstances in the Eastern Anatolian region in Turkey. During harvest time, the two lines in the middle were harvested while each line on the sides and the 0.4 m section on the edges were left as edge effects. Then, the harvested platforms were desiccated.

The measured characteristics of the evaluated soybeans in the study were plant height (PH), first pod height (FPH), branch number (BN), number of nodes (NN), pod number per plant (PNP), seed number per pod (SNP), 1000 seeds weight (SW1000), yield per decare (YD), harvest index (HI).

Descriptive statistics for the dependent and independent variables are given in Table 1 for measuring values of plant characteristics in soybean.

Table 1. Descriptive Statistics of plant's measurements.

Plant characteristics	n	Minimum	Maximum	Mean	Std. Error
PH (cm)	36	66.70	112.80	89.192	1.945
FPH (cm)	36	11.90	35.70	26.217	0.914
BN	36	0.90	5.30	3.042	0.155
NN	36	14.50	21.00	17.114	0.267
PNP	36	18.10	47.80	27.603	1.219
SNP	36	2.00	3.10	2.611	0.044
SW1000 (g)	36	96.97	133.35	113.971	1.888
YD (kg)	36	74.17	113.55	91.580	1.943
HI (%)	36	22.35	58.03	33.528	1.516

PH: Plant height, FPH: First pod height, BN: Branch number, NN: Number of nodes, PNP: Pod number per plant, SNP: Seed number per pod, SW1000: 1000 seeds weight, YD: Yield per decare, HI: Harvest index.

Five different MARS models have been evaluated. These models are as follows.

- Dependent variable: PH. Independent variables: FPH, BN, NN, PNP, SNP, SW1000, YD, HI.
- Dependent variable: FPH. Independent variables: PH, BN, NN, PNP, SNP, SW1000, YD, HI.
- Dependent variable: PNP. Independent variables: PH, FPH, BN, NN, SNP, SW1000, YD, HI.
- Dependent variable: YI, Independent variables: PH, FPH, BN, NN, PNP, SNP, SW1000, YD.
- Dependent variable: YD, Independent variables: PH, FPH, BN, NN, PNP, SNP, SW1000.

METHODS: Multivariate adaptive regression splines (MARS) is a data mining technique that can be used for solving regression-type problems (Hastie *et al*., 2001).

As a widening of classification and regression tree (CART) algorithm, MARS is an effective machine learning algorithm that define the relation between a dependent variable and a set of independent variables (Celik *et al*., 2019).

It is a non-parametric procedure, for invention adaptive regressions that uses piecewise basis functions to define relationships between a dependent variable and a set of estimations. MARS allows for the capture of linear and additive relationships and for the separation in excess of all nodes at each step, rather than just the terminal ones. Hence, MARS compose a bended regression line to fit the data from subgroup to subgroup and from spline to spline. (Friedman, 1991).

In every spline, MARS splits the data anymore inside many subgroups. Several knots are constituted by MARS. These knots can be established between distinct input variables or distinct intervals in the same input variable, to separate the subgroups. The data of each subgroup are called basis function (BF). The model takes the form of an expansion in product spline basis functions, where the number of basic functions as well as the parameters associated with each one (product degree and knot locations) are automatically determined by the data(Friedman, 1991; Sephton 2001).

The MARS algorithm constructs models from two sided functions of the predictors (x) of the form:

$$(x-t)_+ = \begin{cases} x-t, & x>t \\ 0, & otherwise \end{cases}$$

These serve as basis functions for linear or nonlinear expansion that approximates some true underlying function *f(x)*.

The MARS model for a response variable *y*, and M terms, can be given in the sequent equation:

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m H_{km}(X_{v(k,m)})$$

where the aggregate is over the M terms in the model, and β_0 is an intercept, β_m is a coefficient of basis functions, $H_{km}(X_{v(k,m)})$ is a basis function, here $v(k,m)$ is an index of a predictor for an mth component of a kth product (Hastie *et al*., 2001). Function H is defined as:

$$H_{km}(X_{v(k,m)}) = \prod_{k=1}^{K} h_{km}$$

where *xv(k,m)* is the predictor in the k'th of the m'th product. Here, k is a parameter interaction order. For order of interactions *K*=1, the model is additive and for *K=2* the model pairwise interactive (Friedman, 1991).

During forward step, a number of basis functions are added to the model according to a pre-determined maximum which should be considerably larger (twice as much at least) than the optimal (best least-squares fit) (Hastie *et al*., 2001).

A backward procedure is applied in which the model is pruned by removing those basis functions that are associated with the smallest increase in the goodness-of-fit. Generalized Cross Validation error is a measure of the goodness of fit that takes into account both the residual error and the model complexity as well. It is formulated by (Koronacki and Ćwik 2005).

$$GCV = \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\left[1-\frac{C}{n}\right]^2}$$

with,

C=1+cd

Where n is the number of cases in the data set, d is the effective degrees of freedom, which is equal to the number of independent basis functions. The quantity C is the penalty for adding a basis function (Hastie *et al*., 2001).

To comparatively test the estimate criteria of MARS, the following goodness of fit criteria were used (Willmott and Matsuura, 2005; Liddle, 2007; Takma *et al*., 2012; Eyduran *et al*., 2019):

1. Pearson correlation coefficient (r) between the actual and predicted dependent variable values,
2. Coefficient of Determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

3. Adjusted Coefficient of Determination

$$Adj.R^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

4. Root-mean-square error (RMSE) given by the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

5. Standard deviation ratio (SD$_{ratio}$):

$$SD_{ratio} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

6. Akaike Information Criterion (AIC):

$$AIC = n\log\sum_{i=1}^{n}\left(\frac{\left(Y_i - \hat{Y}_i\right)^2}{n}\right) + 2k$$

7. Corrected Akaike Information Criterion (AICc):

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

Where k is the number of selected terms and n is the sample size (Hu, 2007).

Here, Y_i: the observed dependent variable value of ith plant, $\hat{Y}_i$: the predicted dependent values of ith plant $\bar{Y}$: average of the dependent variable values of the plant, ε_i: the residual value of ith plant, $\bar{\varepsilon}$: average of the residual values, k: number of the selected terms in the model, and n: total plant number. The residual value of each observation is expressed as $\varepsilon_i = Y_i - \hat{Y}_i$.

The MARS analysis was performed using the earth package of R software (Milborrow, 2011; Milborrow, 2018; R Core Team, 2014; Eyduran *et al*., 2019).

RESULTS

In this study, five different MARS models were developed to predictive five different dependent variables such as PH, FPH, PNP, YD and HI. The goodness-of-fit statistics (r, R^2, Adj. R^2, SDratio, AIC, AICc and GCV) were calculated to measure predictive performances of the developed MARS models. Results of predictive performances of the MARS models are reported in Table 2. It was understood that the fitted MARS models had high predictive accuracy (Table 2). Grzesiak and Zaborski (2012) reported that the model having SD ratio less than 0.40 had a good fit, as also reported by Eyduran *et al*. (2019). For instance, it was determined that 90.2% of total variability in PH was explained.

Table 2. Goodness of fit criteria for MARS algorithm.

Variables	r	R^2	Adj. R^2	RMSE	SDratio	AIC	AICc	GCV
PH	0.950	0.902	0.878	3.596	0.313	106	110	10.716
FPH	0.961	0.924	0.908	1.494	0.276	41	44	7.712
PNP	0.974	0.949	0.929	1.623	0.225	55	64	15.201
HI	0.993	0.987	0.980	1.029	0.115	26	40	22.203
YD	0.999	0.998	0.994	0.553	0.048	5	115	2.757

Plant height (PH): Results of the MARS algorithm for plant height of soybean are presented in Table 3. All the coefficients for plant height were found very significant (P<0.01). Soybean plants with FPH > 29.2 cm (BF2) would be expected to produce higher PH and the effect of BF1 on PH was masked for those with FPH > 29.2 cm. For soybean plants whose PNP value was more than 28.5, no adverse effect of coefficient (-0.717) of BF3 on PH was found. For soybean plants with YD < 83.88, more increase in PH would be expected as YD decreased to its own lowest value (BF4, P<0.01). In other words, YD had a positive effect on PH for the plants with YD < 83.88

However, HI had the adverse effect on PH for soybean plants with HI > 28.77 (BF5, P<0.001). However, the effect of HI on PH for the plants with HI < 28.77 was masked (BF 5 and BF6). Also, the effect of YD on PH for the plants with HI < 28.77 was also masked by HI when BF6 (the first degree interaction term) was considered.

Table 3. Results of the MARS algorithm for PH of soybean.

	Basic functions	Coefficients	Significance level (p)
	Intercept	96.412	< 2e-16 ***
BF1	max(0, 29.2-FPH)	-0.916	0.003578 **
BF2	max(0, FPH-29.2)	1.859	0.001258 **
BF3	max(0, 28.5-PNP)	-0.717	0.004012 **
BF4	max(0, 83.88-YD)	1.025	0.002097 **
BF5	max(0, HI-28.77)	-8.196	< 2e-16 ***
BF6	YD * max(0, HI-28.77)	0.083	0.004578 **

**p<0.01, ***p<0.001

First pod height (FPH): Results of the MARS algorithm for first pod height of soybean are presented in Table 4. All coefficients in Table 4 were very significant (P<0.01). For the plants with PH > 84.3, PH had a roughly threefold decrease effect on FPH compared with those having PH < 84.3 (BF1 and BF2). For the plants with BN < 3.1, BN had a positive effect on FPH (BF3), whereas the adverse effect of BN on FPH was masked in BF4.

For the plants with BN < 3.1, the positive effect of SW1000 on FPH was masked (BF6), meaning that the effect of SW1000 on FPH could be based on BN. When BF7 was considered, no effect of PNP on FPH was found for the plants with NN < 16.8. The adverse effect of PH on FPH in BF1 and BF2 converted into a small increment for the plants with YD in BF5.

Table 4. Results of the MARS algorithm for FPH of soybean.

	Basic functions	Coefficients	Significance level (p)
	Intercept	24.523	< 2e-16 ***
BF1	max(0,84.3-PH)	-0.431	0.003357 **
BF2	max(0,PH-84.3)	-1.319	0.003743 **
BF3	max(0,3.1-BN)	4.105	3.20e-07 ***
BF4	max(0,BN-3.3)	-26.240	< 2e-16 ***
BF5	max(0,PH-84.3) * YD	0.016	0.004133 **
BF6	max(0,BN-3.1) * SW1000	0.195	0.003729 **
BF7	max(0,NN-16.8) * PNP	-0.052	0.003708 **

**p<0.01, ***p<0.001

FPH prediction equation in terms of the 8 BFs for the MARS model are presented below.
FPH = 24.523 - 0.431 * max(0.84.3-PH)-0.319 * max(0.PH-84.3) + 4.105 * max(0.3.1-BN) -26.240 * max(0.BN-3.3) +0.016* max(0.PH-84.3) * YD + 0.195 * max(0.BN-3.1) * SW1000 - 0.052 * max(0.NN-16.8) * PNP

Pod number per plant (PNP): Results of the MARS algorithm for bean number of soybean are given in Table 5. All the coefficients were very significant (P<0.01) for MARS model with ten selected terms. The second interaction effect of the constructed MARS predictive model for PNP was significant, meaning that the effect of PH on PNP could vary based on values of SW1000 and HI (BF9). For soybean varieties with PH > 84.3 cm, PH had a positive effect on PNP, but the effect of PH on PNP could change according to SW1000 and HI when BF9 was considered. For example, the effect of BF1 and BF9 on PNP was masked for the soybean varieties with SW1000 < 115.95. For those having FPH > 25.6 cm. no significant effect of FPH was found for PNP, but PNP would be expected to be on the increase as FPH value decreased from 25.6 to the possible smallest value.

Table 5. Results of the MARS algorithm for PNP of soybean.

	Basic functions	Coefficients	Significance level (p)
	Intercept	13.633	< 2e-16 ***
BF1	max(0,PH-84.3)	0.561	0.003144 **
BF2	max(0,25.6-FPH)	1.657	1.05e-07 ***
BF3	max(0,SW1000-115.95)	2.130	2.27e-08 ***
BF4	max(0,YD-83.88)	2.499	3.52e-08 ***
BF5	max(0,91.75-YD)	1.058	1.137e-07 ***
BF6	max(0,YD-91.75)	-2.876	7.09e-06 ***
BF7	max(0,35.2-HI)	-1.055	0.000109 ***
BF8	max(0,HI-35.2)	-0.914	0.002785 **
BF9	PH * max(0,SW1000-115.95) * HI	-0.001	0.004013 **

**p<0.01, ***p<0.001

The PNP prediction equation in terms of the 10 BFs for the MARS model are presented in below.
PNP = 13.633 + 0.561 * max(0,PH-84.3) + 1.657 * max(0,25.6-FPH) + 2.130 * max(0,SW1000-115.95) + 2.499 * max(0,YD-83.88) + 1.058 * max(0,91.75-YD) - 2.876 * max(0,YD-91.75) - 1.055 * max(0,35.2-HI) - 0.914 * max(0,HI-35.2) - 0.001 * PH * max(0,SW1000-115.95) * HI

Harvest index (HI): Results of the MARS algorithm for harvest index (HI) of soybean are presented in Table 6. All coefficients in the MARS prediction equation were found very significant for MARS model with 12 selected terms (P<0.01).

Table 6. Results of the MARS algorithm for HI of soybean.

	Basic functions	Coefficients	Significance level (p)
	Intercept	29.895	< 2e-16 ***
BF1	max(0,PNP-26.4)	0.651	0.000385 ***
BF2	max(0,SNP-2.7)	11.485	< 2e-16 ***
BF3	max(0,SW1000-109.17)	-1.452	0.000025 ***
BF4	max(0,83.88-YD)	0.922	0.000036 ***
BF5	max(0,YD-83.88)	1.124	0.000009 ***
BF6	max(0,89.4-PH) * FPH	-0.028	0.004822 **
BF7	max(0,89.4-PH) * PNP	-0.053	0.000035 ***
BF8	max(0,89.4-PH) * YD	0.034	0.002723 **
BF9	FPH * max(0,SW1000-109.17)	0.042	0.001936 **
BF10	SNP * max(0,109.17-SW1000)	-0.175	0.000183 ***
BF11	SNP * max(0,YD-83.88)	0.490	0.000034 ***

**p<0.01, ***p<0.001

The HI prediction equation in terms of the 12 BFs for the MARS model is presented below.
HI = 29.895 + 0.651 * max(0, PNP-26.4) + 11.485 * max(0,SNP-2.7) - 1.452 * max(0, SW1000-109.17) + 0.922 * max(0,83.88-YD) + 1.124 * max(0,YD-83.88) - 0.028 * max(0,89.4-PH) * FPH - 0.053 * max(0,89.4-PH) * PNP + 0.034 * max(0, 89.4-PH) * YD + 0.042 * FPH * max(0,SW1000-109.17) - 0.175 * SNP * max(0,109.17-SW1000) + 0.490 * SNP * max(0,YD-83.88)

Yield per decare (YD): Results of the MARS algorithm for yield per decare (YD) of soybean are reported in Table 7. All the coefficients of the MARS prediction equation were found very significant (P<0.01) for MARS model with 24 selected terms (P<0.01).

All the interacting and non-interacting variables affecting yield in soybean plants are clearly shown in detail (Table 7). In addition, the R codes of the MARS algorithm of the effect of the variables affecting the decar yield are given in the Appendix. The MARS model R commands that examine other dependent variables are similar, except that the dependent variable is different.

Table 7. Results of the MARS algorithm for YD of soybean.

	Basic functions	Coefficients	Significance level (p)
	Intercept	79.533	< 2e-16 ***
BF1	max(0,25.2-FPH)	10.463	< 2e-16 ***
BF2	max(0,FPH-25.2)	20.343	< 2e-16 ***
BF3	max(0,BN-2.7)	40.581	< 2e-16 ***
BF4	max(0,BN-2.9)	-84.060	< 2e-16 ***
BF5	max(0,3.1-BN)	113.115	< 2e-16 ***
BF6	max(0,BN-3.1)	45.415	< 2e-16 ***
BF7	max(0,SNP-2.6)	88.867	< 2e-16 ***
BF8	max(0,2.7-SNP)	-165.410	< 2e-16 ***
BF9	max(0,SNP-2.7)	-20.673	< 2e-16 ***
BF10	max(0,115.95-SW1000)	11.908	< 2e-16 ***
BF11	max(0,SW1000-115.95)	-2.878	4.33e-09 ***
BF12	max(0,SW1000-123)	-6.767	3.58e-07 ***
BF13	max(0,35.2-HI)	5.649	2.46e-06 ***
BF14	PH * max(0,3.1-BN)	-1.278	5.99e-04 ***
BF15	FPH * max(0,PNP-26.4)	0.048	0.003581 **
BF16	max(0,25.2-FPH) * SNP	-4.221	1.72e-05 ***
BF17	max(0,FPH-25.2) * SNP	-7.511	6.34e-10 ***
BF18	FPH * max(0,HI-35.2)	-0.036	0.006718 **
BF19	NN * max(0,2.7-SNP)	9.633	< 2e-16 ***
BF20	SNP * max(0,115.95-SW1000)	-5.801	< 2e-16 ***
BF21	max(0,SW1000-123) * HI	0.328	0.000425 ***
BF22	BN * PNP * max(0,35.2-HI)	-0.038	0.002519 **
BF23	SNP * max(0,115.95-SW1000) * HI	0.034	0.000124 **

**p<0.01, ***p<0.001

DISCUSSION

In the present study, the plant height varied from 66.70 cm to 112.80 cm, with an average plant height of 89.192 cm. This result was in agreement with those reported by Öz (2002), Çopur *et al*. (2009) and Souza *et al*. (2016), but disagreement with those found by Karasu *et al*. (2002) and Yetkin and Arioglu (2009) and higher than those determined by Yari *et al*. (2013) and Naidu *et al*. (2017). These differences are attributable to the differences in varieties and different planting dates and climatic and growing conditions.

First pod height varied from 11.90 cm to 35.70 cm. The present results obtained for the first pod height were found different from the results obtained by Karasu *et al*. (2002) and Yetkin and Arioglu (2009), as a result of the differences in climate and region.

The average harvest index value was determined to be 33.528%. This value was different from the value determined by Öz (2002), higher than the value determined by Yetkin and Arioglu (2009) and lower than the values determined by Daraz *et al*. (2014), Kundu *et al*. (2016) and Aboutalebian and Malmir (2017). The differences in the values may be ascribed to the use of different varieties and differently reduced irrigation levels.

The average pod number per plant was determined to be 3.042 (with the range of 0.90 to 5.30), which was lower than that reported by Souza *et al*. (2016).

The average 1000-seed weight was 113.971 g. This present value was lower than the values determined by Öz (2002), Yetkin and Arioglu (2009), Çopur *et al*. (2009), Yari *et al*. (2013), Souza *et al*. (2016) and Aboutalebian and Malmir (2017), but higher than that obtained by Kundu *et al*. (2016 The average yield per decare was determined to be 91.58 kg/da, which was in disagreement with those reported by some previous authors (Yılmaz and Efe, 1998; Karasu (2002), Unal and Onder. 2008). The difference in yield values is due to the use of a different hybridization method by the researchers.

There are limited number of studies using the MARS method in agricultural sciences (Aytekin *et al*., 2018; Celik *et al*., 2018; Eyduran *et al*., 2017). In a previous study by Chavan *et al*. (2016), path analysis was performed to determine the direct and indirect effects of various characters in soybean. The characters number of pods per plant (0.2919), 100 seed weight (0.4996), days to 50% flowering (0.2068), oil content (0.2176) and days to maturity (0.0531) had high positive direct effect on seed yield per plant (R^2=0.5669).

In another study, path analysis was used to estimate the direct and indirect effects of various characters in soybean yield (Silva *et al*., 2015). Through path analysis, it was determined that the number of seeds/plant was the component of a more direct influence on the yield, and the influence of the number of pods/plant in the productivity is based on indirect influence of the number of seed plant (R^2=0.74). Board (2002), investigated a regression model to predict soybean cultivar yield performance at late planting dates. In the regression model, it was shown that plant height, seed-filling period and total dry matter affect the yield of soybean. El-Mohsen *et al*. (2013) showed simple linear regression analysis of important relationships of irrigation regimes and seed, oil and protein yield. The obtained models were Y=2976-258.9X (R^2=0.62), Y=568.1-50.55X (R^2=0.54), Y=1187-96.96X (R^2=0.64). Where X is the irrigation regime, Y is seed yield, oil yield and protein yield for each model, respectively. The present R^2 results were found much better than those reported by some authors (Board, 2002; Silva *et al*., 2015; Chavan *et al*., 2016). The wide variation may be attributed to differences in plant materials, ecological conditions, and the constructed statistical models. As well-known, the reliability of the obtained results are dependent on choosing influential independent variables and powerful statistical approaches i.e. Artificial Neural Networks (ANNs) and MARS algorithm (Eyduran *et al*., 2018). The present study provided good evidence on the superiority of MARS data mining algorithm.

In agricultural sciences, student t test, one-way ANOVA, two-way ANOVA, multiple linear regression analysis have been widely used (Agaoglu *et al*., 2007; Atila *et al*., 2006a,b; Eyduran and Agaoglu, 2007; Eyduran *et al*., 2007a,b; Eyduran *et al*., 2008a,b,c; Eyduran *et al*.,2015a,b,c; Akin *et al*., 2016a; Eyduran *et al*., 2016; Gecer *et al*., 2016; Eyduran *et al*., 2018; Eyduran *et al*., 2019). Also, much more sophistical approaches i.e. data minings have been adopted recently (Grzesiak and Zaborski, 2012; Akin *et al*., 2016b; Akin *et al*., 2017a,b,c; Akin *et al*., 2018; Gozuacik *et al*., 2018).

Although there are a large number of studies on soybean, no studies were found investigating the use of the MARS algorithm in connection with plant characteristics. Hence, we were unable to provide a further discussion of the matter.

Conclusion: MARS predictive models with the first and second degree interaction effects were developed using the MARS algorithm to estimate the plant height. The first pod height, pod number, yield per decare, harvest index and the yield per decare x harvest interaction affected the plant height. The explanatory power of the established models was 0.902, 0.924, 0.949, 0.987 and 0.998, respectively. The MARS algorithms were determined to be good predictors of the plant characteristics and relationship between these characteristics in agriculture. It can be suggested that the use of MARS algorithm will be beneficial in future studies in agriculture.

REFERENCES

Aboutalebian, M. A., and M. Malmir (2017). Soybean yield and yield components affected by the mycorrhiza and bradyrhizobium at different rates of starter nitrogen fertilizer. Semina: Ciências Agrárias, Londrina, 38(4): 2409-2418.

Agaoglu Y. S., S. P. Eyduran, and E. Eyduran (2007). Comparison of Some Pomological Characteristics of Blackberry Cultivars Growth in Ayaş Conditions. Ankara Universitesi Tarım Bilimleri Dergisi, 13(1): 69-74.

Akin, M., E. Eyduran, and B. M. Reed (2016a). Using the CHAID data mining algorithm for tissue culture medium optimization. In: In vitro cellular and developmental biology-animal. 52, Spring ST, New York, NY 10013, USA, pp. 233.

Akin, M., S. P. Eyduran, S. Ercisli, V. Kapchina-Toteva and E. Eyduran (2016b). Phytochemical Profiles of Wild Blackberries, Black and White Mulberries from Southern Bulgaria. Biotechnology and Biotechnological Equipment 30(5): 899-906

Akin, M., E. Eyduran, and B. M. Reed (2017a). Use of RSM and CHAID data mining algorithm for predicting mineral nutrition of hazelnut. Plant Cell Tissue Organ Cult. 128: 303-316.

Akin, M., E. Eyduran and B. M. Reed (2017b). Developing of optimal tissue culture medium for Hazelnuts. IX International Congress on Hazelnut, 15-19 August, Atakum, Samsun, Turkey.

Akin, M., E. Eyduran, R. P. Niedz and B. M. Reed (2017c). Developing hazelnut tissue culture medium free of ion confounding. Pl. Cell Tissu. Organ Cult., **130**: 483-494. https://doi.org/10.1007/s11240-017-1238-z

Akin, M., C. Hand, E. Eyduran and B. M. Reed (2018). Predicting minor nutrient requirements of hazelnut shoot cultures using regression trees. Pl. Cell Tissu. Organ Cult., 132: 545-559. https://doi.org/10.1007/s11240-017-1353-x

Aksoy, A., Y. E. Ertürk, E. Eyduran and M. M. Tariq (2018). Comparing predictive performances of MARS and CHAID algorithms for defining factors affecting final fattening live weight in cultural beef cattle enterprises. Pakistan J. Zoology, 50(6): 2279-2286.

Arslan, M., and H. Arıoğlu (2003). Determining of soybean (*Glycine max (L.) Merr*.) cultivars and ideal plant type as a double crop for Amik Plain.

J. Agriculture Faculty Çukurova University, 18(3): 39-46.

Atila, S. P., Y. S. Agaoglu, and M. Celik (2006a). A Research on the Adaptation of Some Raspberry Cultivars in Ayas (Ankara) Conditions. Pakistan J. Biological Sciences; 9(8): 1504-1508.

Atila, S. P., Y. S. Agaoglu, and M. Celik (2006b). A Research on the Adaptation of Some Blackberry Cultivars in Ayas (Ankara) Conditions. Pakistan J. Biological Sciences; 9(9): 1791-1794.

Aytekin, I., E. Eyduran, K. Koksal, R. Akşahan, and I. Keskin (2018). Prediction of fattening final live weight from some body measurements and fattening period in young bulls of crossbred and exotic breeds using MARS data mining algorithm. Pakistan J. Zoology, 50(1): 189-195.

Bingöl Municipality (2015). Bingöl Belediyesi Stratejik Planı (2015-2019). Bingöl Belediyesi.

Board, J. E. (2002). A regression model to predict soybean cultivar yield performance at late planting dates. Agronomy J., 94:483-492.

Celik, S., E. Eyduran, A. Tatliyer, K. Karadas, M. K. Kara, and A. Waheed (2018). Comparing predictive performances of some nonlinear functions and Multivariate Adaptive Regression Splines (MARS) for describing the growth of Daera Dın Panah (DDP) goat in Pakistan. Pakistan J. Zoology, 50(3): 1187-1190.

Celik, S., E. Eyduran, K. Karadas and M. M. Tariq (2017). Comparison of predictive performance of data mining algorithms in predicting body weight in Mengali rams of Pakistan**.** Brazilian J. Anim. Science, 46(11): 863-872.

Celik, S., M. Akın, P. Aliyev, S. P. Eyduran, and E. Eyduran (2019). A hybrid approach of combining factor analysis scores with MARS predictive model for regression problems. 1. International Erciyes Agriculture, Animal and Food Sciences Conference (AGANFOS). 24-27 April, Kayseri, Turkey.

Chavan, B. H., D. V. Dahat, H. J. Rajput, M. P. Deshmukh and S. L. Diwane (2016). Correlation and path analysis in soybean. International Research J. Multidisciplinary, 2(9): 1-5.

Çalışkan S., M. Arslan, I. Üremiş, and M. E. Çalışkan (2007). The effects of row spacing on yield and yield components of full season and double cropped soybean. Turkish J. Agriculture and Forestry, 31: 147-154.

Çevre ve Şehircilik İl Müdürlüğü (2011). Bingöl İl Çevre Durum Raporu. T.C. Bingöl Valiliği Çevre ve Şehircilik İl Müdürlüğü.

Çopur, O., M. A. Gur, U. Demirel, and M. Karakus (2009). Performance of some soybean [*Glycine max (L.) Merr.*] genotypes double cropped in semi-arid conditions. Notulae Botanicae Horti Agrobotanic Cluj-Napoca, 37(2): 85-91.

Daraz, G., M. Hameed, F. Ahmad, and Waheedullah (2014). The response of different soybean varieties yield and yield components to different reduced irrigation levels in District Swat of Pakistan. J. Biology, Agriculture and Healthcare, 4(6): 6-10.

El-Mohsen, A. A. A., G. O. Mahmoud and S. A. Safina (2013). Agronomical evaluation of six soybean cultivars using correlation and regression analysis under different irrigation regime conditions. J. Plant Breeding and Crop Science, 5(5): 91-102.

Eyduran, E. (2016). The possibility of using data mining algorithms in prediction of live body weights of small ruminants. Canadian J. Applied Sci. 1:18-21.

Eyduran, E., M. Akin, and S. P. Eyduran (2019). Application of multivariate adaptive regression splines in agricultural sciences through R Software. Nobel Bilimsel Eserler Sertifika No: 20779, Ankara. ISBN: 978-605-2149-81-2.

Eyduran, E., O. Akkus, M. K. Kara, C. Tirink, and M. M. Tariq (2017). Use of Multivariate Adaptive Regression Splines (Mars) in predicting body weight from body measurements in Mengali rams. International Conference on Agriculture, Forest, Food, Sciences and Technologies, ICAFOF, 15-17 May 2017, Cappadocia-Turkey.

Eyduran, S. P. and Y. S. Ağaoğlu (2007). Some pomological and plant characteristics of currant varieties cultivated in Ankara (Ayaş) Condition. Ankara Üniversitesi Tarım Bilimleri Dergisi, 13(3): 293-298.

Eyduran, S. P., Y. S. Ağaoğlu, E. Eyduran, and T. Özdemir (2007a). Comparison of some raspberry cultivars' herbal features by repeated random complete design statistic technique. Pakistan J. Biological Sciences, 10(8): 1270-1275.

Eyduran, S. P., T. Özdemir, and Y. S. Ağaoğlu (2007b). Ankara (Ayaş) koşullarında yetiştirilen böğürtlen çeşitlerinin bazı bitkisel özellikleri. Alatarım. 6 (1): 18-25.

Eyduran, S. P., E. Eyduran, and Y. S. Ağaoğlu (2008a). Estimation of fruit weight by cane traits for various raspberries (*Rubus ideaus L.*) cultivars. African J. Biotechnology, 7(17): 3044-3052.

Eyduran, S. P., Eyduran, E., and Y. S. Ağaoğlu (2008b). Estimation of fruit weight by cane traits for eight American Blackberries (*Rubus fructicosus L.*) cultivars. African J. Biotechnology, 7(17): 3031-3038.

Eyduran, S. P., E. Eyduran, K. M. Khawar, and Y. S. Agaoglu (2008c). Adaptation of eightAmerican

Blackberry (*Rubus fructicosus L.*) Cultivars for Central Anatolia. African J. Biotechnology, 7(15): 2600-2604.

Eyduran, S. P., M. Akin, S. Ercisli, E. Eyduran, and D. Maghradze (2015a). Sugars, organic acids, and phenolic compounds of ancient grape cultivars (*Vitis vinifera* L.) from Igdir Province of Eastern Turkey. Biol Res. 48(2): 1-8.

Eyduran, S. P., M. Akin, S. Ercisli, and E. Eyduran (2015b). Phytochemical profiles and antioxidant activity of some grape accessions (*Vitis* spp.). Native to Eastern Anatolia of Turkey. J. Appl. Bot. Food Qual., 88: 5–9.

Eyduran, S. P., S. Ercisli and M. Akin (2015c). Organic acids, sugars, vitamin C, antioxidant capacity, and phenolic compounds in fruits of white (*Morus alba* L.) and black (*Morus nigra* L.) mulberry genotypes. J Appl Bot Food Qual. 88: 134-138.

Eyduran, E., H. Sevgenler, M. Akın and S. P. Eyduran (2018). Usage multivariate adaptive regression splines for predicting continuous responses. Animal and Plant Sciences. International Agricultural Science Congress. 9-12 May, Van, Turkey.

FAO (2017). Food and Agriculture Organization of the United States. Production statistics, crops. http://www.fao.org/faostat/en/#data/QC.

FAO (2015). Food and Agriculture Organization of the United States. The role of soybean in fighting world hunger. http://www.fao.org/3/a-bs958e.pdf.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines, Annals of Statistics, 19(1): 1-67.

Gecer, M. K., M. Akin, M. Gundogdu and S. P. Eyduran (2016). Organic Acids, Sugars, Phenolic Compounds, and Some Horticultural Characteristics of Black And White Mulberry Accessions from Eastern Anatolia. Can J Plant Sci. 96(1):27-33.

Gozuacik, C., E. Eyduran, H. Cam and M. K. Kara (2018). Detection of infection preferences of the alfalfa seed chalcid, Bruchophagus roddi Gussakovskiy, 1933 (Hymenoptera: Eurytomidae) in alfalfa (Medicago sativa L.) fields of Igdir, Turkey. Legume Res., 41: 150-154.

Grzesiak, W. and D. Zaborski (2012). Examples of the use of data mining methods in animal breeding. In: Data mining applications in engineering and medicine (ed. A. Karahoca). InTech. https://doi.org/10.5772/2616.

Gür, M. A., O. Çopur, M. Karakuş, and U. Demirel (2004). Determination of yield and yield components for some soybean (*Glycine max. L. Merill.*) Genotypes under the Harran Plain conditions. GAP IV. Tarım Kongresi. 21–23 Eylül. Şanlıurfa.

Hastie, T., R. Tibshirani, and J. Friedman (2001). The elements of statistic al learning; Data mining, inference and prediction. Springer Verlag, New York.

Hu, S. (2007). Akaike Information Criterion. Center for Research in Scientific Computation North Carolina State University Raleigh, NC (http://www4.ncsu.edu/~shu3/Presentation/AIC.pdf).

Karadas, K., M. Tariq, M. M. Tariq, and E. Eyduran (2017). Measuring predictive performance of data mining and Artificial Neural Network algorithms for predicting lactation milk yield in indigenous Akkaraman sheep. Pakistan J. Zool., 49(1):1-7

Karasu, A., M. Öz, and A. T. Göksoy (2002). A study on the adaptation of some soybean (*Glycine max (L.) Merill*) cultivars under Bursa Conditions. Uludağ Üniversitesi Ziraat Fakültesi Dergisi 16(2): 25-34.

Kleinschmidt, A. (2009). Soybean seeding rate and row width effect on yield. http://agvanwert.wordpress.com

Kornacki, J. and J. Ćwik (2005). Statistical learning systems (in Polish). WNT Warsaw.

Kovalchuk, I.Y., Z. Mukhitdinova, T. Turdiyev, G. Madiyeva, M. Akin, E. Eyduran, and B. M. Reed (2017). Modeling some mineral nutrient requirements for micropropagated wild apricot shoot cultures. Plant Cell Tiss Organ Cult (2017). doi:10.1007/s11240-017-1180-0.

Kundu, P. K., T. S. Roy, S. H. Khan, K. Parvin, and H. E. M. Khairul Mazed (2016). Effect of sowing date on yield and seed quality of soybean. J. Agriculture and Ecology Research International, 9(4): 1-7.

Liddle, A. R. (2007). Information criteria for astrophysical model selection. Monthly Notices of the Royal Astronomical Society: Letters 377: L74-L78.

Milborrow, S. (2011). Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate adaptive regression splines, 2011. R package.

Milborrow, S. (2018). Notes on the earth package. http://www.milbo.org/doc/earth-notes.pdf pp:1-64.

Naidu, C. R., G. K. Reddy, V. Sumathi, and P. V. M. Reddy (2017). Response of soybean varieties to different sowing times. J. Pharmacognosy and Phytochemistry, 6(5): 1092-1095.

Nazlıcan, A. N. (2018). Soya yetiştiriciliği. https://www.foodelphi.com/2015/02/24/page/8/ (Accessed to: 20.07.2017).

Oz, M. (2002). The effect of different plant populations and nitrogen doses on the yield and yield components in soybean under Bursa, Mustafa Kemal Paşa conditions. Uludağ Üniversitesi Ziraat Fakültesi Dergisi 16: 165-177.

R Core Team. (2014). R*:* A language and environment for statistical computing. R Foundation for Statistical Computing, 2014. Available at: http://www.Rproject.org.

Sall, K. and T. R. Sinclair (1991). Soybean genotypic differences in sensitivity of symbiotic nitrogen fixation to soil dehydration. Plant and Soil, 133: 31-37.

Sephton, P. (2001). Forecasting recessions: Can we do better on MARS(™)?Review (Federal Reserve Bank of St. Louis), (March/April), pp.39-49.

Silva, A. F., T. Sediyama , F. C. S. Silva , A. R. G. Bezerra and L.V. Ferreira (2015). Correlation and path analysis of soybean yield components. International J. Plant, Animal and Environmental Sciences, 5(1): 177-179.

Souza, R., I. Teixeira, E. Reis, and A. Silva (2016). Soybean morphophysiology and yield response to seeding systems and plant populations. Chilean J. Agricultural Research 76(1): 3-8.

Takma. C., H. Atil, and V. Aksakal (2012). Comparison of multiple linear regression and Artificial Neural Network models goodness of fit to lactation milk yields. Kafkas Üniversitesi Veteriner Fakültesi Dergisi 18: 941-944.

Unal, I., and M. Onder (2008). Determination of some agricultural characteristics of the soybean (*Glycine Max (L.)* lines developed by hybridization method. Selçuk Üniversitesi Ziraat Fakültesi Dergisi, 22(45): 52-57.

Willmott, C. and K. Matsuura (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30: 79–82.

Worku, M., and T. Astatkie (2015). Effects of row spacing on productivity and nodulation of two soybean varieties under hot sub-moist tropical conditions in south-western Ethiopia. J. Agriculture and Rural Development in the Tropics and Subtropics, 116(2): 99-106.

Yari, V., A. Frnia, A. Maleki, M. Moradi, R. Naseri, M. Ghasemi, and A. Lotfi (2013). Yield and yield components of soybean cultivars as affected by planting date. Bulletin of Environment, Pharmacology and Life Sciences, 2(7): 85-90.

Yetkin, S. G., and H. Arioglu (2009). Determination of yield and important Plant Characteristics of Some Soybean Varieties and Genotypes Grown as a Main Crop in the Çukurova Region. Çukurova University Fen Bilimleri Enstitüsü, 20(1): 29-37.

Yılmaz, H. A., and L. Efe (1998). Possibilities of growing of some soybean (*Glycine max* (*L.) Merill*) cultivars as a double crop under Kahramanmaraş conditions. Tr. J. of Agriculture and Forestry, 22: 135-142.

Yosmaoğlu, M. (2002). Soya fasulyesi raporu T. C. Tarım ve Köyişleri Bakanlığı, Araştırma Planlama ve Koordinasyon Kurulu Başkanlığı, Ankara.

Appendix. Codes of the package "earth" of R software for statistical analysis of MARS algorithm for YD in soybean.

```
install.packages("earth")
d=read.table("C:/soybean.txt", header=T)
  str(d)
  library(earth)
  m3=earth(YD~., data=d, penalty=-1, nprune=100, degree=3, pmethod="backward", nfold=5,
  nk=300, keepxy=T)
  summary(m3, digits=4)
  evimp(m3)
  n<-length(d$YD)
  n ## sample size
  k= length(m3$selected.terms)
  k ## number of terms in the MARS predictive model
  cor.test(d$YD, predict(m3))
  Pearsoncorr=round(cor(d$YD, predict(m3)), digits = 3)
  Pearsoncorr ## Correlation coefficient
  error=d$YD-predict(m3)
  sdratio=round(sd(error)/sd(d$YD), digits=3)
  sdratio
  RMSE=round(sqrt(mean(error^2)), digits=3)
RMSE
  Rsq=round(1-(sum(error^2)/(var(d$YD)*(n-1))), digits = 3)
Rsq
  AdjRsq=round(1-((1- Rsq)*(n-1)/(n-k-1)), digits=3)
AdjRsq
AIC=round(n*log(mean(error^2), base=exp(1))+2*k, digits=0)
AICc
  AICc=round(n*log(mean(error^2), base=exp(1))+(2*k)+(2*k*(k+1)/(n-k-1)), digits=0)
AICc
  plot(d$BW, predict(m3))
```