Original Research Article   https://doi.org/10.20546/ijcmas.2018.712.339

Arimax Models for Cotton Yield Forecasting in Haryana

Alisha, Sanjeev, Puneet Verma and Urmil Verma*

*Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar-125004, India

*Corresponding author:

A B S T R A C T

Crop yield models are abstract presentation of interaction of the crop with its environment and can range from simple correlation of yield with a finite number of variables to the complex statistical models with predictive end. The pre-harvest forecasts are useful to farmers to decide in advance their future prospects and course of action. ARIMAX models have been fitted for cotton yield forecasting in Hisar, Fatehabad and Sirsa districts of Haryana. The models have been fitted using the time-series cotton yield data for the period 1980-81 to 2010-11 of Hisar and Sirsa districts and 1997-98 to 2010-11 of Fatehabad district. The fortnightly weather data have been utilized as input series from 1980-81 to 2016-17 for fitting/testing ARIMA with weather input i.e. ARIMAX models. Models have been validated for six-steps ahead i.e. 2011-12 to 2016-17, not included in the development of the models. The predictive performance of ARIMAX models were observed in terms of the percent deviations of cotton yield forecasts in relation to the observed yield(s). The ARIMAX models performed well in all time regimes for cotton yield forecasting in the districts under consideration.

Introduction

Generally, univariate Autoregressive Integrated Moving Average (ARIMA) time series models, mainly due to the contributions of Box and Jenkins (1976) 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 or may be called as dynamic regression 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 exogenous variable(s) as input. The ARIMA procedure supports seasonal, subset and factored ARIMA models; intervention or interrupted

time series models, multiple regression analysis with ARMA errors and rational transfer function models of any complexity (Pankratz, 1983). Just to mention a few work; Verma *et al.,* (2011), Chadsuthi *et al.,* (2012), Sanjeev and Verma (2016) and Arya *et al.,* (2015) etc. have used ARIMAX models for forecasting purpose.

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 (2013-14), 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. Almost the entire cotton production is concentrated in nine major cotton growing states; Punjab, Haryana, Rajasthan, Gujarat, Maharashtra, Madhya Pradesh, Karnataka, Andhra Pradesh and Tamil Nadu. In Haryana State, Hisar and Sirsa are the two major cotton producing districts, accounting for about 80% of the acreage and 86% of the cotton production in the state. Cotton is the dominant crop grown in these districts during *kharif* season and occupies almost 40% of the geographical area. It is generally sown before the onset of monsoon (May-June) and is harvested during early winter (Nov.-Dec.).

Materials and Methods

Data description and Statistical methodology adopted

The total geographical area of the Haryana state comprised of 22 districts is 44,212 sq. km. The districts Hisar, Fatehabad and Sirsa have been considered for the purpose under study. The Department of Agriculture (DOA)

cotton yield estimates for the period 1980-81 to 2016-17 were collected from Statistical Abstracts of Haryana. 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 1st 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. The time-series yield/weather data from 1980-81 to 2010-11 have been used for the training set and 2011-12 to 2016-17 have been used for the post-sample validity checking of the developed ARIMAX models.

ARIMA models with exogenous variable(s) as input

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. 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 i.e. the noise series of the system that is independent of input series

$v(B)X_t$ is impulse response function, which allows X to influence Y via distributed lag(s)

B is backshift operator, thus $v(B)X_t = (v_0 + v_1B + v_2B^2 +...) X_t$

Theoretically, the function $v(B)X_t$ has an infinite number of coefficients. Thus, this function can be written as the rational polynomial distributed lag model of finite order.

Results and Discussion

The cotton yield series 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. After experimenting with different lags of moving average and autoregressive processes; ARIMA (0,1,1) for Hisar, Fatehabad and Sirsa districts were found suitable for district-level cotton yield estimation. 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 Sirsa districts were finalized as ARIMAX models for pre-harvest cotton yield forecasting. Marquardt algorithm (1963) was used to minimize the sum of squared residuals and Bayesian Information Criterion guided to select the final models.

Table.1 Parameter estimates of ARIMAX models for cotton yield (kg/ha) of Hisar district

Model		Estimate	S.E.	Sig.
ARIMA(2,1,0) with RH4 and RF7	Constant	-113.09	81.93	0.18
	AR(1)	-0.37	0.18	0.04
	AR(2)	-0.53	0.16	<0.01
	RH4	2.76	1.46	0.07
	RF7	-0.56	0.36	0.12

Table.2 Parameter estimates of ARIMAX models for cotton yield (kg/ha) of Fatehabad district

Model		Estimate	S.E.	Sig.
ARIMA(2,1,0) with TMAX5	Constant	709.85	276.13	0.01
	AR(1)	-0.34	0.18	0.07
	AR(2)	-0.46	0.17	0.02
	TMAX5	-18.94	7.47	0.01

Table.3 Parameter estimates of ARIMAX models for cotton yield (kg/ha) of Sirsa district

Model		Estimate	S.E.	Sig.
ARIMA(0,1,1) with SSH4	Constant	134.70	66.66	0.05
	MA(1)	0.49	0.17	<0.01
	SSH4	-19.48	10.35	0.07

Table.4 District-specific estimated cotton yield(s) (est. yield) based on ARIMAX models and their associated percentage deviations (RD%)=100´(obs. yield-est. yield)/ obs. yield)

District/ Model	Year	Observed yield (kg/ha)	Estimated yield (kg/ha)	RD%
ARIMA (2,1,0) with RH4 and RF7	2011	734.63	669.34	8.89
	2012	608.28	591.73	2.72
	2013	501.39	543.40	-8.38
	2014	368.23	423.20	-14.94
	2015	276.02	240.86	12.73
	2016	623.13	677.12	-8.66
ARIMA (2,1,0) with TMAX5	2011	818.86	739.66	9.67
	2012	715.13	759.32	-6.18
	2013	766.09	720.39	5.97
	2014	653.51	678.20	-3.78
	2015	209.04	226.71	-9.52
	2016	686.03	793.03	15.60
ARIMA (0,1,1) with SSH4	2011	769.83	707.14	8.14
	2012	767.61	701.39	8.63
	2013	697.16	686.87	1.48
	2014	619.21	677.61	-9.43
	2015	295.23	260.43	11.72
	2016	**694.14**	**638.84**	**7.97**

The residual acfs along with the associated Chi-squared test (Ljung-Box, 1978) were used for the checking of random shocks to be white noise. The results obtained in this regard are expressed in Tables 1 to 4.

The level of accuracy achieved by ARIMA model(s) with weather was considered adequate for estimating cotton yield(s) i.e. ARIMA models with weather variable(s) as input could better explain the cotton yield data. Six- steps ahead (out–of-model development period i.e. 2011-12, 2012-13, 2013-14, 2014-15, 2015-16 and 2016-17) estimated values favour the use of ARIMAX models to get the short- term forecasts of cotton yield in Hisar, Fatehabad and Sirsa districts of Haryana. Summarizing the above results, it is inferred that the regression equations with an apparently high degree of fit, as measured by the coefficient of multiple correlation R^2, but with a low value of the Durbin-Watson statistic, couldn't provide adequate predictive accuracy because of the non- stationary behaviour of the data series. However, the time-series ARIMAX models consistently showed the superiority in capturing the percent relative deviations pertaining to pre-harvest cotton yield forecasts in Hisar, Fatehabad and Sirsa districts of Haryana.

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. *International Journal Agricultural and Statistical Sci*ences, 11(2), 381-386.

Box, G.E.P. and Jenkins, G. M. (1976) Time

series analysis: Forecasting and Control. *Holden Day, San Francisco*.

Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S. and Triampo, W. (2012) Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. *Asian Pacific Journal of Tropical Medicine*, 539-546.

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. *Journal of the Society for Industrial and Applied Mathematics,* 2, 431-441.

Pankratz, A. (1983) Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. *John Wiley and Sons, New York*.

Sanjeev and Verma (2016) Arima vs Arimax modeling for sugarcane yield prediction in Haryana. *International Journal of Agricultural and Statistical Sciences*, 12 (2), 327- 334.

Verma, U., Dabas, D.S., Grewal, M.S., Singh, J.P., Hooda, R. S., Yadav, M., Kalubarme, M. H., Sharma, M. P. & Prawasi, R. (2011) Crop yield forecasting in Haryana: 1986 to 2010. *Summary Report, pp.1-148, Department of Soil Science, CCS HAU, Hisar (Haryana)*.