Tropical Agricultural Research Vol. 26 (3): 569 – 578 (2015)
Short Communication

Weighted Modelling and Forecasting of Cocoa Production in Ghana: A Multivariate Approach

Sampson Ankrah[*], B.L. Peiris[1] and R.O. Thattil[1]

Postgraduate Institute of Agriculture
University of Peradeniya
Sri Lanka

ABSTRACT: *In this paper, we develop models for forecasting the annual cocoa production in Ghana. Instead of using the 'best' model for forecasting; a weighted scheme was applied to all competing models, to obtain a weighted model. The weighted scheme used in this paper is the weighted ranking procedure. Annual production, export earnings, exchange rate and domestic processing of cocoa data from 1970 to 2012 from Ghana were used for this study. Forecast accuracy measured from the weighted vector error correction model (VECM) and that of the "best" vector error correction model was used to validate the model. The forecast value from the weighted forecast approach performed better than that of the "best" model. The weighted predicted values were regressed on the real production values to show whether the weighted VECM was adequate to explain the variations in the annual cocoa production. The adjusted R^2 was 0.952 indicating that, the weighted VECM model explained 95.2% of the annual production variability. Hence, the weighted vector error correction model is a better statistical technique in forecasting cocoa production in Ghana.*

Keywords: *ARIMA, weighted ranking, vector error correction model and validation*

INTRODUCTION

Cocoa is one of the most important crops in the economy of Ghana. It contributed about 3.4% to total gross domestic product annually and an average of 29% to total annual export revenue between 1990 and 1999 (Anonymous, 2001). In terms of employment, the cocoa sector employs about 60% of the national agricultural labour force in the country. In volume of production, Ghana is reported to be the second largest cocoa producer in the world, accounting for about 21% of the total production (International Cocoa Organization, ICCO, 2006).

Production levels of Ghana's cocoa have consistently declined from 568,000 (Mt) in 1965 to its lowest level of 160,000 (Mt) in 1983, (Abekoe *et al.,* 2002). Since the mid-1980s, production levels have risen gradually to an average of 400,000 (Mt) during the late 1990's (Abekoe *et al.,* 2002), which still is relatively less than the production levels attained in the mid-1960s. Generally, productivity of cocoa (yield per hectare) 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^{-I}) while it is 360 kg ha^{-I} in Ghana (Abekoe *et al.,*

[1] Department of Crop Science, Faculty of Agriculture, University of Peradeniya, Sri Lanka
[*]. Corresponding author: sampson.ankrah@yahoo.com

2002). Thus, a study that will be able to forecast cocoa production in Ghana 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). However, all these studies are for the cocoa supply function (Bulir, 2002). None of these researchers has developed a model to show the entire Ghanaian cocoa sector. Bulir (1998) made these interesting remarks about these models. He reported that "Most of the researches to date suffer from the problem associated with the estimation of non-stationary time series and arbitrary choice of lag structures; so, these models have been unable to explain the massive decline in recorded cocoa output".

In time series analysis, one is faced with a challenge of choosing the 'best" model among many candidate models for forecasting. Usually, one has to go through a series of testing to get the "best" model. Our preliminary analysis and available literatures show that, the model preferred by a test or information criterion does not necessarily do better than other competing models in terms of prediction risk. Chatfield (2004) and Hoeting *et al.* (1999) have used the term 'model uncertainty' to capture the difficulty in identifying the best model. In addressing this challenge, combining forecast was introduced over the past three decades (Bates and Granger, 1969; Clemen, 1989). Various methods have been proposed. Thus, when there is a substantial uncertainty in finding the best model, alternative method, such as combined model should be considered.

Most often, the following weighting schemes have been distinguished: equal weights, Akaike weights, optimized and constrained weights; and Bayesian weights. The weighted scheme used in this paper is the weighted ranking procedure. Economic growth occurs along many dimensions with one single cause often not enough to explain growth (Armah, 2008). Thus, the aim of this study is to forecast cocoa production by considering other influential factors. In this study, we compare the forecast of cocoa production from weighted and single "best" model approaches.

METHODOLOGY

Data source

Annual data of cocoa production, export earnings, exchange rate and domestic processing spanning from 1970 to 2012 were obtained from the Ghana Cocoa Board, Accra.

Error correction model

The error correction model is used when the time series are not stationary and are cointegrated. The concept of configuration is explained below.

Cointegration

In univariate time series models, time series that have a unit root need to be modelled in first differences. In multivariate models, things become more interesting. It is possible for two time series that are non stationary with unit roots to have a linear relationship that produces a stationary disturbance. That is, in a multivariate situation it is possible to remove unit roots without taking differences. It turns out this has important implications for model specification.

Recall that a time series is said to be *I (d)* if it must be differenced *d* times to become stationary and invertible. We will restrict our study to *I (0)* and *I (1)* time series.

Definition

Two *I (1)* time series $y_{1,t}$ and $y_{2,t}$ are said to be cointegrated if there exists a linear relationship of the form $Z_t = \beta_1 y_{1,t} + \beta_2 y_{2,t}$ such that Z_t is *I (0)*. If we define the vectors

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, y_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} \tag{1}$$

so that the cointegrating relationship is written $Z_t = \beta' y_t$, then β is called the cointegrating vector. The cointegrating vector is not unique. Therefore, it is common to choose of the variables to have a coefficient of one in the cointegrating vector, which then uniquely identifies the rest of the vector. This choice of variable is referred to as the normalization of the co-integrating vector.

The cointegrating relationship is often interpreted as being a long run or equilibrium relationship between the variables. Statistically, the idea is that the variables are *I (1)* and therefore tend to wander randomly over time. However, the cointegrating relationship means there is some relationship from which the variables deviate from only in a stationary manner. In many applications, such statistical relationships are equated with economic equilibrium.

Cointegration can exist in a multivariate time series setting. Suppose

$$y_t = \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{\mathrm{m},t} \end{pmatrix} \tag{2}$$

If there exists a vector $\beta = \left(\beta_1, ..., \beta_m\right)'$ such that $Z_t = \beta' y_t$ then y_t is cointegrated with cointegrating vector β. However, in multivariate time series it is possible that there is more than one cointegrating vector. These cointegrating vectors are linearly independent, meaning that one is not a linear function of the other. The number of linearly independent cointegrating vectors is called the cointegrating rank. The two common tests to determine the cointegrating rank are the trace and the maximum eigenvalues tests. The hypothesis of the test is

H_0: the number of cointegrating vectors is r,
H_1: the number of cointegrating vectors is (r+1)

The two statistics are:

$$\begin{aligned} \lambda_{Trace}(r) &= -T \sum_{i=r+1}^{g} \ln(1-\hat{\lambda}_i) \\ \lambda_{Max}(r, r+1) &= -T \ln(1-\hat{\lambda}_{r+1}) \end{aligned} \tag{3}$$

Where $\hat{\lambda}_i$ is the estimated value for the i^{th} ordered eigenvalue and T is the sample size.

Vector error correction models

The appropriate model for cointegrated time series is called a Vector Error Correction Model (VECM) and is a rearranged restricted form of a VAR. An error correction model is parameterized so that the variables tend to revert back to the equilibrium relationship that is specified by the cointegrating vector.

In general, a VAR (*p*) model

$$y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t \tag{4}$$

is rearranged to give a VECM of the form

$$\Delta y_t = \alpha\beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{5}$$

Note that a VAR of order *p* translates to a VECM with *p - 1* lagged differences of y_t. A VECM thus consists of a mixture of variables in levels and first difference form. If we applied the univariate modeling strategy of taking first differences of any *I (1)* time series, and hence fitting a VAR in first differences, the resulting model would be misspecified because of the omitted error correction term. Conversely we cannot use a VAR in levels to model cointegrated time series because the resulting inference in the presence of the nonstationarity would not be valid. In the presence of cointegration, a VECM is required.

Estimation of weighted ranking procedure

The weighted ranking procedure performs better than weighting schemes such as Akaike weight and equal weight. Thus, we considered the weighted ranking procedure in this paper for better forecast.

The basis of the weighted ranking procedure is that, each competing model has the potential of relatively predicting the future value of a series, since the true model is unknown. Thus, we allow each model in the competing set of models to forecast. We therefore, rank each model based on their predictive performance, by ranking the model with the lowest forecast accuracy measure as first and assign the highest rank to that model; and in that order. The weighted ranking procedures are indicated below:

1. Fit a set of competing models to a dataset. The selection criterion of a model into the entire set of competing models is $p < 5$, (*that is, the lag length of the model should be less than 5*). This selection criterion is somewhat subjective, however, its basis is founded on the principle of parsimonious (*i.e., model with fewer estimates is desirable*).
2. Forecast each model in the entire set of models based on the ‘out-of-sample’ or ‘in-of-sample’ data.
3. Calculate their respective forecast accuracy measure, e.g., MSFE, MAPE etc.
4. Rank models in the entire set by their forecast accuracy measure. Thus, the lowest forecast accuracy measure model receives the highest rank.
5. Sum the ranks and respectively divide the individual rank by the total of the ranks to get the corresponding model weights.

Thus, we can express the proposed weight as:

$$w_i = \frac{\psi_i}{\sum_{i=1}^{s} \psi_i}, \quad i = 1, 2, \ldots, \mathrm{s} \tag{6}$$

where ψ_i is the rank for model *i* forecast accuracy measure, MSFE; and $\sum_{i=1}^{s} \psi_i$ is the sum of ranks of forecast accuracy measure, MSFE, in the entire set of models (s = last model in the entire set).

Weighted VECM model

Once the weights have been derived, we combine the parameter estimates of the entire set of models by applying their respective model weights. The weighted parameter estimates can be defined as

$$\phi_i^* = \frac{\sum_{i=1}^{R} w_i I_i(g_i) \hat{\phi}_{i,\mathrm{h}}}{\sum_{i=1}^{R} w_i I_i(g_i)}, \tag{7}$$

where

$$I_i(g_i) = \{ \begin{matrix} 1, & if \quad \phi_i \quad is \quad in \quad the \quad \text{model} \quad g_i, \\ 0, & otherwise \end{matrix}$$

Here, $\hat{\phi}_{i,\mathrm{h}}$ denotes the estimator of $\hat{\phi}_i$ based on model g_h .

RESULTS AND DISCUSSION

Based on the correlation analysis, the exogenous variables for the production variable are export earnings, exchange rate and domestic processing. None of these variables are stationary (Table A.1 and Table A.4). However, these variables became stationary after differencing once.

Test of cointegration

Since these variables are not stationary, an unrestricted cointegration rank test was performed on these variables, in order to know whether they are cointegrated in the long run. The results are reported in Table 1.

Table 1. Unrestricted cointegration rank test for production

Ho: No. of CE(s)	Trace Statistics	P-value	Max-Eigen Statistics	P-value
None	67.07	0.00	44.66	0.00
At most 1*	22.41	0.27	10.77	0.67
At most 2	11.64	0.17	7.66	0.41

At most 3	3.98	0.042	3.98	0.046

In Table 1, both the trace statistic and the maximum eigenvalue statistic indicate that there is one cointegration equation at the 5% level of significance. The two conditions for using the vector error correction model are met, thus, a VEC model is fitted to the production variable.

Estimation of weights

As indicated earlier, the selection criterion of a model into the entire set of competing models is $p < 5$, (*that is, the lag length of the model should be less than 5*). Thus, the entire set of competing models will have four VECM, (i.e., VECM (*p*), where *p* = 1, 2, 3, 4). We allow each competing model to make a forecast for production; then we rank their performance based on their respective forecast accuracy measure, MAPE. This is illustrated in Table 2.

Table 2. Derived weights for VECM models of production

Model	MAPE (%)	Rank (ψ)	Weights (Equation 2.6)
VECM (1)	12.83	1	0.1
VECM (2)*	8.49	2	0.2
VECM (3)	6.15	3	0.3
VECM (4)	5.43	4	0.4
Sum		**10**	**1**

**'Best' model according to conventional method*

It is obvious that, forecast accuracy measure improves as the lag length of model increases. Several diagnostic testing were performed on each of the four models in Table 2. The various diagnostic tests considered in this section are inverse roots of AR characteristic polynomial, granger causality or block exogeneity Wald tests, normality test and serial correlation. Although, VECM (3) and VECM (4) have lower MAPE but they could not pass other diagnostic tests. Thus, the model VECM (2) was selected as the "best" model according to the conventional method (Table A.3).

Estimation of the weighted model

We multiply the parameter coefficient estimates of the four competing models with their corresponding model weights; these are called the weightage estimates. Thus, we add the weightages across the competing models where they are present and divide by their respective model weights.

Forecast accuracy measure

We derived predicted values for all the observations, that is, from 1970 to 2013. However, we considered observations from 2000 to 2013, for the forecast accuracy measure calculation. Here, we compared the mean absolute percentage error (MAPE) of the combined forecast to that of the 'best' model from the conventional approach. The combined forecast and best model forecast accuracy measures for production are given in Table 3.

The percentage of error of the 'best' model varied from 0.4% to 21.5%; while the percentage error of the combined forecast varied from 0.73% to 20.14% between 2000 and 2012.

However, the combined forecast has an overall minimum forecast accuracy measure, MAPE of 6.6%, which is desirable. The combined forecast value for 2013 is 1,136,353.88, which is relatively higher than the "best" model forecast (i.e., 1,097,086). Thus, combined forecast method which is based on the weighted ranking approach is recommended for forecasting production series in the multivariate modelling.

Table 3. Combined forecast and best model forecast accuracy measures for production

Year	Actual	Best model forecast	Error	MAPE (%)	Combined forecast	Error	MAPE (%)
2000	436947	420269.8	16677.2	4	403151.24	33795.76	7.73
2001	389772	320773.6	68998.4	21.5	362471.59	27300.41	7.00
2002	340562	409621.8	-69059.8	16.9	409161.29	-68599.29	20.14
2003	496846	558543.2	-61697.2	11	531777.43	-34931.43	7.03
2004	736975	623939.1	113035.9	18.1	648454.1	88520.9	12.01
2005	599318	667000.3	-67682.3	10.1	654867.61	-55549.61	9.26
2006	740458	694270.5	46187.5	6.7	699137.75	41320.25	5.58
2007	614532	644825.1	-30293.1	4.7	638007.98	-23475.98	3.82
2008	680781	705413.1	-24632.1	3.5	697168.24	-16387.24	2.40
2009	710642	713260.6	-2618.6	0.4	700702.62	9939.38	1.39
2010	632037	657687.4	-25650.4	3.9	627414.83	4622.17	0.73
2011	102455	958266.2	66287.8	6.9	968872.85	55681.15	5.43
2012	879348	903658.3	-24310.3	2.7	908361.98	-29013.98	3.29
2013	NA	**1097086**			**1136353.88**		
Average				**8.5**			**6.6**

Combined long run relationship

The cointegration test revealed that there was a single long run relationship between production and export earnings, exchange rate, domestic processing. However, since the cointegration equation is not unique among the VECM models, a weighted cointegration equation from all the VECM models is desirable. We derived the weighted cointegration equation by multiplying the coefficient estimates of each VECM model with their corresponding model weights. The weighted cointegration equation for production is given in Table 4.

Table 4. Weighted cointegration equation for production

Parameters	Weighted Estimates				Combined Estimates	t-stats
	VECM(1)	VECM(2)	VECM(3)	VECM(4)		
Pro(-1)	0.1	0.2	0.3	0.4	1.00	
Dom(-1)	-0.487896	-0.875605	-1.301969	-0.910452	-3.575922	-4.99
S.E	0.072746	0.137054	0.133533	0.372876	0.716209	
Erate(-1)	-31979.14	-75660.02	-102122.0	-77030.04	-286791.2	-4.74
S.E	6213.01	10291.76	10617.18	33364.36	60486.31	
Exe(-1)	0.0000301	0.0000574	0.0000768	0.0000704	0.0002347	4.94
S.E	0.0000053	0.0000088	0.000009	0.0000244	0.0000475	
Constant	-16896.9	-33994.9	-49254.4	-116043.5	-216189.8	

** *Where* Dom = Domestic processing, Erate = Exchange Rate and Exe=Export earnings.

It is obvious that, the previous domestic processing, exchange rate and export earnings are significantly associated with production in the long run. Thus, the combined long run relationship equation of these variables is given as:

$$\begin{aligned}&\text{Production} = -216190 - 286791 * \text{Exchange} \quad Rate(-1) + 0.000235 * Export \quad Earnings(-1)\\&(s.e) \qquad\qquad\qquad (60486.31) \qquad\qquad\qquad\qquad (0.0000475)\\&\qquad\qquad - 3.576 * \text{Domestic} \quad \text{Processing}(-1)\\&\qquad\qquad (0.716)\end{aligned}$$

Weight Adjusted R^2=75.3%

In Eviews, each VECM model has two estimates as output: cointegration equation and error correction, with their corresponding R^2 values. Since the above cointegration equation is based on weightage; we therefore produce a weighted adjusted R^2 by multiplying the R^2 adjusted of each VECM model with their respective model weight. Here, domestic processing, exchange rate and export earnings coefficients are all statistically significant at 5%. The adjusted R^2 is 0.753; this means that, in the long run, the model was possible to explain 75.3% of the annual production variability by the variation in exchange rate, domestic processing and export earnings. The partial regression coefficient results suggest that: 1 unit increase in domestic processing leads to 3.576 output decrease in production; again, 1unit increase in export earnings leads to 0.000235 output increase in production; however, 1 unit increase in exchange rate leads to 286791 output decrease in production; given that the other variables are held constant. Thus, the results indicated that, the annual production of cocoa is characterized by the annual exchange rate, export earnings and domestic processing.

Regression model for production

Here, our focus is to construct a regression model using the actual production values against the weighted predicted values of production. The regression model will establish whether the weighted predicted values can explain the actual production values.

Thus, we apply the weights of each model to their corresponding forecast values and then sum the weighted forecast values as the predicted variable. The regression model is given as:

$$\begin{aligned}&\text{Production} = 1526.797 \quad + \quad 0.997 * \text{Pr}\,edicted \quad Value\\&(s.e) \qquad\qquad (16630.98) \qquad (0.0367)\end{aligned}$$

$[Adjusted \quad R^2 = 0.952, F - Statistic = 736.6, p - value = 0.00, Durbin - Watson \quad Stats = 2.13, n = 38]$

Here, the predicted value is significantly associated with the actual production of cocoa. The adjusted R^2 is 0.952; this means that, the model was possible to explain 95.2% of the annual production variability by the variation in the predicted value (which is given by exchange rate, domestic processing and export earnings). Again, the model is statistically significant, since the p-value associated with the F-statistic is 0.00.

The Breusch-Godfrey serial correlation LM Test, [F-statistic = 0.167, p-value = 0.995], failed to reject the null hypothesis that, the residual are uncorrelated; which is a good indication. Again, the heteroscedasticity test: Breusch-Pagan-Godfrey, [F-statistic = 0.00252, p-value = 0.96], suggested that the variance of the residuals are constant, which is good for our model.

CONCLUSION

In this study, the production of cocoa in Ghana was modelled using the weighted ranking procedure and the 'best' model from a multivariate time series approach. It was shown that forecast from the weighted VECM out-performed that of the 'best' VECM. Again, the predicted value of the weighted VECM was possible to explain 95.2% of the annual production variability by the variation in the predicted value. Thus, for accurate forecasting of cocoa production in Ghana, we recommend the use of the weighted ranking procedure.

REFERENCE

Abekoe, M.K, Oben-Ofori, D. and Egyir I.S. (2002). Technology of Cocoa on the Forest Zone of Ghana. Report presented at the "Convergence of Science" International Workshop, 23-29 March 2002.

Anim-Kwapong, G.J. and Frimpong, E.B. (2005) Vulnerability of Agriculture to climate change: impact of climate change. New Tafo Akim: cocoa research institute of Ghana.

Anonymous, (2001). The State of the Ghanaian Economy in 2000. Institute of Statistical Social and Economic Research (ISSER), University of Ghana, Legon, 162 pp.

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.

Bates, J. and Granger C. (1969). The Combination of Forecasts, Operational Research Quarterly, *20,* 451 - 468.

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

Bulir, Ales (1998). "The Price Incentive to Smuggle and the Cocoa Supply in Ghana, 1950-1996," IMF Working Papers 98/88, International Monetary Fund.

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.

Burnham, K.P. and Anderson, D.R. (2002). Model selection and multimodel inference: a practical information-theoretic approach, 2nd Ed, Springer-Verlag New York, Inc.

Chatfield, C. (2004). The Analysis of Time Series: An Introduction. Chapman & Hall/CRC: London, UK.

Clemen, R.T. (1989). Combining Forecasts: A Review and an Annotated Bibliography, International Journal of Forecasting, 5, 559 - 583.

Engle, R.F. and Granger, C.W.J. (1987). Co-integration and Error Correction Representation, Estimation and Testing, Econometrica, *55,* 251 - 276.

Engle, R.F. and Granger, C.W.J. (1991). Long-Run Economic Relations: Readings and Cointegration. Oxford University Press, Oxford.

Hannan, E.J. (1970). Multiple Time Series. Wiley, New York.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, *Vol. 14, No. 4* (Nov., 1999), pp. 382 - 401

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

ICCO. (2005). Annual Report 2003/04: International Cocoa Organization.

ICCO. (2006). Annual Report 2004/05: International Cocoa Organization.

Lütkepohl, H. (1993). Introduction to Multiple Time Series Analysis, New York: Springer-Verlag.

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

Stock, J.H., and Watson M.W. (2002): Macroeconomic Forecasting Using Diffusion Indexes, Journal of Business and Economic Statistics, *20(2),* 147 - 162.

Watson, M.W. (1994). Vector autoregressions and cointegration. In Engle, R.F., and McFadden, D.L. (eds.), Handbook of Econometrics, Vol. IV, 2843-2915. Elsevier, Amsterdam, The Netherlands.