# Comparison of the Results of Factorial Experiments, Fractional Factorial Experiments, Regression Trees and MARS for Fuel Consumption Data

BETÜL KAN
Anadolu University
Faculty of Science, Dept. of Statistics
26470 Eskişehir
TURKEY
bkan@anadolu.edu.tr


BERNA YAZICI
Anadolu University
Faculty of Science, Dept. of Statistics
26470 Eskişehir
TURKEY
bbaloglu@anadolu.edu.tr

*Abstract:-* In this study effects that are significant on the fuel consumption of F-4 aircrafts are analyzed by using different statistical methods and the results of those different methods are compared. The response variable of every possible combination of factors and interactions are analyzed to provide information about main effect(s) and interaction effect(s). MARS is applied in order to find the relationship and a tree model in order to summarize the relationship of main effects and the interactions on a plot. Also a $2^5$ factorial experiment with 4 replications is performed to determine the significant main effects and interactions. The analysis is repeated for one-half fractional factorial experiment. A comparison of performance of the methods is made in order to investigate their applicability.

*Key-Words:* Regression trees, MARS, factorial experiments, fractional factorial experiments, basis function, GCV (Generalized Cross Validation), recursive partitioning, AID (Automatic Interaction Detection).

## 1 Introduction

Classification and modeling are commonly used statistical methods in application. One of the recent methods for classification in regression analysis is regression trees. Several algorithms are proposed and extensively studied for building regression trees by adaptive recursive partitioning of the data sets. The origin of recursive partitioning regression is the AID (Automatic Interaction Detection) program written by Morgan and Sonquist in 1963 [1]. A regression tree is a piecewise constant or piecewise linear estimate of a regression function, constructed by recursively partitioning the data and sample space [2]. The innovations were made by Breiman et al. in 1984 [3]. Loh studied that regression tree models can provide simpler and more intuitive interpretations of interaction effects as differences between conditional main effects. He used regression trees on small data sets including replicated and unreplicated factorial experiments

[4]. Kurematsu et al. studied the classification of a data set on human speech [5]. Enescu et al. studied the numerical investigation of the polynomial regression models [6]. Johanna [7] worked on the problem of determinants of the auction prices of the works of art using three statistical methods: analysis of variance (Anova), multiple regression models and classification trees. Chaudhuri [8] defined the smooth and unsmooth piecewise polynomial regression trees (SUPPORT) algorithm. The latter explanation of recursive partitioning may be considered as MARS. It is recent statistical method presented in Hastie et al. [9,10] and can be considered as generalization of classification and regression trees. MARS solves the problem of high dimensions. The MARS algorithms choose basis functions for the approximation of response by partitioning the domain region of regressors into subregions [11]. On the other hand, Woods and

Lewis in 2005 studied on all-bias designs for polynomial spline regression models [12]. Grove et al. [13] studied multifactor B-spline mixed models in designed experiments for the engine mapping problem.

# 2 Material and Method

## 2.1  $2^k$ Factorial Experiments

Many experiments involve the study of the effects of two or more factors with two levels. Factorial designs are the most efficient for this type of experiment. By a factorial design we mean that in each complete trial or replication of experiment all possible combinations of the levels of the factors are investigated.

The results for the two-factor factorial design may be extended to the general case where *a* levels of factor A, *b* levels of factor B, *c* levels of factor C, and so on, arranged in a factorial experiment. In general, there will be abc…n total observations if there are n replicates of the complete experiment. Once again, note that at least two replicates must be done to determine a sum of squares due to error if all possible interactions are included in the model.

The levels of a factor may be quantitative, or qualitative with two or more levels. In some experiments, we may find that the difference in response between the levels of one factor is not the same at all levels of the other factors. That means there is an interaction between the factors. A complete replicate of such a design requires $2^k$ observations and is called a $2^k$ factorial design [14].

## 2.2  The One-Half Fraction of the $2^k$ Design

Under the assumption that higher-order interactions are negligible in an experiment, low-order interactions and main effects might give the information by running only a fraction of the complete factorial experiment. The objective of using fractional factorial experiments is to identify the factors that have large effects.

The design is formed by selecting a generator. One-half fraction of the $2^k$ design may be constructed by writing down a basic design for a full $2^{k-1}$ factorial. The *k*th factor by identifying its plus and minus levels with the plus and minus signs of the highest-order interaction ABC…(K-1). Any interaction effect could be used to generate the column for the *k*th factor. Hence, the number of treatment combinations is reduced [15].

## 2.3 Recursive Partitioning

The response variable **y** depend on some unknown way on a vector of p predictor variables, $\mathbf{x} = (x_1, x_2, ..., x_p)$, that is modeled with (1).

$$\mathbf{y} = f(x_1, x_2, ..., x_p) + \varepsilon \qquad (1)$$

Assume V be the input space and there are n samples of $\mathbf{y}, \{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n$. Let $\{R_j\}_{j=1}^S$ be a set of disjoint subregions of V, $V \subset \Re^p$ such that

$$V = \bigcup_{j=1}^S R_j$$

Recursive partitioning estimates the unknown function f(**x**) at **x** with

$$\hat{f}(\mathbf{x}) = \hat{f}_j(\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in R_j \qquad (2)$$

where the function $\hat{f}_j(\mathbf{x})$ estimates the true but unknown function f(**x**) over the $R_j$th subregion of V. In recursive partitioning, $\hat{f}_j(\mathbf{x})$ is taken to be the constant function [1,11].

Firstly, the space V is split into two regions, and the response is modeled by the mean of **y** in each region. The variable and split-point to achieve the best fit is chosen. Then one or both of these regions are split into two more regions, and this process is continued, until a stopping rule is applied.

The corresponding regression model predicts **y** with a constant $c_j$ in region $R_j$. The algorithm needs to automatically decide on the splitting variables and split points, and also what topology (shape) the tree should have.

The partitioning of V is accomplished through recursive splitting of subregions by an algorithm. A subregion $R_j$ is split into two regions at a knot k for a covariate $x_v$ as below:

if $\mathbf{x} \in R_j$ then,

    if $x_v \leq k$, then $\mathbf{x} \in R_l$

    else $\mathbf{x} \in R_r$

end if

Here $x_v$ alone represents the values of the vth covariate.

The subregion $R_j$ is deleted and replaced by $R_l, R_r$. The main problem is to choose the values of k. Indeed those are prespecified knots from a set of the covariate v. Generally, this knot set consists of every distinct data value or some subset of these values.

Initially, we start the space region V with the corresponding constant basis function $B_1(\mathbf{x}) = 1$, taking into account all of the data. Then, a splitting variable j and split point k, and replacement of V by the pair of half-planes (two subregions) is considered.

$$R_1(j,k) = \{X \mid X_j \le k\}$$
$$R_2(j,k) = \{X \mid X_j > k\} \tag{3}$$

The basis function $B_1(\mathbf{x}) = 1$ is split into two step functions are $I(x_v \le k)$ and $I(x_v > k)$. If the one of the subregions, $R_1$ is split at knot $k^*$ for a covariate $v^*$, then we get the interaction term for $x_v x_{v^*}^*$. Those step functions are represented by the two basis functions $I(x_v \le k)I(x_{v^*} \le k^*)$ and $I(x_v \le k)I(x_{v^*} > k^*)$.

Suppose we have a partition into M regions $R_1, R_2, ..., R_M$, and we model the response as a constant $c_j$ in each region:

$$f(\mathbf{x}) = \sum_{j=1}^{M} c_j I(\mathbf{x} \in R_j) \tag{4}$$

where $\hat{c}_j = \text{ave}(y_i \mid x_i \in R_i)$.

A key advantage of the recursive binary tree is its interpretability. The feature space partition is fully described by a single tree.


## 2.4 MARS

MARS was first proposed by Friedman as a nonparametric modeling technique. Its main purpose is to predict the values of a continuous dependent variable, $\mathbf{y}_{n \times 1}$ from a set of independent explanatory variables, $\mathbf{x} = (x_{1i}, ..., x_{pi})_{i=1}^{n}$. The MARS model is represented as:

$$y = f(\mathbf{x}) + \varepsilon \tag{5}$$

where $\varepsilon$ is an error vector whose expected value is assumed to be zero.

MARS includes a two step procedures to construct the final model. First it partitions the space of explanatory variables in several possible subregions and fits truncated spline functions in each subregion by a forward stepwise procedure. This implements a very complex and overfitted model therefore a second procedure needs to be introduced. Then the redundant basis functions are removed one by a backward stepwise procedure. A truncated spline function consists of a left-sided and, a right-sided segment separated by a so-called knot location as given below:

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases} (t-x)^q & , x < t \\ 0 & \text{otherwise} \end{cases}$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases} (x-t)^q & , x < t \\ 0 & \text{otherwise} \end{cases}$$
$$\tag{6}$$

where $b_q^-(x-t)$ and $b_q^+(x-t)$ are the spline functions describing the regions right and left of the knot location t, respectively, and q the power to which the spline is raised. The subscript "+" indicates that the result of the function is 0 when the argument is not satisfied. For each of the descriptive variables in the data set MARS selects the pair of spline functions and the knot location that best describes the response variable. In the following step all the spline functions are combined in a complex non-linear model, describing the response as a function of the descriptive variables. The model has the form:

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(\mathbf{x}) \tag{7}$$

where $\hat{y}$ is the predicted value for the response variable, $a_0$ the coefficient of the constant term, M the number of spline basis functions, and $B_m$ the mth spline basis function which may be a single spline function or a product (interaction) of two or more spline basis functions and $a_m$ its coefficient, respectively.

MARS determines which basis functions are included to the final model by a generalized cross-validation criterion (GCV). The GCV is the mean squared residual error divided by a penalty dependent on the model complexity. The GCV criterion is defined as in Eq.(8):

$$GCV(M) = \frac{\sum\limits_{i=1}^{N} (y_i - \hat{f}_M(x_i))^2}{(1 - C(M)/N)^2} \qquad (8)$$

Here, where $C(M)$ is a complexity penalty that increases with the number of basis functions in the model and which is defined as:

$$C(M) = (M+1) + dM \qquad (9)$$

where M is the number of basis functions in Eq. (7), and the parameter d is a penalty for each basis function included into the model [10].

## 3. Application

In the application part of the study, the factors affecting the fuel consumption of F-4 aircrafts using factorial experiments, fractional factorial experiments, regression trees and MARS are obtained. Note that, for a $2^{k-1}$ factorial design, the half of the treatment combinations is used for analysis of variance. Therefore, the half of the treatment combinations is analyzed to reach one-half fractional factorial design. The same half is used to obtain a regression tree and MARS model. Each method is used on 128observations and 64 observations, respectively.

The capital letters A, B, C, D, and E denote the names of the variables, Compressor Input Temperature (CIT), Power Lever Angle (PLA), Revolutions of the Engine(RPM), Input Flow, Air pressure as well as their main effects [15].

**Table 1**. The factors related with the performance of the main fuel control unit

| Factors | | 1 (-) | 2 (+) |
|---|---|---|---|
| Compressor Input Temperature(CIT) | A | 0.0605 | 0.296 |
| Power Lever Angle (PLA) | B | 40 | 100 |
| Revolutions of the Engine(RPM) | C | 1250 | 3160 |
| Input Flow | D | 8000 | 18000 |
| Air pressure | E | 100 | 150 |

Moreover, AB, BC,…, ABC,…, ABCD,…, ABCDE are named to represent the interaction effects. The levels of each factor are indicated in two way by signs as (-) and (+). There are five independent factors each has two levels affecting the main fuel control unit's performance as shown in Table 1.

Below there are two analyses of variance for the fuel consumption of F-4 aircrafts derived from a $2^5$ factorial design and $2^{5-1}$ fractional factorial design.

The significant main effects and interactions are shown in Table 2. According to the results, factor A, D, and E are statistically significant. Then interaction terms AC, AD, CD, and DE are found significant. Third order interactions as ABC, ACD, ACE and ADE are found significant by the analysis of variance. Finally, fourth order interaction term ACDE is found significant at 95% significance level.

The experiment is replicated for four times. Therefore the design matrix is adopted according to the replications. The observations of four experiments are stacked into one column and the corresponding design matrix is constructed for MARS. Hence a model including all possible variables and their interactions are fitted to the data.

**Table 2**. Anova for $2^5$ factorial experiments

| Source | df | Mean Square | F-test | Sig. |
|---|---|---|---|---|
| A | 1 | 7082907.031 | 2201.641 | 0.00 |
| B | 1 | 11514.031 | 3.579 | 0.62 |
| C | 1 | 10011.125 | 3.112 | 0.81 |
| D | 1 | 906531.125 | 281.785 | 0.00 |
| E | 1 | 186853446.125 | 58081.255 | 0.00 |
| AB | 1 | 19306.125 | 6001 | 0.16 |
| AC | 1 | 616882.781 | 191751 | 0.00 |
| BC | 1 | 12129.031 | 3770 | 0.55 |
| ABC | 1 | 17860.500 | 5.552 | 0.20 |
| AD | 1 | 680069.531 | 211.392 | 0.00 |
| BD | 1 | 3341.531 | 1.039 | 0.31 |
| ABD | 1 | 5304.500 | 1.649 | 0.20 |
| CD | 1 | 42486.125 | 13.206 | 0.00 |
| ACD | 1 | 70218.781 | 21.827 | 0.00 |
| BCD | 1 | 1696.531 | 0.527 | 0.47 |
| ABCD | 1 | 3528.000 | 1.097 | 0.30 |
| AE | 1 | 10694.531 | 3.324 | 0.71 |
| BE | 1 | 7290.281 | 2.266 | 0.14 |
| ABE | 1 | 4900.500 | 1.523 | 0.22 |
| CE | 1 | 2.000 | 0.001 | 0.98 |
| ACE | 1 | 13489.031 | 4.193 | 0.43 |
| BCE | 1 | 2945.281 | 0.916 | 0.34 |
| ABCE | 1 | 3828.125 | 1.190 | 0.28 |
| DE | 1 | 378015.125 | 117.502 | 0.00 |
| ADE | 1 | 429896.281 | 133.628 | 0.00 |
| BDE | 1 | 3894.031 | 1.210 | 0.27 |
| ABDE | 1 | 6160.500 | 1.915 | 0.17 |
| CDE | 1 | 5253.125 | 1.633 | 0.20 |
| ACDE | 1 | 20553.781 | 6.389 | 0.13 |
| BCDE | 1 | 1140.031 | 0.354 | 0.55 |
| ABCDE | 1 | 4140.500 | 1.287 | 0.26 |
| Error | 96 | 3217.104 | | |
| Total | 128 | | | |

*Total sum of squares=5188542328.000
*df=degree of freedom

In the next part, a one-half fraction of the $2^5$ design is examined. The $2^{5-1}$ design with four replications is considered to explore the relationship between the effects.

The $2^{5-1}$ design is formed by selecting only the treatment combinations that have a plus in the ABDCE column. In this study, the ABCDE interaction factor is taken as generator. The fractional factorial experiment is replicated for four times. The results of Anova of $2^{5-1}$ design are shown in Table 3. The main effects and third order interaction factors such as A, D, E, BDE, BCE, ADE, ACD, ABE, and ABC are found significant at 95% significance level.

**Table 3**. Anova for $2^{5-1}$ fractional factorial experiment

| Source | df | Mean Square | F-test | Sig. |
|--------|----|-------------|--------|------|
| A | 1 | 3452164.000 | 572.416 | 0.000 |
| B | 1 | 650.250 | 0.108 | 0.744 |
| C | 1 | 15939.062 | 2.643 | 0.111 |
| D | 1 | 514089.000 | 85.243 | 0.000 |
| E | 1 | 94240410.063 | 15626.352 | 0.000 |
| CDE | 1 | 2209.000 | 0.366 | 0.548 |
| BDE | 1 | 261376.563 | 43.340 | 0.000 |
| BCE | 1 | 386262.250 | 64.048 | 0.000 |
| BCD | 1 | 10455.063 | 1.734 | 0.194 |
| ADE | 1 | 148803.063 | 24.674 | 0.000 |
| ACE | 1 | 15129.000 | 2.509 | 0.120 |
| ACD | 1 | 61380.063 | 10.178 | 0.003 |
| ABE | 1 | 38122.563 | 6.321 | 0.015 |
| ABD | 1 | 2550.250 | 0.423 | 0.519 |
| ABC | 1 | 115770.063 | 19.196 | 0.000 |
| Error | 48 | 6030.865 | | |
| Total | 64 | | | |

*Total sum of squares= 2590512982.000
 df=degree of freedom

MARS results for 128 observations are the following below:

20 of 20 terms including constant term, and 19 of 31 predictors are selected by MARS final model. MARS significantly reduced the number of predictors. The importance of variables is sorted in D, BCD, ABCDE, ACD, C, E, BDE, CDE, B, ABD,..., BC, and BCDE (see the details in Table 4)

The model has a GCV value of 0.12 and $R^2$ is obtained as 0.99. Then, 19 basis functions such as BF1=pmax(0, A-(-1)), BF2=pmax(0 ,B-(-1)) , BF3=pmax(0 , C-(-1)), BF4=pmax(0, D-(-1)), …, , BF18= pmax(0,BCDE  -(-1)), BF19=pmax(0, ABCDE-(-1))  are obtained in MARS analysis and low levels of factors for each main effect and each interaction effect are determined as significant factors as the following:

Basis functions of F-4 aircrafts consumption data:

response   = 5.305297
        +0.113*pmax(0,A-(-1))
        - 0.133*pmax(0,B-(-1))
        +0.459*pmax(0,C-(-1))
        +0.520*pmax(0,D-(-1))
        +0.405*pmax(0,E-(-1))
        - 0.051*pmax(0,AB -(-1))
        +0.048*pmax(0,BC-(-1))
        +0.126*pmax(0,ABC-(-1))
        +0.130*pmax(0,ABD -(-1))
        - 0.460*pmax(0,ACD -(-1))
        +0.508*pmax(0,BCD -(-1))
        +0.080*pmax(0,ABE -(-1))
        +0.070*pmax(0,ACE -(-1))
        +0.080*pmax(0,ADE -(-1))
        - 0.242*pmax(0,BDE -(-1))
        - 0.224*pmax(0,CDE -(-1))
        - 0.064*pmax(0,ACDE -(-1))
        +0.044*pmax(0,BCDE -(-1))
        - 0.470*pmax(0,ABCDE-(-1))

**Table 4.** GCV and RSS scores of each variable in each subset

| Variable | Order | Subsets | GCV | RSS |
|----------|-------|---------|--------|---------|
| D | 4 | 19 | 100.000 | 100.000 |
| BCD | 15 | 18 | 85.000 | 82.387 |
| ABCDE | 31 | 17 | 69.857 | 65.611 |
| ACD | 14 | 16 | 56.314 | 51.235 |
| C | 3 | 15 | 42.491 | 37.453 |
| E | 5 | 14 | 27.761 | 23.738 |
| BDE | 26 | 13 | 15.673 | 13.047 |
| CDE | 28 | 12 | 11.436 | 9.243 |
| B | 2 | 11 | 7.580 | 5.971 |
| ABD | 12 | 10 | 6.318 | 4.819 |
| ABC | 9 | 9 | 5.035 | 3.723 |
| A | 1 | 8 | 3.730 | 2.683 |
| ABE | 19 | 7 | 2.650 | 1.860 |
| ADE | 25 | 6 | 2.127 | 1.445 |
| ACE | 21 | 5 | 1.560 | 1.031 |
| ACDE | 29 | 4 | 1.103 | 0.710 |
| AB | 6 | 3 | 0.707 | 0.445 |
| BC | 8 | 2 | 0.453 | 0.275 |
| BCDE | 30 | 1 | 0.211 | 0.125 |

GCV:Generalized Cross Validation
RSS: Residual Sum of Squares

The results are evaluated for both GCV and residual sum of squares (RSS) for each subset in Table 4.

The rows of the table are given in the ascending order respect to their number of subsets they include.

In Table 4, the rows are sorted on the number of subsets criterion that means the number of subsets that include the variable [16]. Variables that are included in more subsets are considered more important. However, there are two more criteria showing the importance of each factor, namely GCV and RSS. A variable's importance is a measure of the effect that observed changes to the variable have on the observed response. Therefore, factor D is the most important variable and then third order interaction term BCD comes, and so on.

**Figure 1.** Plots of each factor from MARS for $2^5$ design
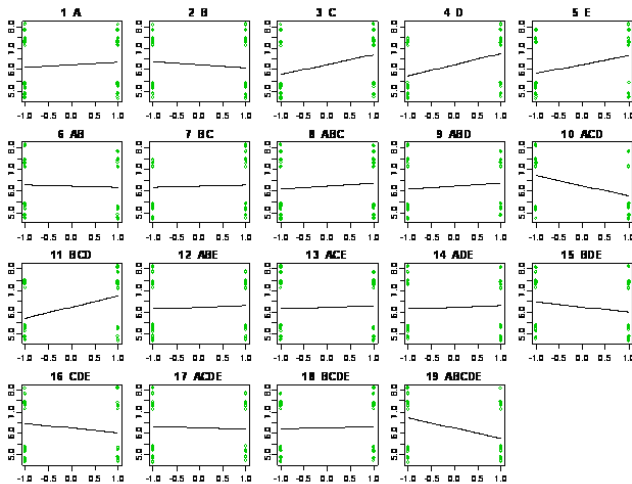


In Figure 1 it is represented that the factors which were selected by the final model of MARS are plotted.

MARS results for the $2^{5-1}$ design are given below:

12 of 12 terms, and 11 of 31 predictors are selected by the MARS final model. According to the results, the factors which taken into the model are AD, E, D, BCD, A, ABC, BD, CD, C, BC, and so on(see Table 5 for details).

The model has a GCV value of 0.13 and $R^2$ is obtained as 0.99. Then, 11 basis functions for $2^{5-1}$ design are selected such as $BF1_f=pmax(0,A-(-1))$, $BF2_f=pmax(0,C-(-1))$, $BF3_f=pmax(0,D-(-1))$, $BF4=pmax(0, D-(-1))$ ,…, $BF10=pmax(0, CD-(-1))$, $BF11_f=pmax(0,BCD-(-1))$ in MARS analysis and low levels of factors for each main and interaction effects are determined as significant factor.

Basis functions of F-4 consumption data obtained from $2^{5-1}$ design:

$response_f$ = 5.3586563
-0.149 *pmax(0,A-(-1))
+0.067 *pmax(0,C-(-1))
+0.609 *pmax(0,D-(-1))
+0.647 *pmax(0,E-(-1))
-0.041 *pmax(0,AB-(-1))
-0.049 *pmax(0,BC-(-1))
+0.125 *pmax(0,ABC-(-1))
-0.655 *pmax(0,AD-(-1))
-0.114 *pmax(0,BD-(-1))
-0.077 *pmax(0,CD-(-1))
+0.517 *pmax(0,BCD-(-1))

**Table 5.** GCV and RSS score of each variable in each subset for $2^{5-1}$ design

| Variable | Order | Subsets | GCV | RSS |
|---|---|---|---|---|
| AD | 10 | 11 | 100.000 | 100.000 |
| E | 5 | 10 | 77.042 | 72.313 |
| D | 4 | 9 | 51.501 | 45.337 |
| BCD | 15 | 8 | 25.900 | 21.438 |
| A | 1 | 7 | 5.168 | 4.213 |
| ABC | 9 | 6 | 3.590 | 2.775 |
| BD | 11 | 5 | 2.399 | 1.766 |
| CD | 13 | 4 | 1.291 | 0.935 |
| C | 3 | 3 | 0.795 | 0.556 |
| BC | 8 | 2 | 0.368 | 0.262 |
| AB | 6 | 1 | 0.157 | 0.109 |

In Table 5, there are the factors within the importance of ascending order obtained by MARS analysis. Consequently, AD interaction term is observed more than the other factors in $2^{5-1}$ design.

**Figure 2.** Plots of each factor from MARS for $2^{5-1}$ design
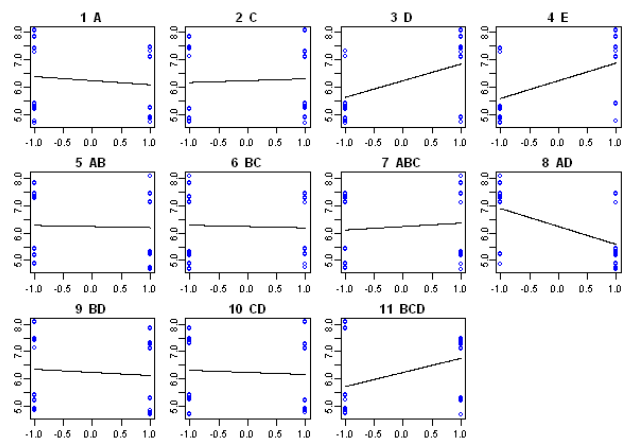


In Figure 2, the factors that are selected in $2^{5-1}$ design data by MARS are plotted.

Tree models are useful for finding the interactions in design of experiments. If we split on one variable and then split on another variable within the partitions of the first variable, we are finding interaction between these two variables.

It is that the first split (node 2 and 3) is on factor D. In detail 64 observations of factor D have low level (-1) with a mean response value of 5.72. On the contrary, factor D has 64 observations with high level (+1) with a mean response value of 6.76.

RSS is observed as 197.54 for the root. After the first split it is reduced from 197.54 to 76.01 and 85.97. RSS at the first split is calculated as 161.98. Factor D gives a large reduction in RSS.

In the same way, we conclude that the factors C and BC are on the second split (node 4, 5, 6, and 7, respectively).

```
 1) root 128 197.5383000 6.244375
 2) D< 0 64  76.9067500 5.723969
   4) C< 0 32  19.9982400 5.251594
     8) ABE>=0 16   0.5819624 4.882812
     9) ABE< 0 16  15.0642900 5.620375 *
   5) C>=0 32  42.6276700 6.196344
    10) E< 0 16  18.9144600 5.565812 *
    11) E>=0 16  10.9909700 6.826875 *
 3) D>=0 64  85.9662300 6.764781
   6) BC< 0 32  35.4025700 6.208719
    12) B>=0 16  12.8299000 5.616375 *
    13) B< 0 16  11.3448000 6.801062 *
   7) BC>=0 32  30.7745100 7.320844
    14) BE>=0 16  22.7795400 6.850062 *
    15) BE< 0 16   0.9026458 7.791625 *
```

Although the relevant information can be viewed from the text-based output, a graphical display is nicer and easier to interpret as in Figure 2 from MARS for $2^{5-1}$ design.

In Figure 3, the depth of the branches is proportional to the reduction in error due to the split. Figure 3 maps the regression tree of relationship among factors [16].

The display always starts at the root (by taking account all of the treatment combinations), and reports the splits in the order that they occurred. Therefore, after examining all possible values of all the variables, recursive partitioning finds that the factor D does the best job to split the observations into the different nodes.

**Figure 3.** Tree model of F-4 Aircrafts Fuel Consumption for $2^5$ design
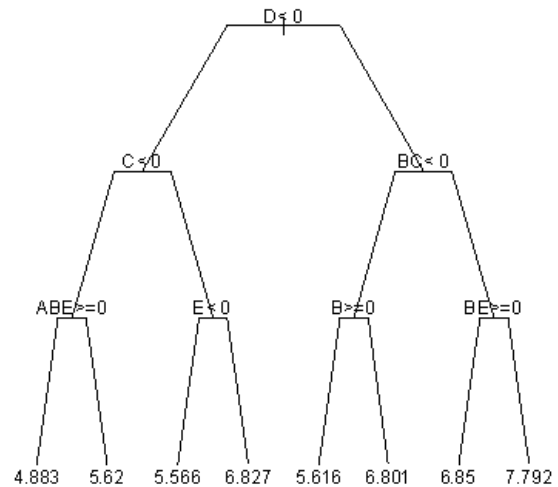


As seen, the first split is on AD interaction terms different from the first tree model obtained by a $2^5$ factorial design. 32 observations of the interaction AD term have high level (+1) with a mean response value of 5.58. On the other hand, the rest of the observations for AD interaction term has low level (-1) with a mean response of 6.89.

Then, the total residual sum of squares is observed as 99.555 from the root for $2^{5-1}$ design. After the first split it is reduced from 99.555 to 32.81 and 39.28. The total RSS at the first split is calculated as 72.19. The interaction term AD gives a large reduction in the RSS.

Another tree model gives the factors for $2^{5-1}$ design of F-4 aircraft fuel consumption, as below:

```
 1) root 64 99.554790 6.238688
 2) AD>=0 32 32.813870 5.583625
   4) ABC< 0 16  1.217838 4.941875 *
   5) ABC>=0 16 18.417050 6.225375 *
 3) AD< 0 32 39.278080 6.893750
   6) A>=0 16 19.255950 6.135875 *
   7) A< 0 16  1.642150 7.651625 *
```

In Figure 4, the depth of the branches is related to the reduction in error due to the split. Figure denotes the regression tree of relationships for $2^{5-1}$ design.

**Figure 4.** Tree model of F-4 Aircrafts Fuel Consumption $2^{5-1}$ design
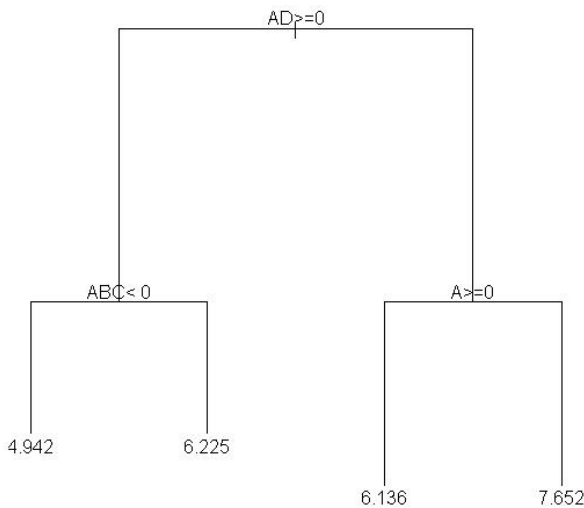


Variables actually used in tree construction are A, ABC, and AD.

The pruning part of regression tree analysis based on a complexity parameter is shown in Table 6.

**Table 6.** Pruning of regression tree model for F-4 Aircrafts Fuel Consumption

| | Control parameter | Number of split | Relative error | Cross validation error | Standard error |
|---|---|---|---|---|---|
| 1 | 0.175 | 0 | 1.000 | 1.024 | 0.044 |
| 2 | 0.100 | 1 | 0.825 | 1.089 | 0.081 |
| 3 | 0.072 | 2 | 0.724 | 1.189 | 0.107 |
| 4 | 0.064 | 3 | 0.652 | 1.249 | 0.120 |
| 5 | 0.057 | 4 | 0.588 | 1.218 | 0.124 |
| 6 | 0.036 | 5 | 0.531 | 1.201 | 0.128 |
| 7 | 0.022 | 6 | 0.495 | 1.125 | 0.115 |
| 8 | 0.010 | 7 | 0.473 | 1.020 | 0.099 |

Here, cross validation error is the smallest cross validation error and standard error is the corresponding standard error. The relative error is $1-R^2$. It is observed from Table 6 that the best tree has eight terminal nodes based on cross-validation. This table is obtained from the smallest tree to the largest one. For any value between 0.175 and 0.100, the best tree model has one split. For any control parameter (as known the ratio of number of terminal nodes to the RSS of the root) between 0.100 and 0.072 the best model is with 2 splits and so on. The splitting criterion is the difference between the sum of squares for the node $(SS_T = \sum (y_i - \overline{y})^2$ ) and the total of the sums of the

squares for the right and left ($SS_R$, $SS_L$), respectively. It is identical to the analysis of variance to maximize the sum of squares between the groups [16].

**Table 7.** Pruning of regression tree model for F-4 Aircrafts Fuel Consumption for $2^{5-1}$ design

| | Control parameter | Number of split | Relative error | Cross validation error | Standard error |
|---|---|---|---|---|---|
| 1 | 0.276 | 0 | 1.000 | 1.032 | 0.063 |
| 2 | 0.185 | 1 | 0.724 | 1.370 | 0.144 |
| 3 | 0.132 | 2 | 0.540 | 1.272 | 0.137 |
| 4 | 0.010 | 3 | 0.407 | 0.737 | 0.065 |

In Table 7, the values between 0.276 and 0.185, the best tree model has one split for $2^{5-2}$ design. For any control parameter between 0.185 and 0.132 the best model is with 2 splits and so on.

Five different factors and their all possible interactions examined using factorial experiments, one-half fractional factorial experiments, regression trees and MARS for both $2^5$ and $2^{5-1}$ designs.

As the results of Anova for $2^5$ design, the factors A, D and E are found significant whereas regression tree finds the factors D, C, E, and B important respectively. MARS gives the basis functions of 5 main effects.

Among the second order interaction terms AC, AD, CD and DE are found significant as a result of Anova, whereas BC and BE are found important as a result of Regression trees. MARS gives the basis functions for AB and AC. Among the third order interactions ABC, ACD, ACE and ADE are found significant as a result of Anova, and only ABE is found important as a result of regression trees. MARS gives the basis functions of ABC, ABD, ACD, BCD, ABE, ACE, ADE, BDE, and CDE among the third order interactions. ACDE is the only significant fourth order interaction according to Anova results. MARS also gives the basis function for the same interaction term and BCDE and also ABCDE.

According to the results of $2^{5-1}$ fractional factorial, the factors A, D, and E are found significant whereas regression tree explores the factors AD, ABC, and A graphically. MARS gives only 4 basis functions of 5 main effects for $2^{5-1}$ design.

The interaction terms BDE, BCE, ADE, ACD, and ABC are found statistically important. Nevertheless, regression tree explores the ABC interaction term. MARS selects both the second order interactions such as AB, BC, AD, BD, CD and third order interactions; ABC and BCD as basis functions for $2^{5-1}$ design.

# 5 Conclusion

Among four methods, $2^5$ factorial and $2^{5-1}$ fractional factorial experiments are used in order to determine the significant factors and their interactions. Regression trees is the better way of denoting the relationships of factors and their interactions on a plot and easily interpreting of those relationships. MARS is used to examine the functional relationship by a model with splines. The results of MARS and $2^{5-1}$ fractional factorial experiment are similar. In this study, all four methods are performed in order to see the parallelism of the results by combining the results to evaluate the findings from different point of views.

Results show that according to all four different methods the main factors D and E, in other words input flow and air pressure are found statistically significant. In case of taking the low levels of both factors the fuel consumption of F-4 aircrafts can be reduced. In other words, input flow should be 8000 instead of 18000 and air pressure should be 100 instead of 150 in order to reduce the fuel consumption. Also, it has been observed that after one-half of the treatment combinations is removed, the main factor A, compressor input temperature (CIT), is occurred significant as it should. On the other hand, most of the interaction terms which those two terms with the low levels interact with the others are also found significant as the results of all four methods.

*References:*

[1] Morgan J.N. and Sonquist, J.A., *Problems in the Analysis of Survey Data and a Proposal*. J. Amer. Statist. Assoc. , 58, 1963, pp. 415–434.

[2] Loh, W.Y.*, Regression Trees with Unbiased Variable Selection and Interaction Detection.* Statistica Sinica, Vol.12*,* 2002, pp.361-386.

[3] Breiman, L., Friedman, J., Olshen, R., and Stone, C., *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.

[4] Loh, W.Y., *Regression Tree Models for Designed Experiments*. In Second E. L. Lehmann Symposium, J. Rojo, Ed., Institute of Mathematical Statistics Lecture Notes-Monograph Series, Vol. 49, 2006, pp. 210-228.

[5] Kurematsu M., Hakura J., Fujita H., *A extraction of emotion in human speech using synthesize and each classifier for each emotion,* Proceedings of the 7th WSEAS International Conference on Applied Computer Science - Computer Science Challenges, ITALY, 2007, pp.385-389.

[6] Enescu D., Coand H.G., Virjoghe E.O., Caciula I., *Numerical Investigation by means of polynomial regression method for determining the temperature fields in a medium with phase transition*, Proceedings of the 8th WSEAS International Conference on Systems Theory And Scientific Computation (Istac'08) - New Aspects of Systems Theory And Scientific Computation, GREECE, 2008, pp.88-93.

[7] Johanna B.B., *Statistical Methods Used for Identification of Art Prices Determinants*, Proceedings of the 10th WSEAS International Conference on Mathematics and Computers in Business and Economics Praque, CZECH REPUBLIC, 2009, pp.36-41.

[8] Chaudhuri, P., Huang, M.C., Loh, W.Y., and Yao, R. *Piecewise-Polynomial Regression Trees*. Statistica Sinica, Vol.4, 1994, pp.143-167.

[9] Lewis, P.A.W and Stevens J.G., *Nonlinear Modeling of Time Series Using MARS,* Journal of the American Statistical Association, Vol.86, No.416, 1991, pp.864-877.

[10] Hastie, T., Tibshirani, R., Friedman J., *The Elements of Statistical Learning-Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2001.

[11] Friedman, H.J, *Multivariate Adaptive Regression Splines*, The Annals of Statistics, Vol.19, No.1, 1991, pp.1-67.

[12] Woods, David C. and Lewis, Susan M., *All-bias designs for polynomial spline regression models,* Australian and New Zealand Journal of Statistics, Vol.48, No.1, 2006, pp.49-58.

[13]  Grove, D.M., Woods, D.C. and Lewis, S.M. *Multifactor B-spline mixed models in designed* experiments for the engine mapping problem,J.Qual.Technol.,Vol.36, No.4, 2004, pp.380–391.

[14]  Montgomery, D.C., *Design and Analysis of Experiments,* 5th.Ed., New York: J. Wiley, 2001.

[15]  Yazıcı, B. and Kasap, Ş., *Determining The Factors That Affect The Fuel Consumption In F-4 Aircrafts By $2^k$ Experiments And Taguchi Method*, JSM 2009, Washington, DC, pp. 3105-3116.

[16]  Milborrow S, *earth: Multivariate Adaptive Regression Spline Models,* R package version: 2.0-2, 2007, URL http://CRAN.R-project.org/package=earth

[17]  Therneau T.M., Atkinson B., *rpart: Recursive Partitioning and Regression Trees,* R Package Version: 3.1-46, 2009, URL http://CRAN.R-project.org/package=rpart