Universal Journal of Agricultural Research 2(7): 264-271, 2014 http://www.hrpub.org
DOI: 10.13189/ujar.2014.020706

Modeling the Causal Effect of World Cocoa Price on Production of Cocoa in Ghana

Sampson Ankrah [1], Kwadwo A.Nyantakyi[1,*], E.Dadey[2]

[1]Postgraduate Institute of Agriculture, University of Peradeniya, Sri Lanka
[2]SSNIT, Research Department, Accra, Ghana
*Corresponding Author: knyantakyi@yahoo.com

Abstract In this paper, models are developed to explain and forecast the effect of world cocoa price on production of cocoa in Ghana by using regression model with time series errors. The focus of the investigation was to find out whether the world cocoa price can assist in forecasting the future behavior of the cocoa production in Ghana. Annual data from 1961 to 2010 were used in fitting the model while 2011 and 2012 were used as out-of-sample data. Based on the behavior of several model adequacy techniques, the regression model with ARIMA(2,2,0) errors was considered as the 'best' model for the production variable. The mean absolute percentage error (MAPE), as a forecast accuracy measure, was used to validate the model. Thus, the MAPE of the regression model with ARIMA (2,2,0) errors was 7.97%. However, the conventional 'best' ARIMA model fitted to the production variable indicated an MAPE of 16%. This shows that, the production variable has smaller MAPE, when it was modeled together with world price using regression with ARIMA errors. Hence, regression model with ARIMA (2,2,0) errors is a better statistical technique in forecasting production of cocoa in Ghana than the conventional ARIMA method.

Keywords ARIMA, Regression with ARIMA Errors, Forecast and MAPE

1. Introduction

Cocoa is one of the most important crops in Ghana's economy. It contributed about 3.4% to total gross domestic product annually and an average of 29% to total export revenue between 1990 and 1999 (Anon., 200l). In terms of employment, the cocoa sector employs about 60% of the national agricultural labour force in the country (Ghana Cocoa Board; COCOBOD, 2007). In volume of production, Ghana is reported to be the second largest producer in the world, accounting for about 21% of the total production (World Cocoa Foundation, 2006).

However, production levels of Ghana's cocoa have consistently declined from 568,000 (Mt) to 160,000 (Mt) in 1965 and 1983 respectively (Anon., 1999; Abekoe*et al,* 2002). But, since the mid-1980s, production levels have risen gradually to an average of 400,000 (Mt) during the late 1990's (Anon., 1999; Abekoe*et al,* 2002), which still is relatively less than the production levels attained in the mid-1960s. Generally, productivity of cocoa (yields per hectare of land) in the country is among the lowest in the world (ICCO 2005). The highest productivity of cocoa is Malaysia (1800 kg ha^{-1}) followed by Ivory Coast (800 kg ha^{-l}) while it is 360 kg ha^{-l} in Ghana (Anon., 1999; Abekoe et al., 2002).

Instability of the world cocoa price creates significant risks to producers, suppliers, government, consumers and other parties that are involved in the cocoa sector of Ghana, (Poku, 2009). The problem of price volatility is particularly decisive for Ghana due to her heavy dependence on cocoa exports for foreign exchange earnings. Price is an incentive to these farmers and Ghana's cocoa sector is a price-taker. In other words, world cocoa price goes a long way to influence the farmer's behaviour towards farming practices which to an extent affect production, producer's price, export earnings and many other factors. Thus, a study that will examine the influence of the world cocoa price on production will be very useful to policy making and decisions.

In literature, there are several econometric models that have been developed for the Ghanaian Cocoa sector since the 1960's, (Bulir, 1998). Examples are;Awudu and Reider, (1995), Bulir, (1998, 2002), King, French and Minami, (1985), Bateman, (1976, 1990), Brempong and Gyimah, (1992), Teal and Vigneri, (2004), Zeitlin, (2005) and Armah, (2008) have all analyzed Ghana Cocoa. However, all these models are for the cocoa supply function, (Bulir, 2002). Bulir(1998), made this interesting remarks about these models. He said, "Most of the research to date suffer from the problem associated with the estimation of non-stationary time series and arbitrary selection of lag structures, accordingly, these models have been unable to explain the massive decline in recorded cocoa output".

Economic growth is typically a complicated process that occurs along numerous dimensions with one single factor

often not enough to explain growth (Armah, 2008). Thus, the objective of this study is to enhance the understanding of the effect of world cocoa price on the production of cocoa in Ghana. The appropriate statistical technique that can handle this objective is regression model with time series errors.

2. Materials and Methods

2.1. Data Source

Secondary data on production and world price of cocoa were obtained from the Ghana Cocoa Board, spanning from 1961 to 2012.

2.2. Multiple Linear Regression

The general form of a multiple regression is:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + e_i \qquad (1)$$

Where y_i is the variable to be forecast and $x_{1,i},\ldots,x_{k,i}$ are the k predictor variables. The coefficients $\beta_1,\ldots,\beta_k$ measure the effect of each predictor after taking account of the effect of all other predictors in the model. Thus, the coefficients measure the *marginal effects* of the predictor variables.

When forecasting, the general assumption we require for the errors ($e_1,\ldots,e_n$) are as follows:

a) the errors have mean zero;
b) the errors are uncorrelated with each other;
c) the errors are uncorrelated with each predictor $x_{j,i}$.

It is also useful to have the errors normally distributed with constant variance in order to produce prediction intervals, but this is not necessary for forecasting.

2.3. Regression Model with ARIMA Errors

In regression analysis using time series variables, the challenge is that, it is possible for the errors (residuals) to have a time series structure. Normally, this violates the usual assumption of independent errors made in ordinary least squares regression. The consequence is that the estimates of coefficients and their standard errors will be wrong if the time series structure of the errors is ignored. Thus, we will allow the errors from a regression to contain autocorrelation. Due to this change in perspective, we will replace e_t (error in regression equation) by n_t in the equation. The error series n_t is assumed to follow an ARIMA model.

2.3.1. Estimation

Suppose that y_t and x_t are time series variables, and if n_t follows an ARIMA(1,1,1) model, it can be written as:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t} + n_t, \\ (1-\phi_1 B)(1-B) n_t &= (1+\theta_1 B) e_t, \end{aligned} \qquad (2)$$

Where e_t is a white noise series.

Notice that the model (**2**) has two error terms; that is the error from the regression model which we denote by n_t and the error from the ARIMA model which we denote by e_t. Only the ARIMA model errors are assumed to be white noise.

An important consideration in estimating a regression with ARMA errors is that all variables in the model must first be stationary. If we estimate the model while any of these are non-stationary, the estimated coefficients can be incorrect. One exception to this is the case where non-stationary variables are co-integrated. Thus, when series are differenced for stationarity, regression model (**2**) with ARIMA(1,1,1) errors becomes:

$$\begin{aligned} y'_t &= \beta_1 x'_{1,t} + \ldots + \beta_k x'_{k,t} + n'_t, \\ (1-\phi_1 B)(1-B) n'_t &= (1+\theta_1 B) e_t, \end{aligned} \qquad (3)$$

Where

$$y'_t = y_t - y_{t-1}, \quad x'_{t,i} = x_{t,i} - x_{t-1,i}, \quad \text{and} \quad n'_t = n_t - n_{t-1}$$

which is a regression model in differences with ARMA errors.

2.3.2. Forecasting

To forecast a regression model with ARIMA errors, we need to forecast the regression part of the model and the ARIMA part of the model, and combine the results.

2.4. ARIMA Models

In order to identify an appropriate model for a series, the series must be checked if it is stationary. A non-stationary time series will have a time-varying mean or a time-varying variance or both. Why are stationary time series so important? This is because if a time series is non-stationary, we can study its behavior only for the time period under consideration. As a consequence, it is not possible to generalize it to other time periods. Thus, for the purpose of forecasting, such non-stationary time series may be of little practical value. Then it is always appropriate to transform a non-stationary time series to a stationary series before doing any meaningful analysis. The unit root test is a formal way of testing the stationarity of a series.

Unit Root Test

This has become widely popular over the past years. Among the various methods of unit root test, the test developed by Dickey and Fuller, known as the augmented Dickey-Fuller (ADF) test, is commonly used. The hypothesis of the test is:

H_0: series has a unit root or not stationary

H_1: series does not have a unit root or stationary

The ADF test consists of estimating the following regression model:

$$\Delta Y_t = \beta_1 + \beta_1 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-1} + \varepsilon_t \quad (4)$$

where ε_t is a pure white noise error term and the ADF test follows an asymptotic distribution.

Moving Average (MA) model

A process is an MA, if and only if a finite number of ψ weights are non-zero, i.e., $\psi_1 = -\theta_1, \psi_2 = -\theta_2, \ldots, \psi_q = -\theta_q$ and $\psi_k = 0, k > q$, then the resulting process is said to be a moving average process or model of order *q* and is denoted as MA(*q*).

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$
$$or \quad (5)$$
$$y_t = \theta(L)\varepsilon_t,$$

and ε_t is a white noise sequence (Box and Jenkins, 1994). MA processes are useful in describing phenomena in which events produce an immediate effect that only lasts for short periods of time (Slutzky, 1927). The ACF and PACF are used to determine the process that a series is following. In MA processes: (1) the number of spikes of the ACF determines the order of MA and (2) while the PACF decay exponentially depending on the sign.

Autoregressive (AR) Model

A series has an AR representation, if and only if a finite number of $\prod$ weights are non-zero, i.e, $\prod_1 = \phi_1, \prod_2 = \phi_2, \ldots, \prod_p = \phi_p$, and $\prod_k = 0$ for $k > p$, then the process is said to be an autoregressive process of order *p [i.e., AR(p)]*. It is defined by

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$
$$or \quad (6)$$
$$\phi(L) y_t = \varepsilon_t,$$

AR processes are useful in describing situations in which the present value of a time series depends on its preceeding values plus a random walk (Slutzky, 1927). In AR processes: (1) the ACF decay exponentially depending on the sign and (2) number of spikes of the PACF determines the order of AR.

Autoregressive Integrated Moving Average model (ARIMA)

Suppose we have a non-stationary ARMA($p + d$, q) process of the form $\phi'(L) y_t = \theta(L)\varepsilon_t$, such that *d* roots of $\phi'(L)=0$ lie on the unit circle. In such situations we can write it as a stationary process ω_t such that $\phi(L)\omega_t = \theta(L)\varepsilon_t$, where $\omega_t = \nabla^d y_t$. We can say y_t is an ARIMA(p, d, q) and ω_t is an ARMA(p, q). If both the ACF and the PACF tail off or exponentially decay, then it indicates a mixed ARMA model.

2.4.2. Model Selection Criteria

The information criteria used in this study are given below:

- The Akaike (1973) information criterion (AIC):

$$AIC = -2\ln L\left(\hat{\theta}_k\right) + 2k \quad (7)$$

- The corrected Akaike information criterion (AICc) (Hurvich and Tsai,1989):

$$AIC_c = -2\ln L\left(\hat{\theta}_k\right) + \frac{2kn}{n-k-2} \quad (8)$$

where $L\left(\hat{\theta}_k\right)$ is the likelihood of the fitted model, $k = p + 1$ (which is the model size), *p = number of parameters* and *n = number of observation*, (McQuarrie and Tsai, 1998) and Box and Jenkins, 1994).

2.5. Measure of Forecast Accuracy

Let y_i denote the i^{th} observation and $\hat{y}_i$ denote a forecast of y_i.

Scale-dependent errors

The forecast error is simply $e_i = y_i - \hat{y}_i$, which is on the same scale as the data. The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

Mean Absolute Error:

$$MAE = mean\left(|e_i|\right) \quad (9)$$

Root Mean Squared Error:

$$RMSE = \sqrt{mean\left(e_i^2\right)}$$

Percentage errors

The percentage error is given by $pi = 100e_i/y_i$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance between different data sets. The most commonly used measure is:

Mean absolute percentage error:

$$MAPE = mean(|pi|). \quad (10)$$

2.6. Diagnostic Checking

Once the model is identified and fitted to an observed

series, the next stage is to check the model for possible discrepancies. Residuals obtained from the fitted model are important for investigating the possible discrepancies in the model and also to further suggest some modifications to the model. Residuals are analyzed and checked to see if they satisfy the model assumptions. Any significant differences from the model assumptions mean we fail to prove that our fitted model is correct. Test of autocorrelations provides an important diagnostic tool. Any autocorrelation (partial autocorrelation) which is significant indicates some non-randomness in the residuals. Instead of testing the significance of individual autocorrelations, the Ljung-Box, Q test is used for the first *m* autocorrelations. The hypothesis is given as:

H_0 : residuals are not auto-correlated

H_1 : residuals are auto-correlated

The Ljung-Box, Q test is defined as:

$$Q = n(n+2)\sum_{h=1}^{m}\left(\frac{\hat{\rho}_h^2}{n-h}\right) \qquad (11)$$

$\hat{\rho}_h$ is estimated autocorrelation at lag *h*.

2.7. Statistical Software

The R software, with the package '*forecast*' was used in fitting both the regression with ARIMA errors and conventional ARIMA.

3. Results and Discussions

3.1. Estimation of Regression Model with ARIMA Error

A graphical representation of production and world price variables indicated that the variables were not stationary, (see appendix A1). The Augmented Dickey-Fuller test also revealed that, the variables, production [*W=-0.7579, p-value = 0.9594*] and the world cocoa price [*W = -1.9396, p-value = 0.5985*] were not stationary at level. And again, these variables were not co-integrated. Thus, the variables became stationary after differencing twice, that is, the variables are integrated of order 2, *I*(2).

The autocorrelation test on the conventional regression model of production on world price revealed that, the residuals are correlated [X-squared = 103.569, df = 4, p-value < 2.2e-16], (see appendix A3). This suggests that, the regression model (or the variables) have a time series structure in the errors. Thus, regression model with ARIMA errors techniques is appropriate for modeling when the residuals of a regression model are correlated, or in other words when residuals have a time series structure.

In selecting the 'best' ARIMA errors for the regression model, five models were fitted. The result is given in Table 1.

The 'best' model is the model with the minimum AIC, thus in Table 1 the 'best' model is (2,2,0). The regression model with ARIMA (2,2,0) errors estimates are given in Table 2.

Table 1. Competing Models of ARIMA Errors with their Corresponding AIC

No.	Models	AIC
1	(1,2,0)	1186.98
2	(2,2,0)**	1184.45
3	(1,2,1)	1188.07
4	(0,2,1)	1185.43
5	(0,2,2)	1189.81

Table 2. Parameter estimates of regression model with ARIMA errors for Production

Parameters	Coefficient	Standard Errors
World Price	52.51	25.23
AR(1)	-0.92	0.13
AR(2)	-0.42	0.13

In Table 2, all the parameter estimates are statistical significant. For model diagnostic checking, the Box-Ljung test, (x^2=*8.81, p-value = 0.12*), indicated that the residuals are now uncorrelated and the Shapiro-wilk test (*w = 0.974, p-value = 0.34*), indicated that the residuals are normally distributed. This suggests that the model fits very well.

Thus, the regression model with ARIMA(2,2,0) errors for production can be written as:

$$\text{Production}'_t = 52.51 * world \quad price'_{1,t} + n'_t,$$

$$(25.23)$$

$$\left(1-0.92B-0.42B^2\right)\nabla^2 n'_t = e_t,$$

Here the world cocoa price has a positive sign; this indicates that a unit change in world cocoa price leads to a 52.51 unit increase in production, with other factors held constant.

Forecasting of the Regression Model with ARIMA errors

For forecasting, we need to forecast the regression part and the ARIMA part of the model separately and add the two forecast values to get the overall forecast value for this model. We present this in Table 3.

In Table 3, the forecast accuracy measure is good**.** Thus, the regression model with ARIMA(2,2,0) errors is a better predictive model for production.

Table 3. Forecast Accuracy Measure of the Regression Model with ARIMA errors

Period	Actual Value	Regression Forecast	ARIMA Forecast	Overall Forecast	Error	MAPE (%)
2011	1024554	355453.5	668419.1	1023872.6	681.4	0.0666
2012	879348	370604.5	648284.6	1018889.1	139541.1	15.867
Average						7.97

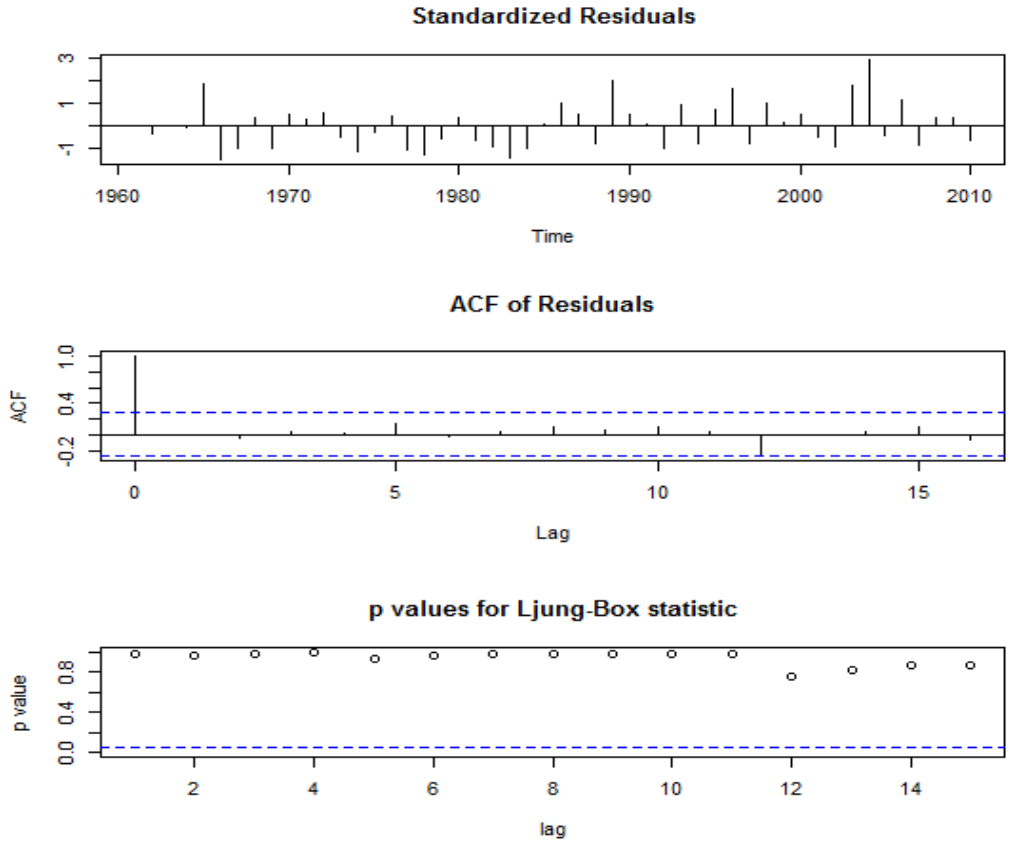


Figure 1. Plots of the standardized residuals, the ACF of the residuals, and the value of the Q-statistic

3.2. ARIMA Model for Production Variable

The focus of this paper is not to construct ARIMA models; however, we will like to evaluate the results of the regression model with ARIMA error with that of the conventional ARIMA model. According to the sample ACF and PACF (see Appendix A4), the production variable follows an MA(1). However, it is always appropriate to examine other competing models. It should be noted that the production series became stationary after the second difference. Thus, the lowest value of the AICc and statistical significance of the parameter estimates were used to select the 'best' model among competing models. The results are presented in Table 4 below.

In Table 4, the selection criterion, AICc indicates model (0,2,1) as the "best" model. Now, we focus on the ARIMA (0, 2, 1) which is the 'best' model. In Figure 1, we use a plot of the standardized residuals, the ACF of the residuals (note that R includes lag zero which is one), to explain the diagnostic checking of the 'best' model.

Inspection of the time plot of the standardized residuals in Figure 1 shows no obvious patterns. Notice that there is no outlier, however, with two value approaching 3 standard deviations in magnitude. The ACF of the standardized residuals shows no apparent departure from the model assumptions, and the Q-statistic is never significant at the lags shown, indicating that the residuals are not correlated. This means that the 'best' model ARIMA(0,2,1) fits the production series well.

The estimates of the 'best' model for the production variable with the forecast accuracy measure, MAPE, value are given in Table 5. The 'best' model passed the residual diagnostic tests.

Table 4. Competing Models with corresponding Selection Criteria

Models	AICc
1. (1,2,0)	278.16
2. (2,2,0)	279.71
3. (2,2,1)	282.19
*4. (0,2,1)***	*277.83*
5. (0,2,2)	279.85
6. (1,2,2)	280.08

**indicates the "best" model

Table 5. Best Model Estimates and Accuracy Measure

Cocoa Variable	'Best' Model & Estimates ARIMA(0,2,1)	MAPE (%)
Production	$\theta_1 = -0.3072\ [0.1356]$ $\delta = 0.2658\ [0.4419]$	16

Standard Errors in squared brackets

In Table 5, it is obvious that the parameter estimate, θ_1 is statistically significant but the drift is not significant. The MAPE of the 'best' model for production is relatively higher; however the model fitted well. We present in Figure 2, the forecast graph of the 'best' ARIMA model (0,2,1).

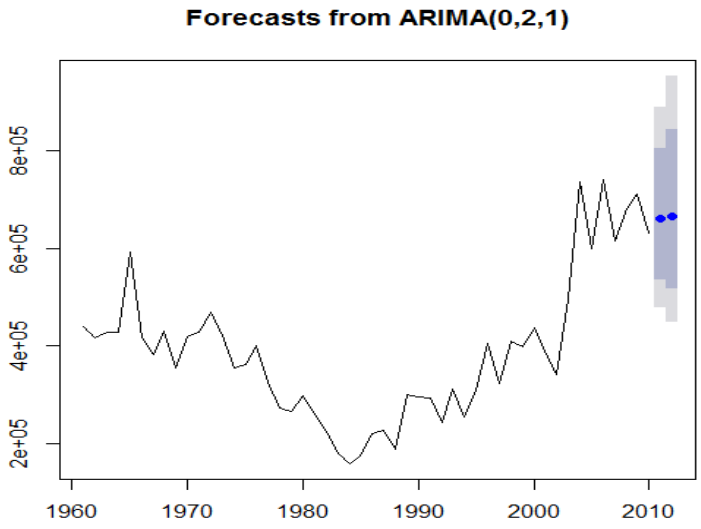


Figure 2. Forecast Plot of ARIMA (0,2,1) model

In Figure 2, a forecast value is given for one lead period with an accompanying 80% (dark grey) and 95% (light grey) prediction interval. It is obvious that forecast value increases as we go from 2011 to 2012.

4. Conclusions

The objective of this study is to investigate whether the world cocoa price can assist in forecasting the future behavior of cocoa production in Ghana. The production data was analyzed with the world cocoa price data using regression model with ARIMA errors.

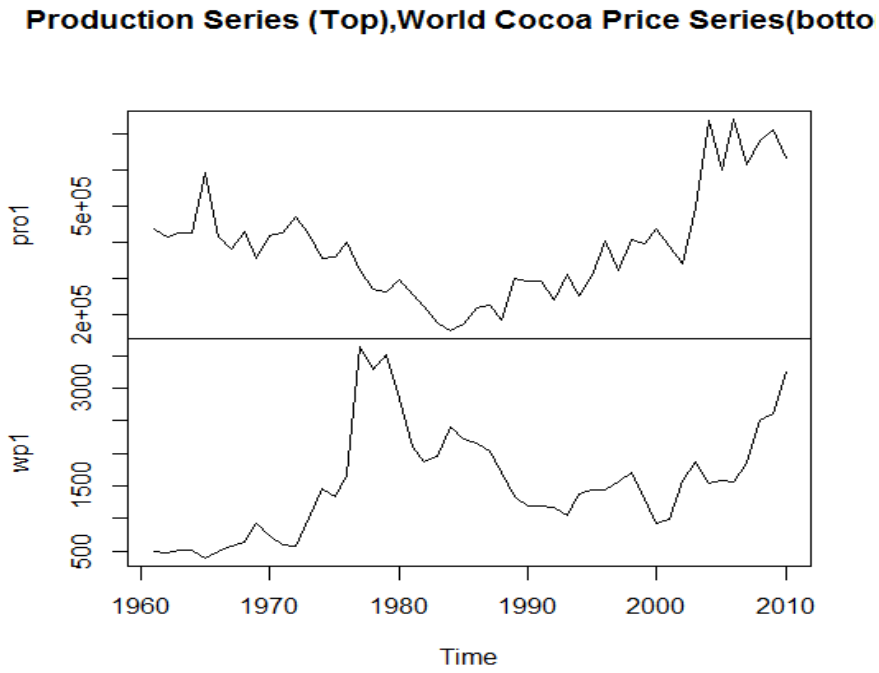
Base on the behavior of several model adequacy techniques the regression model with ARIMA(2,2,0) was identified as the most parsimonious and 'best' suited for predicting cocoa production. The model adequacy was confirmed by the Ljung-Box test of uncorrelated errors and normality test. The model was validated by testing for 2011 to 2012 data yielding a forecast accuracy measure, MAPE of 7.97% .In addition, the conventional 'best' ARIMA model was identified and fitted to production series and its MAPE was 16%.

This indicates that, production variable has a smaller forecast accuracy measure (MAPE), when it is modeled together with world price using regression with ARIMA errors. Thus, world cocoa price has a significant impact on explaining the future behavior of production. Hence, regression model with ARIMA errors is a better statistical technique in forecasting production of cocoa in Ghana than the conventional ARIMA method.

It is clear that the world cocoa price has a positive relationship with the production of cocoa in Ghana. In other words, a decrease in the world cocoa price will demotivate cocoa farmers, which will indirectly affect cocoa production. Thus, for the purpose of policy making, government can give subsidy or give flexible loans to the cocoa farmers when the world cocoa price decreases in order to motivate or improve their farming practices, which will eventually increase production level of cocoa.

Appendix

A1: Time Series plot of Production series and World Cocoa Price series



A2: Test of Stationarity

1. Augmented Dickey-Fuller Test for the Production series

   Dickey-Fuller = -0.7579, Lag order = 3, p-value = 0.9594
   alternative hypothesis: stationary

2. Augmented Dickey-Fuller Test for World Cocoa Price

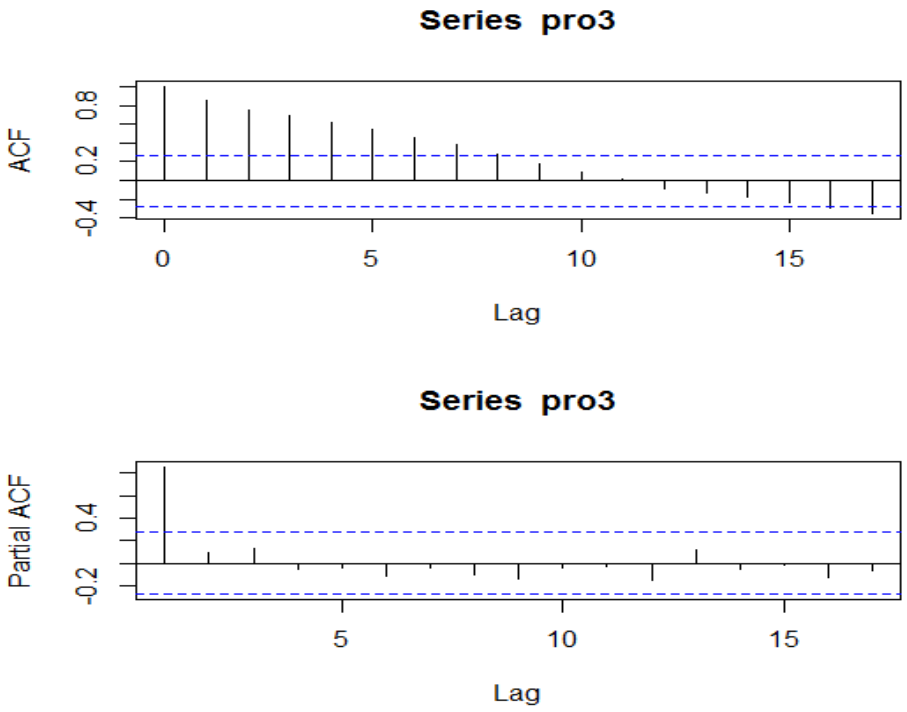
   Dickey-Fuller = -1.9396, Lag order = 3, p-value = 0.5985

alternative hypothesis: stationary

A3: Autocorrelation Test on Regression Model

Box-Ljung test for the conventional regression model
data: fit7$residuals
X-squared = 103.569, df = 4, p-value < 2.2e-16
alternative hypothesis: residuals are correlated

A4: ACF and PACF of ARIMA model for Production



REFERENCES

[1] Abdulai, A., and Reider, P. (1995) The Impact of Agricultural Policy on Cocoa Supply in Ghana: Error-Correction Estimation. Journal of African Economies, 4, pp 315-335

[2] Anim-Kwapong, G.J. and E.B. Frimpong (2005) 'Vulnerability of Agriculture to climate change: impact of climate change'. New TafoAkim: cocoa research institute of Ghana.

[3] Armah, E. S., (2008). Explaining Ghana's Recent Good Cocoa Karma: Smuggling Incentive Argument, a poster at the American Agricultural Economics Association Annual Meeting, Orlando, FL.

[4] Box, G. and Jenkins, G. (1994). Time Series Analysis, Forecasting and Control. John Wiley& Sons: New Jersey, USA.

[5] Bulir, Ales (2002). Can Price Incentive to Smuggle Explain the Contraction of the Cocoa Supply in Ghana? Journal of African Economies, 11 (3), pp 413-436.

[6] COCOBOD (2000) Ghana Cocoa Board Handbook (8th ed.). Accra: Jamiesnon's Cambridge Fax books Ltd, Accra.

[7] COCOBOD (2009) Overview of Ghana Cocoa Industry. Unpublished document, Ghana Cocoa Board.

[8] Hyndman, R. J., Athanasopoulos, G., (2013). Forecasting: Principles and Practice. Online Book: https://www.otexts.org/book/fpp, accessed period: October, 2013- January, 2014

[9] Poku, A., (2009). Agricultural Production and Pricing Policy Nexus: A Reflection of the Ghana Cocoa Industry. A dissertation submitted the Graduate school of Development Studies, the Hague, the Natherlands.

[10] Shumway, R. H. and D. S. Stoffer (2011). Time series analysis and its applications: with R examples. 3rd ed. New York: Springer.