Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 10 [7] June 2021 : 33-40

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD

ORIGINAL ARTICLE **OPEN ACCESS**

Development of Weather based Models and Simulation for pre-harvest Mustard Yield Forecasting in Haryana

Ajay Kumar* and Urmil Verma
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana
*Corresponding Author E-mail: ajaystatistics@gmail.com

ABSTRACT
The study aims at developing weather-yield models following multiple linear regression for mustard yield forecasting in Bhiwani, Fatehabad, Hisar, Sirsa, Gurugram, Jhajjar and Mahendragarh districts of Haryana. The fitted models are based on weather parameters viz., maximum temperature, minimum temperature and rainfall along with crop condition term (CCT) as categorical/dummy regressor (s). The post-sample validity assessment of the developed model has been verified by determining prediction errors using root mean square errors and average absolute percent relative deviations. Incorporating CCT as categorical covariate along with weather parameters enhanced the predictive accuracy of weather-yield models inspite of showing lower adj. R^2 and higher standard error of estimate. Student's t copula procedure in SAS is used to simulate the mustard yield achieved from weather+CCT based regression model. The forecasts obtained from regression based weather+CCT model being remarkably close to the forecasts obtained through the simulation process indicate the preference of using developed models for pre-harvest mustard yield forecasting in Haryana.
Keywords*: Root mean square errors (RMSEs), Crop Condition Term (CCT), Simulation, Dummy variable and Mustard yield forecasts*

Received 04.05.2021 Revised 12.05.2021 Accepted 11.06.2021

INTRODUCTION

Regression analysis is the most frequently used statistical technique for investigating and modeling the relationship between variables. Building a regression model is an iterative process. Usually several analyses are required as improvement in the model structure and flaws in the data are discovered [1]. Today, agriculture has become much more cost-intensive and highly input. In the changed scenario today, it is significantly important to predict certain aspects relating to agriculture. The present situation is far from satisfactory despite the clear need for accurate and timely forecasts. However, a few months after actual harvesting of the crop the final forecasts are issued. Thus, the timeliness and accuracy of data are some of the drawbacks of conventional methods. So, there is always a considerable scope of improvement in the conventional system.

Rapeseed is the world's third-largest vegetable oil source and the second-largest protein meal source. Indian mustard (*Brassica juncea* (L.) Czern & Coss.) is predominantly cultivated in Gujarat, Uttar Pradesh, Haryana, Rajasthan and Madhya Pradesh. The crop requires temperature between 10°C to 25°C and is grown in the area receiving 25 to 40 cm of rainfall. It is primarily a winter crop during the *rabi* season in Haryana and grown during September-October to February-March.

Multiple regression analysis plays an important role in forecasting a variable's unknown value from the known value of two or more variables. It is commonly used to forecast crop production and to analyze the effect of weather variables on crop yield. Kumar and Bhar [4] used multiple linear regression to forecast Indian mustard production in Hisar district of Haryana. The earliest and latest forecast was done 4-5 weeks before harvesting. Garde *et al.* [2] derived multiple linear regression equations for estimating wheat productivity using weather parameters for Ghazipur district in eastern Uttar Pradesh. They observed that the forecasting model produced the most accurate forecast in 15th week of the crop growing season and the relationship between actual and forecast wheat yield was highly significant bearing R^2 from 0.72 to 0.89.

Verma *et al.* [7, 8] and Goyal and Verma [3] have used agromet/spectral indices in the context of pre-harvest yield forecasting of cotton, sugarcane, mustard and wheat crops in Haryana. Ravita and Verma [6] applied multivariate statistical technique to achieve district-level mustard yield estimation in Haryana. The weather-yield models having crop condition term as dummy regressor had the desired forecast accuracy by

showing 5-10 percent mean deviations in most of the mustard-growing districts in the State. Niedbała *et al*. [5] developed a model for prediction and simulation of winter rapeseed yield using the multiple regression method (MLR) based on meteorological data (air temperature and precipitation) and information about mineral fertilization (2008-2017) based on determining prediction errors using RAE, RMS, MAE, and MAPE error.

MATERIAL AND METHODS

The study aims at developing weather-yield models based on multiple linear regression for mustard yield forecasting in Bhiwani, Fatehabad, Hisar, Sirsa, Gurugram, Jhajjar and Mahendragarh districts of Haryana. The State Department of Agriculture and Farmers Welfare (SDOA) mustard yield data compiled for the period 1980-81 to 2015-16 of Bhiwani, Sirsa, Hisar, Mahendragarh and Gurugram and 1997-98 to 2015-16 of Jhajjar and Fatehabad districts were utilized for the purpose. The mustard yield data from 1980-81 to 2012-13 along with weather data (collected from Indian Meteorology Department (IMD), Delhi and different meteorological stations in Haryana) of the same period were used for the training set. The weather-yield data of post-sample period, *i.e.*, 2013-14 to 2015-16 have been used for validity testing of the developed mustard yield forecast models. The fortnightly weather data (rainfall and temperature) were prepared from daily data as shown below:

Average Maximum Temperature (Tmx) $=\frac{\sum_{i=1}^{15} \mathrm{Tmx}_i}{15}$

Average Minimum Temperature (Tmn) $=\frac{\sum_{j=1}^{15} \mathrm{Tmn}_j}{15}$

Accumulated Rainfall (Arf) $=\sum_{k=1}^{15} \mathrm{Arf}_k$

where Tmx_i= i^{th} day maximum temperature

Tmn_j= j^{th} day minimum temperature

Arf_k= k^{th} day rainfall

i,j,k = 1,2,.....,15 (daily weather data)

Weather-yield model building

The linear time-trend forecast model(s) fitted for all the districts may be expressed as $T_r = a+bt$, where T_r = Yield (q/ha), a = Intercept, b = Slope and t= Year. Predictions T_r based on this model yielded a predictor variable and that has been denoted as 'trend yield'.

The purpose of standard regression model is to explore an association between dependent and independent variables, to identify the impact of these covariates on the response that further helps in predicting the future values of the dependent variable. Weather variability both within and between seasons is major uncontrollable source of variability in yield. Weather variables affect the crop differentially during different stages of development. This increases the number of variables in the model and in turn, a large number of parameters are to be evaluated from the time series data for precise estimation. Thus, a technique based on relatively smaller number of manageable parameters and at the same time, taking care of entire weather distribution is always preferred to solve the problem. For quantitative forecasting, regression models have been fitted by taking weather variables and trend yield as regressors and crop yield as regressand by following stepwise regression method.

The multiple linear regression model considered may be expressed as follows:

$$b_j Tmn_j + \sum_{k=1}^{12} b_k Arf_k + e$$

$$b_i Tmx_i + \sum_{j=1}^{10}$$

$$Y = a + cT_r + \sum_{i=1}^{10}$$

where,

Y - Mustard yield (q/ha)

T_r - Trend yield (q/ha)

a- Overall mean effect

c - Regression coefficient of trend yield

b_i, b_j, b_k - Regression coefficients of weather variables

(i, j, k - weather fortnights, *i.e.*1,2,3…10/12 over crop growth period)

e - Error term with assumption NID (0, σ^2)

The weather-yield models have been fitted to relate crop yield to average maximum temperature, average minimum temperature calculated for 10 fortnights covering the crop growth period i.e. 1st fortnight of October to 2nd fortnight of February, and accumulated rainfall obtained for 12 fortnights over the period 1st fortnight of September to 2nd fortnight of February.

The best subsets of weather variables are selected using the stepwise regression method [1] in which all the variables were first included in the model and eliminated one at a time with decisions at any particular step conditioned by the results of previous step. The best supported weather variables in the model are retained if they had the highest adjusted R^2 and lowest standard error of estimate at a given stage. Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of estimated parameters. Commonly used checks of goodness of fit include R^2, analysis of the pattern of residuals and hypothesis testing. Statistical significance is checked by an F-test of the overall fit, followed by t-test of individual parameters. The weather-yield models based on regression analysis were compared on the basis of per cent relative deviations and root mean square errors to obtain pre-harvest mustard yield forecasts.

To further enhance the predictive performance, the weather-yield models were again fitted by taking crop condition term (CCT) as categorical/dummy variable along with weather variables as repressors and DOA mustard yield as regressand. The CCT being an indicator variable is generated by splitting the trend yield data into different non-overlapping classes.

Simulation

Simulation is a method of solving decision making problems by designing, constructing and manipulating a model of real system. Simulation duplicates the essence of a system or activity without actually obtaining reality. Most simulations are random number driven. For each application of random numbers in a simulation, a distribution must be chosen. The distribution determines the likelihood of different values occurring. A distribution is uniquely specified by the name of its family (such as uniform, exponential, or normal etc.) and its parameter values (such as the mean and standard deviation).

The statistical simulation technique like Normal Copula/T-copula provide approximate solutions to the problems expressed mathematically. It utilizes a sequence of random numbers to perform the simulation. Proc Copula (SAS) is the most powerful tool for analyzing complex problems and uses random numbers to solve problems which involves conditions of uncertainty. It gives a solution which falls very close to the optimal but not necessarily the exact solution. As the number of simulated trials leads to infinity, the solution converges to the optimal solution.

Let $\Theta = \{(v, \Sigma): v \in (1, \infty), \Sigma \in R^{mxm}\}$ and let t_v be the univariate t distribution with v degrees of freedom. The Student's t copula can be written as

$$C_\Theta(u_1, u_2, \ldots\ldots, u_m) = t_{v,\Sigma}\left(t_v^{-1}(u_1), t_v^{-1}(u_2), \ldots\ldots, t_v^{-1}(u_m)\right)$$

where $t_{v,\Sigma}$ is the multivariate Student's t distribution with a correlation matrix Σ with degrees of freedom. The input parameters for the simulation are (v, Σ). The t copula can be simulated by the following two steps:

- Generating a multivariate vector$Xt_m(v, 0, \Sigma)$ following the centered t distribution with v degrees of freedom and correlation matrix Σ .
- Transforming the vector X into $U = \left(t_v(X_1), t_v(X_2), \ldots\ldots t_v(X_m)\right)^T$, where t_v is the distribution function of univariate t distribution with v degrees of freedom.

RESULTS AND DISCUSSION

Time-trend analysis often reflects an underlying pattern/activity in a time series which would otherwise partially or almost fully obscure by noise. By taking time(year) as an independent variable and regressed against yield to get the trend predicted yield. The weather-yield models have been developed by stepwise regression method using trend-based yield and weather parameters (rainfall, minimum temperature and maximum temperature) computed over different fortnights of crop growth period, as regressors. The yield forecasts based on finally selected weather-yield models shown in Table-1 had higher percent relative deviations from real-time mustard yield(s) in most of the districts, sometimes even too high than to be considered acceptable. Thus, adding linear time-trend to the model, along with the selected weather variables couldn't satisfactorily favour the forecast accuracy of district-level mustard yield in Haryana.

Further, an attempt was put to improve the predictive accuracy of weather-yield models by identifying and adding additional covariate to the model, along with the selected weather variables. In particular, CCT/dummy variable was incorporated in weather-yield models by repeating the stepwise regression analysis and that substantially improved the predictive accuracy of the models. The CCT is a categorical covariate obtained by dividing the trend predicted yield series into three non-overlapping classes

reflecting high, normal and low yield. Incorporating CCT as dummy variable along with weather parameters enhanced the predictive accuracy of weather-yield models. The fitted models along with Adj. R^2 and standard error of estimate are shown in the following table. District-specific model predicted yield(s) along with observed yield(s) and per cent relative deviations based on above models are presented in Table-2. A perusal of the results indicate the preference of using model 2 for mustard yield forecasting in the districts under consideration.

Regression diagnostics of weather-yield model incorporating CCT as categorical covariate

Residual Diagnostics are intended with testing the goodness of fit of a model and if possible, to recommend suitable modifications. Thus, residual histogram and normal-probability plots for the best fitted model was prepared for examining normality assumptions of the residuals. Histograms show approximate behaviour with slight deviation from normality. The P-P plots also infer the same. Standardized residual plots appear fine.

Student's t copula simulation based district-level mustard yield

Student's t copula procedure in SAS has been used to simulate the mustard yield achieved from model 2 (i.e. weather+CCT based models). The copula approach to formulating a multivariate distribution provided a way to isolate the description of the dependence structure from the marginal distributions. The copula function combined the marginal distributions of variables into a specific multivariate distribution. The results pertaining to simulation for the post-sample period are described in Table-3.

The forecast performance(s) of finally selected model-2 based on weather +CCT have been observed in terms of per cent relative deviations of yield(s) prediction from real time yield(s) and root mean square errors (RMSEs). The overall results indicate the preference of using Crop Condition Term (CCT) as categorical covariate along with weather variables in capturing lower percent relative deviations. It is inferred that the mustard yield(s) prediction based on simulated weather + CCT model are quite consistent to the yield(s) prediction obtained from fitted weather + CCT model in relation to real-time yield(s) for all the districts under consideration. Moreover, the developed regression based weather + CCT model is capable of providing the district-level mustard yield forecasts well in advance of the crop harvest in the state.

Table 1: Weather-yield forecast models incorporating trend yield and CCT as categorical/dummy regressor (s)

Types	Fitted Models	Adj. R^2	SE
Model 1	Y_{est} = -18.60 +1.02 T_r + 0.01 Arf_2 – 0.04 Arf_3 + 0.49 Tmx_2 + 0.31 Tmn_8 + 0.20 Tmx_9 – 0.43 Tmn_{10}	0.71	1.90
Model 2	Y_{est} = -16.72 + 2.70 CCT + 0.01 Arf_1 + 0.02 Arf_2 + 0.48 Tmx_3–0.41 Tmn_4 + 0.59 Tmx_2	0.62	2.18
Model 3	Y_{est} = -9.54–5.71 D_1 - 2.62 D_2 + 0.02 Arf_2 + 0.54 Tmx_3 + 0.41 Tmx_5–0.29 Tmn_4 + 0.01 Arf_1	0.61	2.20

Where,

Y_{est}- Model predicted yield (q/ha)

T_r- Trend yield (q/ha)

Tmx - Maximum temperature

Tmn - Minimum temperature

Arf- Accumulated rainfall

CCT- Crop condition term

D - Dummy variable

R^2- Coefficient of Determination

SE-Standard error of estimate

Model 1 - Weather parameters & trend yield as regressors

Model 2 - Weather parameters & CCT as regressors

Model 3 - Weather parameters & dummy variables as regressors

Table 2: Comparative view of post-sample district-specific mustard yield(s) prediction based on fitted models

District/ Forecast Years Bhiwani	Observed Yield (q/ha)	Model 1		Model 2		Model 3	
		Fitted Yield (q/ha)	RD (%)	Fitted Yield (q/ha)	RD (%)	Fitted Yield (q/ha)	RD (%)
2013-14	15.16	18.26	-20.45	15.73	-3.76	15.35	-1.25
2014-15	13.98	13.24	5.29	15.10	-8.01	15.89	-13.66
2015-16	14.61	17.26	-18.14	14.96	-2.40	14.12	3.35
Av. Abs. percent dev.			14.63		4.72		6.09
Fatehabad							
2013-14	18.53	20.27	-9.39	15.72	15.11	15.35	17.16
2014-15	15.37	15.30	0.46	15.09	1.76	15.89	-3.38
2015-16	13.55	19.37	-42.95	14.96	-10.41	14.12	-4.21
Av. Abs. percent dev.			17.60		9.09		8.25
Hisar							
2013-14	16.26	18.97	-16.67	15.73	3.26	15.35	5.60
2014-15	14.17	13.96	1.48	15.10	-6.56	15.89	-12.14
2015-16	18.16	17.98	0.99	14.96	17.62	14.12	22.25
Av. Abs. percent dev.			6.38		9.15		13.33
Sirsa							
2013-14	17.37	19.31	-11.17	15.73	9.44	15.35	11.63
2014-15	15.00	14.33	4.47	15.10	-0.67	15.89	-5.93
2015-16	17.09	18.39	-7.61	14.96	12.46	14.12	17.38
Av. Abs. percent dev.			7.75		7.52		11.65
Gurugram							
2013-14	15.94	14.73	7.59	16.28	-2.13	16.40	-2.89
2014-15	12.69	15.55	-22.54	16.28	-28.29*	17.07	-34.52
2015-16	16.52	17.80	-7.75	16.32	1.21	16.36	0.97
Av. Abs. percent dev.			12.63		10.54		12.79
Jhajjar							
2013-14	15.49	14.50	6.39	16.28	-5.10	16.39	-5.87
2014-15	13.66	15.32	-12.15	16.28	-19.18*	17.07	-24.96
2015-16	15.80	17.57	-11.20	16.32	-3.29	16.36	-3.54
Av. Abs. percent dev.			9.92		9.19		11.46
Mahendragarh							
2013-14	16.99	15.57	8.36	16.28	4.18	16.40	3.47
2014-15	14.99	16.42	-9.54	16.28	-8.61	17.07	-13.88
2015-16	15.73	18.71	-18.94	16.32	-3.75	16.36	-4.01
Av. Abs. percent dev.			12.28		5.51		7.12

***** As per IMD 2015, forty percent of the state's net sown area (2.24 million ha) was affected by unseasonal rainfall and hailstorm in March (India, Ministry of Agriculture, 2015c)

Percent Relative Deviation= 100 × (observed yield – estimated yield) / observed yield); measures the deviation (in percentage) of forecast yield from the actual yield

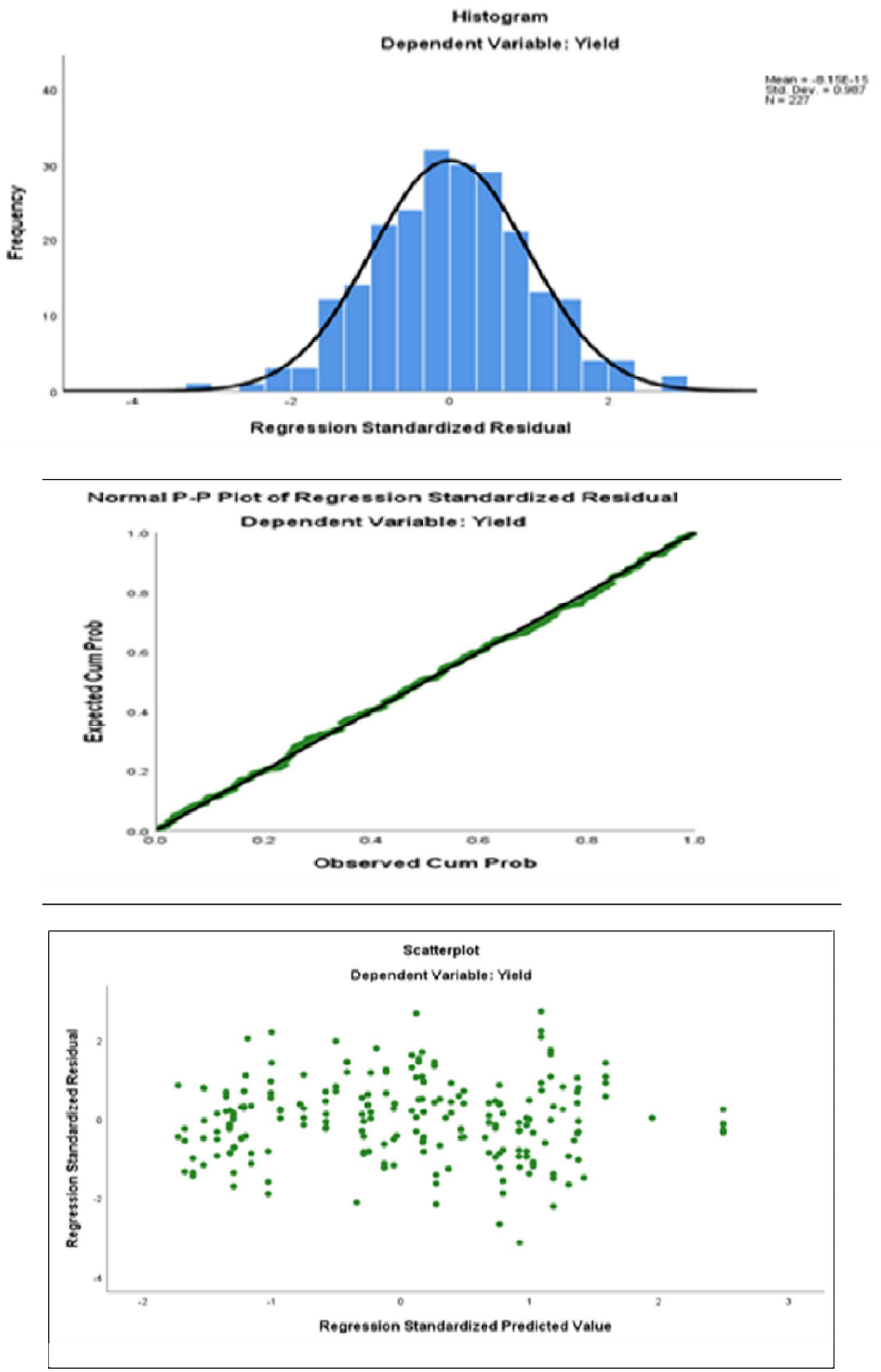

Figure 1: Regression diagnostics of the fitted model (CCT + weather variables)

Table 3: Post-sample mustard yield(s) along with simulated yield(s) and percent relative deviations based on regression based weather+CCT model for all the districts

District(s)	Forecast Years	Observed Yield (q/ha)	Simulated Yield (q/ha)	Percent Relative Deviation
Bhiwani	**2013-14**	15.16	15.84	-4.49
	2014-15	13.98	15.84	-13.30
	2015-16	14.61	15.80	-8.15
Av. Abs. percent dev.		**8.65**		
Fatehabad	**2013-14**	18.53	15.99	13.69
	2014-15	15.37	15.95	-3.78
	2015-16	13.55	15.92	-17.53
Av. Abs. percent dev.		**11.66**		
Hisar	**2013-14**	16.26	15.84	2.58
	2014-15	14.17	15.84	-11.79
	2015-16	18.16	15.80	13.01
Av. Abs. percent dev.		**9.12**		
Sirsa	**2013-14**	17.37	15.84	8.81
	2014-15	15.00	15.84	-5.60
	2015-16	17.09	15.80	7.56
Av. Abs. percent dev.		**7.32**		
Gurugram	**2013-14**	15.94	16.11	-1.09
	2014-15	12.69	16.06	-26.57
	2015-16	16.52	16.08	2.65
Av. Abs. percent dev.		**10.10**		
Jhajjar	**2013-14**	15.49	16.29	-5.16
	2014-15	13.66	16.30	-19.33
	2015-16	15.80	16.30	-3.17
Av. Abs. percent dev.		**9.22**		
Mahendragarh	**2013-14**	16.99	16.11	5.15
	2014-15	14.99	16.06	-7.15
	2015-16	15.73	16.08	-2.24
Av. Abs. percent dev.		**4.85**		

Table 4: Comparative view in terms of average absolute percent relative deviations and root mean square error of mustard yield forecasts with real time yield(s) for all the districts

District(s)	Absolute percent relative deviations		RMSEs	
	Regression model with weather & CCT	Simulated model with weather & CCT	Regression model with weather & CCT	Simulated model with weather & CCT
Bhiwani	4.72	8.65	0.75	1.33
Fatehabad	9.09	11.66	1.82	2.03
Hisar	9.15	9.12	1.95	1.69
Sirsa	7.52	7.32	1.55	1.25
Gurugram	10.54	10.10	2.09	1.96
Jhajjar	9.19	9.22	1.61	1.62
Mahendragarh	5.51	4.85	0.92	0.82

REFERENCES

1. Draper NR, Smith H (2003) Applied Regression Analysis, 3rd Edition. John Wiley & Sons, New York
2. Garde YA, Singh S, Mishra GC, Singh T (2012) Weather based pre-harvest forecasting of wheat at Ghazipur (U.P.). International Journal of Agricultural Sciences **8**(2): 325-328
3. Goyal M, Verma U (2015) Spectral-weather-crop yield forecasting: Discriminant function analysis. Journal of Applied Probability and Statistics **10**(1): 1-14
4. Kumar A, Bhar L (2005) Forecasting model for yield of Indian mustard using weather parameter. Indian Journal of Agricultural Science **75**(10): 688-90
5. Niedbała G, Piekutowska M, Adamski M (2018) Multiple regression analysis model to predict and simulate winter rapeseed yield. Journal of Research and Applications in Agricultural Engineering**63**(4): 139-144
6. Ravita, Verma U (2017) Use of crop condition based dummy regressor and weather input for parameter estimation of mustard yield forecast models. Journal of Applied and Natural Science**9**(3): 1703-1709

7. Verma U, Dabas DS, Hooda RS, Kalubarme MH, Yadav M, Sharma MP (2011) Remote sensing-based wheat acreage and spectral-trend-agrometeorological yield forecasting: Factor Analysis Approach. Society of Statistics, Computers and Applications**9** (1&2): 1-13
8. Verma U, Piepho HP, Ogutu JO, Kalubarme MH, Goyal M (2014) Development of agromet models for district-level cotton yield forecasts in Haryana state. International Journal of Agricultural and Statistical Sciences**10**(1): 59-65

CITATION OF THIS ARTICLE

A Kumar and U Verma. Development of Weather based Models and Simulation for pre-harvest Mustard Yield Forecasting in Haryana**.** Bull. Env. Pharmacol. Life Sci., Vol10[7] June 2021 : 33-40