Journal of Pharmacognosy and Phytochemistry
Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2018; 7(5): 1695-1700
Received: 18-07-2018
Accepted: 20-08-2018

Suman Verma
Ph.D. (Statistics), Department of Mathematics, Statistics and Physics CCS HAU, Hisar, Haryana, India

Correspondence
Suman Verma
Ph.D. (Statistics), Department of Mathematics, Statistics and Physics CCS HAU, Hisar, Haryana, India

Modeling and forecasting maize yield of India using ARIMA and state space models

Suman Verma

Abstract
Time series modeling using Autoregressive Integrated Moving Average (ARIMA) and state space (SS) models, was developed for individual univariate series of maize yield in India. In ARIMA modeling, the underlying parameters are assumed to be constant however the data in agriculture are generally collected over time and thus have the time-dependency in parameters. Such data can be analyzed using SS procedures by the application of Kalman filtering technique. The aim of this study was to evaluate univariate time series methods to forecast the maize yield in India. ARIMA (0, 1, 1) model was found to be appropriate but the SS model with lower error metrics showed the superiority over ARIMA model for this empirical study. The performances of the models were validated by comparing with the observed values.

Keywords: Autocorrelation function, Kalman filtering technique, State space modeling, Akaike's information criterion, Maize yield forecast

Introduction
Agriculture plays a crucial role in the overall growth of any country and so it is necessary to ensure its development. For example, the major population of India is working as farmers accounting for around 16% of the total GDP. This ratio is enough to say that India is highly dependent on its agriculture as a huge amount of land is used for it. Maize is one of the most important cereal crops of the world and contributes to food security in most of the developing countries. It is grown in 70 countries of the world. The major maize growing countries are USA, China, Brazil, Mexico, Indonesia, India, France and Argentina. India is at 6th position in maize production and fifteenth position in its productivity in the World.
In India, maize is emerging as third most important crop after rice and wheat and is grown over 4 per cent of the net area sown of the country. The major maize producing states are Karnataka, Andhra Pradesh, Madhya Pradesh, Bihar, Rajasthan, Tamil Nadu, Telangana and Uttar Pradesh. In India its importance lies in the fact that it is not only used for human food and animal feed but at the same time it is also widely used for corn starch industry, corn oil production and baby corns etc. The increasing use of maize as feed, increasing interest of the consumers in nutritionally enriched products and rising demand for maize seed are the core driving forces behind emerging importance of maize crop in India.
Maize is mainly a rainfed kharif crop which is sown just before the onset of monsoon and is harvested after retreat of the monsoon. However, despite the production strength, Indian corn yields are significantly below the yields in major corn producing countries. There is immense scope for an increase in India's corn production by increasing area under hybrids, adoption of better genetics and improved agronomic practices.
ARIMA time series models could be regarded as means of transforming the data to white noise that is to an uncorrelated sequence of errors. ARIMA models are widely used in practice for forecasting, mainly due to the contributions of Box and Jenkins (1976) [6]. Pindyck and Rubinfeld (1981) [18] and Makridakis *et al.* (1982) [12] have also emphasised the use of ARIMA models. Badmus and Ariyo (2011) [5] focused on forecasting the cultivated area and production of maize in Nigeria using ARIMA model. Further ARIMA modeling technique have been employed by Verma *et al.* (2011) [22] for wheat, sugarcane, cotton and mustard crops operational yield for forecasting purpose in Haryana, Mishra *et al.* (2014) [14] to analyse and forecast fertilizer statistics in India and Ali *et al.* (2015) [3] for forecasting the production and yield of sugarcane crop in Pakistan. Recent developments in time series modeling offer further scope in improving these models and also for developing ARIMA models under multivariate framework.
At the national level, not much work has been done on SS modeling in the field of agriculture.

SS models are time varying parameters models as they allow for known changes in the structure of the system over time. All Box-Jenkins models can be restated in the SS form. The SS model originated in the field of engineering (Kalman, 1960) [10] and was later applied into economics by Rosenberg (1973) [19]. Expositions of the state space approach to multivariate forecasting can be found in studies of Akaike (1976) [1], Kitagawa and Gersh (1984) [11], Aksu and Narayan (1990) [2] and Hyndman *et al*. (2002) etc. A good account on SS modeling is given by Many authors (Durbin, 2002; Piepho and Ogutu, 2007; Yusof and Kane, 2012; Yemitan and Shittu 2015; Omekara *et al*., 2016, Suman *et al*., 2017) [7, 17, 24, 23, 15, 21] and in the book by Aoki (1987) [4]. Keeping in view the above subject matter, the maize yield estimates of India have been obtained with the emphasis to compare the forecasting performance of the models developed on the basis of ARIMA and SS procedures.

Data and Methodology

This study is based on secondary data of cash crop for forecasting purpose. The yearly yield (kg/hc) data of maize crop have been taken from Food and Agriculture Organization of India (FAOSTAT) for modeling and forecasting the maize yield in India. The emphasis has been given in predicting the future value(s) on the basis of previous time series observations. Data from 1961 to 2011 were used for model building and 2012 to 2016 were used to check the forecasting performance of the model. The main objective of this study was to compare the both in-sample and out-of-sample forecasts of maize yield obtained by using ARIMA and SS modeling techniques.

Box-Jenkins Autoregressive Integrated Moving Average methodology

The existing study applies Box-Jenkins ARIMA modeling technique, which is an extrapolation method for forecasting. It requires historical time series data of underlying variable and applicable to both discrete data as well as continuous data. However, the data should be available at equally spaced discrete time intervals. The data has to be made stationary in order to choose an appropriate ARIMA model for forecasting. One of the simplest transformations called ‘differencing’ can be applied when the mean of a series changes over time and log transformation is used when the variance of a series changes through time. The main stages in setting up a Box-Jenkins forecasting model are: Identification, Estimating the parameters, Diagnostic checking and Forecasting.

The stationarity of the data series can be tested both through graphics and other formal techniques i.e. Autocorrelation Function (Acf) and Partial Autocorrelation Function (Pacf), Augmented Dickey-Fuller test (ADF) of unit root and Kwiatkowski– Phillips–Schmidt–Shin (KPSS) unit root test. By considering the patterns of the Acfs and the Pacfs, we can guess a reasonable model for the data. The general functional form of ARIMA model i.e. ARIMA (p, d, q) used for the present study is expressed as:

$\phi_p(B)\, \Delta^d\, Y_t = c' + \theta_q(B)\, e_t$, where $c' = 0$ if Y_t was adjusted for its mean

where Y = Variable under forecasting, t = time subscript, B = Lag operator, e = Error term $(Y - \hat{Y})$, where $\hat{Y}$ is the estimated value of Y, $\phi_p(B)$ = non-seasonal AR process, $(1\text{-}B)^d$ = non-seasonal difference, $\theta_q(B)$ = non-seasonal MA process, ϕ’s and θ’s = the parameters to be estimated (Pankratz, 1991) [16].

Further, at the estimation stage, an attempt was made to obtain the precise estimates of a small number of parameters of the model. Linear least-squares can be used to estimate only pure auto regressive models and non-linear least squares (NLS) method for all other models. Furthermore, the diagnostic tests were performed to check if the random shocks were independent or not. The residuals were analyzed using Box-Ljung Statistic and “Histogram with normal plot”. The accuracy of post-sample forecasts were tested using the following tests such as Relative Deviation in percentage (RD %) and Root Mean Square Error (RMSE).

The state space model

SS modeling consists of a measurement (observation) equation and a state (transition) equation where the state equation formulates the dynamics of the state variables while the measurement equation relates the observed variables to the unobserved state vector. Let $\mathbf{y}_t$ be the $r \times 1$ vector of observed variables after differencing if needed and subtracting the sample mean. Let $\mathbf{z}_t$ be the state vector of dimension s, $s \geq r$, where the first r components of $\mathbf{z}_t$ consist of $\mathbf{y}_t$. Various forms of the SS model are in use but the model fitted with the help of STATESPACE procedure in SAS for this study is based on Akaike (1976) [1]. The SS model defined by the state transition equation is

$$z_{t+1} = F\, z_t + G\, e_{t+1}$$

where, $\mathbf{z}_t$ is a state vector of dimension s, whose first r elements are $\mathbf{y}_t$ and whose last s-r elements are conditional prediction of future $\mathbf{y}_t$. $\mathbf{F}$ is an $s \times s$ transition matrix and $\mathbf{G}$ is an $s \times r$ input matrix; for model identification, the first r rows and r columns of $\mathbf{G}$ are set to an $r \times r$ identity matrix. $\mathbf{e}_t$ is a sequence of independent normally distributed random vectors of dimension r with mean $\mathbf{0}$ and covariance matrix Σ_{ee}.

In addition to the state transition equation, SS models usually include a measurement or observation equation that gives the observed values $\mathbf{y}_t$ as a function of the state vector $\mathbf{z}_t$.

The measurement equation used by the STATESPACE procedure is

$y_t = H\, z_t$, $H = [I_r\, 0]$ and I_r is an $r \times r$ identity matrix and 0 is an $\{r \times (s\text{-}r)\}$ zero matrix.

The methods used by the SS procedure also assume the input series to be stationary. Therefore, the first step is to examine the data and test the requirement of differencing. SS procedure employs canonical correlation analysis for the identification of SS model. The identification of the canonical SS model is accomplished in two steps. The first step involves the determination of the amount of past information to be used in the canonical correlation analysis. This is achieved by fitting successively higher order vector autoregressive (VAR) models and computing Akaike information criterion (AIC) for each fitted model. The optimum lag (p) into the past is chosen as the order of VAR model for which AIC is minimum.

The second step involves the selection of state vector via canonical correlation analysis between the set of present and past values and the set of present and future values. The canonical correlation coefficients are computed for the sample covariance matrices of the set of successively increasing number of present and future values and the fixed set of present and past values. If the smallest canonical correlation coefficient of the sample covariance matrix that corresponds to the component being evaluated for inclusion in the state vector is non-zero, then that particular component is included in the state vector. Once the state vector is determined, the SS model is fitted to the data. The parameters in F, G and Σ_{ee} are estimated using maximum likelihood procedure.

The SS forecasts are obtained through the Kalman filtering (Harvey, 1984) [8]. Kalman filter (that updates the knowledge of the system each time a new observation is brought) minimizes the error terms. The m-step ahead forecast of z_{t+m} i.e. $z_{t+m/t}$ denotes the conditional expectation of $z_{t+m/t}$ given the information available at time t i.e.

$$y_{t+m/t} = H\, z_{t+m/t}$$

where the matrix **H= [I$_r$ 0]**

The m-step ahead forecast error is

$$z_{t+m} - z_{t+m/t} = \sum_{i=0}^{m-1} \Psi_i\, e_{t+m-1}$$

and its variance is

$$V_{z,m} = \sum_{i=0}^{m-1} \Psi_i\, \Sigma_{ee} \Psi_i'$$

Letting $V_{z,0} = 0$, $V_{z,m}$ can be computed recursively using Kalman filter as

$$V_{z,m} = V_{z,m-1} + \Psi_{m-1} \Sigma_{ee} \Psi_{m-1}'$$

Thus, the variance of m-step ahead forecast error of y_{t+m} obtained is

$$V_{y,m} = H\, V_{z,m} H'$$

Results and Discussion

The most common method to check stationarity through examining the Acfs and Pacfs graphs of maize yield shown in Figure 1 indicates that the data series was non-stationary. It may be observed from Figure 2 that differencing of order one was sufficient for making an appropriate stationary series.

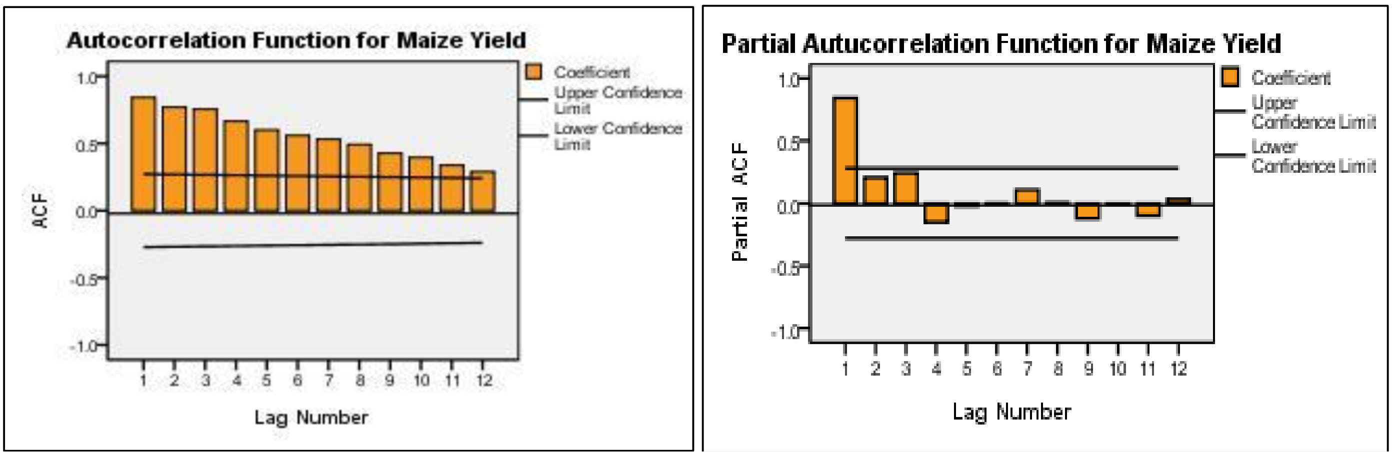

Fig 1: Acf and Pacf for maize yield (with 5% significance limits for autocorrelations)

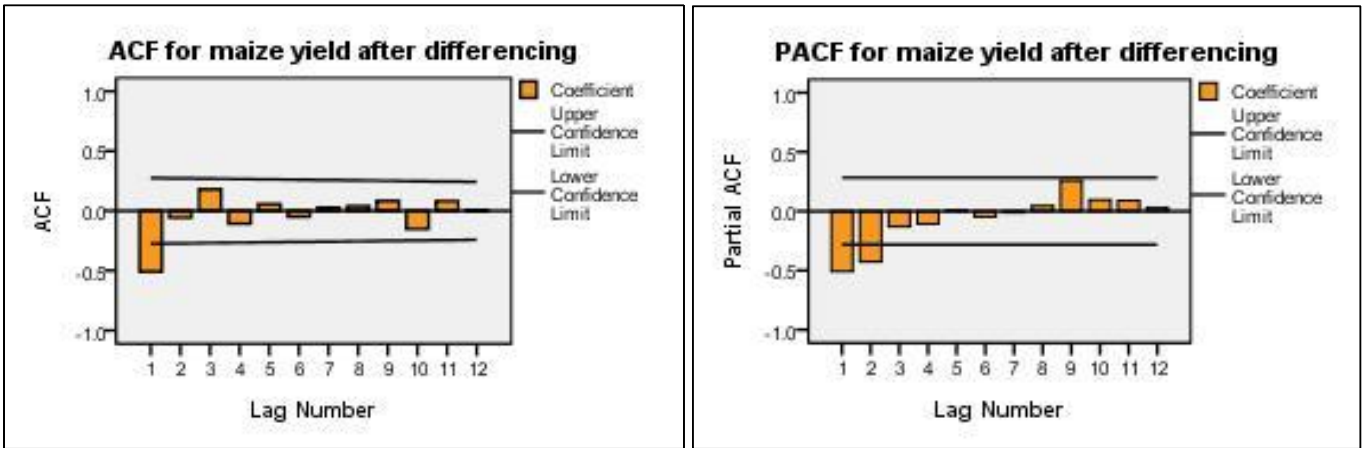

Fig 2: Acf and Pacf for maize yield after differencing (with 5% significance limits for autocorrelations)

Besides the graphical method two formal tests were used to check the stationarity condition for the data series, ADF and KPSS unit root tests. The null hypothesis (H0) in the ADF-test was that the time series data is non-stationary while alternative hypothesis (Ha) was the series is stationary. But for KPSS unit root test the null hypothesis (H0) was that the series is stationary against the alternative (Ha) of non-stationary data series. The hypothesis was then tested for original data and by performing first differencing. The ADF test with at the usual 5% level of significance adequately declared that the data series is stationary after first differencing and suggest that there is no unit root. The KPSS test of the data was unable to reject the null hypothesis of stationarity after applying the first differencing on the data series. The results, as obtained are shown below (Table 1):

Table 1: Results of ADF and KPSS tests.

Test	Test statistic	Lag order	p-value	Decision
ADF at level	-1.225	3	0.89	Data Non-stationary
ADF at first difference	-5.682	3	0.01	Data Stationary
KPSS level	2.389	1	0.01	Data Non-stationary
KPSS at first difference	0.116	1	0.10	Data Stationary

The appropriate orders of AR and MA polynomials i.e. the values of p and q were determined from the Acfs and Pacfs of the stationary series. Marquardt algorithm (1963) [13] was used to minimize the sum of squared residuals. Log Likelihood, Schwarz's Bayesian Criterion (1978) and residual variance decided the criteria for the selection/estimation of AR and MA coefficients in the model. After experimenting with different lags of the moving average and the autoregressive processes, ARIMA (0, 1, 1) was fitted for obtaining maize yield forecasts.

The fitted ARIMA (0, 1, 1) model may be elaborated as below:

$$Y_t = Y_{t-1} - \theta_1 e_{t-1} + e_t$$

The presence of lagged values of dependent variable and random shocks in above equation indicates the presence of autoregressive and moving average components both. The parameter estimates of fitted ARIMA models were presented in Table 2. Ljung-Box tests statistics for this model was greater than 0.05 at 5% level of significance and strongly suggested to accept that there is no autocorrelation among the residuals of the fitted ARIMA (0, 1, 1) model (Table 3**)**. Also, here "Histogram with Normal Curve" was used to check the normality assumption for the residuals of the fitted model. The curve represented in Figure 3 suggests to accept the normality assumption that the residuals of the fitted ARIMA (0, 1, 1) model are normally distributed. Therefore, it is clear that the fitted ARIMA (0, 1, 1) model is the best fitted model and adequately used to forecast the maize yield in India.

Table 2: Parameter estimates of fitted ARIMA models.

Model		Estimate	Standard Error	p-value
ARIMA (0,1,1)	Constant	29.09	7.24	<0.01
	Difference	1		
	MA lag 1	0.69	0.12	<0.01

Table 3: Diagnostic checking of residual autocorrelations of maize yield.

Model	Ljung-Box Q statistic(s)	
	Statistic	Sig.
ARIMA (0, 1, 1)	9.41	0.93

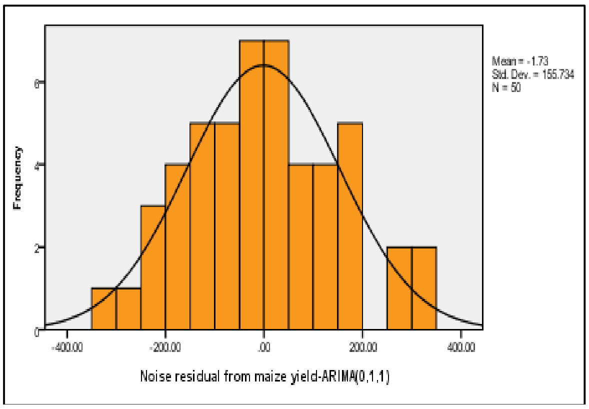

Fig 3: Histogram of residuals from maize yield ARIMA (0, 1, 1) model with Normal Curve.

The SS model assumes that the time series is stationary. Hence, the data was checked for stationarity. Here, y_t, the $r \times 1$ vector of observed variables after differencing and subtracting the sample mean from Y_t, can be expressed as follows:

$$y_t = (1-B)\, Y_t - 31.35$$

The smallest AIC value, in this case is 517.68 at lag 2, determines the number of autocovariance matrices analyzed in the canonical correlation phase. Next, the Yule-Walker estimates of the selected AR model was obtained as Lag1=-0.721 and Lag2=-0.414. After the autoregressive order selection process of determining the number of lags used in canonical correlation analysis, the state vector was selected. Information from the canonical correlation and preliminary autoregression analyses were used to form the preliminary parameter estimates of state space models as shown in Table 4.

Table 4: Parameter estimates of the state space models.

Parameter	Estimate	Standard Error	t Value
F(2,1)	-0.268	0.206	-1.30
F(2,2)	-0.437	0.286	-1.53
G(2,1)	-0.786	0.140	-5.64

The fitted state space model for maize yield can be elaborated as:

$$\begin{bmatrix} y_{t+1} \\ y_{t+2/t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -0.268 & -0.437 \end{bmatrix} \begin{bmatrix} y_t \\ y_{t+1/t} \end{bmatrix} + \begin{bmatrix} 1 \\ -0.786 \end{bmatrix} (23180.83)$$

The two models were compared for their in-sample (Table 5 and Figure 4) as well as their out-of-sample forecast (Figure 5) performance of maize yield in India. The first model was ARIMA (0, 1, 1) identified to fit the yearly maize yield data series and the second model was the SS mode which captures the time varying characteristics of the data series. The predictive performance(s) of the contending models were observed in terms of percent deviations of maize yield forecasts in relation to observed yield(s) and root mean square error(s) as well. The level of accuracy achieved by SS model was considered adequate for estimating the maize yield(s) i.e. the SS model consistently showed the superiority over ARIMA model in capturing percent relative deviations pertaining to maize yield forecasts in India.

Table 5: Maize yield estimates and their associated percent relative deviations based on ARIMA and SS models.

Year	Arima			State Space		
	Obs. Yield (Kg/ha)	Est. yield (Kg/ha)	RD (%)	Obs. Yield (Kg/ha)	Est. yield (Kg/ha)	RD (%)
2012	2555.7	2410.02	5.70	2555.7	2488.21	2.64
2013	2572.6	2439.11	5.19	2572.6	2533.39	1.52
2014	2610.7	2468.21	5.46	2610.7	2575.91	1.33
2015	2597.2	2497.31	3.85	2597.2	2616.74	-0.75
2016	2574.5	2426.4	5.75	2574.5	2640.96	-2.58
RMSE	135.10			49.20		

RD% = 100×((Observed yield–Estimated yield)/ Observed yield)

$$RMSE = \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} (O_i - E_i)^2 \right\} \right]^{\frac{1}{2}},$$

where O_i and E_i were the observed and forecast yield(s) and 'n' being the number of forecast years.

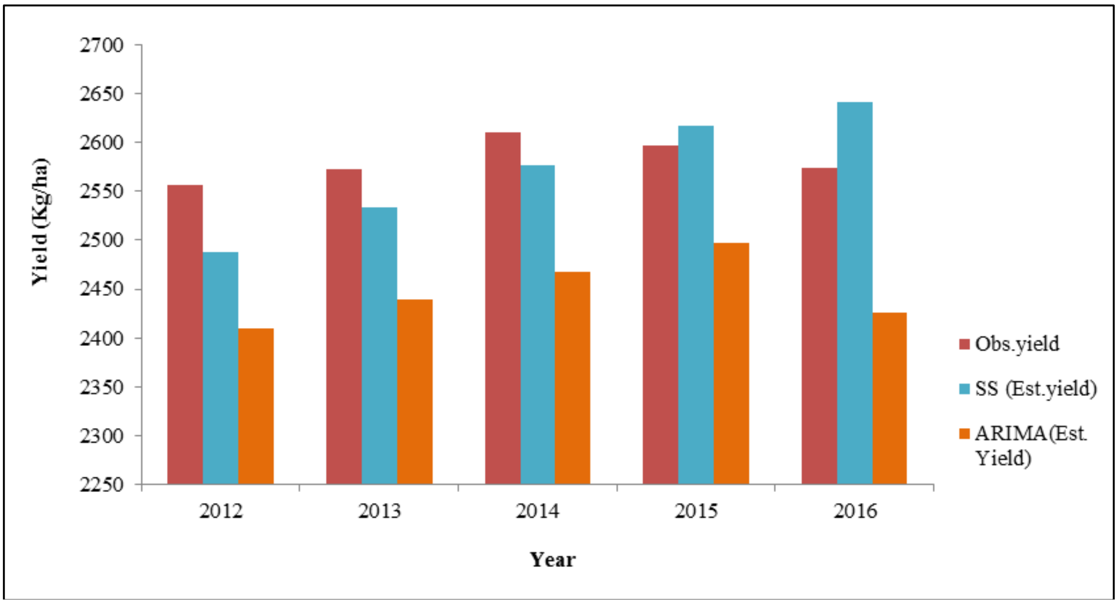

Fig 4: Maize yield estimates along with observed yields based on ARIMA and SS modelling

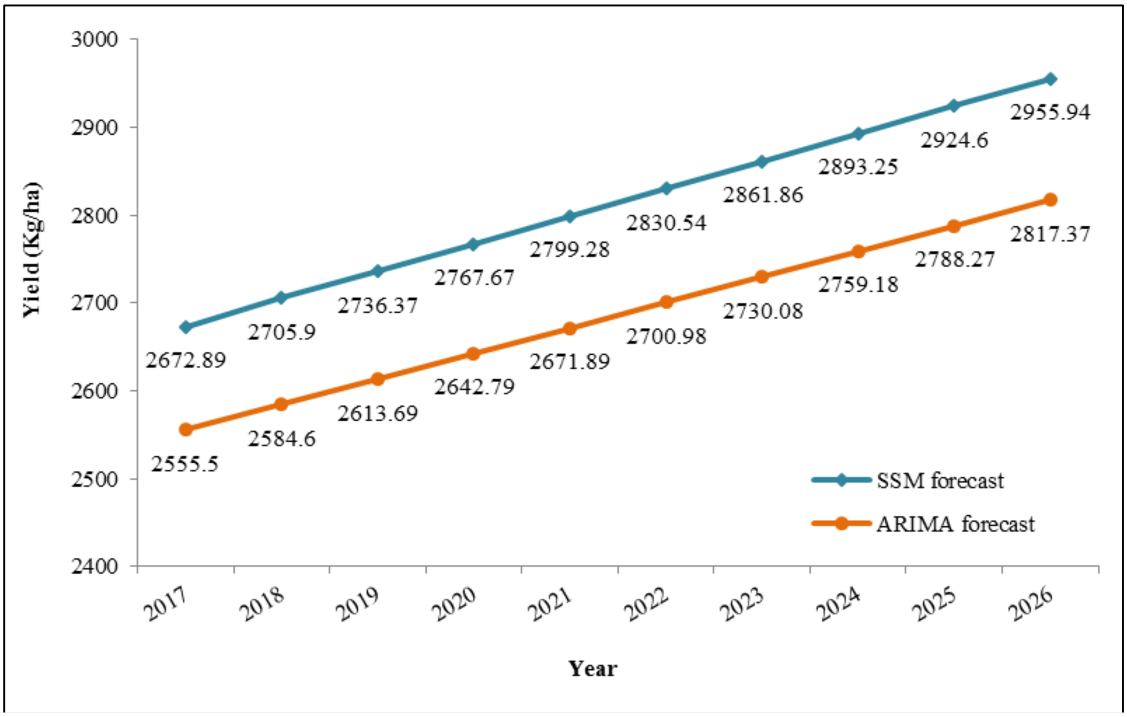

Fig 5: A comparative view of ARIMA and SS modeling for out-of-sample forecast performances

Conclusion

In this study time series analysis of maize yield data from the period of (1961-2016) was carried out. The comparison of the ARIMA and SS approach has been focused from a statistical point of view. The forecasting performance(s) of the contending models were observed in terms of the percent deviations of maize yield forecasts in relation to the observed yield (s) and root mean square error(s) as well. Both the models individually could provide the suitable relationship(s) to reliably estimate the maize yield and found to be stable in nature for both in-sample and out–of-sample forecasts. Comparison between the estimated ARIMA (0, 1, 1) and SS model was made and the result confirmed SS model to be more adequate. Therefore, the parameters being time-dependent, the state space modeling may be effectively used, as it can take into account the time dependency of the underlying parameters which may further enhance the predictive accuracy of the forecast models.

References

1. Akaike H. Canonical correlations analysis of time series and the use of an information criterion in advances and case studies in system identification (R. Mehra and D.G. Lainiotis (Eds.)). Academic Press, New York, 1976.
2. Aksu C, Narayan JY. Forecasting with vector ARMA and state space methods. Working paper (Temple University, Philadelphia, PA), 1990.
3. Ali S, Badar N, Fatima H. Forecasting production and yield of sugarcane and cotton crops of Pakistan for 2013-2030. Research Article. 2015; 31(1):1.
4. Aoki M. State space modeling of time series. Springer. Berlin, 1987.
5. Badmus MA, Ariyo OS. Forecasting cultivated areas and production of maize in Nigeria using ARIMA model. Asian Journal of Agricultural Sciences. 2011; 3(3):171-176.

6. Box GEP, Jenkins GM. Time series analysis: Forecasting and Control. Holden Day. San Francisco, 1996.
7. Durbin J. The Foreman Lecture: The State Space approach to time series analysis and its potential for Official Statistics. Austral. New Zealand J Statist. 2002; 42:1-23.
8. Harvey AC. A unified view of statistical forecasting procedures. J Forecasting. 1984; 3:245-275.
9. Hyndman RJ, Koehlerb AB, Snydera RD, Grosea S. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting. 2000; 18:439-454.
10. Kalman RE. A new approach to linear filtering and prediction problem. J Basic Engineering. 1960; 82(D):35-45.
11. Kitagawa G, Gersch W. A smoothness priors-state space modeling of time series with trend and seasonality. J Amer. Statist. Assoc. 1984; 79:378-389.
12. Makridakis S, Anderson A, Fildes R, Hibon M, Lewandowski R, Newton J *et al*. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. J Forecasting. 1982; 1:111-153.
13. Marquardt DW. An algorithm for least-squares estimation of non-linear parameters. J Soc. Ind. Appl. Math. 1963; 2:431-441.
14. Mishra P, Sahu PK, Uday JPS. ARIMA modelling technique in analyzing and forecasting fertilizer statistics in India. Trends in Biosciences. 2014; 7(3):170-176.
15. Omekara CO, Okereke OE, Ehighibe SE. Time series analysis of interest rate in Nigeria: A com-parison of Arima and state space models. International Journal of Probability and Statistics. 2016; 5(2):33-47.
16. Pankratz A. Forecasting with dynamic regression models. Wiley-Interscience, 1991.
17. Piepho HP, Ogutu JO. Simple state-space models in a mixed model framework. Amer. Statist. 2007; 61:224-232.
18. Pindyck RS, Rubinfeld DL. Econometric Models and Economic Forecasts. McGraw Hill Book Co, Inc. NY, USA, 1981.
19. Rosenberg B. The analysis of a cross-section of time series by stochastically convergent parameter regression. Ann.Eco. Social Measurement. 1973; 2:399-428.
20. Schwarz G. Estimating the dimension of a model, Ann. Stat. 1978; 62:461-464.
21. Suman, Verma U. State space modeling and forecasting of sugarcane yield in Haryana, India. Journal of Applied and Natural Science. 2017; 9(4):2036-2042
22. Verma U, Dabas DS, Grewal MS, Singh JP, Hooda RS, Yadav M *et al*. Crop yield forecasting in Haryana: 1986 to 2010, Summary Report, Department of Soil Science, CCS HAU, Hisar (Haryana). 2011, 1-148.
23. Yemitan RA, Shittu OI. Forecasting Inflation in Nigeria by state space modeling. International Journal of Scientific & Engineering Research. 2015; 6:778-786.
24. Yusof F, Kane IL. Modelling monthly rainfall time series using ETS state space and SARIMA models. International Journal of Current Research. 2012; 4:195-200.