MARS SOLUTION FOR MULTIPLE QUANTITATIVE RESPONSES

Assoc. Prof. Ecevit EYDURAN

Biometry Genetics Unit

Department of Animal Science

Iğdır University, Turkey

THE AIM OF THE CURRENT STUDY

- The aim of this research work is to simultaneously build a regression model of four solar irradiation parameters by using the multivariate adaptive regression splines (MARS) technique at Küllük local region of the Turkey.

THE MAIN AIM OF THE REGRESSION MODELS

- To predict a difficult measured trait from some traits that can measure easily!
- To reduce the differences between the observed values and the predicted values in response variable, that is,
- To reduce error variance
- To built the prediction model from predictors and response variable

REGRESSION MODELS

- A simple linear regression model (SLRM) is a model that has a response variable and a predictor!
- Multiple linear regression model (MLRM) is a model that has a response variable and multiple predictors
- Multivariate Multiple Linear Regression (MMLR) is a model that has multiple response and predictors,

HYBRID REGRESSION MODELS

- Use of Factor Analysis Scores in Multiple Linear Regression

- Use of PCA Scores in Multiple Linear Regression
- Where, Factor analysis and PCA are multivariate statistical techniques which are used to remove multicollinearity problems.

DATA MINING ALGORITHMS

- CART (Classification and Regression Tree)
- CHAID (Chi-Square Automatic Interaction Detection)
- Exhaustive CHAID
- Random Forest
- SVM (Support Vector Machines)
- ANNs (Artificial Neural Networks)
- and MARS (Multivariate adaptive regression splines)

Among these, MARS is a powerful statistical tool!

MULTIPLE RESPONSES FOR MARS ALGORITHM

- In the estimation of global irradiation parameters as response variables:
- Ed: Average daily electricity production from the given system (kWh),
- Em: Average monthly electricity production from the given system (kWh),
- Hd: Average daily sum of global irradiation per square meter received by the modules of the given system (kWh/m^2),
- Hm: Average sum of global irradiation per square meter received by the modules of the given system (kWh/m^2)

PREDICTORS FOR MARS ALGORITHM

- For predicting each of the irradiation parameters, some predictors
- ESTLOSTEMP (estimated losses due to temperature and low irradiance),
- ESTLOSANGREF (estimated loss due to angular reflectance effect), and
- COMPVLOSS (Combined Photo Voltaic system losses)

- Global irradiation parameters were predicted through multivariate adaptive regression splines (MARS) algorithm for multiple response variables (Ed, Em, Hd and Hm) with the help of the package “earth” of R software program.
- In the package “earth”, the command “cbind(Ed, Em, Hd, Hm)” was specialized for simultaneously analyzing multiple response variables.

All the statistical computations were conducted through the R package program (R Core Team 2017, R Foundation for Statistical Computing, Vienna, Austria)

R STUDIO FREE SOFTWARE

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

$$\hat{y} = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} h_{km}(X_{v(k,m)})$$

$\hat{y}$ is the predicted value of the response variable

β_0 is a constant

K_m is described as the parameter that limits the order of interaction in the MARS

$h_{km}(X_{v(k,m)})$ is the basis function

SOME SPECIFICATIONS FOR MARS ALGORITHM

- The maximum number of basis functions in the MARS modeling was defined at the first step as 100 and the MARS model was constructed by using interaction order of 2.
- After building the most complex MARS model, the basis functions which did not make contribution much to the level of the model predictive accuracy were removed from the process of the so-called pruning based on the following generalized cross-validation error (GCV)

GENERALIZED CROSS VALIDATION

$$GCV(\lambda) = \frac{\sum_{i=1}^{n} (y_i - y_{ip})^2}{\left[1 - \frac{M(\lambda)}{n}\right]^2}$$

n is the number of training cases

y_{ip} is the predicted value of a response variable

$M(\lambda)$ presents a penalty function for the complexity of the specified model having λ terms

GOODNESS OF FIT CRITERIA

- The goodness of fit criteria for measuring their predictive performance of the MARS algorithm evaluated statistically here are presented as follows
- Goodness of fit criteria allow analysts to obtain information on degree of predictive accuracy of MARS predictive modeling

GOODNESS OF FIT CRITERIA

- Pearson correlation coefficient between the observed values and the predicted values for each response variable!
- The highest correlation coefficient is unity which means the greatest predictive accuracy for MARS algorithm.

GOODNESS OF FIT CRITERIA

- Coefficient of Determination
- Its highest value is unity!

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

GOODNESS OF FIT CRITERIA

- ✓ Adjusted Coefficient of Determination
- ✓ Its highest value is unity

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

GOODNESS OF FIT CRITERIA

- Standard Deviation Ratio
- Lower than 0.10 for the best fit
- Lower than 0.40 for a good fit
- The best fit was obtained here!

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

TO ENSURE HIGH PREDICTIVE ACCURACY!

- In the MARS model as an alternative to Response Surface Method, some data mining algorithms (CART, CHAID and ANNs) and Multiple linear regression etc.
- The smallest GCV,
- The smallest SD_{RATIO}
- The highest coefficient of determination (R^2), and
- The highest Pearson correlation coefficient (r) between actual and predicted values was considered as the best one!

SOME SPECIFICATIONS FOR A GOOD MARS SOLUTION

- We performed -1 penalty and V three-fold cross-validation definitions in the command "cbind (Ed, Em, Hd, Hm) of the package `earth` of R studio free software to improve predictive accuracy of the MARS algorithm.
- MARS prediction equation was achieved at the smallest estimates of GCV which is calculated as the ratio of RSS to n (sample size) for penalty = -1.

	GCV	RSS	GRSq	RSq	CVRSq
ED	0.00005	0.0005	0.999	0.999	0.994
EM	0.05564	0.5564	0.998	0.998	0.993
HD	0.00006	0.0006	0.999	0.999	0.993
HM	0.08064	0.8064	0.998	0.998	0.990

A SIMULTANEOUSLY MARS SOLUTION WITH NO INTERACTION EFFECT!

	ED	EM	HD	HM
(Intercept)	3.2453846	98.756410	4.4153846	134.256410
h(40-AZIMUTH)	0.0027000	0.070000	0.0027000	0.070000
h(AZIMUTH-40)	-0.0067308	-0.203846	-0.0084808	-0.228846
h(3.1-ESLOANGREF)	0.2461538	7.435897	0.3461538	12.435897
h(ESLOANGREF-3.1)	-0.1807692	-5.846154	-0.2557692	-11.346154

A SIMULTANEOUSLY MARS SOLUTION WITH NO INTERACTION EFFECT!

	GCV	RSS	GRSq	RSq	CVRSq
ED	0.00003	0.0003	0.999	0.999	0.966
EM	0.03020	0.3020	0.999	0.999	0.948
HD	0.00002	0.0002	1.000	1.000	0.946
HM	0.03196	0.3196	0.999	0.999	0.947

A SIMULTANEOUSLY MARS SOLUTION WITH INTERACTION EFFECT!

	ED	EM	HD	HM
(Intercept)	3.241	98.61	4.410	134.06
h(40-AZIMUTH)	0.003	0.07	0.003	0.07
h(AZIMUTH-40)	0.097	3.47	0.125	4.85
h(3.1-ESLOANGREF)	0.286	8.86	0.398	14.41
h(ESLOANGREF-3.1)	-0.022	-0.22	-0.051	-3.56
h(AZIMUTH-40) * COMPVSLOSS	-0.004	-0.14	-0.005	-0.19

PREDICTION EQUATIONS FOR MULTIPLE RESPONSE MARS MODELING

PREDICTION EQUATIONS FOR MULTIPLE RESPONSE MARS MODELING

- Ed= 3.241 + 0.003*max(0,40-AZIMUTH)+ 0.097*max(0, AZIMUTH-40) + 0.286*max(0, 3.1-ESLOANGREF)-0.022*max(0, ESLOANGREF-3.1)-0.004*max(0, AZIMUTH-40)*COMPVSLOSS
- Em= 98.61 + 0.07*max(0,40-AZIMUTH)+ 3.47*max(0, AZIMUTH-40) + 8.86*max(0, 3.1-ESLOANGREF)-0.22*max(0, ESLOANGREF-3.1)-0.14*max(0, AZIMUTH-40)*COMPVSLOSS
- Hd= 4.410 + 0.003*max(0,40-AZIMUTH)+ 0.125*max(0, AZIMUTH-40) + 0.398*max(0, 3.1-ESLOANGREF)-0.051*max(0, ESLOANGREF-3.1)-0.005*max(0, AZIMUTH-40)*COMPVSLOSS
- Hm= 134.06 + 0.07*max(0,40-AZIMUTH)+ 4.85*max(0, AZIMUTH-40) + 14.41*max(0, 3.1-ESLOANGREF)-3.56*max(0, ESLOANGREF-3.1)-0.19*max(0, AZIMUTH-40)*COMPVSLOSS

As ESLOANGREF is smaller/greater than 3.1 at 40^0 azimuth angle, the parameters increase/decrease. In this case, the effect of COMPVLOSS on solar parameters was masked at 40 angle AZIMUTH!

The influence of COMPVSLOSS on the studied parameters was masked at 40^0 or narrower azimuth angle,

	GCV	RSS	GRSq	RSq	CVRSq
ED	0.00003	0.0003	0.999	0.999	0.966
EM	0.03020	0.3020	0.999	0.999	0.948
HD	0.00002	0.0002	1.000	1.000	0.946
HM	0.03196	0.3196	0.999	0.999	0.947

With interaction effect!

With no interaction effect!

	GCV	RSS	GRSq	RSq	CVRSq
ED	0.00005	0.0005	0.999	0.999	0.994
EM	0.05564	0.5564	0.998	0.998	0.993
HD	0.00006	0.0006	0.999	0.999	0.993
HM	0.08064	0.8064	0.998	0.998	0.990

Of course, we select the MARS model with interaction effect

COMPARISON OF THE MARS MODELS WITH/WITHOUT INTERACTION EFFECT

CONCLUSION

- Thus, MARS algorithm as a non-parametric regression is an extraordinary statistical tool that estimates the suitable cut-off values i.e. 40^0 azimuth angle and 3.1 ESLOANGREF values for the influential predictors and their interactions affecting the solar irradiation parameters without requiring any assumptions on the distribution of the predictors.
- As a result, it can be suggested that the procedure of MARS algorithm, producing the greatest predictive accuracy of 100(%) or nearly 100(%) here,
- MARS approach permits researchers to obtain some remarkable hints for ascertaining predictors affecting solar irradiation parameters.

WHY MARS IS BEING USED FOR SOLVING OPTIMIZATION PROBLEMS OR DIFFICULT PROBLEMS?

- MARS is a powerful alternative to Response Surface Approach for multiple continuous responses
- MARS produces more effective results compared to classical approaches like Multiple Linear Regression in the violation of distributional assumptions
- MARS produces more understandable outputs compared to ANNs types i.e. MLP and RBF.
- MARS produces the prediction equations for all the responses
- MARS produces much understandable outputs in the R studio free package program

ALL COMMANDS ON R SOFTWARE

- > ## MARS algorithm for multiple continuous responses in R software##
- > d=read.table("C:/gok.txt", header= T)
- > str(d)
- > install.packages("earth")
- > library(earth)
- > gokhan=earth(cbind(Ed, Em, Hd, Hm)~., data=d, penalty=-1, pmethod="backward", degree = 2, nk=150, nfold=3, keepxy=T)
- > summary(gokhan) ## Summary results of the MARS algorithm
- > evimp(gokhan) ## Relative importance of the influential predictors
- > plot(gokhan, nresponse = 1)
- > plot(gokhan, nresponse = 2)
- > plot(gokhan, nresponse = 3)
- > plot(gokhan, nresponse = 4)

ALL COMMANDS ON R SOFTWARE

- > plotmo(gokhan, nresponse = 1)
- > plotmo(gokhan, nresponse = 2)
- > plotmo(gokhan, nresponse = 3)
- > plotmo(gokhan, nresponse = 4)
- > ## Use the following commands for estimating SD Ratio of each response
- > install.packages("pastecs")
- > library(pastecs)
- > residualvalues <- gokhan$residuals
- > stat.desc(residualvalues) ## descriptive statistics of residuals for each response variable
- > stat.desc(d$Ed)## descriptive statistics of the actual Ed values
- > stat.desc(d$Em)## descriptive statistics of the actual Em values
- > stat.desc(d$Hd)## descriptive statistics of the actual Hd values
- > stat.desc(d$Hm)## descriptive statistics of the actual Hm values

MANY THANKS FOR YOUR INTEREST!