Indian Journal of Agricultural Sciences **83** (2): 180–3, February 2013/Article

Statistical modelling for forecasting of wheat yield based on weather variables

RANJIT KUMAR PAUL[1], PRAJNESHU[2] and HIMADRI GHOSH[3]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 19 May 2012; Revised accepted: 16 January 2013

ABSTRACT

Forecasting of crop yield based on historical data and pertinent external climatic information is considered. To this end, Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) time-series model along with its estimation procedure is studied. In the present investigation, five models at five important stages of wheat growth are developed by including the most important weather variables. The weekly maximum temperature at crown root initiation (CRI) stage, tillering stage, anthesis stage, milk stage and dough stage and evapotranspiration at CRI stage are used for model development. As an illustration, ARIMAX models are employed for forecasting of wheat yield in Kanpur district of Uttar Pradesh. Comparative study of the fitted models is carried out from the viewpoint of Relative mean absolute prediction error (RMAPE). It is demonstrated that the ARIMAX methodology is able to provide pre-harvest forecasts based on weather variables at various stages of wheat crop growth, starting from CRI stage (21 days after sowing) to dough stage (126 days after sowing). It is observed that, as wheat crop grows towards maturity, pre-harvest forecasts get closer to actual values.

Key words: ARIMAX model, Forecasting, Weather variables, Wheat yield

Quantitative understanding of crop responses to climate requires development of statistical models for various characteristics of crop by taking into account its time-series behaviour along with exogenous climate factors. The response of interest (e.g. crop yield) as the response variable and various climate related quantities (e.g. temperature, and humidity) as the predictor variable(s) need to be considered for model building. Kumar *et al*. (2001) studied the effect of different weather variables on wheat yield and found that maximum temperature was negatively correlated with yield of late sown wheat in Tarai region. Lobell *et al*. (2006) developed weather based yield forecast model for 12 California crops. The authors combined weather and yield data in a Linear regression model to test how well yield anomalies could be predicted before harvest based on monthly weather measurements. But the authors did not take care of the time-series behaviour of the data. Maximum temperature plays a very important role in wheat crop development and wheat yield (Pathak and Wassmann 2009). It is well-known that one of the main factors causing yields to change from year to year is climate variability as no two growing seasons experience exactly the same weather (Lobell *et al.* 2010). When forecasting is applied to dynamic behaviour of crop yield, it should be able to take advantage not only of the historical data, but also of the impact of various driving forces, like temperature, relative humidity, and evapotranspiration from the external environment. The key problem for forecasting is how to incorporate pertinent external information into the forecasting process and subsequently to the decision-making process. Asseng *et al.* (2011) reported that the effect of temperature on wheat production is underestimated. The authors observed that variations in average growing-season temperatures of ±2 °C in the main wheat growing regions of Australia can cause reductions in grain production of up to 50%. Kaur *et al.* (2012) reported that the effect of maximum temperature on wheat yield is more negative than that of the minimum temperature.

Wheat is the second most important food crop of India after rice both in area and production. It accounts for 26 percent of the total area and 36.5 percent of the total production of cereals in the country. India stands second in the production of wheat in the World contributing over 13 percent of the total area and 12 percent of the total production. In India, Uttar Pradesh ranks first in area (36.58%) and production (36.27%). In Kanpur district of Uttar Pradesh, wheat crop is generally sown during late October to early November and harvested during end of March or early April every year. In

[1]Scientist (e mail: ranjitstat@gmail.com), Division of Biometrics and Statistical Modelling, [2]Principal Scientist and Head (e mail: prajneshu@yahoo.co.in), Division of Biometrics and Statistical Modelling, [3]Senior Scientist (e mail: hghosh@gmail.com), Division of Biometrics and Statistical Modelling

this paper, we investigated the temporal impact on wheat yield of maximum temperature at different important stages of wheat growth, viz. Crown root initiation, tillering, anthesis, milky, and dough stages appearing respectively after 21,42, 84, 105, and 126 days of sowing.

MATERIALS AND METHODS

The ARIMAX model (Bierens 1987) is a generalization of the ARIMA model, which is capable of incorporating an external input variable (X). Given a (k+1) time-series process $\{(y_t, x_t)\}$, where y_t and k-components of x_t are real valued random variables, the ARIMAX model assumes the form

$$\left(1-\sum_{s=1}^{p}\alpha_s L^s\right)\Delta y_t = \mu + \sum_{s=1}^{q}\beta_s' L^s x_t + \left(1+\sum_{s=1}^{r}\gamma_s L^s\right)e_t \quad (1)$$

where L is the usual lag operator, i.e. $L^s y_t = y_{t-s}$, $\Delta y_t = y_t - y_{t-1}$, $\mu \in R$, $\alpha_s \in R$, $\beta_s \in R^k$ and $\gamma_s \in R$ are the unknown parameters and e_t's are the errors, and p, q and r are natural numbers specified in advance.

The first step in building an ARIMAX model consists of identifying a suitable ARIMA model for the endogenous variable. The ARIMAX model concept requires testing of stationarity of exogenous variable before modelling. The transformed variable is added to the ARIMA model in the second step, in which the lag length r is also estimated. Nonlinear least squares estimation procedure is employed to estimate the parameters of ARIMAX model (Bierens 1987). Fortunately, the ARIMAX model can be fitted to data by using a software package, like SAS, MATLAB, EViews and R. In the present investigation, SAS, Version 9.3 is used for data analysis.

RESULTS AND DISCUSSION

As an illustration, annual wheat yield data of Kanpur district of Uttar Pradesh during 1972 to 2011 comprising 40 data points are obtained from Directorate of Economics and Statistics, Government of India. The first 36 observations, i.e. the data from 1972 to 2007 are used for model building and the remaining 4 data points, i.e. the data from 2008 to 2011 are used for validating the model. The daily climate data on maximum temperature, minimum temperature, evapotranspiration and relative humidity during the same time-period are obtained from India Meteorological Department. The daily data is first converted to weekly data. In consonance with the results of Pathak and Wassmann (2009), exploratory data analysis for present data showed that the correlation coefficients between wheat yield and weekly maximum temperature at various stages, viz. CRI, tillering, anthesis, milk, and dough stages are statistically significant at 5% level of significance. Further, correlation coefficient between wheat yield and evapotranspiration at CRI stage was also found to be significant at 5% level. So, only these variables are used for subsequent model development.

Fitting of ARIMAX model

Evidently, the data set of wheat yield is not stationary. In order to select the order of the ARIMA model, unit root test proposed by Dickey and Fuller (1979) is applied for parameter ρ in the auxiliary regression

$$\Delta y_t = \rho y_{t-1} + \alpha_1 \Delta y_{t-1} + \varepsilon_t \quad (2)$$

where $\Delta y_t = y_t - y_{t-1}$. The relevant null hypothesis is H_0: $\rho = 0$ and the alternative is H_1: $\rho < 0$. For the given data, the estimate of ρ is computed as –0.64 with calculated *t*-statistic as –4.62, which is less than the critical value of *t* at 5% level of significance, i.e. –1.95 (Franses 1998, Page 82). Therefore, H_0 is not rejected at 5% level and so $\rho = 0$. Thus, there is presence of one unit root and so differencing is required. On the other hand, the time-series data of weekly maximum temperature for different stages and evapotranspiration at CRI stage are found to be stationary. On the basis of minimum Akaike information criterion (AIC) and Schwartz-Bayesian criterion (BIC) values, best ARIMAX model was selected. In the present investigation, we developed five models for forecasting wheat yield at five stages of wheat crop, ie CRI stage, tillering stage, anthesis stage, milk stage and dough stage. The model developed at CRI stage includes maximum temperature and evapotranspiration at CRI stage as exogenous variables along with the time-series data of wheat yield. The model developed at tillering stage included maximum temperature at that stage plus maximum temperature and evapotranspiration at CRI stage as exogenous variables along with the time-series data of wheat yield. Similarly, the model developed at anthesis stage included maximum temperature at this stage along with other exogenous variables. The model at milk stage included maximum temperature at that stage along with some other exogenous variables. Finally, the model at dough stage included maximum temperature at that stage along with other exogenous variables. Specifically, the following models were fitted to data:

(i) Model I (At CRI stage):

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \beta_1 x_1 + \beta_2 x_2$$

(ii) Model II (At tillering stage):

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

(iii) Model III (At anthesis stage):

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

(iv) Model IV (At milk stage):

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

(v) Model V (At dough stage):

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

In the above, x_1, x_3, x_4, x_5, x_6 denote respectively the weekly maximum temperatures at CRI, tillering, anthesis, milk, dough stages and x_2 denotes evapotranspiration at CRI stage.

Estimates of the parameters of above models along with their respective standard errrors in brackets () are respectively computed as follows:

Model I:
$$\underset{(7.85)\ (0.13)\ \ (0.27)\ \ (0.13)}{\Delta y_t = 18.42 - 0.67\Delta y_{t-1} - 0.64 x_1 - 0.20 x_2}$$

Model II:
$$\underset{(7.56)\ (0.14)\ \ (0.34)\ \ (0.21)\ \ (0.22)}{\Delta y_t = 20.00 - 0.69\Delta y_{t-1} - 1.03 x_1 - 0.25 x_2 + 0.41 x_3}$$

Model III:
$$\underset{(7.28)\ (0.13)\ \ (0.36)\ \ (0.57)\ \ (0.22)\ \ (0.14)}{\Delta y_t = 20.10 - 0.69\Delta y_{t-1} - 1.34 x_1 - 0.67 x_2 + 0.53 x_3 + 0.27 x_4}$$

Model IV:
$$\underset{(7.27)\ (0.14)\ \ (0.37)\ \ (0.39)\ \ (0.22)\ \ (0.15)\ \ (0.14)}{\Delta y_t = 20.00 - 0.71\Delta y_{t-1} - 1.41 x_1 - 0.47 x_2 + 0.53 x_3 + 0.21 x_4 + 0.16 x_5}$$

Model V:
$$\underset{(9.02)\ (0.13)\ \ (0.37)\ \ (0.26)\ \ (0.23)\ \ (0.15)\ \ (0.13)\ \ (0.02)}{\Delta y_t = 20.68 - 0.71\Delta y_{t-1} - 1.42 x_1 - 0.47 x_2 + 0.53 x_3 + 0.21 x_4 + 0.17 x_5 + 0.03 x_6}$$

To get a visual idea, the fitted model at dough stage along with data points is exhibited in Fig 1.

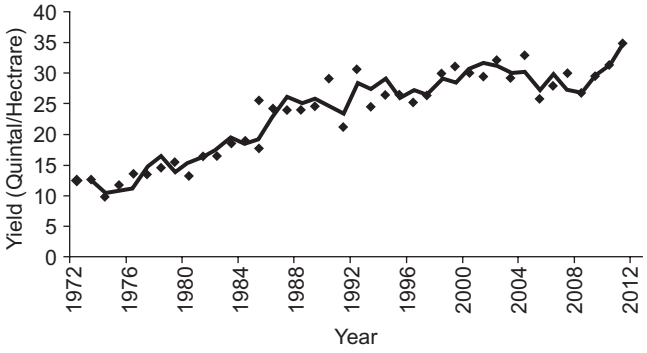

Fig 1 Fitted ARIMAX model at the dough stage along with data points

Validation of models for hold-out data

One-step ahead forecasts of wheat yield for the years 2008 to 2011 in respect of above fitted models are reported in Table 1. A comparative study of forecasts by these models is carried out on the basis of Relative Mean Absolute Prediction Error (RMAPE) values defined as

$$\text{RMAPE} = 1/4\sum_{i=1}^{4}\{|y_{t+i} - \hat{y}_{t+i}|/\, y_{t+i}\}\times 100 \qquad (3)$$

and are also presented in Table 1. It may be noted that all the five models perform quite satisfactorily as the RMAPE values are throughout less than 5%. Further, these values go on decreasing from Model I to Model V, which is logical as the forecasts should improve as the growing season progresses.

Table 1 Forecasts of wheat yield (in Quintals/Hectare) and their performance for hold-out data

Year	Actual	Forecasts by Models				
		I	II	III	IV	V
2008	26.66	27.03	26.56	27.53	26.80	26.80
2009	29.40	28.88	28.35	30.26	29.63	29.57
2010	31.50	29.46	30.06	30.66	31.20	31.40
2011	34.80	31.34	32.74	33.94	34.67	34.69
	RMAPE(%)	4.89	3.60	2.80	0.66	0.43

Pre-harvest forecast for current year

For carrying out pre-harvest forecasts of wheat yield for Kanpur district, all five models were used. For the models I-V, the pre-harvest forecast for the current year 2012 has been computed by using following equations:

Forecast at CRI stage:
$$\hat{y}_{2012}=18.42+y_{2011}(1–0.67)+0.67y_{2010}–0.64x_{1,2011}–0.20\ x_{2,2011}$$

Forecast at tillering stage:
$$\hat{y}_{2012}=20.00 + y_{2011}(1–0.69) + 0.69y_{2010} – 1.03x_{1,2011}–0.25\ x_{2,2011}+ 0.41x_{3,2011}$$

Forecast at anthesis stage:
$$\hat{y}_{2012}=20.00 + y_{2011}(1–0.69) + 0.69y_{2010} – 1.34x_{1,2011}–0.67\ x_{2,2011} + 0.53x_{3,2011} + 0.27x_{4,2011}$$

Forecast at milk stage:
$$\hat{y}_{2012}=20.00 + y_{2011}(1–0.71) + 0.71y_{2010} – 1.41x_{1,2011}–0.47\ x_{2,2011} + 0.53x_{3,2011} + 0.21x_{4,2011}+ 0.16x_{5,2011}$$

Forecast at dough stage:
$$\hat{y}_{2012}= 20.68 + y_{2011}(1–0.71) + 0.71y_{2010} – 1.42x_{1,2011}–0.47\ x_{2,2011} + 0.53x_{3,2011} + 0.21x_{4,2011} + 0.17\ x_{5,2011} + 0.03\ x_{6,2011}$$

In the above, $\hat{y}_{2012}$ denotes forecast of wheat yield in year 2012, y_{2011} and y_{2010} denote respectively the wheat yields in years 2011 and 2010. The variables $x_{1,2011}$, $x_{3,2011}$, $x_{4,2011}$, $x_{5,2011}$, $x_{6,2011}$ denote respectively the weekly maximum temperatures in year 2011 at CRI, tillering, anthesis, milk, and dough stages and $x_{2,2011}$ denotes evapotranspiration at CRI stage.

Forecasts (in Quintals/Hectare) for the year 2012 by using above five models are computed respectively as 32.34, 32.33, 32.07, 33.28 and 34.71. It is observed that at the initial stage of wheat growth, i.e. just 21 days after sowing of wheat crop (CRI stage), the methodology is able to provide reasonably good forecast of wheat yield. Further, with the passage of time, more reliable forecasts can be built up. It may also be pointed out that the ARIMAX methodology can also be used for obtaining more than one-step-ahead forecasts by first computing the multi-step-ahead forecasts of the exogenous variables used in Models I-V through ARIMA approach. Finally, the ARIMAX methodology may also be

employed for obtaining pre-harvest forecasts based on weather variables for other crops.

REFERENCES

Asseng S, Foster I and Turner N C. 2011.The impact of temperature variability on wheat yields. *Global Change Biology* **17**: 997-1 012.

Bierens H J. 1987. ARMAX model specification testing, with an application to unemployment in the Netherlands. *Journal of Econometrics* **35**: 161-90.

Dickey D A and Fuller W A. 1979. Distribution of the estimators for the autoregressive time series with a unit root. *Journal of the American Statistical Association* **74**: 427-31.

Franses P H. 1998. *Time Series Models for Business and Economic Forecasting*. Cambridge University Press, UK.

Kaur H, Jalota S K, Kanwar R and Vashisht B B. 2012. Climate change impacts on yield, evapotranspiration and nitrogen uptake in irrigated maize (*Zea mays*)–wheat (*Triticum aestivum*) cropping system: A simulation analysis. *Indian Journal of Agricultural Sciences* **82**: 213-29.

Kumar S, Mishra H S, Sharma A K and Kumar S. 2001. Effect of weather variables on the yield of early, timely and late sown wheat in the Tarai region. *Journal of Agricultural Physics* **1**: 58-62.

Lobell D B, Field C B, Cahill K N and Bonfils C. 2006. Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. *Agricultural and Forest Meteorology* **141**: 208–18.

Lobell D B and Burkeb M B. 2010. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology* **150**: 1 443–52.

Pathak H and Wassmann R. 2009. Quantitative evaluation of climatic variability and risks for wheat yield in India. *Climate Change* **93**: 157-75.