International Journal of
Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; SP-8(4): 123-127
© 2023 Stats & Maths
https://www.mathsjournal.com
Received: 13-04-2023
Accepted: 16-05-2023

Regression weather-yield models for western-zonal cotton yield prediction of Haryana

Aditi, Pushpa, Chetna and Darvinder Kumar

Aditi
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Pushpa
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Chetna
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Darvinder Kumar
Associate Professor, Department of Statistics, PGDAV College, University of Delhi, Delhi, India

Abstract

The preharvest forecasts are useful for farmers to decide in advance their prospects and course of action. This study focuses on the multiple linear regression method for evaluating yield prediction of cotton in the districts of Bhiwani, Hisar, Fatehabad and Sirsa of Haryana to develop pre-harvest models for cotton yield. The data of cotton yield and weather parameters *viz.*, minimum and maximum temperature, sunshine hours, relative humidity and rainfall of 39 years from the period 1980-81 to 2019-20 was collected from HARSAC (Haryana Space Application Center) and College of Agriculture, HAU, Hisar. The time period from 2014-15 to 2019-20 was excluded from the construction of models used for model validation. In all the phases for the districts during evaluation, zonal yield models including CCT and meteorological factors consistently gave good results for cotton yield prediction and outperformed the other models with reduced error metrics.

Keywords: Multiple linear regression, weather variable, crop condition term (CCT)

Introduction

Highlights

Because cotton is the most significant agricultural crop farmed in Haryana, the current research emphasised weather variables influencing cotton crop yield in the Hisar, Bhiwani, Sirsa and Fatehabad districts. Understanding and measuring the relationships between cotton yield and climate factors have been demonstrated in this study. The trend yield/CCT/dummy variables were important parameters along with fortnightly weather variables used as regressors and cotton yield as a dependent variable. The overall efficiency of fitted model(s) has been investigated using several statistics such as adj R^2, per cent deviation of predicted yield from the measured yield and root mean square error (RMSE).

India has become one of the world's largest cotton producers, accounting for approximately 22% of global cotton production. India has the largest area undergoing cotton farming, accounting for around 37% of the global area under cotton cultivation, ranging from 12.0 million hectares to 13.5 million hectares. Cotton growing regions in India can be categorised into three major zones and nearly 99 percent of cotton production in the country comes from the first three zones with the Eastern zone contributing merely one percent. The northern zone (Panjab, Haryana, Rajasthan) typically produces 35 percent of the total cotton produced in India, the central zone (Gujarat, Madhya Pradesh and Maharashtra) 40 percent, and the zone south (Tamil Nadu, Karnataka and Andhra Pradesh) 24 percent. Accounting for approximately 11.9 per cent of production and 6.77 percent of hectarage, Haryana is India's fifth-largest cotton supplier. Haryana's Hisar, Sirsa, and Fatehabad districts contribute over 80 percent of total production. Rankja *et al.* (2010) [2] evaluated the quantitative relationship between weather variables and cotton yield. The data of 32 years (1975-76 to 2006-07) of the Banaskanath district were collected and to develop a preharvest forecast model with the help of multiple linear models. Kalubrame and Saroha (2016) [3] evaluated district-level agro-meteorology yield models in five major growing cotton-growing districts in Panjab state. The crop condition term was incorporated into yield models and used subdivision-level weather data like minimum and maximum temperature and rainfall from the period 1980-2003. Devi *et al.* (2019) [1] presented the variation pattern of cotton production and used principle component analysis to develop the cotton yield prediction model for Hisar, Haryana.

Corresponding Author:
Aditi
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

The purpose of this research evolved from a perceived need for appropriate usage of advanced estimations or short-term predictions of area and production of important food grain crops, which must be acquired scientifically by developing reliable, resilient, and enhanced statistical forecasting methods. In the present study, seeing the importance of the cotton crop an attempt was made to develop a pre-harvest yield forecasting model by using multiple linear regression methods and incorporating crop condition-based dummy regressor (s) and testing the model's validation.

Materials and Methods

In this study, secondary data was used for the period 1980-81 to 2019-20. The districts Hisar, Bhiwani, Sirsa and Fatehabad pertaining to western agro-climatic zone of the state, have been considered for the model building. The data of the period 1980-81 to 2019-20 of cotton crops collected from Statistical Abstract of Haryana and College of Agriculture, CCS HAU, Hisar have been used to determine the trend yield. The meteorological factors such as minimum and maximum temperature, sunshine hours, relative humidity and rainfall for the same time period were collected from the Department of Agricultural Meteorology, CCS HAU, Hisar. The model was built using weather data spanning 10 fortnights, from the second fortnight of May to the first fortnight of October (crop growth period: May to October). The time-series yield/weather data from 1980-81 to 2013-14 were utilised for training, and the remaining data *i.e.* 2014-15 to 2019-20 was used to validate the constructed zonal yield model.

Modelling procedure

Linear time trend

Years vs. cotton yield scatter plots were created to analyze the fluctuations and trends in cotton yields over time. The crop yield was used to fit the linear time-trend-based model(s)

$T_r = a + bt,$

Where

T_r = Yield (kg/ha),

a = Intercept,

b = Slope, and

t = Year for all districts

Multiple linear regression

The multiple linear regression model was used to relate crop yield to the average maximum temperature, average minimum temperature, average relative humidity, average sunshine hours and accumulated rainfall obtained of the period May to October. The standard linear regression model considered may be written as:

$$Y=a_0 + \sum_{i=1}^{12} b_i TMX_i + \sum_{j=1}^{12} b_j TMN_j + \sum_{k=1}^{12} b_k ARF_k + \sum_{l=1}^{12} b_l RH_l + \sum_{m=1}^{12} b_m ASSH_m$$

$$+ cCCT/dummy + e$$

Where

Y = yield (kg/ha)

a_0 = Overall mean effect

b_i, b_j, b_k, b_l, b_m= Regression coefficients of the weather variables

c = Regression coefficient of dummy variable

TMX_i = ith day maximum temperature

TMN_j = jth day minimum temperature

ARF_K = kth day rainfall

RH_l = lth day relative humidity

SSHm = mth day sunshine hours

i, j, k, l, m = Meteorological fortnights (1,2,3…12)

e = The error term with assumption NID $(0, \sigma^2)$

The climate variables were chosen using a stepwise regression procedure (Draper and Smith, 1981) in which all variables were first included in the model then excluded one at a time, with decisions at any each stage conditioned by the outcome of the preceding phase. Predictions T_r based on this model produced a predictor variable known as 'trend yield'.

CCT as dummy variable in analysis

The crop condition variable(CCT), an indicator variable derived by partitioning the department of agricultural cotton crop yield sequence in eight non-overlapping classes indicated by numbers 1, 2, 3, 4, 5, 6, 7 & 8 and corresponding to observed yields of 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, 700-800, and 800-900 in kg/ha, respectively.

Regression Diagnostics

Examining residuals is a critical component of all statistical modelling. Inference for the general linear model is based on various assumptions, including error independence, normality, and variance homogeneity. Any graph suitable for illustrating the distribution of a collection of data can be used to assess the normality of a group of residuals' distribution. Histograms, normal probability plots, and dot plots are the three most popular forms.

Model comparison and post-sample validation

Zonal weather-yield models have been used to predict cotton yield in the districts under consideration. The model predicted yield(s) have been compared with the corresponding DOA yield for the post-sample period(s) 2014-15 to 2019-20. Finally, a comparison between crop yield forecasts has been made on the basis of MAPE, RMSE and percent relative deviation from observed data.

Results and Discussion

The trend analysis of cotton yield data have been performed to assess the fluctuations and trends in Hisar, Bhiwani, Sirsa and Fatehabad districts for cotton yields over the years 1980-2014 shown in Table 1. Cotton yields were found to be highly fluctuating across the study period, as seen by the comparatively low R^2 values for all districts, depending on climatic conditions and pest/disease incidence. The DOA yield data of cotton for Hisar, Sirsa, Bhiwani and Fatehabad districts were used by considering time (year) as an independent variable and had been regressed against yield to get the trend equation.

Table 1: Computation of Trends in cotton yield of all Districts

Districts	R^2	Intercept(a)	Slop(b)
Hisar	0.33	15.04	7.75
Bhiwani	0.39	13.82	7.09
Sirsa	0.53	23.69	12.10
Fatehabad	0.55	25.53	13.04

Further, the linear trend-based yield(s) along with fortnightly weather data were used for the development of regression models at zonal level. The pre-harvest cotton yield prediction models have been fitted on the basis of crop condition-based dummy regressors and weather data of the period 1980-81 to 2013-14. Six steps ahead *i.e.*, 2014, 2015, 2016, 2017, 2018 and 2019 district-level cotton yield estimation have been achieved and the predictive performance of the developed zonal models along with and without CCT has been comparatively evaluated. Crop forecast models were obtained by step-wise regression method using the statistical software ‘SPSS’. The selected models along with coefficient(s) of determination and standard error(s) are presented in Table 2.

Table 2: Zonal weather-yield models for all four districts of Haryana

Types	Fitted Models	Adj.R^2	SE
Model -1	$Y = -1044.02 + 1.23Tr - 0.73ARF_3 + 30.21TMN_3 + 7.63RH_7 -0.63RH_8 + 18.4SSH_6 - 11.52TMX_1$	0.63	88.3
Model -2	$Y = -1168.08 + 1.27Tr - 0.70ARF_3 + 30.42TMN_3 + 8.84RH_7 -0.84RH_8 + 19.4SSH_6 - 10.47TMX_1 - 0.45ARF_5$	0.64	86.8
Model -3	$Y = 192.66 + 93.2CCT - 3.22TMX_8 - 0.24ARF_3$	0.96	26.6
Model -4	$Y = 223.68 + 92.39CCT - 4.30TMX_8 + 0.13ARF_4$	0.96	26.5
Model-5	$Y = 900 - 77.72D_7 - 168.08D_6 - 552.77D_2 - 661.29D_1- 456.25D_3 - 379.49D_4-291.44D_5-4.50TMN_{10}$	0.96	25.1
Model-6	$Y = 870.55 - 82.715D_7 - 170.80D_6 - 557.19D_2 - 673.19D_1 -459.75D_3 - 381.84D_4 - 293.45D_5-0.82RH_{10}$	0.96	25.8

Zone comprised of Hisar, Bhiwani, Sirsa and Fatehabad districts. All the regressors are significant at $p \leq 0.05$ in above zonal yield models.

Model 1, 2 consisted of Weather parameters and trend yield, Model 3, 4 consisted of Weather parameters and CCT, Model 5, 6 consisted of Weather parameters and dummy

The prediction performance(s) of the zonal yield models were compared using Adj. R^2 and percent deviations of crop yield estimations from the real-time/observed yield(s) and root mean square errors (RMSEs). Due to Cotton Leaf Curl Disease (Yadav *et al.*, 2016) [5], yield estimates for 2015-16 in Hisar, Bhiwani, Sirsa, and Fatehabad districts were considered for ARIMA/ARIMAX did not produce satisfactory results. Percent relative deviations from the observed data were not satisfactory for models 1 & 2 and further efforts were put to see the performance of models 3 to 6 by incorporating CCT and CCT-based dummy regressors along with weather parameters. Table 3 provides a comparative view in terms of per cent deviations of cotton yield estimates from real-time yield(s) based on all four models. Table 4 displays the average absolute percent deviations and RMSEs of the selected zonal yield models.

Table 2: Percent deviations of predicted yield (kg/ha) from the real-time/observed yield using CCT/dummy regressors-based model.

Forecasting Years	District	Relative Deviation (%)			
		Model-3	Model-4	Model-5	Model-6
2014-15	Hisar	4.0	2.1	4.0	1.3
	Bhiwani	-1.2	-2.3	2.4	-0.03
	Sirsa	-2.2	-2.9	-3.6	-5.3
	Fatehabad	3.1	2.4	1.8	0.2
2015-16	Hisar	5.8	2.4	5.6	2.8
	Bhiwani	-5.7	-8.3	-6.8	-9.4
	Sirsa	11.8	8.7	11.7	9.1
	Fatehabad	-15.6	-10.1	-15.8	-19.5
2016-17	Hisar	0.3	1.1	-1.9	-3.6
	Bhiwani	-1.9	-3.8	1.2	-0.9
	Sirsa	10.5	9.2	8.5	6.9
	Fatehabad	9.5	8.1	7.4	5.9
2017-18	Hisar	8.2	1.5	11	9.6
	Bhiwani	6.7	2.1	5.0	4.1
	Sirsa	-10.9	-9.4	-9.6	-11.4
	Fatehabad	1.1	1.0	4.1	2.7
2018-19	Hisar	5.3	2.9	2.8	3.4
	Bhiwani	-3.3	2.4	-0.8	-0.4
	Sirsa	-12.8	-9.7	-7.3	7.1
	Fatehabad	11.5	8.5	5.8	5.4
2019-20	Hisar	-4.4	6.8	3.4	0.9
	Bhiwani	6.4	5.3	5.6	7.9
	Sirsa	8.0	9.1	8.7	8.9
	Fatehabad	10.1	2.1	3.4	-2.8

Table 3: Mean absolute percent error (MAPE) and root mean square error (RMSE) of cotton yield estimates based on finally selected models.

Districts	MAPE		RMSE	
	Model-4	Model-5	Model-4	Model-5
Hisar	2.8	4.8	14.4	25.3
Bhiwani	4.1	4.5	18.5	24.1
Sirsa	8.5	9.4	47.9	49.9
Fatehabad	8.1	8.6	27.4	32.4

The regression models with a sufficiently good fit, as indicated by the coefficient of determination R^2, were unable to achieve adequate predicted accuracy. The cotton yield calculated using weather-yield zonal models exhibited larger percent deviations from the real-time yield(s), *i.e.* too high for reliable yield prediction in all the districts under analysis. Based on the findings, incorporating CCT as a categorical covariate in addition to meteorological parameters is proposed to improve the prediction accuracy of zonal weather–yield models. The results demonstrated that the district-level yield(s) projection correlates well with the DOA yield predictions in Haryana's western agro-climatic zone. Model 4 was finally chosen for pre-harvest cotton yield forecast in the districts under consideration, based on the pattern of different statistic(s) and computational ease. The model's goodness-of-fit was evaluated using residual diagnostics, specifically histograms and normal-probability plots of residuals, as well as a plot of residuals against fitted values (Fig. 1). The per cent deviations within reasonable limits favour the adoption of developed models for cotton yield prediction in Haryana's western zone. The average absolute percent deviations of post-sample period estimates ranging from 4 to 9 percent support the utility of established models for cotton yield prediction in Haryana's western zone. The selected model could be utilised for pre-harvest 4 weeks before harvesting of cotton crop. Figures 2–5 show the predicted yield(s) as well as the observed yield(s) of all districts obtained using the selected model.

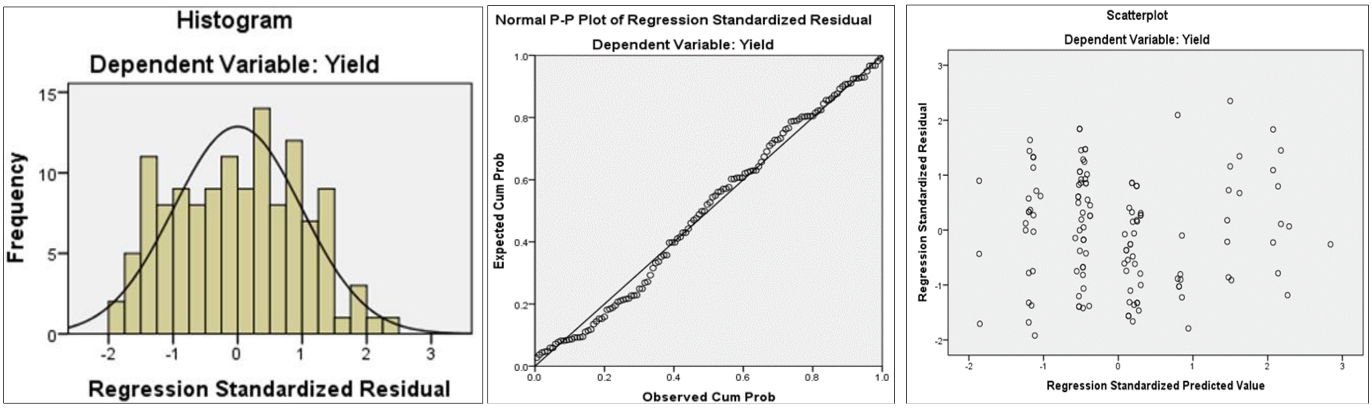

Fig 1: Regression diagnostics of the selected zonal yield forecast model. (CCT+ weather variables)

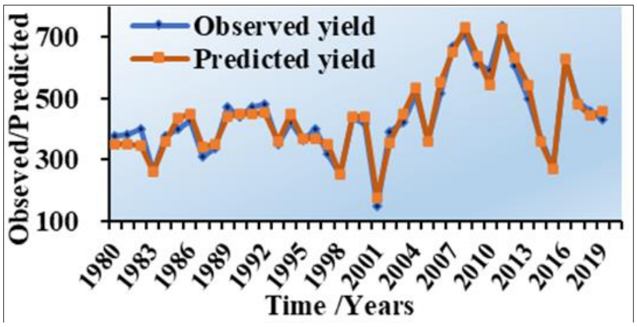

Fig 2: Observed and predicted values of cotton yield in Hisar district

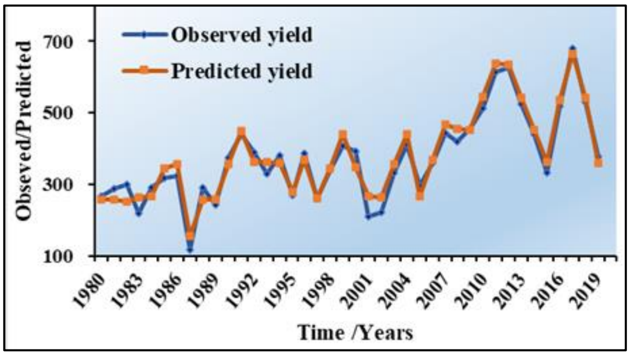

Fig 3: Observed and predicted values of cotton yield in Bhiwani district

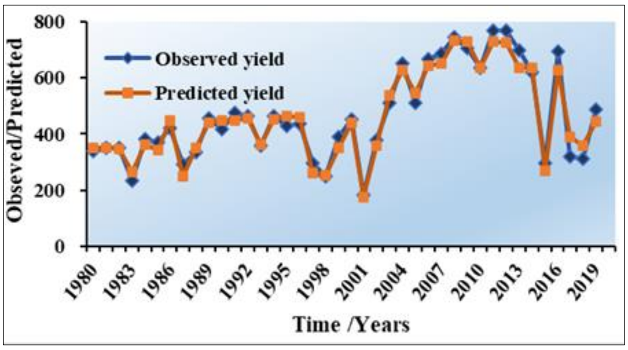

Fig 4: Observed and predicted values of cotton yield in Sirsa district

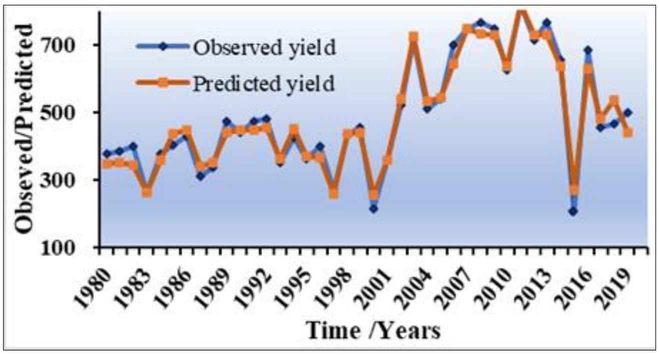

Fig 5: Observed and predicted values of cotton yield in Fatehabad district

Conclusion

The study was conducted in four major cotton-growing districts Hisar, Bhiwani, Sirsa and Fatehabad which account 80% production of Haryana state. The weather yield modelling approach was used in this study to create district-level cotton yield models utilizing agro-meteorological data from the previous 39 years. The weather variable was found statistically significant as predictors where R^2 couldn't provide satisfactory predictive accuracy. Yields predicted by the weather-yield zonal models had larger percent deviations from real-time yield(s), which was counted as very high for effective yield prediction. The goal of cotton yield modelling, on the other hand, was to test the predicted accuracy of the established model (s). As a result, an attempt was made to increase the predictive accuracy of the proposed model by the addition of trend yield-based CCT to the weather-yield model, which greatly improved forecast model predictive accuracy. The level of accuracy obtained by zonal yield model(s) adopting CCT as a categorical covariate in associated with weather variables was deemed appropriate for estimating district-level cotton yield(s) at least 4-5 weeks before crop harvest. Multiple criteria were used to evaluate the predictive performance(s) of the zonal weather-yield models, including adjusted R^2, percent deviations of expected yields from DOA yield estimations, RMSE and MAPE. Based on the pattern of different statistic(s) and computational ease, Model 4 was finally chosen for the pre-harvest cotton yield forecast in the districts under consideration.

Acknowledgement

The weather data received from Haryana Space Applications Centre, Hisar and Department of Agrometeorology, CCS HAU, Hisar, India are gratefully acknowledged.

References

1. Devi M, Mishra P, Malik DP, Mehala V, Mehta VP, Bhardwaj N. Study of Climatic Factors Affecting the Productivity of Cotton and its Instability. Economic Affairs. 2019;64(4):761-767. http://dx.doi.org/10.30954/0424-2513.4.2019.11
2. Rankja NJ, Upadhyay SM, Pandya HR, Parmar BA, Varmora SI. Estimation of cotton yield based on weather parameters of Banaskantha district in Gujarat state. Journal of Agrometeorology. 2010;12(1):47-52. https://doi.org/10.54386/jam.v12i1.1268
3. Kalubarme Manik H, Saroha GP. Development of district-level Agro-meteorological Cotton Yield Models in Punjab. International Journal of Environmental Research and Development. 2016;6(1):17-32.
4. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis, 3rd edition, John Wiley & Sons. 2014, p. 310-316.
5. Yadav NK, Kumar D, Nain J, Beniwal J. Occurrence and severity of cotton leaf curl disease in Haryana. International Journal of Agriculture Sciences. 2016;8(52):2450-2452.