Environment and Ecology 42 (3) : 948—955, July—September 2024
Article DOI: https://doi.org/10.60151/envec/WUMC4415
ISSN 0970-0420

Sugarcane Yield Prediction in Bihar using Biometric Characteristics

Muhammed Irshad M., Mahesh Kumar, D. N. Kamat

Received 27 November 2023, Accepted 15 May 2023, Published on 3 July 2024

ABSTRACT

This study focuses on creating predictive models for sugarcane yield in Bihar, utilizing various biometrical characteristics. The research involved collecting observations on specific plant biometrical traits from 50 farmers' fields of Samastipur, West Champaran and East Champaran Districts of Bihar. Simple random sampling was employed for field selection. Various regression analyses were conducted to identify the optimal combination. The accuracy of the model was assessed by comparing the actual yield from 10% of the observations not used in model development with their predicted values. The results demonstrated a close resemblance, with a margin of error ranging from 5.91% to 8.36%. The forecasted sugarcane yield for Bihar, based on the proposed model, is 847.82 q/ha.

Keywords Multiple linear regression, Yield forecasting, Biometrical characters, Sugarcane yield.

INTRODUCTION

Sugarcane (*Saccharum officianarum*) stands out as a crucial commercial crop in India, particularly in Bihar, where it spans 0.21 million hectares, yielding a total production of 12.06 million tonnes and boasting a productivity rate of 57 tonnes/ha (Ministry of Agriculture and Farmers Welfare 2022). The application of forecasting techniques holds paramount importance in the realm of research within agriculture and related sectors. Such methodologies play a pivotal role in aiding governments, policymakers, agricultural scientists, farmers and various agencies in planning their future operations.

The challenge lies in achieving a reliable and timely forecast of crop production before the actual harvest, a task that has become increasingly significant. In India, the responsibility for forecasting crops before harvesting or for the upcoming years rests with the Director of Economics and Statistics at the Ministry of Agriculture in New Delhi. However, the existing forecasting methods carry a subjective nature, relying on eye-estimates and the personal judgment of agricultural officials. Moreover, the final crop production estimates, derived from objective crop-cutting experiments, have limited utility due to their delayed availability post-harvest. Consequently, there is a pressing need for the development of an objective methodology for pre-harvest crop forecasting. Such an approach would not only enhance

Muhammed Irshad M.[1]*, Mahesh Kumar[2], D. N. Kamat[3]

[2, 3]Associate Professor-Cum-Sr. Scientist

[1]Department of Agricultural Statistics, Palli Siksha Bhavana, Visva Bharati, Sriniketan, Birbhum, West Bengal 731236, India

[2]Department of BS and L (Statistics, Mathematics, Computer Application, Physics and Languages), Dr Rajendra Prasad Central Agricultural University, Pusa, Samastipur, Bihar 848125, India

[3]Sugarcane Research Institute, Dr Rajendra Prasad Central Agricultural University, Pusa, Samastipur, Bihar 848125, India

Email : irsh375@gmail.com
*Corresponding author

the accuracy of predictions but also contribute to more informed decision-making among stakeholders in the agricultural sector. In recent years, there has been a significant surge in the application of Multiple Linear Regression (MLR) in crop yield modelling, particularly with a focus on forecasting purposes. Researchers have increasingly turned to MLR as a statistical tool to analyze and predict crop yields based on various influencing factors. The essence of MLR lies in its ability to explore the relationships between multiple independent variables, such as weather conditions, soil properties, management practices, and pest occurrences, and a dependent variable, which is typically the crop yield. By collecting and analyzing data on these factors over time, researchers can develop MLR models that provide valuable insights into how changes in these variables affect crop productivity.

Agrawal and Mehta (2007) summarized crop yield modelling using weather indices and extended it to predict pests and diseases in various crops across different regions. They found that this approach provides reliable forewarnings at least one week in advance. Aditya and Das (2012) used discriminant function analysis to create a wheat yield forecasting model in Kanpur district, UP, experimenting with different methods to compute discriminant scores. Sisodia *et al.* (2014) replicated this approach for wheat yield forecasting in Faizabad district. The India Meteorological Department (IMD) collaborates with 46 Agromet Field Units (AMFUs) nationwide to develop operational yield forecasts for 13 major crops during *kharif* and *rabi* seasons as part of the FASAL project, employing statistical models (Ghosh *et al.* 2014).

Annu *et al.* (2015) developed statistical models for predicting pre-harvest wheat yield based on biometric characteristics under normal and late sowing conditions. They found that the linear multiple regression model, using biometric characters in their original form, consistently outperformed other models. This model exhibited smaller percent standard errors for wheat yield forecasts and achieved the maximum R-squared adjusted value, ranging from 49% to 51%.

Annu *et al.* (2016) utilized principal component analysis (PCA) to create statistical models for predicting preharvest rice yields based on biometric characteristics. The forecasted yields have a maximum standard error of nearly 10%. This study marks the first application of PCA as a statistical tool to develop a pre-harvest forecast model using experimental data.

Banakara *et al.* (2018) applied Multiple Linear Regression (MLR) techniques and Discriminant Function Analysis to estimate average rice production in Surat district, South Gujarat. The results from Surat district indicated that MLR techniques out performed Discriminant Function Analysis in forecasting rice crop yield before harvest.

Irshad *et al.* (2023) utilized a Multiple Linear Regression Model to predict sugarcane wilt in Bihar using weather parameters. They achieved an adjusted R-squared value of 90.7, leading them to recommend this model as the most effective forewarning tool for sugarcane wilt in Bihar.

MATERIALS AND METHODS

The present study is to develop yield forecast model of sugarcane in Bihar using biometrical characters. The biometrical characters are as follows (Table 1).

Number of millable canes per 100 m^2 (X_1)
Average plant height in cm (X_2)

Table 1. Measurable and non-measurable characters.

Sl. No.	Variables	Codes of variables	Unit of measurement	Type of characters
1	Yield	Y	q/ha	Measurable
2	Number of millable canes	X_1	per m^2	Measurable
3	Average plant height	X_2	cm	Measurable
4	Average cane girth	X_3	cm	Measurable
5	Average length of third leaves	X_4	cm	Measurable
6	Average width of third leaves	X_5	cm	Measurable
7	Average cane perimeter	X_6	cm	Measurable
8	Single cane weight	X_7	kg	Measurable
9	Average plant population	X_8	Numbers	Measurable
10	Number of irrigations	X_9	Numbers	Measurable

Table 1. Continued.

Sl. No.	Variables	Codes of variables	Unit of measurement	Type of characters
11	Average number of tillers	X_{10}	per m^2	Measurable
12	Nitrogen (N)	X_{11}	kg/ha	Measurable
13	Phosphorus (P_2O_5)	X_{12}	kg/ha	Measurable
14	Potassium (K_2O)	X_{13}	kg/ha	Measurable
15	Disease infestation	X_{14}	percentage	Measurable
16	Average plant condition	X_{15}	Eye estimate	Non-measurable

Average cane girth in cm (X_3)
Average length of third leaves cm (X_4)
Average width of third leaves in cm (X_5)
Average cane perimeter in cm (X_6)
Single cane weight in kg (X_7)
Average plant population per 100 m^2 (X_8)
Number of irrigations in entire crop season (X_9)
Average number of tillers per 100 m^2 (X_{10})
Application of nitrogen (N) in kg/ha (X_{11})
Application of phosphorus (P_2O_5) in kg/ha (X_{12})
Application of potassium (K_2O) in kg/ha (X_{13})
Disease infestation in percentage (X_{14})
Average plant condition (X_{15})

Planning done to record observations from Samastipur, West Champaran and East Champaran district of Bihar from where samples were obtained randomly. Selections of these districts were because of those covers three major sugarcane growing districts of Bihar. Total numbers of observations were 50, out of which 30 are collected from Samastipur district, 10 observations are from West-Champaran and 10 observations are from East Champaran. 10% observation will be kept for modal validation purpose where 90% observation will be used for developed linear model.

Out of 50 observations obtained from sampling, 10% of recorded observations were kept for model validation purpose and 90% of observations were used for developing forecast model.

Multiple regression

The variations in regressors (X'S) cause variation in y. We fit multiple regression of y on X's to account for this variation. Multiple regression of y on X's is denoted as

$$Y=\beta_0 + \beta_1X_1 + \beta_2X_2 + . . . + \beta_PX_p + e$$

Where

β_0 denotes intercept

β_i's (i=1, 2..., p) are called partial regression coefficient and e is indicated as random error.

Multiple linear regression fitting

Suppose n observation are made on y and X's. Then, for each observation we have our unobserved error term ei. We make the following assumptions regarding the ei's which are random variables (i) errors are independent (ii) errors have zero mean and constant variance (σ^2). These assumptions can also be written as

$$E (e_i) = 0, V (e_i) = \sigma^2 \text{ for all } i = 1, 2, ., n.$$
$$Cov. (e_i, e_j) = 0$$

Analysis of variance

A regression ANOVA (Analysis of variance) table is a statistical table used in the context of linear regression analysis to assess the significance of the regression model and its individual parameters. The table is a key component in understanding the variance in the data and evaluating the overall fit of the regression model. The table typically includes the following components :

Source of variation : This column identifies the sources of variation in the data. It usually includes "Regression" and "Residual" (error).

Degrees of freedom (DF) : This column indicates the degrees of freedom associated with each source of variation. Degrees of freedom for regression are usually equal to the number of predictors (independent variables) in the model, and for residual, it is equal to the total number of observations minus the number of predictors.

Table 2. Analysis of variance (ANOVA) table.

Source of variation	df	SS	MS
Regression	P	$\Sigma b_i Sx_{iy}$	MSR
Deviation from regression (residual)	n-p-1	SSE	S^2 = SSE/ (n-p-1) = MSE
Total (corrected mean)	n-1	S_{yy}	

Sum of squares (SS) : This column provides the sum of squares associated with each source of variation. It represents the sum of the squared differences between the observed values and the predicted values for each variable.

Mean square (MS) : This is calculated by dividing the sum of squares by the degrees of freedom. It provides a measure of the average squared deviation from the mean.

F-statistic : The F-statistic is the ratio of the mean square for regression to the mean square for residual. It is used to test the overall significance of the regression model.

p-value : The p-value associated with the F-statistic helps determine whether the regression model is statistically significant. A low p-value (typically below 0.05) suggests that the overall model is significant.

R-squared (R^2) : This column provides the coefficient of determination, representing the proportion of the total variability in the dependent variable that is explained by the regression model.

Adjusted R-squared : This is a modified version of R-squared that accounts for the number of predictors in the model. It is useful for comparing models with different numbers of predictors.

The regression ANOVA table is a valuable tool for assessing the goodness of fit of a regression model and understanding the contributions of different sources of variation. It helps researchers and analysts make informed decisions about the relevance and significance of the model (Table 2).

Variance inflating factor

The VIF of predictor variable Xi is defined VIFi =1/TOLi and is major of the amount by which the variance of the standardize regression co-efficient is inflated by multicollinearity generally TOL lesser than 0.1 or equivalently VIF>10 I is regarded as a sign of multicollinearity in respect of the predictor and similar behavior of the average TOL or VIF a sign of multicollinearity of the entire model.

In this data said it is evident that there is not multicollinearity on all count's because VIF is < 10.

TOL = Tolerance of X_i is defined to be $TOL_i = 1- R_i^2$.

$$R^2 = \frac{\text{Regression sum of square}}{\text{Total sum of square}}$$

Where Total sum of square = Regression sum of square + Residual sum of square

$$R^2 = \frac{\text{1-Residuale sum of square}}{\text{Total sum of square}}$$

R^2 gives the percentage of variation explained by the predictor and hence is a useful indicator of the usefulness of the fitted regression.

Selection of best subset of regression analysis

Selection of variable through step up procedure and step-down procedure does not give unique result whereas selection of variable through all possible regression requires heavy computation and it is possible only through high-speed computing facilities. Since computational facilities are available for all possible regression equation, the best subset has to be chosen on the basis of following criteria.

R^2 criteria
adj R^2 criteria
Root mean square criteria
Coefficient of variation

Validity test for forecast model

Using the unused 10% observations of biometrical data of sugarcane, we calculate Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE). These are used for model validation.

RESULTS AND DISCUSSION

The following stages have been carried for developing forecast model for yield of sugarcane.

Sample selection

Use all possible regression models for developing forecast model

Fit model for Bihar and testing the validity of fitted model

The basic statistical measure and moments of all variable are presented in the Table 3 and found that average number of millable canes per 100 m^2 is 481, average plant height is 214.47 cm, average length of third leaves is 99.07 cm, single cane weight is 1.51 kg, population are 16.31, average plant height is 118.76 cm, average no. number of tillers are 15.62., minimum CV is 73.9% for X_4 variable, minimum SD is 0.23 for X_7.

Table 3. Basic statistical measures of biometrical characteristics.

Variable	Mean	SD	Sum	Variance	CV
Y	713.20	117.71	32094.00	13855.44	16.50
X_1	481.47	95.65	21666.00	9149.16	19.87
X_2	214.47	15.46	9651.00	238.86	7.21
X_3	2.43	0.42	109.11	0.18	17.41
X_4	99.07	7.32	4458.00	53.53	7.39
X_5	3.47	0.62	156.30	0.39	17.92
X_6	7.62	1.33	342.80	1.78	17.49
X_7	1.51	0.23	67.85	0.05	15.39
X_8	274.31	27.55	12344.00	759.16	10.05
X_9	3.73	1.64	168.00	2.70	44.01
X_{10}	15.62	1.84	703.00	3.38	11.76
X_{11}	123.87	28.52	5574.00	813.41	23.03
X_{12}	76.80	19.29	3456.00	371.95	25.11
X_{13}	63.86	18.58	2874.00	345.02	29.09
X_{14}	15.64	4.73	704.00	22.33	30.20
X_{15}	3.27	0.72	147.00	0.52	22.04

Table 4. Best five models among all possible regression.

Model	Number of variables	R^2	Adj R^2	RMSE	CV
X_1, X_7, X_9, X_{13}	**4**	**0.9229**	**0.9152**	**34.2806**	**4.8066**
X_1, X_5, X_7, X_9, X_{13}	5	0.9295	0.9205	33.1958	4.6545
X_1, X_5, X_7, X_9, X_{13}, X_{15}	6	0.9335	0.9231	32.6507	4.5781
X_1, X_2, X_5, X_7, X_9, X_{13}, X_{15}	7	0.9360	0.9239	32.4667	4.5523
X_1, X_2, X_4, X_5, X_7, X_9, X_{13}, X_{15}	8	0.9385	0.9248	32.2694	4.5246

Forecasting through all possible regressions

Out of 50 observation, 5 observation were kept for model validation and 45 observations were put for developing forecast model. The all-possible regression analysis was computed for 45 observations through R software. The best five model were selected on the basis of R^2, AdjR^2, RMSE, and CV and R^2 value has been presented (Table 4.) also regression analysis of proposed model is presented (Table 5) and analysis of variance is presented further (Table 6).

Table 5. Parameter estimates of first model.

Variable	Parameter estimate	Standard error	t value	Pr> |t|	Variance inflation
Intercept	-480.07542	62.6209	-7.67	<.0001	0
X_1	1.28392	0.06224	20.63	<.0001	1.32698
X_7	365.01454	25.5849	14.27	<.0001	1.32001
X_9	-5.15932	3.17116	-1.63	0.1116	1.01661
X_{13}	0.68923	0.28175	2.45	0.0189	1.02544

Table 6. Analysis of variance (ANOVA).

Sources of variation	DF	Sum of squares	Mean square	F- value	Pr> F
Model	4	562633	140658	119.69	<.0001
Error	40	47006	1175.16		
Corrected total	44	609639			

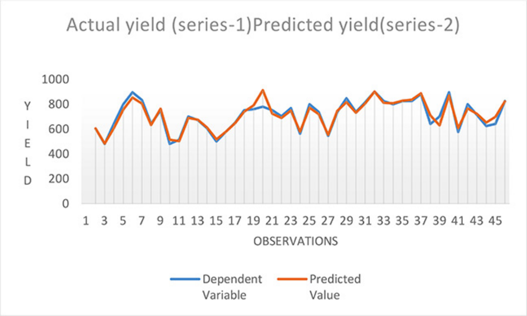

Fig. 1. Line diagram of actual and predicted yield (Model 1).

Table 7. Model validation among five selected models.

MODEL	MSE	MAPE	MAE
X_1, X_7, X_9, X_{13}	**5.855**	**2.749**	**22.014**
X_1, X_5, X_7, X_9, X_{13}	5.7616	2.755	21.834
X_1, X_5, X_7, X_9, X_{13}, X_{15}	5.7141	2.5703	20.57
X_1, X_2, X_5, X_7, X_9, X_{13}, X_{15}	5.698	2.463	19.81
X_1, X_2, X_4, X_5, X_7, X_9, X_{13}, X_{15}	5.6806	2.382	19.274

$\hat{Y}$= -480.07542+ 1.28392X_1 + 365.01454X_7 – 5.15932X_9 + 0.68923X_{13}

Where,

X_1 is number of millable canes per 100 m^2

X_7 is single cane weight in kg ()

X_9 is number of irrigations in entire crop season and

X_{13} is the application of potassium (K_2O) in kg/ha

A predicted vs. observed plot, also known as a fitted vs. actual plot or predicted vs observed scatter-plot, is a graphical tool used in regression analysis to visually assess how well the model's predictions align with the actual observed value (Fig. 1). If graph are superimposed /coincides to each other model, can be says well fitted. Even in some cases R^2 is more than 90% but graph is not superimposed to actual yield and estimated yield, model is not good fit. In this research, graph is superimposed / coincide to actual yield and estimated yield, it indicates model is good

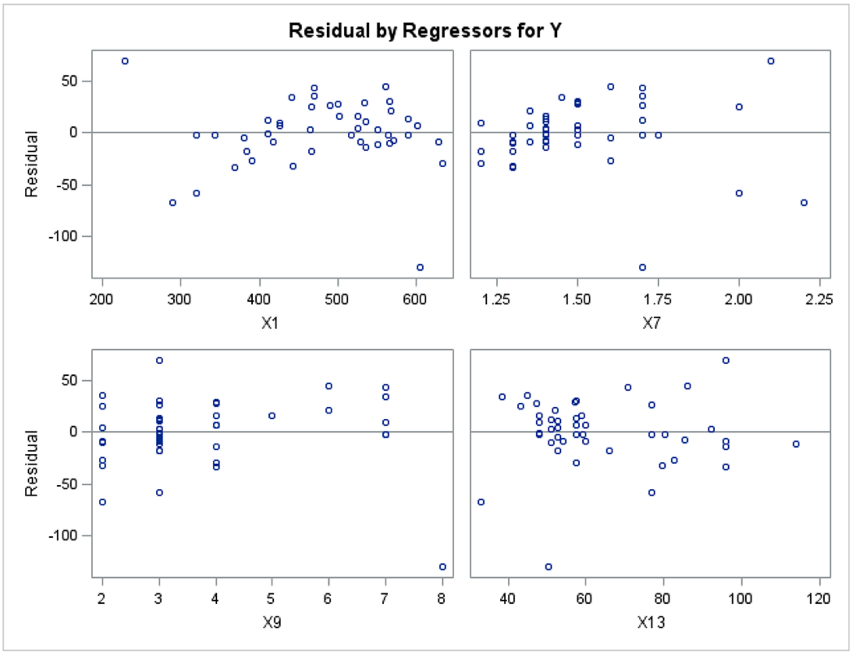

Fig. 2. Residuals vs individual predictor variables.

fitted. The value of standard error mean predicted almost very low in comparison to another selected model for Bihar. The root mean square error is also very low in comparison to other selected five models. The actual yield and predicted value are also very close to each other. This indicated that, the selected regression subset consisting of X_1, X_7, X_9 and X_{13} could be considered the best subset for prediction purpose.

On the basis of above fact first model is best model for forecasting sugarcane yield for Bihar. The analysis of variance also satisfied that F-value indicates that it is significant at 1% level. The value variance inflating factor is less than 10, so that there is no sign of multicollinearity, value in the Table 5 consisting of regression subset X_1, X_7, X_9 and X_{13}, i.e., number of millable canes, number of irrigation and Potash (K_2O), respectively are contributing characters of sugarcane yield.

Residual analysis

Residual analysis in regression is a critical step in assessing the goodness of fit of a regression model. Residuals are the differences between the observed values and the values predicted by the regression model. Analyzing these residuals helps to identify any patterns or deviations from the assumptions of the regression model. Create scatterplots of residuals against each individual predictor variable. These plots help identify relationships between residuals and specific predictors. Patterns in these plots may suggest non-linear relationships or the presence of outliers associated with particular predictor values. In this study residual plots also confirms that there is no pattern left in the residuals (Fig. 2).

Validity test for proposed forecast model

The 5 set of observations correspond to the variables included in the model has been given. These observations have not been used in model building. For each observation set, the estimated deviation and per-cent error of forecast has been analyzed and it was observed that per cent error of forecast is below 8.18%. This indicated that the model could be used to forecast sugarcane yield in Bihar with good accuracy. Model validation on kept five observations were done and results are described which also confirm the model (Table 7). Since the first model and the rest of the models does not show significant differences in the MSE, MAPE and MAE we conclude that first model is sufficient and valid for forecasting sugarcane yield in Bihar.

CONCLUSION

Residual analysis and residual plot did not indicate any model violation for first model (as discussed in result and discussion). The forecast error computed on the basis of 10% of observations (not included in model building) were found within the permissible limit i.e., 8.18%. Thus, the purpose models are expected to perform better. The R^2 of proposed model for Bihar is equal 92.29%. This indicate that these characters explained 92.29% of variation in sugarcane yield of Bihar. Also, on the basis of above fact first model is sufficient model for forecasting sugarcane yield in Bihar. The analysis of variance also satisfied that F-value and indicates that it is significant at 1% level. Thus, pre harvest forecasted yield of sugarcane has been worked out i.e. 847.82 q/ha for Bihar with the help of proposed model.

ACKNOWLEDGMENT

We would like to express my sincere gratitude to the Department of SMCA, RPCAU Pusa and Sugarcane Research Institute, Pusa who have played a significant role in the completion of this research paper. So grateful for their support and encouragement throughout this research endeavor.

REFERENCES

Aditya K, Das S (2012) Crop yield forecasting using discriminant function analysis. LAP Lambert Academic Publishing.

Agrawal R, Mehta SC (2007) Weather based forecasting of crop yields, pests and diseases-IASRI models. *Journal of Indian Society of Agricultural Statistics* 61 (2) : 255—263.

Agriculture statistics at a glance (2022) Ministry of agriculture and farmers welfare, Government of India.

Annu A, Sisodia BVS, Kumar S (2015) Pre-harvest forecast models for wheat yield based on biometrical characters.

Economic Affairs 60 (1): In press.

Annu A, Sisodia BVS, Rai VN (2016) An application of principal component analysis for pre-harvest forecast model for rice crop based on biometrical characters. *Journal of Applied and Natural Science* 8 (3) : 1164—1167. https://doi.org/10.31018/jans.v8i3.935.

Banakara KB, Garde YA, Pisal RR, Bhatt BK (2018) Pre-harvest forecasting of rice yield for effective crop planning decision in Surat District of South Gujarat. *International Journal of Current Microbiology and Applied Science* 7 (06) : 3410—3422. http://dx.doi.org/10.20546/ijcmas.2018.706.400.

Brock WA, Scheinkman JA, Dechert WD, LeBaron B (1996) A test for independence based on the correlation dimension. *Econometric Reviews* 15 (3) : 197—235. https:// doi.org/10.1080/07474939608800353.

Ghosh K, Balasubramanian R, Bandopadhyay S, Chattopadhyay N, Singh KK, Rathore LS (2014) Development of crop yield forecast models under FASAL-a case study of *kharif* rice in West Bengal. *Journal of Agrometeorology* 16 (1) : 1—8. https://doi.org/10.54386/jam.v16i1.1496.

Irshad MM, Kumar M, Ray M, Sattar A, Paswan S, Minnattulla M (2023) Effect of meteorological elements on sugarcane wilt in Bihar. *International Journal of Statistics and Applied Mathematics* SP-8 (4) : 128—133.

Sisodia BVS, Yadav RR, Kumar S, Sharma MK (2014) Forecasting of pre-harvest crop yield using discriminant function analysis of meteorological parameters. *Journal of Agrometeorology* 16 (1) : 121—125. https://doi.org/10.54386/jam.v16i1.1496.