Indian Journal of Agricultural Sciences 87 (1): 87–91, January 2017/Article

Wheat yield forecast using detrended yield over a sub-humid climatic environment in five districts of Uttar Pradesh, India

BISHAL GURUNG[1], SANJEEV PANWAR[2], K N SINGH[3], RAHUL BANERJEE[4], SISIR RAJ GURUNG[5] and ABHISHEK RATHORE[6]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 17 September 2015; Accepted: 8 September 2016

ABSTRACT

A study was carried out to forecast the yield of the wheat crop for five districts of Uttar Pradesh namely Lucknow, Kanpur, Banda, Jhansi and Faizabad. The daily weather data on variables such as maximum temperature, rainfall, minimum temperature, and relative humidity were arranged week wise from sowing to harvesting and the relations between the weather variables and yield was worked out using statistical tools like correlation and regression. The yield has been detrended by obtaining the parameter estimates of the model and subsequently the detrended yield was used to forecast the yield of the crop using ARIMA model. The proposed method of obtaining pre-harvest forecasting of yield of crops was compared with the traditional approaches of forecasting and the proposed method was evaluated in terms of criteria's such as goodness of fit of the model. It was observed that in all the districts the proposed model performed better as compared to the traditional method both in terms of goodness of fit as well as forecasting performance. Thus it can be concluded that the proposed approach is better and more suitable as compared to the traditional approach for forecasting the wheat yield in the five districts of Uttar Pradesh.

Key words: ARIMA, Detrended yield, Forecasting performance, Goodness of fit, Long term weather data, Yield forecast

In such a country where agriculture is the life blood of the country's economy and the livelihood of the people the importance of crop yield forecasting well in advance of harvest is undeniable.In a country like that of India reliable and a routine forecast of crop production will be very advantageous for advance planning, formulation and implementation of a number of policies dealing with food procurement its distribution, pricing structure, import and export decisions and for exercising several administrative measures related to the storage and marketing of the agricultural commodities. Losses may be faced by farmers due to uncertainties of weather, government policies relating to price, etc.

The statistical techniques which are to be employed for the forecasting purposes should be able to provide an objective, consistent and a very comprehensive forecast of the yield of the crop with reasonable precision well before the crop is harvested. Generally there are two basic approaches for predicting crop yield which are the simulation models and the models based on multiple regression approach. It has been found that crop yield is mainly influenced by two factors the first is weather and the second are inputs. There is an approach for forecasting crop production which is based on utilizing the information on those factors. The effect of Weather on crop growth varies with the different stages of crop growth. It has been found that the influence of weather on the yield of a crop depends on the magnitude of weather variables as well as on the manner in which the weather gets distributed over the different growth stages of the crop because different growth stages of the crop growth have different sensitivities towards weather parameters few of them are very sensitive to weather fluctuations, whereas others are less sensitive. Hence for precise forecasting we need to divide the entire crop growth phase into very fine interval. As a result of which there will be an increase in the number of variables in the model and consequently because of the fact more parameters have to be evaluated from the data. Hence, in such situations a very long series data will be a prerequisite for precise estimation of the parameters, which may be practically very difficult to obtain. Hence, the solution to this situation lies in the fact that one has to seek a model that is based on less number of parameters that could be easily evaluated and at the same time it should also take into account the parrern or the manner in which the weather is distributed over the entire crop growth phase.

We have used a regression approach in this study because of the well-established fact that the simulation models are much more complex and has higher data

[2]e mail: scientist1775@gmail.com, ICAR, Krishi Bhawan, New Delhi. [5]Jain University, Bengaluru. [6]ICRISAT, Hyderabad.

requirement. Approaches based on various production functions that could capture the effect of climate variables on crop yields was proposed by Yang *et al.* (1992), Dixon *et al.* (1994), Kandiannan *et al.* (2002), and Chen and Chang (2005). It was proved by Tannura *et al.* (2008) and studies conducted by others that, the explanatory power of the multiple regression models are much more and they can very well express how weather conditions and crop yield are related to one another.

In this study, an attempt is made to remove trend effect, i.e. detrended yield through statistical approach. Regression models has been developed in which the weather parameters were used as regressors whereas the detrended yield has been used as the explanatory varaiable.

MATERIALS AND METHODS

The study has been conducted for five different districts of the state of Uttar Pradesh, India namely Jhansi, Lucknow, Kanpur, Banda and Faizabad, which has a latitudinal extent of 30° 20′ N to 23° 53′ N and a longitudinal extent of 77° 4′ E to 84° 39′ E. In terms of area, Uttar Pradesh is the fourth largest state of India and in terms of population Uttar Pradesh ranks first in being the most populous state of the country and is located in the north central part of the Indian sub-continent. It usually ranges over a wide geographical area and the plains are also very different from the mountainous regions that are usually located in the northern part of the state. There is also a huge diversity in the climatic range of the state. It can be as high as 47°C in the summer to as low as -1°C in the winter. A major portion of the Uttar Pradesh lies in the central part which chiefly comprises the alluvial deposits which are being carried down from the Himalayas by the mighty rivers like the Ganges and its tributaries. These alluvial soils are extremely fertile and they usually vary from sandy to clayey soils. The texture of the soil varies from medium to medium heavy textured and are very easily ploughable. The favourable climate of the state, presence of ample irrigation facilities and the soil type makes the cultivation of rice and wheat a natural choice for the state. Wheat crop *(Triticum* sp.*)* is mainly cultivated during the *rabi* season as because this period provides congenial condition for the growth of the wheat crop. In the present study forecast models have been developed for the wheat crop, for different districts of UP state for the period from 1970-2010.Weekly data (1970-2010) on the weather variables of different districts of Uttar Pradesh has been collected during the different growth stages of the wheat crop. It has been obtained from Central Research Institute of Dryland Agriculture (CRIDA) Hyderabad and Indian Metrological Department, Pune.

Data has been mainly collected on weather variables that includes, maximum temperature, minimum temperature, relative humidity, rainfall. Sixteen weeks data from 40th Standard Meteorological Week to 3rd Standard Meteorological Week. Hereafter, these 16 weeks will be referred as period 1, 2, 3, ...,16.

Suppose we have a simple linear regression equation:

Mathematically,

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, i = 1,2,..., n$$

where, y_i is the dependent variable and the variable under study, x_i is the explanatory variable and is the error component. After estimating the parameters θ_0 and θ_1 of the regression equation through different parameter estimating techniques we can obtain the predicted value of the dependent variable as:

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i, i = 1,2,...,n$$

$\hat{y}_i$ obtained from above models has been used (as y_i) for the further development of forecast models and for further prediction respectively.

Suppose we assume that m denotes weeks (w=1(1)m) at which the pre-harvest forecast of the crop yield need to be released. If we use the weekly data on m weeks in p variables, now new weather variables and interaction components can be generated with respect to each of the weather variables using the below mentioned procedure. Forecast model has been developed by considering all the generated variables simultaneously including the time trend (T).

In order to study the individual effect of each weather variables, two new variables from each variable can be generated as follows:

Let X_{iw} be the value of the ith (i=1(1)p) weather variable at the wth week (w=1(1)n), r_{iw} be the simple correlation coefficient between weather variable X_i at the wth week and yield over a period of k years. The generated variables are given by:

$$Z_{ij} = \frac{\sum_{w=1}^{n} r_{iw}^{j} x_{iw}}{\sum_{w=1}^{n} r_{iw}^{j}}; j = 0,1$$

For j=0 we have unweighted generated variable as:

$$Z_{i0} = \frac{\sum_{w=1}^{n} X_{iw}}{n}$$

And weighted generated variables as:

$$Z_{i1} = \frac{\sum_{w=1}^{n} r_{iw} X_{iw}}{\sum_{w=1}^{n} r_{iw}}$$

For each year [1].

The following model is then fitted to study the effect of individual weather variables.

$$Y = a_0 + a_1 Z_{i0} + a_2 Z_{i1} + cT + \varepsilon$$

where, Y is yield, T is variable expressing time effect, a_0, a_1, a_2 and c are constant entities known as the parameters that needs to be evaluated from the model and ε is the error term

which is supposed to be distributed with a null expectation and a constant dispersion σ^2. Thus for each of the weather parameter two variable will be obtained which along with the time component and the intercept term makes a total of ten parameters thus ten parameters have been estimated in order to detrend the yield.

ARIMA is simply a generalization that incorporates a differencing term in the model as a result of this it includes a wide range of non-stationary time-series models. Random walk provides a best example of a non-stationary time-series that gets converted to a stationary time-series model just by the inclusion of a differencing term. A process $\{y_t\}$ is said to follow an Integrated Autoregressive Moving average, denoted by ARIMA (p, d, g), if $\nabla^d y_t = (1\text{-}B)^d \varepsilon_t$ is ARMA (p, g). The model is written as:

$$\phi(B)(1-B)^d y_t = \delta(B)\varepsilon_t$$

where, $\varepsilon_t \sim WN(0,\sigma^2)$. The letters *WN* represent White Noise. The range of the integration parameter is always greater than equal to zero, hence it is a positive integer. If $d = 0$, ARIMA (p, d, q) is equivalent to ARMA (p, q). The ARIMA process occurs at three steps, the first stage includes identification, the second stage involves estimation and the third stage involves diagnostic and checking. In the estimation stage one estimates the parameters of an ARIMA model that has been selected provisionally. In the diagnostic-checking stage one looks for the adequacy of the model selected tentatively. The three stages are continued in the same fashion till one attains satisfactory results, in case the results obtained are inadequate. Programmes for fitting of ARIMA is available in almost all standard software packages like SAS, SPSS, MATLAB, S-Plus and E-Views. ARIMA model usually has the notation *(p, d, q)* where, *p, d* and *q* denote the orders of auto-regression, integration (differencing) and moving average respectively

Suppose, if we have a time series of data X_t, where t is an integer index and X_t are real numbers then an ARMA *(p, q)* model is defined as

$$\left(1-\sum_{i=1}^{p}\varphi_i F^i\right)X_t = \left(1+\sum_{i=1}^{q}\delta_i F^i\right)\varepsilon_t$$

where F is the lag operator, ϕ_i are the model parameters that occur in the autoregressive part and δ_i are the model parameters that occurs in the moving average part. And ε_t are the error terms and it is assumed that $\varepsilon_t \sim iidN(0,\sigma^2)$. Now, on considering the polynomial:

$$\left(1-\sum_{i=1}^{p}\varphi_i F^i\right)$$

If it possesses a root of value 1 and of multiplicity d then it can be also be written as:

$$\left(1-\sum_{i=1}^{p}\varphi_i F^i\right) = \left(1-\sum_{i=1}^{p-d}\phi_i F^i\right)(1-F)^d$$

Now, An ARIMA *(p', d, q)* process is based on this polynomial factorisation property with $p' = p\text{-}d$ and can be written as:

$$\left(1-\sum_{i=1}^{p'}\phi_i F^i\right)(1-F)X_t = \left(1+\sum_{i=1}^{q}\delta_i F^i\right)\varepsilon_t$$

Hence, in general ARIMA may be considered as a special case of ARMA (p+d, q) process that contains the autoregressive polynomial with d unit roots.

Usually in the traditional approach the observed values of response variable are in general used for forecasting the crop yield. Here, in this study we have mainly used the detrended value, i.e. $\hat{y}_i$ that has been considered as the basis for forecasting. Here detrended yield basically means the predicted values of the study variable that will be obtained after estimating the parameters of the regression equation based on weather indices, The estimates of the parameters of the model has been obtained for all the five districts of the state and this estimates have been used to detrend the yield. Now the detrended yield has been fitted in an ARIMA model with AR(1), AR(2) and MA(1) and the forecasts and their respective standard error has also been calculated for all the five districts of the state. The residuals have also been obtained and they have been checked for the existence of correlation between them, i.e. to check if the residuals are distributed independently.

The squared residuals were also checked in terms of autocorrelation and partial auto correlation to check if there exists volatility in the data, i.e. if there are evidences of heteroscedasticity in the data for all the five districts.

We compared our model with the traditional models available in the litreature using various statistical measures. Goodness of fit as well as the forecasting performance of the models developed were evaluated using appropriate statistical measures. The goodness of fit of the model developed using Detrended yield approach were measured by statistical measures like the Mean Square Error.

RESULTS AND DISCUSSION

The distribution of weather variables spread over 16 weeks from First October up to Third week of January is given in (Table 1) particularly for the four weather variables namely Maximum temperature (X1), Minimum temperature (X2), Rainfall (X3) and Morning relative humidity (X4). The first week's data correspond to 40th standard meteorological week (SMW), second week's data correspond to 41st standard meteorological week (SMW) and so on and the 16th week's data correspond to 3rd standard meteorological week (SMW) of the next year. Hereafter these data (40th SMW to 52nd SMW and 1st to 3rd SMW) will be denoted as data corresponding to week number 1(1)16.

Computation of the parameter estimates of the weather indices model

The estimates of the parameters of the weather indices models have been obtained namely for the districts of Lucknow, Kanpur, Banda, Faizabad and Jhansi of the Uttar Pradesh State for detrending the yield. There are total four

Table 1 Mean values of the weather variables during crop growth phase

Growth phase	Week No.	Maximum temp. (°C)	Minimum temp. (°C)	Rainfall (mm)	Relative humidity (%)
Pre-sowing and germination phase	1	32.66 (1.65)	21.91 (1.98)	16.82 (32.23)	76.36 (13.72)
	2	32.58 (2.12)	20.88 (1.96)	18.69 (56.24)	73.79 (10.29)
	3	32.21 (1.59)	19.10 (2.22)	7.30 (21.54)	70.28 (9.99)
	4	31.39 (1.56)	17.47 (1.81)	0.03 (0.17)	69.81 (8.88)
Crown root initiation phase	5	30.61 (1.36)	15.60 (1.59)	0.86 (2.88)	67.85 (8.00)
	6	29.40 (1.75)	14.12 (1.52)	1.78 (6.49)	67.14 (11.29)
	7	28.38 (1.70)	13.15 (1.62)	0.70 (3.66)	69.70 (11.75)
Tillering phase	8	27.08 (1.01)	12.09 (2.17)	1.49 (4.59)	70.34 (11.91)
	9	25.36 (1.52)	10.47 (2.02)	3.32 (10.35)	71.37 (11.82)
	10	24.28 (1.29)	9.26 (1.48)	2.30 (11.02)	74.83 (7.46)
Jointing and reproductive phase	11	23.85 (1.71)	8.82 (1.93)	1.45 (5.53)	76.65 (9.60)
	12	22.51 (1.96)	8.34 (1.77)	0.76 (1.78)	80.10 (8.51)
	13	21.25 (2.49)	7.77 (1.61)	3.84 (7.82)	82.31 (7.90)
	14	20.02 (2.80)	6.96 (1.86)	2.78 (10.68)	83.24 (6.98)
	15	20.28 (2.82)	7.50 (2.03)	2.82 (6.13)	84.18 (8.49)
	16	20.47 (2.57)	7.14 (2.21)	3.41 (11.01)	83.57 (7.89)

Table 2 Estimates of parameters for Lucknow district

	Coefficients	Standard error	t-statistic
Intercept	37.94	1.39	27.21
P-1	17.06	3.55	4.81
P-2	8.18	1.69	4.85
P-3	-0.94	0.68	-1.37
P-4	-20.54	1.85	-11.07
P-5	0.42	0.02	19.82
P-6	-0.63	0.07	-9.38
P-7	-0.01	0.01	-2.93
P-8	-0.04	0.00	-9.44
Trend	0.48	0.01	39.02

Table 3 Estimates of parameters for Faizabad district

	Coefficients	Standard error	t-statistic
Intercept	12.36	2.51	4.91
P-1	-38.63	5.14	-7.51
P-2	9.62	1.91	5.02
P-3	-11.85	0.52	-22.60
P-4	-37.57	3.29	-11.41
P-5	0.05	0.01	2.81
P-6	-0.59	0.08	-7.33
P-7	-0.03	0.01	-11.52
P-8	0.01	0.00	3.85
Trend	0.48	0.01	64.38

Table 4 Estimates of parameters for Kanpur district

	Coefficients	Standard error	t-statistic
Intercept	40.62	2.02	20.09
P-1	7.56	0.64	11.80
P-2	15.45	3.68	4.19
P-3	-0.56	1.32	-0.43
P-4	-51.26	2.93	-17.44
P-5	-0.18	0.01	-10.00
P-6	-0.593	0.11	-5.39
P-7	-0.02	0.01	-4.53
P-8	0.04	0.01	5.35
Trend	0.49	0.01	52.48

weather variables, viz. Maximum temperature, Rainfall, Relative humidity and Minimum temperature. For each of the weather variable we would again get two indices one weighted and another unweighted index. Therefore, there are 10 parameters that need to be estimated including the intercept term. The parameters have been represented as the first weather variable is Maximum temperature for which we have two index that are Unweighted generated maximum temperature index (P-1) and the Weighted generated maximum temperature index (P-2), Second weather variable

Table 5 Estimates of parameters for Banda district

	Coefficients	Standard error	t-statistic
Intercept	3.06	0.99	3.02
P-1	-37.72	2.85	-12.54
P-2	-54.82	5.73	-9.56
P-3	-5.08	1.29	-3.95
P-4	-21.86	1.99	-11.01
P-5	-0.01	0.04	-0.39
P-6	1.98	0.19	10.19
P-7	-0.02	0.01	-4.44
P-8	0.08	0.01	10.51
Trend	0.14	0.01	12.54

Table 6 Estimates of parameters for Jhansi district

	Coefficients	Standard error	t-statistic
Intercept	29.66	1.25	23.78
P-1	-77.16	3.72	-20.71
P-2	17.44	3.37	5.18
P-3	-23.42	2.25	-10.40
P-4	-55.29	2.55	-21.69
P-5	0.64	0.07	-2.49
P-6	-0.22	0.09	-2.48
P-7	-0.03	0.01	-7.53
P-8	0.04	0.00	4.95
Trend	0.28	0.01	22.76

Table 7 The parameter estimates after using detrended yield time series data obtained for different districts of Uttar Pradesh using ARIMA, i.e. AR(2) and MA(1) are illustrated in the following table

	AR(1)	AR(2)	MA(1)
Banda	1.47	-0.47	-0.95
	(0.17)	(0.17)	(0.06)
Lucknow	1.23	-0.23	-0.84
	(0.22)	(0.22)	(0.15)
Faizabad	0.12	0.82	0.42
	(0.26)	(0.24)	(0.33)
Kanpur	0.66	0.34	-0.47
	(0.25)	(0.25)	(0.23)
Jhansi	0.22	0.78	0.68
	(0.19)	(0.18)	(0.24)

is Minimum temperature for which we have two index that are Unweighted generated minimum temperature index (P-3) and the Weighted generated minimum temperature index (P-4), Third weather variable is rainfall for which we have two index that are Unweighted generated rainfall index (P-5) and the Weighted generated rainfall index (P-6), Fourth weather variable is Relative humidity for which we have two index that are Unweighted generated relative humidity index (P-7) and the Weighted generated relative humidity index (P-8), along with trend and the intercept there are ten parameters that has been estimated.

Fitting of ARIMA to the detrended yield variable

After obtaining the estimates of the parameter the yield is detrended and after detrending the yield, the detrended yield variable is fitted using ARIMA AR(1) AR(2) and MA(1) and the corresponding values computed for each district using the ARIMA model.

Computation of Autocorrelation, Partial Autocorrelation and Q-statistic of the Residuals

The Autocorrelation, Partial Autocorrelation as well as the Q-statistic were calculated separately for each of the districts and from the probability figures for all the districts it is quite clear that the residuals are random.

Computation of Autocorrelation, Partial Autocorrelation and Q-statistic of the Squared Residuals

The Autocorrelation, Partial Autocorrelation as well as the Q-statistic were calculated separately for each of the districts based on the Squared Residuals and from the probability figures for all the districts it is quite clear that there does not exist volatility in the data.

The proposed approach was then compared with the traditional method of forecasting which employs forecasting on the basis of the observed values of the detrended variable as well as simple ARIMA based on only the detrended. The proposed model was compared for the goodness of fit as well as forecasting performance in the five districts of Uttar Pradesh. It was observed that in all the five districts the proposed model performed better as compared to the traditional method as well as simple ARIMA in terms of goodness of fit as well as forecasting performance of the model. Hence, it may be concluded that the proposed method is performing relatively better than the other traditional methods and is more suitable for forecasting the yield of the wheat crop in the districts of Jhansi, Lucknow, Kanpur, Banda and Faizabad for the state of Uttar Pradesh.

REFERENCES

Agrawal R, Jain R C and Jha M P. 1986. Models for studying rice crop-weather relationship, *Mausam* **37**(1): 67–70.

Agrawal R, Jain R C and Jha M P. 2001. Yield forecast based on weather variables and agricultural input on agro climatic zone basis. *Indian Journal of Agricultural Sciences* **71**(7): 487–90.

Agrawal R, Jain R C Jha, M P and Singh D. 1980. Forecasting of rice yield using climatic variables. *Indian Journal of Agricultural Sciences* **50**(9): 680–4.

Bannayan C E. 2003. Corn yield prediction using climatology. *Journal of Climatology and Applied Meteorology* **25**: 581–90.

Bannayan M Crout N M J and Hoogenboom G. 2003. Application of the CERES-Wheat Model for Within-Season prediction of winter wheat yield in the United Kingdom. *Agronomy Journal* **95**: 114–25.

Chandrahas, Agrawal R and Walia S S. 2010. Use of discriminant function and principal component technique for weather based crop yield forecasts. IASRI publication.

Kandiannan K, Karithikeyan R, Krishnan, Kailasam, R C and Balasubramanian T N. 2002. A Crop-Weather Model for prediction of rice yield using an Empirical- Statistical Tehnique. *Journal of Agronomy and Crop Science* **188**: 59–62.

Mehta S C, Agrawal R and Singh V P N 2000. Strategies for composite forecasts. *Journal of the Indian Society of Agricultural Statistics* **53**(3): 262–72.

Mehta S C, Pal S and Kumar V. 2010. Weather based models for forecasting potato yields in Uttar Pradesh. IASRI Publication.

Singh D. 2006. Forecasting technique under the FASAL Scheme. *Proceedings of the Annual Review Meeting*, Pune.

Singh P K, Singh A K, Baxla A K, Kumar B, Bhan S C and Rathore L S. 2014. Crop yield prediction using CERES-Rice vs 4.5 model for the climate variabilility of the different agroclimatic zone of south and north-west plain zone of Bihar. *Mausam* **65**: 529–38.