JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

Forecasting of Sugarcane Productivity Estimation in India - A Comparative Study with Advanced Non-Parametric Regression Models

Kiran Kumar Paidipati [1], Arjun Banik [2], Bhavin Shah [3], and Narpat Ram Sangwa [4]

[1]Assistant Professor, Area of Decision Sciences, Indian Institute of Management (IIM), Sirmaur-Himachal Pradesh-173025, India kkpaidipati@iimsirmaur.ac.in

[2]Department of Mathematics and Statistics, University of Victoria, British Columbia, Canada abanik@uvic.ca

[3]Assistant Professor, Area of Operations and Supply Chain Management, Indian Institute of Management (IIM), Sirmaur- Himachal Pradesh-173025 bhavin.shah@iimsirmaur.ac.in

[4]Assistant Professor, Area of Operations and Supply Chain Management, Indian Institute of Management (IIM), Sirmaur- Himachal Pradesh-173025 narpat.sangwa@iimsirmaur.ac.in

ABSTRACT

Purpose: In recent times, sugarcane production and area under cultivation have been fluctuating from year to year depending on climate and price policy, adversely affecting sugarcane growers' decisions to invest in cultivation and their livelihood. The declining trend of productivity may affect the future competitiveness, and therefore it needs to be investigated.

Design/Methodology/Approach: In this study, prediction of sugarcane yielding through regression analysis is performed with the help of Multivariate Adaptive Regression Splines (MARS), Support Vector Regression (SVR), Partial Least Square Regression (PLSR), Elastic-Net Regression, and Multiple Linear Regression (MLR) on the basis of the historical data of sugarcane cultivation from 1971-72 to 2018-19. The prediction is done by training all the regression models with 80% of the data, by taking the overall Indian sugarcane productivity as a dependent variable and other major sugarcane producing states as independent variables.

Findings: As a main result, the non-parametric regression model MARS is found to be much better than other well-fitted models. All of these models' performances are cross-validated using the Root Mean Square Error (RMSE), Mean Absolute Percentage Error

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

(MAPE), and the Wilcoxon Signed-Rank test. Also, the MARS model is found to be a more flexible and accurate model in predicting the behavior of sugarcane yielding in India.
Practical Implications: The practitioners and farmers facilitate the model comparisons to achieve more profits through accurate estimation. The research outcome implicates the agricultural industry to improve the sugarcane cultivation and productivity under uncertain environments.
Originality/Value: The study suggests best management practices can be developed to increase the large potential of sugarcane production in India towards greater sustainability and food security modelling.

Keywords: Regression Models; Agriculture Security Modelling; Sustainable Modelling; Sugarcane Productivity; Yield Estimation; Machine Learning

1. Introduction and Literature

Agriculture plays an inevitable role in the development of a country's economy and way of life as it is the predominant occupation of its population. In countries like India, where the demand for food is increasing at an alarming rate due to the rapid population increase, agriculture is a tremendous bolster to meet future needs. But, in this twenty-first century, global food security is one of the most pressing issues. The population proliferates and it leads to a decline in the area under cultivation. Prediction is one of the most important statistical tools for forecasting future demand for major food crops in terms of total and irrigated areas. In India, agriculture generates employment for almost 54.6% of the total population (Census of India). Sugarcane itself supports more than 50 million farmers and their families (Solomon, 2016), because it occupies an important place among all commercial food crops.

In India, besides being the largest consumer and second largest producer of sugarcane in the world after Brazil, the sugar industry is the second largest agro-based industry (Solomon, 2014). The sugar industry in India generates approximately Rs. 80,000 crores in annual revenue and directly employs 5.5 lakh skilled and semi-skilled workers in sugar mills (Amaladoss, 2015). The Indian sugar industry every year generates approximately 6–8 million tons of jaggery and khandsari, followed by 27–28 Mt of white sugar and also around 350–370 Mt of cane (Solomon, 2011), to meet the domestic consumption of sweeteners. As the crop production and price rates are closely associated, the market surplus and the earnings of the farmers decrease with the unforeseen decrement in production, and this leads to a price hike.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

Besides sugar, many by-products and co-products are obtainable from the sugarcane crop and can also provide organic fertilizer, fibres, and biofuel (Singh et al. 2018). Sugarcane has become a highly strategic and crucial commercial crop in India, mainly in the South and South-Western states such as Maharashtra, Tamil Nadu, and Andhra Pradesh. Especially in Tamil Nadu, as (Balanagammal et al. 2000) forecasted and discussed the productivity of sugarcane. Prediction of productivity as the ratio of output to input volume is an intrinsic parameter for establishing a supporting policy decision concerning revenue generation, management practices, environmental issues, effective land use allocation, and so on (Suryani et al. 2020). The proposed study concentrated on forecasting the sugarcane area, production, and productivity of Tamil Nadu by fitting univariate ARIMA models for subsequent years (Suresh et al. 2011). The authors focused on forecasting the sugarcane area, production, productivity, and sugar production of India and major sugarcane producing states. Different ARIMA models were applied to India as a whole and to major states, and the results were validated by comparing actual values. The study was useful for making policy decisions on the production scenario in the country (Vishawajith et al. 2016). The study proposed predicting three leading years through the yearly time series data of sugarcane production in India by ARIMA models and validating them with standard statistical techniques (Mandal, 2005). The authors fitted Box-Jenkins's ARIMA model to forecast sugarcane production in India for the situation in the next five years and an average growth rate of approximately 11% per year (Kumar and Anand, 2014). Some researchers outside India applied the ARIMA model to predict the sugarcane production in Pakistan by validating it with error analysis, and the results were beneficial to governments, sugar mills, and farmers as well (Mehmood et al. 2019). The researchers used a linear trend regression model to determine the growth and production of sugarcane in India as well as major contributing states over five-year plans from 2000 to 2010 (Nandhini et al. 2017). Also, the study proposed applying multiple regressions to auto-correlated sugarcane production data to predict alcohol production as well as renewable energy sources like ethanol (Pedroso et al. 2014). A subsequent part is to focus on reviewing emerging machine learning technologies on big data from agricultural production systems. The paper thoroughly explores regression models and machine learning techniques applied to agricultural production and productivity by many researchers across the globe (Liakos et al. 2018).

The researchers focused on developing three machine learning regression techniques such as Support Vector Regression (SVR), Random Forest Regression (RFR) and Partial Least Squares Regression (PLSR) algorithms for agricultural management schemes for nature protection and areas without land use incentives (Schwieder et al. 2014). The researchers proposed supervised machine learning approaches such as support vector regression and the Naïve-Bayes algorithm for sugarcane yield prediction in Karnataka using long-term time series data (Medar et al. 2019). The authors developed data mining models

such as random forest, boosting, and support vector machines to understand the factors which influence the prediction of sugarcane yield in Brazil and validated the estimated values with RMSE (Hammer et al. 2020).

Similarly, demand forecasting models for agriculture performance analysis were tested by Priyadarshi et al. (2019). Recently, Suryani et al. (2020) presented a model to estimate productivity and improve corn production under uncertain environmental dynamics. However, such a regression modelling of sugarcane yielding through prediction and machine learning model comparison would assist practitioners to estimate and improve the productivity. The objective of this research work is to connect gap by addressing the research question as, how can we improve the sugarcane productivity by comparing different estimation models? It also explores factors affecting cultivating trends under Indian environmental contexts.

In this study, MARS and some robust regression models such as SVR, PLSR, Elastic-Net Regression, and MLR are employed. Initially, explanatory analysis for sugarcane yields in India and some major sugarcane-producing states has been done. Further, the models are trained with over 80% of the data of sugarcane yielding in India, selected randomly from 1971-72 to 2018-19. Finally, the comparison of all the regression models is carried out through the error analysis and with the 20% test data of sugarcane yielding in India. The last section concludes the findings along with future scope and limitations of the study.

2. Methodology

This section highlights data collection and model building process followed by results comparisons.

2.1 Collection of Data

The data used in this study were collected from the Ministry of Agriculture and Farmers' Welfare, Govt. of India. The dataset includes sugarcane yield, production, and area under cultivation for each state as well as for India as a whole from 1971-72 to 2018-19. Then, 80% of the data were selected randomly to train the regression models, playing a key role in model building and prediction. The rest of the unselected data points were used for testing and validating the well-fitted regression models.

2.2 Model Building and Fitting

2.2.1 MARS Model

The concept of the Multivariate Adaptive Regression Splines (MARS) model was popularized by (Friedman, 1991). The feature of this model is that it automatically models the nonlinearities and the interactions between the variables. It belongs to the family of non-parametric regression models.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

The MARS model is governed by an estimated function of the following form:

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x) \quad \ldots (1)$$

where c_i is the constant coefficient and $B_i(x)$ is the weighted sum of basis function. Each of the basis function takes at least one of the following forms:

i. A constant term, which is the intercept of the model.
ii. A hinge function: The MARS model automatically selects the cut-off points as knots and the value of the variable as knots as hinge functions.
iii. A product of two or more hinge functions. The MARS model has the degrees of interactions between two or more variables.

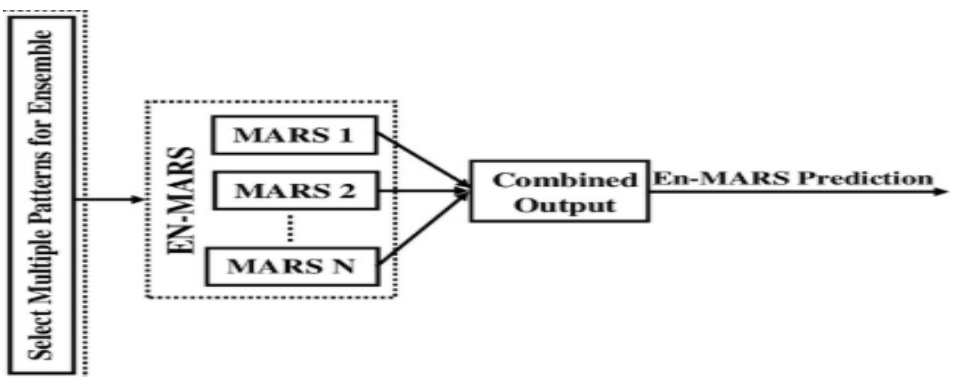

Figure 1: The process of Ensemble for the MARS model

2.2.1.1 Hinge Function

In the core concept of building the MARS model, hinge function plays the major role. The hinge function can be of the form

$$\max(0, x - c)$$

or

$$\max(0, c - x) \quad \ldots (2)$$

where c is a constant known as knot.

2.2.1.2 Model Building Process

In the model building process of MARS model, there exist two phases, such as:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

i. Forward Pass

The model building of MARS starts with just an intercept term. Then, it repeatedly adds the basis function in the form of pairs in the model. At each step, it finds the pairs of basic functions that help to provide the maximum reduction in the sum of squares due to error. This process of adding continues until there is no change in residual error and the maximum number of terms is reached.

ii. Backward Pass

The backward plays an important role in overcoming the problem of over-fitting caused by forward passing. It helps to prune the model to build a model with much better generalization ability. It removes the terms one by one, at each step, to find the best sub-model. Then, the subsets of models are compared and examined with the help of a special validation criterion presented below.

2.2.1.3 Generalized Cross Validation (GCV)

The Generalized Cross Validation (GCV) helps to compare the performance of the subsets of models, for choosing the best subset. The formula for this GCV is

$$GCV = RSS/(N \cdot (1 - (Effective\ number\ of\ parameters)/N)^2) \quad \ldots (3)$$

where N is the number of observations and RSS is the residual sum of squares.

$$Effective\ number\ of\ parameters = (Number\ of\ Mars\ terms) + (Penalty) \cdot ((Number\ of\ Mars\ terms) - 1)/2 \quad \ldots (4)$$

where the value for penalty is either 2 or 3.

If the GCV values are lower, then the model can be interpreted as a better model.

2.2.2 PLSR Model

The Partial Least Square Regression (PLSR) model is a statistical method useful for finding the hyperplanes of maximum variances between the dependent and independent variables. It is essentially a machine learning technique that can be used to solve both single-level and multi-level learning problems. It was introduced by Herman O. A. Wold, and the main objective of this model is that it tries to find the multidimensional direction in a space that explains the maximum multidimensional variance direction in another space. Because of its computational efficiency, along with its ability to simultaneously operate on dimension reduction and model training, PLSR is always a salient choice for the purpose of prediction. The purpose of the PLSR is to predict a set of response variables from a set of explanatory variables. This prediction can be attained by extracting a set of orthogonal factors from the predictors having the best predictive power, known as latent variables. Moreover, the quality of the prediction can be obtained by evaluating the PLSR model with

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

the help of cross-validation techniques such as bootstrapping and jackknife. Moreover, the potential of PLSR in the field of agriculture remains to be explored, especially for the purpose of prediction in sugarcane agriculture.

2.2.3 SVM Model

Support Vector Machines (SVM) are supervised machine learning models that are used to analyse the data for classification and regression. In this study, the SVM was used for regression analysis. This non-parametric regression model is very useful for prediction when the data is affected by nonlinearities and it also plays a significant role in the presence of outliers. Figure 2 illustrates the work of Support Vector Regression (SVR) dealing with outliers and the linear parameterization of SVM regression.

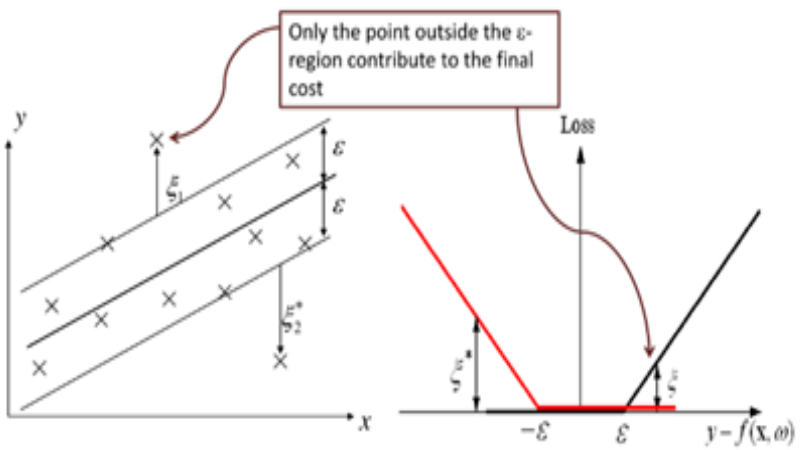

Figure 2: Geometrical representation of linear SVR

The SVR model can be expressed as a functional equation of the following form:

$$f(x) = \{w, \phi(x)\} + b \quad \ldots (5)$$

where w is the weight vector of inputs, b is the bias and $\phi(x)$ is a kernel function, where a non-linear function is used to transform a non-linear input into a linear input.

The goal is to find the function f(x) that has maximum ε-deviations from the obtained targets y_i, for all the training data. The errors are neglected, as long as they lie inside the ε-insensitive band. The insensitive loss function to SVR ε, as introduced by Vapin, can be expressed as

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

$$L_{\varepsilon} = (f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon & if\ |f(x) - y| \geq \varepsilon, \\ 0 & Otherwise \end{cases} \quad \ldots (6)$$

where ε is the region ε-insensitivity. When the predicted value falls outside the band, then the difference between the predicted value and the margin becomes equal to the loss; whereas, if the predicted values are inside the band area, then there is no loss. The objective function and the constraints can be expressed as

$$\begin{aligned} &\min \ \frac{1}{2}(w,w) + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \\ &Subject\ to \quad ((w, \phi(x_i) + b) - y_i \leq \varepsilon + \xi_i, \\ &\qquad y_i - \left(\left(w, \phi(x_i)\right) + b\right) \leq \varepsilon + \xi_i^*, \\ &\qquad \xi_i, \xi_i^* \geq 0, i = 1, 2, ..., n \end{aligned} \quad \ldots (7)$$

where n is the number of training data, ξ_i and ξi^* are the slack variables and ($\xi_i + \xi_i^*$) is the empirical risk, C is the modifying coefficient, which gives the trade-off between model complexity and training error. After, selecting a band width (ε), kernel function (φ) and modifying coefficient (C), Lagrange function is used to obtain the optimum value of each parameter.

2.2.4 Elastic-Net Model

The Elastic–Net regression model is one of the most useful models among all the non-parametric regression models. It is a regularized regression method that combines the Ridge and Lasso regressions. This Elastic-Net model overcomes the limitation of Lasso regression which uses the penalty function based on

$$\|\beta\|_1 = \sum_{j=1}^{p} \left|\beta_j\right| \quad \ldots (8)$$

To overcome the limitations of the saturation of the Lasso model in the case of large numbers and of highly correlated variables, where Lasso selects only one variable in a group and ignores the others, the elastic net adds a quadratic part to the penalty. Then the estimates of this model are defined by

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1) \quad \ldots (9)$$

In some cases, the Elastic-Net Regression gets reduced to SVR.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

2.2.5 Multiple Linear Regression

Multiple linear regression is a statistical technique, which takes into account several independent variables to predict the dependent variable. It is based on some assumptions list as follows:

- The response and explanatory variables are linearly related.
- The explanatory variables are not highly correlated with each other.
- Residuals are normally distributed with mean 0 and variance σ^2.
- Y_i's are selected independently and randomly from the population.

The goal of multiple regressions is to model the linear relationship between the independent and dependent variables. It can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon \quad \ldots (10)$$

$where\ for\ i = n\ \ observations:$

$y_i = \text{dependent variable}$

$x_i = \text{explanatory variable}$

$\beta_0 = y - \text{intercept (constant term)}$

$\beta_p = \text{slope coefficients for each explanatory variable}$

ε=the model's error term (also known as the residuals)

2.3 Model Evaluation and Accuracy Measures

The evaluation of models can be performed using some robust accuracy measures such as the root mean square error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD). They are defined by

$$MAPE = 100 * \frac{\left(\sum_{i=1}^{n} \left| F_i - O_i \right| / O_i \right)}{n} \quad \ldots (11)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (F_i - O_i)^2}{n}} \quad \ldots (12)$$

where O_i is the actual variable, F_i is the predicted variable and n is the number of variables, and $MAD = \frac{\sum_{i=1}^{n} \left| x_i - y_i \right|}{n}$ … (13)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

where x_i is the actual values, y_i is the predicted values and n is the number of values in each model. Also, the performance of the regression models depends upon the lowest MAPE value, lowest RMSE and the lowest MAD values.

Lastly, the Wilcoxon Signed-Rank test, which is one of the powerful non – parametric statistical hypothesis tests, is used to compare the difference between predicted values depending on the corresponding model and the actual values.

Let, N is the sample size. Then, there will be a total of 2N data points. Let, x_{1i} and x_{2i} be the measurements for pairs i = 1, 2... N.

Then the test statistic W is given as,

$$W = \sum_{i=1}^{N_r} \left[\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i \right] \quad \ldots (14)$$

where R_i denotes the rank. If the value of the test statistic W gets larger than the critical value, the Null hypothesis, which states the difference between two samples equals to 0, gets rejected.

3. Results and Discussions

The detailed study of parametric and some non–parametric regression models shows that the non–parametric regression models are the widely used methodologies for expressing the characteristics of the response variable on several independent variables.

3.1 Overview of the Data

An overview of sugarcane cultivation in India and the major sugarcane producing states is shown in Table 1. Among all the sugarcane producing states, Tamil Nadu yields the most as well as it is negatively skewed and leptokurtic, which is also reflected in the distribution of sugarcane yields for overall India. The states with the minimum deviation in sugarcane yielding are Andhra Pradesh and Maharashtra, which are also negatively skewed and platykurtic, whereas Uttar Pradesh and Punjab are positively skewed and leptokurtic.

Table 1: Descriptive statistics of Yielding in Sugarcane Production of India and Other Major states

Parameters	Mean	Median	Standard Deviation	Skewness	Kurtosis
A.P.	74.57	75.18	5.63	-0.268	-0.817
U.P.	53.59	55.68	8.78	0.07	0.396
Maharashtra	82.93	83.19	8.46	-0.84	-0.06
Punjab	60.70	60.81	8.63	0.20	0.66
Tamil Nadu	100.42	101.50	7.98	-0.988	0.86

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

Overall India	63.35	65.50	7.88	-0.39	0.27

The graphical view of the sugarcane yields of India and the major producing states is shown in Figure 3. From this figure, it can be interpreted that the sugarcane yielding in Punjab is almost similar to the overall sugarcane yielding in India. The sugarcane yield in Uttar Pradesh is the lowest, whereas Tamil Nadu is the leading sugarcane yielding state among all other major sugarcane producing states in India. The overall yield of sugarcane in India is increasing significantly.

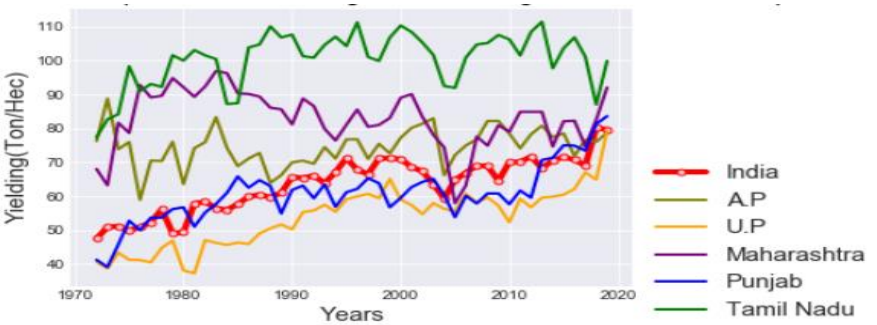

Figure 3: Comparison between Sugarcane Yielding of Overall India and major producing states of India (1971-72 to 2018-19)

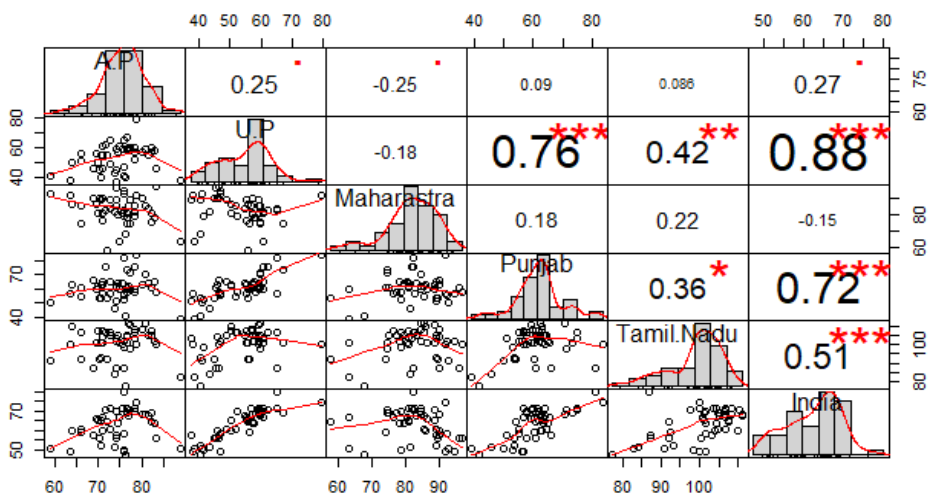

Figure 4: Performance graph of India and other important states in sugarcane yielding (1971-72 to 2018-19) along with the correlations between them.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

The performance graph of India and major producing states, as shown in Figure 4, shows that the yielding of sugarcane in Uttar Pradesh and overall India is highly positively correlated, as well as with Punjab also. The sugarcane yield in U.P is only negatively correlated with Maharashtra. The yield of sugarcane for overall India is also positively correlated with Punjab.

3.2 Fitting of MARS model

The well–fitted MARS model of degree 3 is obtained to demonstrate the nonlinear relationship between the dependent variable India and major producing Indian states such as Andhra Pradesh, Uttar Pradesh, Maharashtra, Punjab, and Tamil Nadu as independent variables. This degree 3 MARS model acquired all potential knots across all supplied features and, based on the expected change in R^2, it pruned to the optimal number of knots (cut points). In this study, the obtained MARS model used five of the 15 terms. So, there will be 5 terms in the model, and it includes the produced hinge functions from the original 5 predictors.

The model selection plot graphs the GCV R^2 based on the number of terms retained in the model that are assembled from the original predictors, as illustrated in Figure 5.

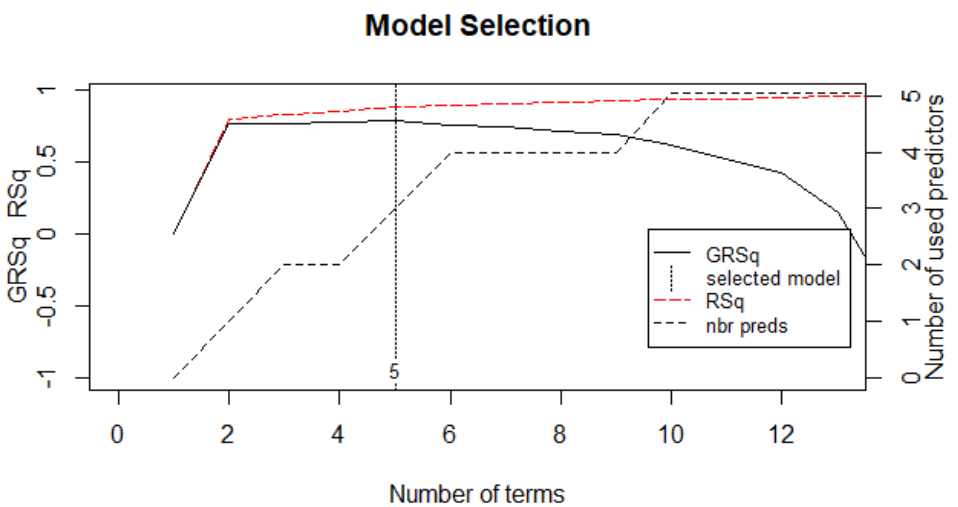

Figure 5: Model selection plot for the MARS model of degree 3

The line dashed vertically to 5 in Figure 5 shows the retained optimal number of non-intercept terms for which the marginal increase in GCV R^2 is not more than 0.001. The optimal combination includes 3rd degree interactions and retains 12 terms with an R^2 value of 0.87, which concludes that the model fits quite well. Then, the 10-fold cross validation

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

is performed, and the cross validated RMSE for the models of degrees 1, 2, and 3 is obtained as illustrated in Figure 6. The optimal model's cross-validated RMSE lies between 0 and 3.

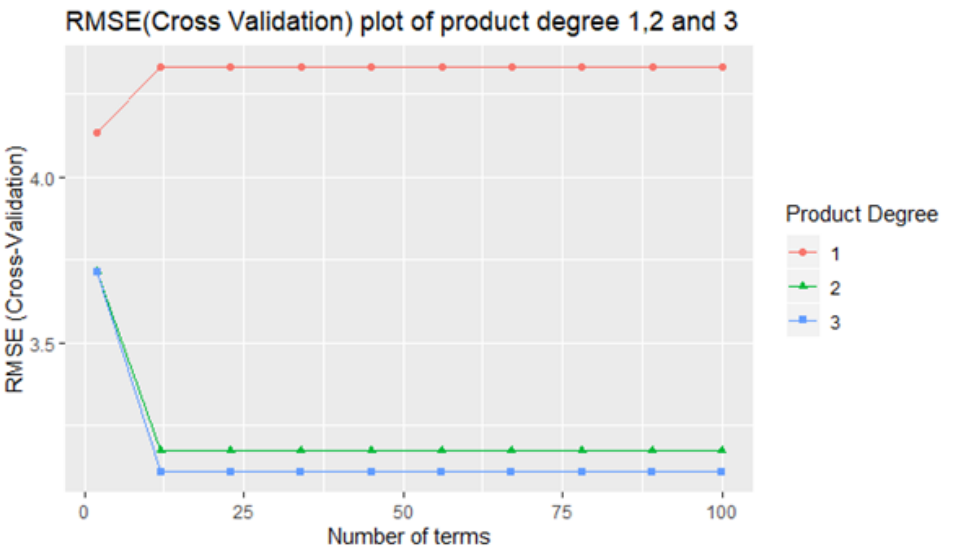

Figure 6: RMSE (Cross Validation) plot for MARS model up to 3rd degree interaction

The minimum RMSE (cross validation) is obtained for the 3rd degree interaction of the MARS model, whereas for the 1st degree the RMSE (cross validation) value is much higher. So, the model with optimized error is obtained as MARS with degree 3. The model includes interaction terms between multiple hinge functions.
Therefore, the model is obtained as:

$$\begin{aligned} India \;=\; & 68.43 - 0.76 * (59.11 - U.P) - 0.2 * (106.76 - Tamil\ Nadu) \\ & -0.07 * (A.P. - 78.24) * U.P. - 0.04 * (A.P. - 76.27) * Maharashtra \\ & + 0.000505 * (59.583 - U.P.) * Punjab * Tamil\ Nadu \end{aligned} \quad \ldots (13)$$

where the intercept is 68.43 and following are the multiple hinge functions with their coefficient values, which defines the interaction between the variables. For instance, in the obtained model, there is a knot point at 59.11 for U.P with a coefficient of 0.76, there is also another knot point for Tamil Nadu at 106.76, with a coefficient of 0.2, and so on with the interaction between the independent variables. These knot points are the points where there are significant changes in the characteristics and behavior of the curve.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

3.3. Model Building using MARS, SVR and Some Other Regression Models

Due to its inflexibility and the tendency to get highly affected in the presence of non-linearity in the data, the multiple linear regression model fails to provide predicted values with better accuracy in comparison with other robust regression models. The predicted values for the yield of sugarcane production in India are shown in Table 2, where the regression models are trained on the basis of 80% of the data randomly selected, and then the predicted values are tested on the rest of the data. Among all the regression models, the MARS of degree 3 and the SVR model predict the sugarcane yield much closer to the actual values.

Table 2: Predicted Values of Sugarcane Yielding in India with the fitted Regression Models

Years	Actual values	MLR	Elastic Net	SVM	PLSR	MARS
1977	52.353	52.63	53.199	52.02	53.38	51.42
1978	56.16	53.73	54.86	54.17	54.66	55.52
1987	60.443	63.8	62.03	61.32	63.33	60.31
1999	71.205	68.799	69.29	71.13	67.49	69.42
2005	64.754	60.5	62.74	63.33	59.85	62.66
2008	68.879	65.73	68.506	69.31	67.95	67.13
2018	70.72	67.42	71.08	71.89	69.72	68.45
2019	79.6	79.91	82.17	82.82	76.46	79.07

Then, the predicted values that are obtained from the different regression models are plotted as shown in Figure 7, which visualizes the closeness of the predicted values obtained from various regression models to the actual value.

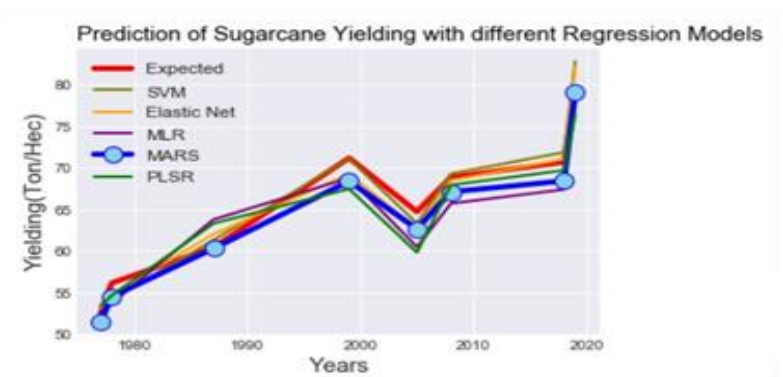

Figure 7: Prediction of Sugarcane Yielding in India with various Regression Models

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 760 - 778
https://publishoa.com
ISSN: 1309-3452

3.4. Comparison on the well - fitted robust Regression Models

Among all the fitted regression models, the MARS model predicts much better than all the other models. The root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute deviation (MAD) values are 1.72, 0.013 and 1.209, respectively, which is much less than the accuracy measure values of all the other regression models. The prediction using SVM regression comes closer to the actual, but the values of the accuracy measures are a bit higher than those of the MARS model.

Table 3: Accuracy Measures of all the fitted Regression Models for the Sugarcane Yielding of India (1970-71 to 2018-19)

Regression Models					
Accuracy Measures	**MLR**	**Elastic-Net**	**SVM**	**PLSR**	**MARS**
RMSE	2.72	1.805	1.75	2.68	1.72
MAPE	0.04	0.024	0.0201	0.037	0.013
MAD	2.51025	1.4455	1.264875	2.427625	1.209125

Table 4: Values of p-value obtained from Wilcoxon Signed - Rank Test

Regression Models					
Models	MLR	Elastic Net	SVM	PLSR	MARS
p-value	0.4688	0.9375	0.9546	0.6875	0.9865

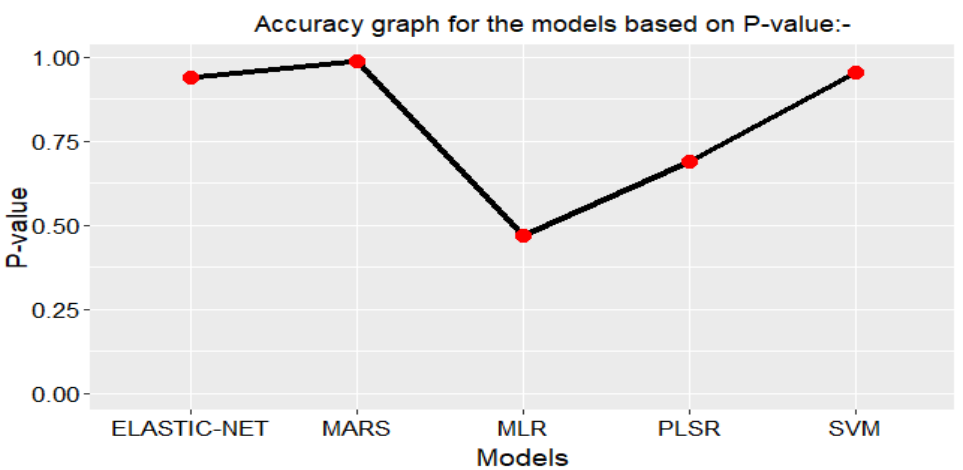

Figure 8: Accuracy graph of different regression models based on p-value

Finally, the Wilcoxon Signed-Rank test is used to compare the predicted values obtained from each of the fitted regression models with the actual values. In this study, the p-values obtained shown in Table-4 indicate that all the null hypotheses of the difference between actual and predicted value being zero get accepted. Because the p–value obtained by the MARS model is 0.9865, which is greater than the p–values obtained by the other models, it can be concluded that the difference between the predicted value using the MARS model and the actual values is nearly close to zero, implying that the prediction using the MARS model is better than all other regression models. Moreover, the graph of the p-value obtained by all the fitted regression models is shown in Figure 8, which elucidates the accuracy of the regression models more prominently.

4. Conclusions and Future Work

As a result of their modeling flexibility, non–parametric regression models are generally superior and are widely used in most fields. The MARS model, in particular, flourished in model building and prediction due to its capability of handling complex situations. In this study, MARS, SVR, and some other regression models are used for predicting in the field of sugarcane cultivation in India. The robustness of these modeling techniques is tested, which are largely for non-linear modeling. In this study, the MARS model is found to be more flexible and better than other regression models because it aids in the development of a more accurate model. The study suggests best management practices can be developed to increase the large potential of sugarcane production in India towards greater sustainability. It also addresses the global food security issues raised by (Aguilar-Rivera, 2022).

In the future, the MARS model can also be applied in various agricultural fields such as rice, tobacco, oilseeds, and corn (Chami et al. 2020). Due to the flexibility of the MARS model in building piecewise linear regression models, it will be very useful in making short-term and long-term forecasting models. This particular study is restricted to the overall sugarcane productivity of India and the major sugarcane producing states of India. Further, it can be extended to deal with areas under cultivation as well as the production of sugarcane. Also, the MARS model can be used to forecast the future values, and then it will give a clearer picture of the accuracy and flexibility of the model to suit non-linear patterns of agricultural production.

References:

1. Aguilar-Rivera, N. (2022). Bioindicators for the Sustainability of Sugar Agro-Industry. *Sugar Tech*, 1-11.

2. Amaladoss, A. X. (2015). A Study of Industrial Relations At M/S Perambalur Sugar Mills Ltd, Eraiyur, Perambalur Dt (Doctoral Dissertation, Vinayaka Missions University).

3. Balanagammal, D., Ranganathan, C. R., & Sundaresan, K. (2000). Forecasting of agricultural scenario in Tamilnadu: A time series analysis. *Journal of Indian Society of Agricultural Statistics*, *53*(3), 273-286.

4. Census of India, 2011; https://censusindia.gov.in/2011 common/censusdata2011.html

5. El Chami, D., Daccache, A., & El Moujabber, M. (2020). What are the impacts of sugarcane production on ecosystem services and human well-being? A review. *Annals of Agricultural Sciences*, *65*(2), 188-199.

6. Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.

7. Hammer, R. G., Sentelhas, P. C., & Mariano, J. C. (2020). Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. *Sugar Tech*, *22*(2), 216-225. https://doi.org/10.1007/s12355-019-00776-z

8. Kumar, M., & Anand, M. (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. *Studies in Business and Economics*, *9*(1), 81-94.

9. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

10. Mandal, B. N. (2005). Forecasting sugarcane production in India with ARIMA model. *Inter Stat, October*.

11. Medar, R. A., Rajpurohit, V. S., & Ambekar, A. M. (2019). Sugarcane Crop Yield Forecasting Model Using Supervised Machine Learning. *International Journal of Intelligent Systems and Applications*, *11*(8), 11. DOI: 10.5815/ijisa.2019.08.02

12. Mehmood, Q., Sial, M. H., Riaz, M., & Shaheen, N. (2019). Forecasting the Production of Sugarcane in Pakistan for the year 2018-2030, using Box-Jenkin's Methodology. *JAPS, Journal of Animal and Plant Sciences*, *29*(5), 1396-1401.

13. Nandhini, T. S. K. D., & Padmavathy, V. (2017). A Study on Sugarcane Production in India. *International Journal of Advanced Research in Botany*, *3*(2), 13-17.

14. Pedroso, G. T., Silva-Mann, R., Camargo, M. E. I., & Russo, S. L. (2014). Applied multiple regression for autocorrelated sugarcane data. *African Journal of Agricultural Research*, *9*(10), 914-920. https://doi.org/10.5897/AJAR2013.7282

15. Priyadarshi, R., Panigrahi, A., Routroy, S. and Garg, G.K. (2019), "Demand forecasting at retail stage for selected vegetables: a performance analysis*"*, *Journal of Modelling in Management*, Vol. 14 No. 4, pp. 1042-1063.

16. Schwieder, M., Leitão, P. J., Suess, S., Senf, C., & Hostert, P. (2014). Estimating fractional shrub cover using simulated EnMAP data: A comparison of three machine learning regression techniques. *Remote Sensing*, *6*(4), 3427-3445. https://doi.org/10.3390/rs6043427

17. Singh, P., & Tiwari, A. K. (Eds.). (2018). *Sustainable sugarcane production*. CRC press.

18. Solomon, S. (2011). The Indian sugar industry: an overview. *Sugar Tech*, *13*(4), 255-265. https://doi.org/10.1007/s12355-011-0115-z

19. Solomon, S. (2014). Sugarcane agriculture and sugar industry in India: at a glance. *Sugar Tech*, *16*(2), 113-124. https://doi.org/10.1007/s12355-014-0303-8

20. Solomon, S. (2016). Sugarcane production and development of sugar industry in India. *Sugar Tech*, *18*(6), 588-602. https://doi.org/10.1007/s12355-016-0494-2

21. Suresh, K. K., & Priya, S. K. (2011). Forecasting sugarcane yield of Tamilnadu using ARIMA models. *Sugar Tech*, *13*(1), 23-26. https://doi.org/10.1007/s12355-011-0071-7

22. Suryani, E., Dewi, L.P., Junaedi, L. and Hendrawan, R.A. (2020), "A model to improve corn productivity and production", *Journal of Modelling in Management*, Vol. 15 No. 2, pp. 589-621.

23. Vishawajith, K. P., Sahu, P. K., Dhekale, B. S., & Mishra, P. (2016). Modelling and forecasting sugarcane and sugar production in India. *Indian Journal of Economics and Development*, *12*(1), 71-80.