J. Cotton Res. Dev. **35**(2), 280-285 (July, 2021) **ISSN No. 0972-8619**

Cotton yield forecasting in districts of Haryana using fortnightly weather variables: A time series approach

ALISHA MITTAL AND BAISHALI MISHRA*

Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar-125004

**Email: baishalimishra@hau.ac.in*

ABSTRACT : The study has been categorized into three parts *i.e.* the fitting of Random Walk, ARIMA and ARIMAX models for cotton yield forecasting in Hisar, Fatehabad and Sirsa districts of Haryana. The Random Walk and ARIMA models have been fitted using the time series cotton yield data for the period 1980-1981 to 2010-2011 of Hisar and Sirsa districts and 1997-1998 to 2010-2011 of Fatehabad district. The fortnightly weather data have been utilized as input series from 1980-1981 to 2016-2017 for fitting/testing the Random walk/ARIMA with weather input *i.e.* ARIMAX models. Models have been validated using the data on subsequent years *i.e.*, 2011-2012 to 2016-2017, not included in the development of the models. Random Walk *i.e.* I(1) and ARIMA (0, 1, 1) for Hisar, Fatehabad and Sirsa districts have been fitted for cotton yield forecasting. Alternatively, Random Walk models with exogenous input were tried by utilizing the fortnightly weather variables (*viz.*, TMIN1, RF11, SSH3 and SSH4 over the crop growth period). Lastly, ARIMA (2,1,0) for Hisar and Fatehabad and ARIMA (0,1,1) for Sirsa districts along with fortnightly weather variables (*viz.*, TMAX5, RF7, SSH4 and RH4 over the crop growth period) as input were finalized as ARIMAX models for cotton yield forecasting. The predictive performance (s) of the contending models were observed in terms of the per cent deviations of cotton yield forecasts in relation to the observed yield (s) and root mean square error (s) as well. The ARIMAX models performed well with lower error metrics as compared to the Random Walk and ARIMA models in all time regimes.

Keywords: Fortnightly maximum temperature, minimum temperature, rainfall, relative humidity, sunshine, time series analysis

Availability of an efficient crop forecasting methodology is essential to provide information on food supply situation which is useful for export/import planning, procurement operation and price determination. Forecast of any crop is usually made on the basis of crop input variables. Therefore, selection of a set of predictor variables to give an adequate basis for forecasting is very important. Time series models are useful in certain situations as they can be used more easily for forecasting purposes because the historical sequence(s) of observations upon study variables are readily available at equally spaced intervals over discrete point of time. The Box-Jenkins (1976) methodology is a powerful tool for time-series analysis, when the time-sequenced observations in a data series may be statistically dependent or related to each other. Generally, univariate autoregressive integrated moving average (ARIMA) time series models, mainly due to the contributions of Box and Jenkins are widely used in practice for forecasting.

However, when the patterns of the time-series under study are affected by some external factors then the forecasting performance of ARIMA model may be affected. Under such situations, the model can be improved by employing some appropriate technique like ARIMA with regressor(s) analysis. When an ARIMA model includes other time series as input variables, the model is sometime referred as an ARIMAX model. In view of the importance of the subject matter, the present study has been undertaken with the objectives, (i) To develop Random Walk, ARIMA and ARIMAX (weather parameters as regressors) models for cotton yield forecasting in western zone of Haryana. (ii) To compare the forecasting performance and post-sample validity testing of the developed models.

Cotton, the "White Gold" or the "King of Fibres", enjoys a predominant position amongst all cash crops in India. In India, cotton occupies an area of nearly 117.27 million hectares, with a production of 398 lakh bales (2013-2014), ranking third in the world after China and USA which accounts for about 18 per cent of the world cotton production. It has also the distinction of having the largest area under cotton cultivation in the world constituting about 27 per cent of the world area under cotton cultivation. The lint productivity of cotton is 577 kg/ha, which is the lowest and far below that of the world average of 756 kg/ha. During last fifty years, production of cotton rose from 30 lakh bales (1 bale = 170 kg) in 1950-1951 to 398 lakh bales in 2013-2014. During the same period, the area under cultivation increased from 56.48 lakh hectares to 112.42 lakh hectares. Significant increase in the area under cultivation of cotton was observed over a period of fifty years. Nearly 65 per cent cotton cultivation is rain dependent and subject to heavy vagaries of monsoon rains.

Arya *et al.*, (2015) fitted ARIMAX time-series model for forecasting the pest population after testing for stationarity. Primary weekly data (2008-2012) for three pests namely Jassids, Whitefly and Thrips in Guntur and Faridkot districts along with weekly maximum temperature, minimum temperature, rainfall, maximum RH and minimum RH were used for the model development. The results showed that maximum temperature and minimum temperature along with maximum relative humidity have a significant role for Whitefly and Thrips. Ravita and Verma (2016) applied ARIMA modeling for mustard yield prediction in Hisar, Bhiwani, Sirsa, Mahendergarh and Gurgaon districts of Haryana. On experimenting with different lags of the moving average and autoregressive processes, the best fitted models were; ARIMA (0, 1, 1) for Hisar, Bhiwani and Sirsa districts and ARIMA (1, 1, 0) for Mahendergarh and Gurgaon districts. The analysis indicated that the per cent deviations of the forecast yield(s) from the observed yield(s) were within acceptable limits and favoured the use of ARIMA modeling to get short-term forecast estimates.

MATERIALS AND METHODS

The Haryana state comprising of 22 districts is situated between 74° 25' E to 77° 38' E longitude and 27° 40' N to 30° 55' N latitude. The total geographical area of the state is 44,212 sq. km. The districts Hisar, Fatehabad and Sirsa have been considered for the model building. The Department of Agriculture (DOA) cotton yield estimates for the period 1980-1981 to 2016-2017 collected from Statistical Abstracts of Haryana have been used for computing the trend based yield. The daily weather data on maximum temperature (TMAX), minimum temperature (TMIN), rainfall (RF), sunshine hours (SSH) and relative humidity (RH) were collected for the same period. Weather data starting from first fortnight of May to 1 month before harvest (*i.e.*11 fortnights) were utilized for the model building (crop growth period: May to October/November).

Keeping in view the targeted objectives, the emphasis has been given in predicting the future values on the basis of previous time-series observations, and along with weather parameters as input series. The time-series yield/weather data from 1980-1981 to 2010-2011 have been used for the training set and the remaining data *i.e.* 2011-2012 to 2016-2017 have been used for the post-sample validity checking of the developed Random walk, ARIMA and ARIMAX models.

Random Walk Model: The classical example of non-stationary time series is the random walk model or I(1) model. Basically, there are two types of walks,

1) random walk without drift (*i.e.*, no constant or intercept term)

A series Yt is said to be a random walk or I(1) model if Y_t= Y_t-1+at *i.e.*, the value of Y at time t is equal to its value at time (t-1) plus a white noise error term at with mean 0 and variance σ2. This can also be seen as a regression of Y at time t on its lagged one period.

2) random walk with drift (*i.e.*, a constant term is present)

If Yt= δ+Y_t-1+at, where δ is known as the drift parameter, then this is I(1) model with drift. The mean as well as the variance increases over time, violating the conditions of stationarity. In short, Random walk model, with or without drift, is a non-stationary stochastic process.

Autoregressive Integrated Moving Average model: Univariate Box-Jenkins ARIMA forecasts are based only on past values of the variable being forecast. They are not based on any other data series and are especially suited to short-term forecasting. The method applies to both discrete data as well as to continuous data. However, the data should be available at equally spaced discrete time intervals. Also, building of an ARIMA model requires a minimum sample size of about 35-40 time-series observations and applies only to stationary time series data. A stationary time series has mean, variance and auto-correlation function essentially constant over time. If a time series is stationary then the mean of any major subset of the series does not differ significantly from the mean of any other subset. Similarly, if a data series is stationary then the variance of any major subset of the series will differ from the variance of any other major subset only by chance. An ARIMA (p, d, q) may be expressed as:

$\phi_p(B)\Delta^d Y_t = c + \theta_q(B)a_t$

where,

Y_t = Variable under forecasting

B = Lag operator

a = Error term (Y_t -Ŷ, where Ŷ is the estimated value of Y_t)

t = the time subscript

$\phi_p(B)$ = Non-seasonal AR

$(1\text{-}B)^d$ = Non-seasonal difference

$\phi_q(B)$ = Non-seasonal MA

ARIMA models with input series/exogenous variable(s): ARIMAX is an acronym for autoregressive integrated moving average with exogenous variables. It is a logical extension of pure ARIMA modeling that incorporates independent variables which add explanatory value. Conceptually, it is a merging of ARIMA and regression modeling. When the AR and MA terms in a pure ARIMA model are not sufficient to provide an acceptably overall explanatory power of a model, it is only natural to look for other driving phenomena whose influence over time is not sufficiently embedded in the historical values of the dependent time series. When an ARIMA model includes other time series as input variables, the model is sometimes referred to as an ARIMAX model *i.e.* in addition to past values of the response series and past errors, the response series is modeled using the current and past values of input series. Assuming two time series denoted as Y_t and X_t which are both stationary, then, the ARIMAX model may be written as follows:

$Y_t = C + v(B)X_t + N_t$

where,

Y_t is the output series (dependent variable)

X_t is the input series (independent variable)

C is the constant term

N_t is the stochastic disturbance

$v(B)X_t = (v0 + v1B + v2B^2 + \ldots + vpB^p)\, X_t$ is impulse response function, which allows X to influence Y via distributed lag(s).

B is backshift operator

RESULTS AND DISCUSSION

The cotton yield(s) data were found to be non-stationary for Hisar, Fatehabad and Sirsa districts. Differencing of order one was sufficient for getting an appropriate stationary series in these districts. Almost all the autocorrelations upto lag 9 significantly different from zero in Tables 1 to 3 confirmed non- stationarity. I(1) with TMIN1 for Hisar, I(1) with SSH3 and RF11 for Fatehabad and I(1) with SSH4 for Sirsa districts were finalized as Random walk models for pre-harvest cotton yield forecasting (Table 4). After experimenting with different lags of the moving average and autoregressive processes; ARIMA (0, 1 ,1) for Hisar, Fatehabad and Sirsa districts were fitted for estimating the district-level cotton yield (Table 5). Also, all Chi-Squared statistic (s) in this concern calculated using the Ljung-Box (1978) formula showed that none of the residual auto correlation functions in any of the districts were significantly different from zero at a reasonable level as has been shown in Table 7. This ruled out any systematic pattern in the residuals. The fitted ARIMA (0,1,1) model for Hisar, Fatehabad and Sirsa districts may be elaborated as below:

$(1\text{-}B)\ Y_t = (1\text{-}\theta 1B)at$

$Y_t - BY_t = at - \theta 1Bat$

$Y_t = Y_t\text{-}1 - \theta 1\ at\text{-}1\text{+}at$

In an effort to improve the predictive performance; the ARIMA models with alternative combinations of weather variables were tried. Consequent upon, ARIMA (2, 1, 0) with RH4 and RF7 for Hisar and ARIMA (2, 1, 0) with TMAX5 for Fatehabad and ARIMA (0, 1, 1) with SSH4 for

Table 1. Autocorrelations of cotton yield (kg/ha) for Hisar district

Lag	Autocorrelation	Std. Error	Box-Ljung Statistic		
			Value	df	Sig.
1	.526	.171	9.449	1	.002
2	.310	.168	12.842	2	.002
3	.260	.165	15.308	3	.002
4	.212	.162	17.003	4	.002
5	.004	.159	17.004	5	.004
6	-.189	.156	18.462	6	.005
7	-.215	.153	20.431	7	.005
8	-.134	.150	21.228	8	.007
9	-.292	.147	25.192	9	.003

Table 2. Autocorrelations of cotton yield (kg/ha) for Fatehabad district

Lag	Autocorrelation	Std. Error	Box-Ljung Statistic		
			Value	df	Sig.
1	.697	.171	16.583	1	<.001
2	.481	.168	24.755	2	<.001
3	.414	.165	31.005	3	<.001
4	.370	.162	36.203	4	<.001
5	.216	.159	38.032	5	<.001
6	.023	.156	38.053	6	<.001
7	-.001	.153	38.053	7	<.001
8	-.162	.150	39.224	8	<.001
9	-.197	.147	41.023	9	<.001

Table 3. Autocorrelations of cotton yield (kg/ha) for Sirsa district

Lag	Autocorrelation	Std. Error	Box-Ljung Statistic		
			Value	df	Sig.
1	.681	171	15.817	1	<.001
2	.504	.168	24.793	2	<.001
3	.396	.165	30.514	3	<.001
4	.335	.162	34.766	4	<.001
5	.104	.159	35.192	5	<.001
6	-.077	.156	35.438	6	<.001
7	-.197	.153	37.085	7	<.001
8	-.185	.150	38.612	8	<.001
9	-.218	.147	40.825	9	<.001

Table 4. Parameter estimates of I(1) models for cotton yield (kg/ha) of Hisar, Fatehabad and Sirsa districts

District (s)	Models		Estimate	Standard error	Approx. probability
	I(1)	Constant	7.08	20.32	0.73
Hisar	I(1) with	Constant	-445.90	227.22	0.06
	TMIN1	TMIN1	20.61	10.30	0.05
	I(1)	Constant	8.19	20.37	0.69
Fatehabad	I(1) with	Constant	155.33	85.41	0.08
	SSH3 and	SSH3	-2.08	1.37	0.14
	RF11	RF11	-18.93	11.66	0.11
Sirsa	I(1)	Constant	9.80	19.88	0.63
	I(1) with	Constant	132.60	66.88	0.05
	SSH4	SSH4	-19.22	10.04	0.06

Table 5. Parameter estimates of the ARIMA models for cotton yield (kg/ha) of Hisar, Fatehabad and Sirsa districts

District (s)	Models		Estimate	Standard error	Approx. probability
Hisar	ARIMA (0, 1, 1)	MA(1)	0.57	0.16	<0.01
Fatehabad	ARIMA (0, 1, 1)	MA(1)	0.55	0.16	<0.01
Sirsa	ARIMA (0, 1, 1)	MA(1)	0.49	0.17	<0.01

Table 6. Parameter estimates of ARIMAX models for cotton yield (kg/ha) of Hisar, Fatehabad and Sirsa districts

District (s)	Models		Estimate	S.E.	Sig.
Hisar	ARIMA (2, 1, 0) with	Constant	-113.09	81.93	0.18
	RH4 and RF7	AR (1)	-0.37	0.18	0.04
		AR (2)	-0.53	0.17	<0.01
		RH4	2.76	1.46	0.07
		RF7	-0.56	0.36	0.12
Fatehabad	ARIMA (2, 1, 0) with	Constant	709.85	276.13	0.01
	TMAX5	AR (1)	-0.34	0.18	0.07
		AR (2)	-0.46	0.18	0.02
		TMAX5	-18.94	7.47	0.01
Sirsa	ARIMA (0, 1, 1) with	Constant	134.70	66.66	0.05
	SSH4	MA (1)	0.49	0.17	<0.01
		SSH4	-19.48	10.35	0.07

Table 7. Diagnostic checking of residual autocorrelations of cotton yield (kg/ha) based on I(1), ARIMA and ARIMAX models for all the districts

District (s)	Models	Ljung-box Q statistic		
		Statistic	df	Sig
Hisar	I(1) with TMIN1	21.39	18	0.26
	ARIMA (0, 1, 1)	19.81	17	0.28
	ARIMAX (2, 1, 0) with RH4 and RF7	18.01	16	0.32
Fatehabad	I(1) with SSH3 and RF11	22.47	18	0.21
	ARIMA (0, 1, 1)	21.57	17	0.20
	ARIMAX (2, 1, 0) with TMAX5	24.12	16	0.09
Sirsa	I(1) with SSH4	20.04	18	0.28
	ARIMA (0, 1, 1)	19.84	17	0.28
	ARIMAX (0, 1, 1) with SSH4	19.47	16	0.30

Sirsa districts were finalized as ARIMAX models for pre-harvest cotton yield forecasting (Table 6). Marquardt (1963) algorithm was used to minimize the sum of squared residuals and Bayesian Information Criterion guided to select the final models. The residual acfs along with the associated Chi-squared test were used for the checking of random shocks to be white noise.

Comparison of the fitted models

Cotton yield forecasts and percent relative deviations for the years 2011-2012, 2012-2013, 2013-2014, 2014-2015, 2015-2016 and 2016-2017 were obtained on the basis of I(1), ARIMA and ARIMAX models (Table 8). The performance(s) of the contending models were compared on the basis of different statistics *i.e.* RMSE and MAPE as shown in Tables 9 and 10. The results indicate the preference of using ARIMAX models over I(1) and ARIMA models for cotton yield forecasting in all the three districts as given below:

It is inferred from the above results that ARIMA with weather variables as input series *i.e.* ARIMAX models consistently showed the superiority over I(1) and ARIMA models in capturing percent relative deviations pertaining to cotton yield forecasts in Hisar, Fatehabad and Sirsa districts of Haryana. The ARIMAX models performed well with lower error metrics as compared to the other models in all time regimes. In addition, the developed models are capable of providing the reliable estimates of cotton yield well in advance of the crop harvest while on the other hand, the DOA yield estimates/real time cotton yield(s) are obtained quite late after the actual harvest of the crop.

REFERENCES

Arya, P., Paul, R.A., Kumar, A., Singh, K.N., Sivaramne, N and Chaudhary, P. 2015. Predicting pest population using weather variables: An ARIMAX time series framework. *Internat. Jour. Agri. Stat. Sci.*, **11** : 381-86.

Box, G. E. P and Jenkins, G. M. 1976. *Time series analysis: Forecasting and Control.* Holden Day, San Francisco.

Ljung, G. M and Box, G. E. P. 1978. On a measure of lack of fit in time series models. *Biometrika*, **65**: 297-303.

Marquardt, D. W. 1963. An algorithm for least-squares estimation of non-linear parameters. *Jour. Soc. Indus. App. Math.*, **2**: 431-41.

Ravita and Verma, U. 2016. Application of ARIMA modeling for mustard yield prediction in Haryana. *Internat. Jour. App. Math. Stat. Sci.*, **5** : 23-28.

Yadav, N. K., Kumar, D., Nain, J and Beniwal, J. 2016. Occurrence and severity of cotton leaf curl disease in Haryana. *Internat. Jour. Agri. Sci.*, **8** : 2450-52.

Received for publication : May 10, 2021
Accepted for publication : June 16, 2021