Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Sch. J. Phys. Math. Stat. 2015; 2(2B):196-204 **ISSN 2393-8064 (Online)**
©Scholars Academic and Scientific Publishers (SAS Publishers)
(An International Publisher for Academic and Scientific Resources)

Application of Time Series Model for predicting Future adoption of sugarcane variety: KEN 83-737

Ong'ala J.O, Mwanga, D.M.
Dept of Economics and Biometrics, Sugar Research Institute, Kenya

***Corresponding Author:**
Ong'ala J.O
Email: ongalajacob@gmail.com

Abstract: Both exponential smoothing and Box-Jenkins' ARIMA models are used in this study as time series modeling approaches to forecast sugarcane variety adoption in Kenya. The accuracy of the two methods are assessed and ARIMA (4,1,1) was found to be the best model to estimate the future prediction of adoption status. Efforts were made to forecast the future adoption of sugarcane variety (KEN 83-737) for two years by fitting ARIMA(4,1,1) model to our time series data. The results indicated a predicted drop in adoption of KEN 83-737 in 2012 and 2013.

Keywords: Exponential smoothing, ARIMA, Sugarcane, Forecasting, Time series**.**

INTRODUCTION

Sugarcane growing was introduced in Kenya in the early 1900's by the Indian laborers engaged in the construction of the Uganda Railway [1]. The Sugar Industry has grown rapidly in Kenya making sugarcane to play a significant role in socio-economic development of the Kenyan economy [2]. The establishment of the Kenya Sugar Research Foundation (KESREF) in 2002 contributed to the growth of the Kenya Sugar Industry. Since the inception of KESREF, more than 21 improved varieties have been generated and released for commercial production; the latest release were in 2014[3]

The aim of generating and releasing improved varieties is to increase performance and enhance productivity of sugarcane given the same environmental conditions. The improved sugarcane varieties in Kenya have a nomenclature with prefix "KEN" to mean Kenyan series. Their major attributes are: high yielding, early maturing, disease, pest and drought resistant and high sucrose. These attributes give high expectation if the KEN series at their release.

Earlier studies done on adoption of improved sugarcane varieties revealed that the most adopted improved sugarcane variety in Nyando Sugar zone was KEN 82-808 though the area under the improved cane varieties was in the decline [4]. On the other hand, a web based analysis model developed by Ong'ala *et. al* [5] indicated that the adoption of the KEN83-737 (released in 2002) has been on average the highest in the Kenya Sugar Industry among the improved varieties despite the high number of the varieties developed so far in Kenya. In this paper we assume that adoption is measured by the area in hectares of land covered by the variety.

The adoption of KEN 83-737 seems to be increasingly having an upward trend. Based on unpublished raw data of a baseline study by Ong'ala *et. al* [6], Kwale International Sugar Company Ltd (KISCOL) based in the Coastal region of Kenya has adopted the KEN 83-737 more than 50% area coverage.

Currently statistical techniques of time series analysis have been widely disseminated in the literature and there is a great variety of circumstances of researches in which they can be used, especially in studies involving time dependent data. In this paper therefore, an effort is made to forecast the adoption status of the KEN 83 - 737 for the four leading years. The model developed here for forecasting are; exponential smoothing [7] and an Autoregressive Integrated Moving Average (ARIMA) model which was introduced by Box and Jenkins in 1960 hence the name Box-Jenkins Model.

Box-Jenkins Models

A time series can be understood as a sequence of data at regular time intervals during a specified period. When analyzing time series, one must first model the studied phenomenon, which from this point can be made describing the behavior of the series, their estimates and finally the evaluation of the factors that influence the behavior of the series, taking into view to establishing the cause and effect relationship.

According to Raymond Y.C. Tse, [8], the sequence for determining the model for the ARIMA family that best represents the series that can be used to make predictions is: model identification, parameter estimation and testing for model validity after which a forecast can be done. It is important to note that for the application of the models of Box-Jenkins, the time series under study must be stationary that is, not present trend or seasonality [8].

The model proposed by Box *et al* [9] , which will be highlighted in this paper is (1):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \qquad (1)$$

Where, α_0 represents a constant in the estimated model, α_1 to α_p are parameters that adjust the past values of y_t from the immediately prior time to the farthest represented by p. The values of ε represent a sequence of random shocks and independent of each other, ε_t is a non-controlled portion of the model, is commonly referred to as white noise. The parameters β_1 to β_q are used to write the series as a function of past shocks. In general each ε_t is considered to have normal distribution, zero mean, constant variance and non-correlation.

Model (1) is stationary if for every t and $t-s$: (i) $E(y_t) = E(y_{t-s}) = \mu$ (constant mean), (ii) $E(y_t - \mu)^2 = E(y_{t-s} - \mu)^2 = \sigma_y^2$ (constant variance), and
(iii) $E((y_t - \mu)(y_{t-s} - \mu)) = E\big((y_{t-j} - \mu)(y_{t-j-s} - \mu)\big) = \gamma_y$ (constant covariance), otherwise if not stationary, (1) has to be transformed to stationary using differentiation through the use of the operator Δ defined by $\Delta_j^d = (1 - L^j)^d$ and $L^j y_t = y_{t-j}$ (backward shift operator). Refer to [9] for the difference version for model (1).

Time series patterns can be due to; trend, seasonality and random effect. These three components have to be identified and decomposed and assess their effects on the time series. In terms of the three components, the time series y_t can be written as

$$y_t = S_t + T_t + E_t \qquad (2)$$

where y_t is the data at period t, S_t is the seasonal component at period t, T_t is the trend component at period t and E_t is the remainder (or irregular or error) component at period t.

When forecasting was introduced as a subject of interest, the method used most widely was the exponential smoothing method which was applicable in Business [10]. These exponential smoothing methods still live on today. Makridakis et.al [11] investigated the predictive ability of various methods of time series forecasting and reported that the exponential smoothing is viable, however generate correlation prediction errors that compromise the long term prediction. However Oliveira *et al* [12] stated that the simple forecasting methods can provide very satisfactory predictions under certain conditions and that adoption of a more complex method should be investigated.

Later, more advanced methods taking seasonality and trend into account were brought forward in the 60's and 70's [10]. As managers later understood that actions such as promotional activities, competitor action and product introduction would shape and create demand, these variables needed to be understood and incorporated into the forecasts. One method to incorporate explanatory variables was the ARIMA-model [10] .With the introduction of computers, more advanced forecasting measures has emerged[13]. The application of ARIMA models started as early as 1970s when the model was developed. To date numerous models have been fitted and used for forecasting in a wide range of fields finance, manufacturing and agriculture production. A few examples are discussed here.

Meyler, et al. [14] considered two alternative approaches of identifying ARIMA models; the Box Jenkins approach and the objective penalty function methods. Their emphasis was on forecast performance that suggested more focus on minimizing out-of-sample forecast errors than on maximizing in-sample 'goodness of fit'. Thus, the approach followed is unashamedly one of 'model mining' with the aim of optimizing forecast performance.

Stergiou [15] analyzed as 17-year record (1964–1980, 204 observations) of monthly catches of pilchard (*Sardina pilchardus*) from Greek waters using Auto Regressive Integrated Moving Average (ARIMA) techniques. Two models were found to be suitable for describing the dynamics of the fishery and for forecasting up to 12 months ahead. He compared his forecasts with the actual data for 1981 that were not used in the estimation of the parameters of either model. The results showed that the mean error was 14.6% and 12 % for the two models respectively. His results suggested that ARIMA procedures are capable of describing and forecasting the complex dynamics of the Greek pilchard fishery, which have hitherto been regarded as difficult to predict owing to the strong influence of year-to-year changes in oceanographic and biological conditions and socio-economic factors (low commercial value and demand, high discard rate).

Kumar *et al*[16] in their paper used the time series ARIMA model technique to predict sugarcane production in India using a 62 sugarcane production data from 1950 to 2012. The model was able to predict an increase in production for the year and 2013 then a fall in 2014 and subsequent year up to 2017. Other studies that have used ARIMA model in fitting and forecasting include work done by; Kaur & Dham [17], Findley *et al* [18] and Han P *et al*[19].

MATERIALS AND METHODS

The data for sugarcane variety adoption status was collected from July 2003 to December 2012 on a quarterly basis (see Table 1). The data was captured using a pre designed questionnaires which was updated from time to time for convenient though not significant to change the data collected. The respondents in this study were the sugarcane milling factory (in this case 11 factories viz; Mumias, Nzoia, West Kenya, Butali, Kibos, Chemilil, Muhoroni, SONY, Transmara, Sukari and KISCOL). Additional data was collected from randomly selected farmers around the sugar factory to validate the data collected from the factory.

The adoption status for KEN 83-737 was then extracted for all the quarters indicated. However, the data for some quarters were missing leaving the data with some gap. The gaps were filled and the data interpolated into monthly using an R software package called '*zoo*'[20] . The zoo package ensures that the gaps in a data set are filled and increases the number of observations without interfering with the trend behavior. Appropriate ARIMA (p,d,q) model was then identified for the data, fitted for the data upto 2011. The data remaining for the one year (2012) will be used to check the adequacy of the forecast.

RESULTS AND DISCUSSION

Seasonal Effect on Time series forecasting

Viewing the time series data as described in Equation (2), the results shown in figure 2 were obtained.

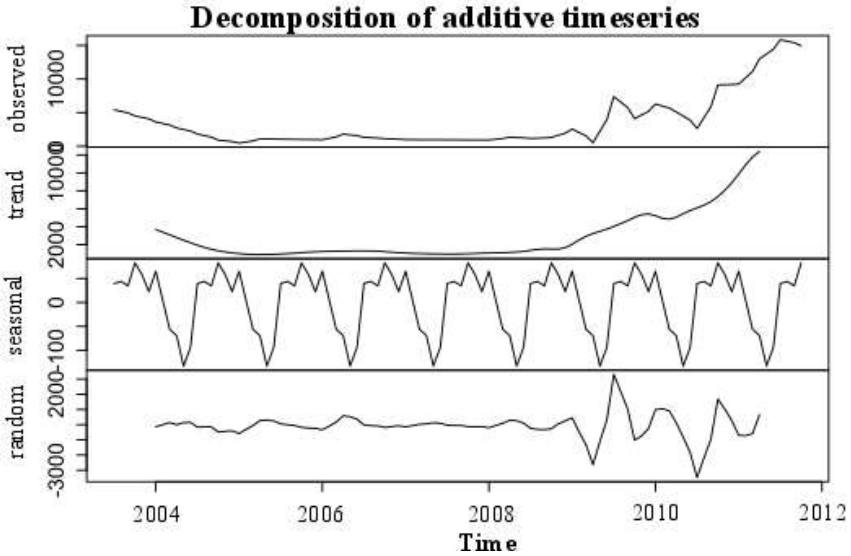

Fig- 1: Decomposition of additive time series using classical approach

Notice that the seasonal component changes very slowly over time (Fig- 1), so that for the consecutive years very similar pattern are seen, but years far apart may probably have different seasonal patterns. The remainder component shown in the bottom panel is the random effect. The random effect experienced more between 2009 and 2011.

When the decomposition is done using the STL method, better time series plots can be obtained. , STL is a very versatile and robust method for decomposing time series. STL is an acronym for "Seasonal and Trend decomposition using Loess", while Loess is a method for estimating nonlinear relationships. The STL method was developed by Cleveland et al. [21] .

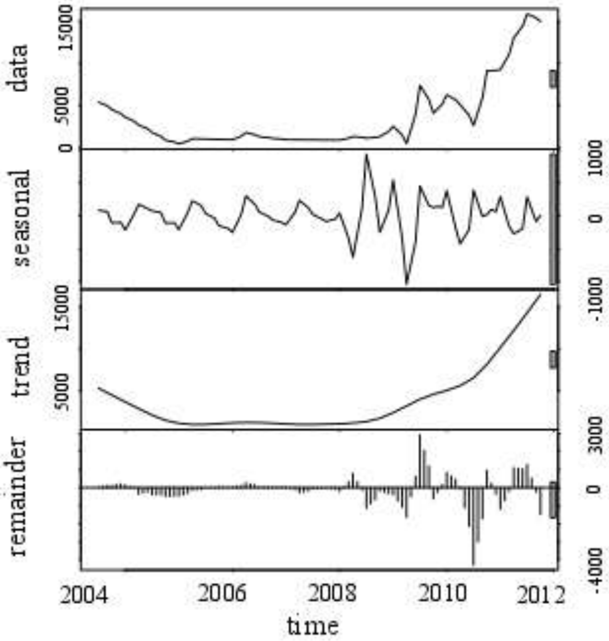

Fig-2: Decomposition of additive time series using STL approach

The grey bars to the right of each panel show the relative scales of the components (Fig-2). Each grey bar represents the same length but because the plots are on different scales, the bars vary in size. The large grey bar in the second panel shows that the variation in the seasonal component is small compared to the variation in the data and trend which has a bar about one quarter the, hence the seasonal variation can be ignored.

The seasonally adjusted time series has almost the same plot as the original data (**Fig-3**). Using a Kolmogorov -Sminov test (D = 0.09, p-value = 0.8127), there is no significant difference between the seasonally adjusted and origin al data.

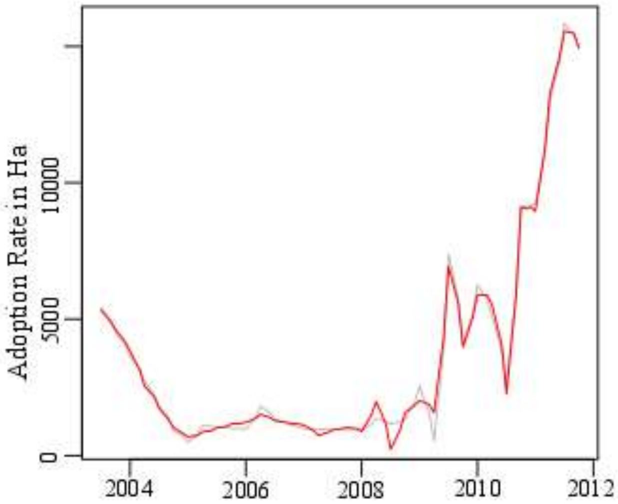

Fig-3: Season adjusted Time series

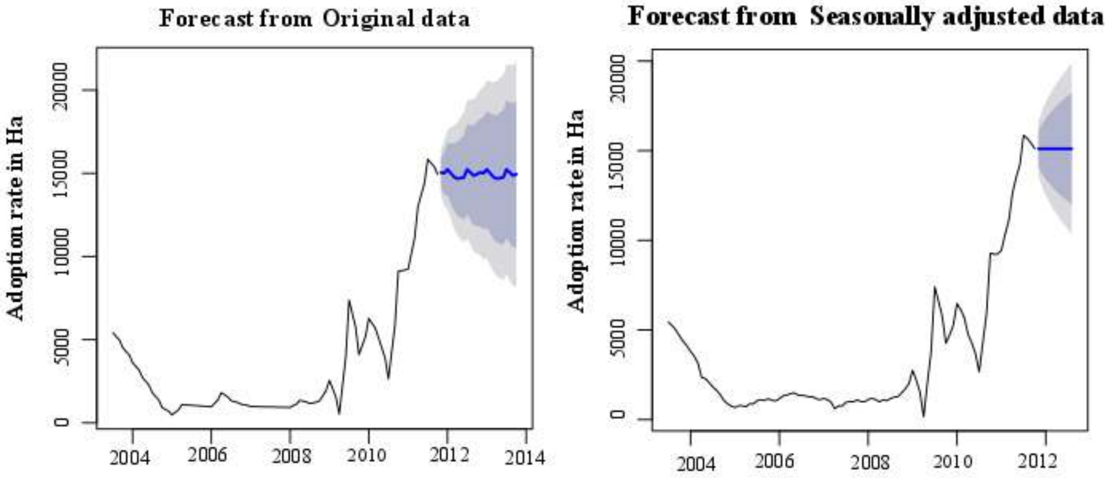

Fig- 4: Forecasts of the adoption of KEN 83-737 data based on a forecast of the seasonally adjusted data and the original time series data.

On comparing and evaluating the forecast made from the two sets of data, we note that when this time series data is adjusted by removing the seasonal variation, the forecasting error is increased (see **Table-1**) hence insignificance of adjusting the data.

Table-1: Forecast accuracy measures.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Forecast Adjusted	97.754	766.541	433.902	-8.246	21.373	0.190	0.354
Forecast from Original	97.111	707.333	403.564	-0.523	13.709	0.177	0.400

As discussed above, a non seasonal ARIMA model is proposed for forecasting the data. In ARIMA forecasting approach, it is required that the following steps are followed: (i) Model Identification, (ii) Parameter Estimation and Selection, and (iii) Diagnostic Checking (or Modal Validation); before we can (iv) use the Model for forecasting application. We, therefore, will first try to identify the model for fitness.

Model Identification

The graph in the upper panel of **Fig- 1** shows the presence of non stationary in the time series and confirmed by the Augmented Dickey-Fuller (ADF) test (Dickey-Fuller = -0.05, Lag order = 4, p-value = 0.99, Ha: stationary). The ARIMA model cannot be build until the series is made stationary. We achieved the stationary by differencing the time series in order to have ARIMA (p,d,q) with 'd' as the order of differencing used. Kumar [16] explained in his paper how to obtain the most accurate order of differencing.

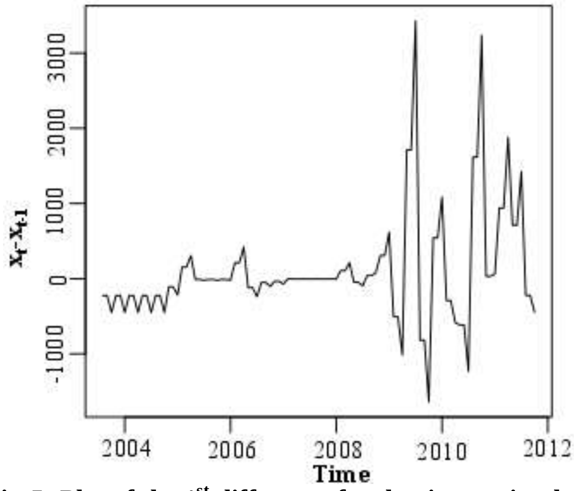

Fig-5: Plot of the 1st difference for the time series data

In Fig-5, it is evident that the data is stationary despite the increased noise between 2009 and 2012. The (ADF) test (Dickey-Fuller = -4.5251, Lag order = 4, p-value = 0.01, Ha: stationary).

Parameter Estimation and Selection

We examine the correlogram and partial correlogram in (Fig-6)of the stationary time series to find the suitable p in AR and q in MA in the ARIMA (p,d,q).

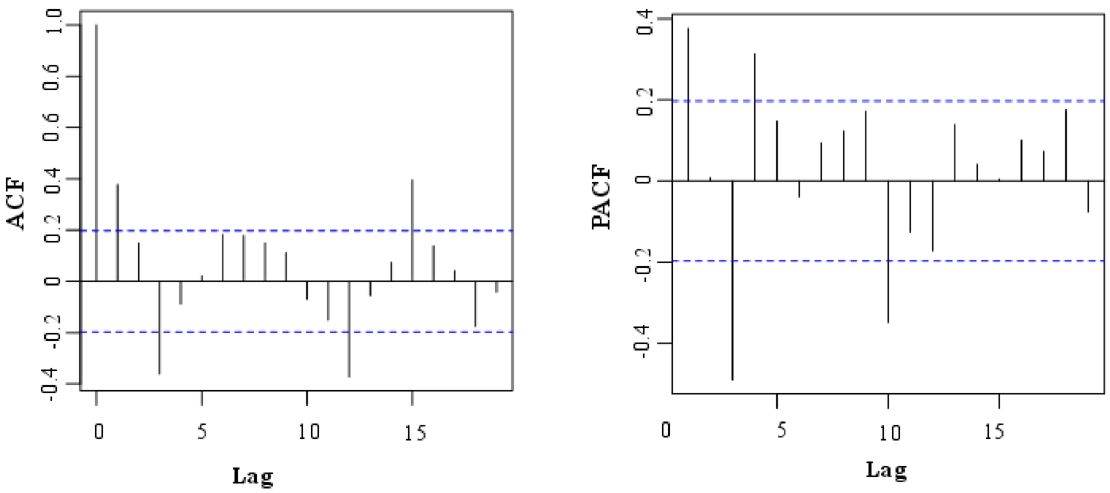

Fig-6: Estimated ACF and PACF for the variety adoption data

The autocorrelation at lag 1 and 3 are significant while the autocorrelation from 4 to 18 are within the limits. Thought at lag 12 and 15 the auctocorrelation crosses the significant lines, they are due error and happens by chance alone. On the other hand in PACF graph of Fig-6, partial autocorrelation at lag 1, 3,4 are significant then the rest are insignificant otherwise by chance.

The following possible ARMA (auto regressive moving average) models for the first differenced time series data of KEN 83737 adoption can therefore be: ARMA(4,0), ARMA(0,3) or . An ARMA(p,q) model with p and q both greater than 0 since autocorrelation and partial autocorrelation both tail off to zero. The candidate for the ARIMA model are therefore listed in the table below and their statistics shown.

Table 2: AIC and BIC values of the fitted ARIMA models

ARIMA	AIC	AICc	BIC
0,1,0	1596.49	1596.54	1599.08
0,1,1	1576.52	1576.65	1581.69
0,1,2	1570.13	1570.38	1577.88
0,1,3	1543.56	1543.99	1553.90
1,1,0	1588.16	1588.28	1593.33
1,1,1	1567.10	1567.35	1574.85
1,1,2	1569.10	1569.53	1579.44
1,1,3	1545.19	1545.84	1558.12
2,1,0	1588.23	1588.49	1595.99
2,1,1	1581.34	1581.77	1591.68
2,1,2	1562.66	1563.31	1575.59
2,1,3	1535.48	1536.40	1550.99
3,1,0	1548.56	1548.99	1558.90
3,1,1	1533.25	1533.90	1546.17
3,1,2	1533.56	1534.48	1549.07
3,1,3	1532.86	1534.10	1550.95
4,1,0	1541.63	1542.29	1554.56
4,1,1	**1532.76**	**1533.68**	**1548.27**
4,1,2	1534.51	1535.75	1552.60
4,1,3	1532.90	1534.51	1553.58

The RMSE for models range between (540- 590)

It is observed clearly that in Table 2, ARIMA (4, 1, 1) is the best predictive model (with the lowest AICc) for making forecasts for future values. The ARIMA(4,1,1) model has RMSE of 555.4907 . The ACF plot in Fig-7 of the residuals from the ARIMA(4,1,1) model shows that some correlations are not within the threshold limits indicating that the residuals are behaving like white noise and Ljung-Box test returns a large p-value (X^2 = 46.46, df = 16, p-value = 0.168276), also suggesting the residuals are white noise.

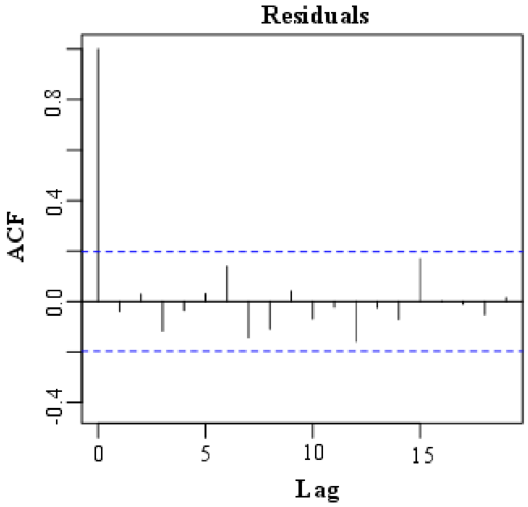

Fig-7: ACF plot for the residuals

Forecasting

When the fitted ARIMA (4,1,1) model is used to forecast (in 80% and 95% confidence interval) the adoption status for two years, The plot in Fig-8 is obtained.

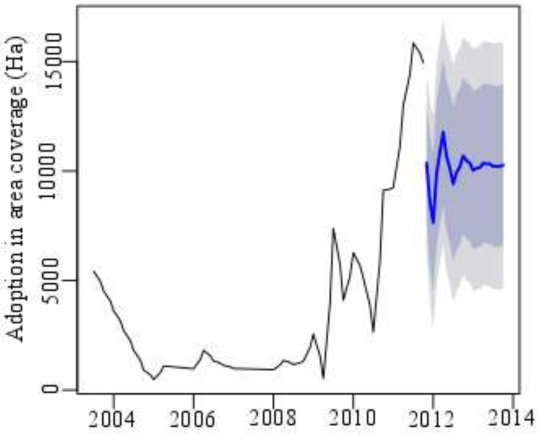

Fig-8: Forecast from ARIMA (4, 1, 1)

The forecasted values when compared with the original data from Nov 2011 to Oct 201, most of the actual dat a were within the confidence interval estimates (see Table 3) indicating some high level of accuracy in the forecasting m ethod.

Table 3: Forecast Estimate from ARIMA(4,1,1)

		80% CI		95% CI						80% CI		95% CI	
		lower	**upper**	**lower**	**upper**	**Actual**				**lower**	**upper**	**lower**	**upper**
Nov	2011	7660.7	13058.2	6232.0	14486.8	12523.9		Nov	2012	6899.9	14041.8	5009.5	15932.2
Dec	2011	5592.6	11592.2	4004.6	13180.2	10097.1		Dec	2012	6777.5	13944.5	4880.5	15841.4
Jan	2012	4466.1	10819.9	2784.3	12501.7	5243.6		Jan	2013	6421.6	13646.5	4509.2	15558.9
Feb	2012	6560.3	13157.6	4814.0	14903.8	8876.8		Feb	2013	6507.9	13746.0	4592.1	15661.8
Mar	2012	7564.5	14191.0	5810.5	15944.9	12510.0		Mar	2013	6527.2	13783.6	4606.5	15704.3
Apr	2012	8441.6	15112.0	6676.1	16877.6	19776.5		Apr	2013	6728.8	13990.5	4806.8	15912.6
May	2012	7250.0	14112.9	5433.5	15929.4	15037.5		May	2013	6683.3	13965.6	4755.8	15893.1
Jun	2012	6647.8	13597.6	4808.2	15437.1	14298.5		Jun	2013	6682.5	13981.2	4750.7	15913.0
Jul	2012	5866.1	12965.3	3987.1	14844.4	10820.5		Jul	2013	6544.8	13873.1	4605.1	15812.8
Aug	2012	6378.3	13477.5	4499.2	15356.6	23552.1		Aug	2013	6548.4	13894.3	4604.0	15838.6
Sep	2012	6647.0	13751.1	4766.6	15631.5	12283.7		Sep	2013	6518.7	13885.3	4568.9	15835.1
Oct	2012	7144.3	14249.2	5263.7	16129.8	19747.0		Oct	2013	6583.9	13963.4	4630.7	15916.6

CONCLUSION

In this study, two methods of forecasting were investigated; the exponential smoothing and the ARIMA methods. Results show that fitting the data using ARIMA models was the accurate than when exponential smoothing is used since the RMSE for ARIMA was way lower than exponential smoothing method. Further the best ARIMA candidate model selected for making predictions for upto 2years for the adoption of KEN 83-737 variety was ARIMA (4,1,1). The model was tested and validated statistically by studying the successive residuals (forecast errors) in the fitted ARIMA and found existence of white noise residuals. Hence, can conclude that the ARIMA (4,1,1) provide an adequate predictive model for the adoption of KEN 83-737 in the Kenya Sugar Industry.

The ARIMA(4,1,1) model predicted an decrease in adoption of KEN 83-737 from November 2011 to 2013 (Table 3). The prediction for 2013 is on average between 4700 Ha to 15000 Ha of the variety coverage. This model can be used to predict the future adoption of sugarcane varieties. As well the method of modeling and forecasting outlined here can be used for any time dependent occurrences in the sugar Industry.

ACKNOWLEDGMENT

I acknowledge Wawire, Kipruto, Shiundu for initiating and maintaining the collection and compilation of sug ar and sugarcane statistics in the Kenya Sugar Industry. I also acknowledge the director SRI for its technical and financial support for the project.

REFERENCES

1. Wawire NW, Kahora FW, Wachira PM, Kipruto KB; Technology adoption study in the Kenya sugar industry. Kenya Sugar Res Found Tech Bull, 2006; 1:51-77.
2. KSB; Strategic Plan 2009-2014. Nairobi : Kenya Sugar Board (KSB), 2009.
3. Kenya Gazette. The Seeds and Plant Varieties Act. Nairobi : Republic of Kenya, 2014. pp. 1704-1706, Gazette Notic e.
4. Odenya JO, Ochia CO, Korir C, Otieno V, Bor GK ; Adoption of improved Sugarcane Varieties in Kenya . 2010.
5. Ong'ala, J; Cane Variety Adoption. Sugarcane WebBased Database. [Online] 2013. http://www.keref.org.
6. Ong'ala, J; Geo-spartial Modeling of Sugarcane Smut. s.l. : Unpublised raw data, 2015.
7. Stevenson WJ, Sum CC; Operations management (Vol. 8). Boston, MA: McGraw-Hill/Irwin. 2009.
8. Tse RY; An application of the ARIMA model to real-estate prices in Hong Kong. Journal of Property Finance, 1997; 8(2):152-163.
9. Box GE; Jenkins GM; Time Series Analysis: Forecasting and Control. Time Series and Digital Processing, 1976.
10. Lapide L; History to demand-driven forecasting. The Journal of Business Forecasting, 2009; 28(2):18.
11. Makridakis S, Wheelwright SC, Hyndman RJ; Forecasting methods and applications. John Wiley & Sons, 2008.
12. Oliveira SCD, Pereira LMM, Hanashiro JTS, Val P D; A study about the performance of time series models for the analysis of agricultural prices. Revista GEPROS, 2012; 7(3):11.
13. Mahmoud E; Accuracy in forecasting: A survey. Journal of Forecasting, 1984; 3(2):139-159.
14. Meyler A, Kenny G, Quinn T; Forecasting Irish inflation using ARIMA models., 1988; 1-48. .
15. Stergiou KI; Modelling and forecasting the fishery for pilchard (Sardina pilchardus) in Greek waters using ARIMA time-series models. Journal du Conseil: ICES Journal of Marine Science, 1989;46(1):16-23.
16. Kumar M, Anand M; An Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India. Studies in Business & Economics, 2014;9(1):81-94.
17. Kaur G, Dham JK; Forecasting of rice exports from Punjab–An application of arima model. Asian Journal of Research in Business Economics and Management, 2013; 3(8):152-164.
18. Findley DF, Monsell BC, Bell WR, Otto MC, Chen BC; New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. Journal of Business & Economic Statistics, 1998; 16(2):127-152.
19. Han P, Wang PX, Zhang SY, Zhu DH; Drought forecasting based on the remote sensing data using ARIMA models. Mathematical and Computer Modelling, 2010; 51(11):1398-1403.
20. Zeileis A, Grothendieck G; zoo: S3 infrastructure for regular and irregular time series. 2005. arXiv preprint math/0505527.
21. Cleveland RB, Cleveland WS, McRae JE, Terpenning I; STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 1990; 6(1):3-73.