Application of the ARIMA Models for Predicting Students' Admissions in the University of Lagos

Josephine N. Onyeka-Ubaka,[1]* Samuel O. N. Agwuegbo[2] and Olaide Abass[3]
[1]Department of Mathematics, University of Lagos, Lagos, Nigeria
[2]Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria
[3]Department of Computer Sciences, University of Lagos, Lagos, Nigeria
**jonyeka-ubaka@unilag.edu.ng*

Abstract
The objective was to assess the performance of the AutoRegressive Integrated Moving Average (ARIMA) models when occasional level shifts occur in the time series under study. The secondary data on the University of Lagos' undergraduates' admissions (1962–2016) were collected and analysed. It is predicted that universities in Nigeria and elsewhere could forecast their enrolment figures and student population growth rate based on the ARIMA models. The Box–Jenkins (B–J) approach provided the theoretical framework for the statistical analysis. The study used the Kalman Filter (KF) algorithm to develop a method using an ARIMA model to overcome and resolve the three main problems of the B–J methodology. The KF estimated the states for dynamic systems in state-variable formulations.
Forecasting university admissions is necessary if student population must match the infrastructural provisions on campuses. The best ARIMA models have been selected by using criteria such as Akaike's Information Criterion (AIC), Schwarz's Bayesian Criterion (SBC), Absolute Mean Error (AME), Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE). To select the best ARIMA model, the data was split into two periods: estimation period and validation period. The results clearly showed a continual increase in the demand for university education in the University of Lagos and, by extension, other universities in Nigeria.

Keywords: AutoCorrelation Functions (ACF), forecast, Kalman Filter, stationarity

Introduction

There is often an increase in the population of Nigerian universities after a matriculation exercise. The influx of students is such that there could be insufficient seats in many lecture rooms, shortage of hostel accommodations and traffic congestion, etc. in every academic session. This necessitates the need to model the number of undergraduate admissions in the University of Lagos so as to streamline students' population in line with the available facilities.

In this paper, the AutoRegressive Integrated Moving Average (ARIMA) model is the stochastic model used as a forecasting tool. An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series. The ARIMA model for forecasting is justified because ARIMA is a forecasting technique that projects the future values of a series based entirely on its own inertia. Its main application is in the area of short term forecasting requiring at least 40 historical data points. It works best when the data exhibits a stable or consistent pattern over time with a minimum amount of outliers. The ARIMA procedure provides a comprehensive set of tools for univariate time series model identification, parameter estimation and forecasting. It offers great flexibility in the kinds of ARIMA models that can be analysed (Hoff, 1983). The ARIMA procedure supports seasonal, subset and factored ARIMA models; intervention or interrupted time series models; multiple regression analysis with ARIMA errors and rational transfer function models of any complexity.

Many scholarly works on time series include the Kalman Filter and state-space models such as Khashei *et al*. (2012), Lee and Ho (2011), Tsay (2010), Javier *et al*. (2003), Hamilton (1994) and Harvey (1989). A vast class of ARIMA models captures short memory processes. The prediction and the effects of shocks in a time series data are very different for long and short memory processes. Stationary ARIMA models can represent series that homogeneously fluctuate around a constant level and non-stationary (Box *et al*., 1994). The structure of the series may change occasionally and the time periods of structural changes are determined by a stochastic process. Modelling the nature of the changing behaviour or outlying observations and deriving methods according to the proposed models have found ways in the works of

Box and Tiao (1968) and Yao (1984). The random-effect formulation is intuitively appealing because economic and environmental time series are affected by many unusual events; the occurrence and impact of which may be described by probability laws. ARIMA and other statistical models (regression method, exponential smoothing, generalised autoregressive conditional heteroskedasticity (GARCH)) are robust and efficient in time series forecasting especially in short-term predictions (Onyeka-Ubaka and Abass, 2013).

Consider the univariate time series y_t satisfying:

$$y_t = \mu_t + e_t \quad e_t \sim N(0,\ \sigma_e^2) \tag{1}$$

$$\mu_{t+1} = \mu_t + \eta_t \quad \eta_t \sim N(0,\ \sigma_\eta^2) \tag{2}$$

where $\{e_t\}$ and $\{\eta_t\}$ are two independent Gaussian white noises at $t = 1, \ldots, T$. The initial value μ_1 is either given or follows a known distribution, and it is independent of $\{e_t\}$ and $\{\eta_t\}$ for $t > 0$. Here, μ_t is a pure random walk with initial μ_1 and y_t is an observed version of μ_t with added noise e_t. In the Literature, μ_t is referred to as the trend of the series, which is not directly observable, and y_t is the observed data with observational noise e_t. The dynamic dependence of y_t is governed by that of μ_t because $\{e_t\}$ is not serially correlated. If there is no measurement error in (1), i.e., $\sigma_e = 0$, then $y_t = \mu_t$, which is an ARIMA (0, 1, 0) model. If $\sigma_e > 0$, i.e., there exist measurement errors, then y_t is an ARIMA (0, 1, 1) satisfying

$$(1 - B)y_t = (1 - \theta B)a_t \tag{3}$$

where a_t is a Gaussian white noise with mean zero and variance σ_a^2. The values of θ and σ_a are determined by σ_e and σ_η. This result can be derived as follows:

From (2), we have

$$(1 - B)\mu_{t+1} = \eta_t \quad \text{or} \quad \mu_{t+1} = \frac{1}{1-B}\eta_t$$

Using this result, (1) can be written as

$$y_t = \frac{1}{1-B}\eta_t + e_t \tag{4}$$

Multiplying (4) by (1 – B) gives

$$(1 - B)y_t = \eta_{t-1} + e_t - e_{t-1} \tag{5}$$

In representing the ARIMA model in the state space form, the ARIMA (p, q) model is

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{6}$$

written in the usual lag operator notation as

$\phi(B)y_t = \theta(B)\varepsilon_t$ where $\varepsilon_t \sim N(0,\ \sigma^2)$.

The possibility that $\phi(B)$ has roots inside or on the unit circle and that the model is non-stationary (ARIMA) is not excluded.

The general state space form is

$$y_t = Z_t\alpha_t + G_t\varepsilon_t \tag{5}$$

$$\alpha_{t+1} = \mathrm{T}_t\alpha_t + \mathrm{H}_t\varepsilon_t, \qquad t = 1,\ \cdots,\ n \tag{6}$$

where $\varepsilon_t \sim (0, \sigma^2\mathrm{I})$, $\alpha_1 \sim (a_1, \sigma^2\mathrm{P}_1)$ and the ε_t and α_1 are mutually uncorrelated. The system matrices $\mathrm{Z_t}$, $\mathrm{T_t}$, $\mathrm{G_t}$ and $\mathrm{H_t}$ are non-random, typically depend on hyper-parameters and, as the notation indicates, may vary over time. For a univariate model with an s x 1 state vector α_t and m x 1 vector of errors ε_t, the matrices $\mathrm{Z_t}$, $\mathrm{T_t}$, $\mathrm{G_t}$ and $\mathrm{H_t}$ are 1 x s, s x s, 1 x m and s x m, respectively. Pearlman (1980) puts forward the following ARIMA state space representation as well known for t =1, ⋯, n, $Z_t = Z$ = (1, 0, ⋯, 0), $G_t = G$ =1

$$\mathrm{T}_t = \mathrm{T} = \begin{bmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ \phi_m & 0 & \cdots & \cdots & 0 \end{bmatrix}, \qquad \mathrm{H}_t = \mathrm{H} = \begin{bmatrix} \theta_1 & + & \phi_1 \\ \theta_2 & + & \phi_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \theta_m & + & \phi_m \end{bmatrix}$$

where m = max(p, q). However, Pearlman gives no references for this form and it does not appear to be dealt with in the Literature. We shall refer to it as the max(p, q) representation. In this representation ε_t in (5) and (6) is the same as ε_t in the original ARIMA model. Note that $G_t H_t' \neq 0$, implying correlated measurement and state noise.

In the Literature (Box, Jenkins and Reinsel, 1994), the max(p, q) representation has been overlooked in favour of one in which the state vector is of length m = max(p, q + 1). In this version, Z and T are as above but G = 0 and H = (1, θ_1, ⋯ , θ_{m-1}). The prevalence of this form may be explained by the fact that the measurement and state noise are uncorrelated; there is no measurement noise. Uncorrelated measurement and state noise fits in with the more usual state space form where $G_t H_t' = 0$ (Anderson and Moore, 1979).

Using the Kalman filter, the observations y_t are transformed to innovations ν_t. In general, for t=1,⋯, n

$$\nu_t = y_t - Z_t a_t, \quad F_t = Z_t \mathrm{P}_t Z_t' + G_t G_t'$$

$$K_t = \left(\mathrm{T}_t \mathrm{P}_t Z_t' + \mathrm{H}_t G_t'\right) F_t^{-1}$$

$$a_{t+1} = \mathrm{T}_t a_t + K_t \nu_t, \quad \mathrm{P}_{t+1} = \mathrm{T}_t \mathrm{P}_t L_t' + \mathrm{H}_t J_t' \tag{7}$$

where $L_t = T_t - K_t Z_t$ and $J_t = H_t - K_t G_t$. The slight simplification of (7) made possible by the max(p, q+1) representation must be balanced against desirable features of the max(p, q) form. It is our contention that the arguments in favour of the max(p, q) version are compelling. First, when $q \geq p$, the state vector is shorter providing a slight computational advantage. Second, the converged quantities in the

max(p, q) representation take on convenient and readily interpretable forms.

Most importantly, the convergence of the Kalman filtering is established using the properties of the underlying ARIMA model.
Put $Y_{t,-\infty} = [y_t, y_{t-1}, y_1, y_0, \cdots]$, the linear space spanned by the entire past of the series. If the series is a pure AR(p) then, for minimum mean square error linear prediction, $Y_{t,-\infty}$ can be replaced by $Y_t = [y_t, \cdots, y_1]$ whenever $t > p$. By appropriate choice of c, an invertible ARIMA (p, q) model can be approximated, to any degree of accuracy, by an AR(c) process. Thus with an invertible ARIMA (p, q), we can find c so that, for purposes of prediction, $Y_{t,-\infty}$ can be replaced by Y_t whenever $t > c$. The size of c depends on how close the roots of the MA polynomial $\theta(B)$ are to the unit disc, that is, the closeness of the model to non-invertibility.

Consider the max(p, q) representation. Let $\alpha_{j,t}$ denote component j of α_t for $j = 1, \ldots, n$.
From (5), $\alpha_{1,t} = y_t - \varepsilon_t$. Using the state equation (6), for $j = 1, \ldots, m$,

$$\begin{aligned}\alpha_{j,\ t+1} &= \phi_j(y_t - \varepsilon_t) + \alpha_{j+1,\ t} + (\theta_j + \phi_j)\varepsilon_t \\ &= \phi_j y_t + \alpha_{j+1,\ t} + \theta_j \varepsilon_t \\ &= \phi_j y_t + \cdots + \phi_m y_{t-m+j} + \theta_j \varepsilon_t + \cdots + \theta_m \varepsilon_{t-m+j}\end{aligned} \tag{8}$$

provided the linear combination in (8) does not extend back to the presample period, that is, provided $t \geq m$. Thus, if the model is invertible, $\alpha_{t+1} \in Y_{t,-\infty}$ for $t > m$ and so, for $t \geq c$, we can assume $\alpha_{t+1} \in Y_t$. By definition of the filter $a_{t+1} = E(\alpha_{t+1}|Y_t)$ so, for $t > c$ we have $a_t = \alpha_t$. This implies that, once the filter has converged, $\mathrm{P}_t = \sigma^{-2}\ \mathrm{var}(\alpha_t - a_t) = 0$ and hence $F_t = 1$, $K_t = \mathrm{H}$, $J_t = 0$ and L_t has the same form as T but with $\theta_j's$ replacing the $\emptyset_j's$. Thus, for $t > c$, the Kalman filter collapses to the prediction error computation

$$\begin{aligned}v_t = y_t - \emptyset_1 y_{t-1} - \cdots - \emptyset_m y_{t-m} - \theta_1 \varepsilon_{t-1} \\ - \cdots - \theta_m \varepsilon_{t-m} = \frac{\emptyset(B)}{\theta(B)} y_t = \varepsilon_t\end{aligned} \tag{9}$$

which is conceptually and computationally convenient. Kalman filtering implicitly inverts the moving average polynomial and this inversion is achieved recursively without assumptions about presample values.

With the max(p, q+1) representation and $\emptyset(B)y_t = \theta(B)\varepsilon_{t-1}$. Hence $\varepsilon_e \in Y_{t+1,-\infty} - Y_{t,-\infty}$ and, for $t > c$, $a_{j,t+1} = \alpha_{j,t+1} - \theta_{j-1}\varepsilon_t$. Thus for $t > c$, $\mathrm{P}_t = \mathrm{HH}'$, $F_t = 1$, $K_t = (\emptyset_1, \cdots, \emptyset_m)'$, L_t is the same as T except for the top left entry where it is 0, $J_t = \mathrm{H}$ while $v_t = \varepsilon_{t-1}$ and $a_t = \alpha_t - H\varepsilon_{t-1}$. Thus, with the max($p$, q+1) representation, the state estimate does not converge to the state and the interpretation of filtered quantities is awkward.

Considering the convergence properties for Kalman filter smoothing, smoothing quantities under the max(p, q) representation also converge to convenient and readily interpretable constructs. The smoothing filter, corresponding to the general state space representation takes the following form.
Put $r_n = 0$, $\mathrm{N}_n = 0$ and for $t = n, \cdots, 1$,

$$u_t = F_t^{-1} v_t - K_t' r_t, \quad \mathrm{M}_t = F_t^{-1} + K_t' \mathrm{N}_t K_t$$
$$r_{t-1} = Z_t' u_t + \mathrm{T}_t' r_t, \quad \mathrm{N}_{t-1} = Z_t' F_t^{-1} Z_t + L_t' \mathrm{N}_t L_t,$$

The smoothations u_t contain information about departures from the model of De Jong and Penzer (2000) and have the interpolation characterization

$$u_t = \left\{\mathrm{var}(y_t | \mathrm{Y}_n^t)\right\}^{-1} \left\{y_t - \mathrm{E}(y_t | \mathrm{Y}_n^t)\right\}$$

where $\mathrm{Y}_n^t = [y_n, \cdots, y_{t+1}, y_{t-1}, \cdots, y_1]$, the punctured space (the striking period in the university). An appealing expression for the smoothations is readily derived using the max(p, q) representation.
First note that $\alpha_t \in Y_{t-1,-\infty} \subset Y_{n,-\infty}$ hence for $t = c + 1, \cdots, n$, $\alpha_t = E(\alpha_t | Y_n)$.
Similarly, $\varepsilon_t \in Y_{t,-\infty} \subset Y_{n,-\infty}$ implying for $t = c + 1, \cdots, n$,

$$\varepsilon_t = \mathrm{E}(\varepsilon_t | \mathrm{Y}_n) = G_t' u_t + \mathrm{H}_t' r_t = u_t + \mathrm{H}' r_t,$$

where the second equality follows from Koopman (1993). Thus, $u_t = \varepsilon_t - \mathrm{H}' r_t$ and iterating this identity using $r_{t-1} = \mathrm{Z}' u_t + \mathrm{T}' r_t$ yields

$$u_t = \varepsilon_t - \phi_1 \varepsilon_{t+1} - \cdots - \phi_m \varepsilon_{t+m} - \theta_1 u_{t+1} - \cdots - \theta_m u_{t+m}$$

or, in the lag polynomial notation, for $t = c + 1, \cdots, n$,

$$u_t = \frac{\phi(\mathrm{B}^{-1})}{\theta(\mathrm{B}^{-1})} \varepsilon_t = \frac{\phi(\mathrm{B}^{-1})\phi(\mathrm{B})}{\theta(\mathrm{B}^{-1})\theta(\mathrm{B})} y_t \tag{10}$$

where ε_t and y_t are interpreted as zero if $t > n$. This expression is exact provided that the Kalman filter has converged so the result only requires invertibility. The expression (10) mirrors the infinite sample Wiener-Kolmogorov interpolation formula, Whittle (1984). The Kalman filter smoothing computes exact finite sample interpolation errors. Provided Kalman filter converges, these interpolation errors coincide with the infinite sample interpolation errors for $t \leq n - m$. There is no requirement for smoothing filter convergence or stationarity.

Methods

Statistical models are built to explain phenomena that involve uncertainty. The choice of suitable models usually depends on the purpose of the practitioners. The design of ARIMA adapted closely follows the Box–Jenkins strategy for time series modelling with

features for the identification, estimation, diagnostic checking and forecasting steps of the Box–Jenkins method. The diagnostic check is applied to see if time series data satisfy stationary conditions and fit (see Shim *et al.*, 1994 and Fuller, 1976).

The first step in applying ARIMA methodology is to check for stationarity. "Stationarity" implies that the series remains at a fairly constant level over time. Without these stationarity conditions being met, many of the calculations associated with the process cannot be computed. At the identification stage, we specify the response series and identify candidate ARIMA models for it. If a graphical plot of the data indicates non-stationarity then the "difference" technique should be applied to the series. Differencing is an excellent way of transforming a nonstationary series to a stationary one. This is done by subtracting the observation in the current period from the previous one. If this transformation is done only once to a series, we say that the data has been "first differenced". This process essentially eliminates the trend if the series is growing at a fairly constant rate. If it is growing at an increasing rate, we can apply the same procedure and difference the data again. The data would then be "second differenced". That is, the non-stationary trending behaviour of the data is achieved by transformation, possibly differencing them, and computes auto-correlations, inverse auto-correlations, partial auto-correlations and cross-correlations.

At the estimation and diagnostic checking stage, we use the estimate statement to specify the ARIMA model to fit to the variable and to estimate the parameters of that model. The estimate statement also produces diagnostic statistics to help us judge the adequacy of the model. Significance tests for parameter estimates indicate whether some terms in the model may be unnecessary. Goodness-of-fit statistics aid in comparing this model to others. Tests for white noise residuals indicate whether the residual series contains additional information that might be utilised by a more complex model. If the diagnostic tests indicate problems with the model, the rule of thumb demand that we try another model, then repeat the estimation and diagnostic checking stage. At the forecasting stage, we use the forecast statement to forecast future values of the time series and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding estimate statement.

The Box–Jenkins methodology of time series analysis being currently one of the most accurate of the historical approaches to forecasting imposes some important limitations to the procedure: (i) it requires an extensive amount of past observations in order to develop an acceptable model, (ii) the model identification process requires a great deal of time and expertise and (iii) the model selected is a constant model; there is no convenient way to modify the coefficients with new observations. The research reported here uses the Kalman filter algorithm to develop a method using an autoregressive moving average model to overcome the three problems mentioned above.

ARIMA model parameters are often estimated using normal based maximum likelihood. With the max(p, q) representation the log-likelihood takes the form, ignoring constants,

$$-\frac{1}{2}\left\{n\log\sigma^2+\sum_{t=1}^{c}\left[\log F_t+\frac{v_t^2}{\sigma^2 F_t}\right]+\sum_{t=c+1}^{n}\frac{v_t^2}{\sigma^2}\right\}$$

$$\text{where } v_t=\varepsilon_t=\left\{\frac{\phi(\mathrm{B})}{\theta(\mathrm{B})}\right\}y_t \text{ for } t>c.$$

Initial conditions are handled exactly and there is no need for such tools as backcasting (Box, Jenkins and Reinsel, 1994). Maximising the log-likelihood with respect to σ^2 yields

$$\hat{\sigma}^2=\frac{1}{n}\left(\sum_{t=1}^{c}\frac{v_t^2}{F_t}+\sum_{t=c+1}^{n}\varepsilon_t^2\right)$$

Substituting back gives the concentrated log-likelihood, which is maximised by minimising

$$n\log\hat{\sigma}^2+\sum_{t=1}^{c}\log F_t$$

with respect to the remaining parameters. This is least squares provided that we can ignore the determinental term $\sum_{t=1}^{c}\log F_t$. Kalman filtering is required only for the initial c observations. These expressions are the similar to those for the max(p, q+1) representation, except that in the max(p, q+1) parametrisation the computation of the converged quantities is less convenient.

Properties of Forecast Error

Specifically, given the parameter estimates, we use the Kalman filter to obtain the 1-step-ahead forecast error $\{v_t\}$, which is useful in many applications. Given the initial values $\Sigma_{1|0}$ and $\mu_{1|0}$, which are independent of y_t, the Kalman filter enables us to compute v_t recursively as a linear function of $\{y_1, \cdots, y_t\}$. Specifically, by repeated substitutions,

$$v_1=y_1-\mu_{1|0}$$

$$v_2=y_2-\mu_{2|1}=y_2-\mu_{1|0}-K_1(y_1-\mu_{1|0})$$

$$v_3=y_3-\mu_{3|2}=y_3-\mu_{1|0}-K_2(y_2-\mu_{1|0})-K_1(1-K_2)(y_1-\mu_{1|0})$$

The transformation can be written in matrix form as

$$v=K(y-\mu_{1|0}I_T) \qquad (11)$$

where $v = (v_1, \ldots, v_T)'$, $y = (y_1, \ldots, y_T)'$, I_T is the T-dimensional vector of ones and K is a lower triangular matrix defined as

$$K = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ k_{21} & 1 & 0 & \cdots & 0 \\ k_{31} & k_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{T1} & k_{T2} & k_{T3} & \cdots & 1 \end{bmatrix}$$

where $k_{i,\,j-1} = -k_{i-1}$ and $k_{ij} = -(1-k_{i-1})(1-k_{i-2})\cdots(1-k_{j+1})k_j$ for $i = 2, \cdots, T$ and $j = 1, \ldots, i-2$. It should be noted that from the definition, the Kalman gain K_T does not depend on $\mu_{1|0}$ or the data $\{y_1, \cdots, y_t\}$; it depends on $\Sigma_{1|0}$ and σ_e^2 and σ_η^2.

The transformation in (11) has several important implications: (i) $\{v_t\}$ is mutually independent under the normality assumption.

Proof

Consider the joint probability density function (pdf) of the data

$$P(y_1, \cdots, y_T) = P(y_1) \prod_{j=2}^{T} P(y_j | F_{j-1})$$

Equation (4) indicates that the transformation from y_t to v_t has a unit Jacobian so that $p(v) = p(y)$. Importantly, since $\mu_{1|0}$ is given, $p(v_1) = p(y_1)$. The joint pdf of v is

$$P(v) = P(y) = P(y_1) \prod_{j=2}^{T} P(y_j | F_{j-1}) = P(v_1) \prod_{j=2}^{T} P(v_j) = \prod_{j=1}^{T} P(v_j)$$

This shows that $\{v_t\}$ are mutually independent.

(ii) The Kalman filter provides a Cholesky decomposition of the covariance matrix of y.

Proof

Let $\Omega = \mathrm{Cov}(y)$.

Equation (11) shows that $\mathrm{Cov}(v) = K\Omega K'$. On the other hand, $\{v_t\}$ is mutually independent with $Var(v_t) = V_t$. Therefore, $K\Omega K' = diag\{V_1, \ldots, V_T\}$, which is precisely a Cholesky decomposition of Ω.

To see the estimation error of the state variable μ_t, define $x_t = \mu_t - \mu_{t|t-1}$ as the forecast error of the state variable μ_t given data F_{t-1}. The variance, $var(x_t|F_{t-1}) = \Sigma_{t|t-1}$.

From the Kalman filter,

$$v_t = y_t - \mu_{t|t-1} = \mu_t + e_t - \mu_{t|t-1} = x_t + e_t$$

$$x_{t+1} = \mu_{t+1} - \mu_{t+1|t} = \mu_t + \eta_t - (\mu_{t|t-1} + K_t v_t)$$

$$= x_t + \eta_t - K_t v_t = x_t + \eta_t - K(x_t + e_t) = L_t x_t + \eta_t - K_t e_t$$

where $L_t = 1 - K_t = 1 - \frac{\Sigma_{t|t-1}}{V_t} = \frac{V_t - \Sigma_{t|t-1}}{V_t} = \frac{\sigma_e^2}{V_t}$.

Consequently, for the state errors, we have

$$v_t = x_t + e_t, \quad x_{t+1} = L_t x_t + \eta_t - K_t e_t, \quad t = 1, \cdots, T \qquad (12)$$

where $x_t = \mu_1 - \mu_{1|0}$.

Equation (12) is in the form of a time-varying state-space model with x_t being the state variable and v_t the observation (Tsay, 2010).

Results and Discussion

Let us assume that $y_1, y_2, \ldots, y_T$ follows the general ARIMA (p, d, q) model that can be written in terms of a linear combination of past values and past errors, ε_t:

$$y_t = \frac{\Theta(L)}{\Phi(L)\Delta^d}\varepsilon_t = \psi_\infty(L)\varepsilon_t = \left(1 + \psi_1 L + \psi_2 L^2 + \cdots\right)\varepsilon_t \qquad (13)$$

If no differencing is done ($d = 0$), the models are usually referred to as ARIMA (p, q) models. The future value $y_{T+\ell}$ is generated by model (13).

Thus $y_{T+\ell} = \varepsilon_{T+\ell} + \psi_1\varepsilon_{T+\ell-1} + \psi_2\varepsilon_{T+\ell-2} + \cdots$

where $y_T(\ell)$, the ℓ-step ahead forecast of $y_{T+\ell}$ made at origin T. The optimal forecast of $y_{T+\ell}$ is the conditional expectation of $y_{T+\ell}$ given the information set denoted by $E[y_{T+\ell}|Y_T]$. The term optimal is used in the sense that minimises the mean squared error (MSE). If the process is normal, the minimum MSE forecast (MMSE) is linear. Therefore, the optimal forecast ℓ-step ahead is

$$\begin{aligned} y_T(\ell) &= E[y_{T+\ell}|Y_T] \\ &= E[\varepsilon_{T+\ell} + \psi_1\varepsilon_{T+\ell-1} + \psi_2\varepsilon_{T+\ell-2} + \cdots |Y_T] \\ &= \psi_\ell\varepsilon_T + \psi_{\ell+1}\varepsilon_{T-1} + \psi_{\ell+2}\varepsilon_{T-2} + \cdots \end{aligned}$$

Since past values ε_{T+j}, for $j \leq 0$, are known and future value of $\varepsilon_{T(j)}$, for $j > 0$ have zero expectation. The ℓ-step ahead forecast error is a linear combination of the future shocks entering the system after time T:

$$\begin{aligned} e_T(\ell) &= y_{T+\ell} - y_T(\ell) \\ &= \varepsilon_{T+\ell} + \psi_1\varepsilon_{T+\ell-1} + \cdots + \psi_{\ell-1}\varepsilon_{T+1} \end{aligned}$$

Since $E[\varepsilon_T(\ell)|Y_T] = 0$, the forecast $y_T(\ell)$ is unbiased with MSE given as

$$\mathrm{MSE}[y_T(\ell)] = Var\left(e_T(\ell)\right) = \sigma_\varepsilon^2(1 + \psi_1^2 + \cdots \psi_\ell^2)$$

Given these results, if the process is normal, the $(1-\alpha)$ forecast interval is

$$\left[y_T(\ell) \pm Z\frac{\alpha}{2}\sqrt{Var(e_T(\ell))} \right]$$

For $\ell = 1$, the one-step ahead forecast error is $e_T(\ell) = y_{T+1} - y_T(1) = \varepsilon_{T+1}$, therefore σ_ε^2 can be interpreted as the one-step ahead prediction error variance.

In computing forecasts, ARIMA (p, d, q) model can be written as

$$\pi_{p+d}(L)y_t = (1-\pi_1 L-\pi_2 L^2-\cdots-\pi_{p+d}L^{p+d}) = \Theta(L)\varepsilon_t \quad (14)$$

where $\pi_{p+d}(L) = \Phi_p(L)(1-L)^d$.

Thus the future value of $y_{T+\ell}$ generated by (14) is

$$y_{T+\ell} = \pi_1 y_{T+\ell-1} + \cdots + \pi_{p+d} y_{T+\ell-p-d} + \varepsilon_{T+\ell} + \theta_1\varepsilon_{T+\ell-1} + \cdots + \theta_q\varepsilon_{T+\ell-q}$$

and the MMSE forecast is given by the expectation conditional to the information set:

$$y_T(\ell) = E[y_{T+\ell}|Y_T] = \pi_1 E[y_{T+\ell-1}|Y_T] + \cdots + \pi_{p+d}E[y_{T+\ell-p-d}|Y_T] + E[\varepsilon_{T+\ell}|Y_T] + \theta_1 E[\varepsilon_{T+\ell-1}|Y_T] + \cdots + \theta_q E[\varepsilon_{T+\ell-q}|Y_T]$$

The forecast $y_T(\ell)$ is computed substituting past expectations for known values and future expectations by forecast values, that is,

$$E[y_{T+j}|Y_T] = \begin{cases} y_{T+j} & j \le 0 \\ y_T(j) & j > 0 \end{cases} \quad (15)$$

The behaviour of the forecast function beyond the reach of the starting values can be characterised in terms of the roots of the autoregressive operator. It may be assumed that none of the roots of $\alpha(L) = 0$ lie inside the unit circle; for if there were roots inside the circle then the process would be radically unstable. If all of the roots are less than unity, then $\hat{y}_{t+\ell}$ will converge to zero as ℓ increases. If one of the roots of $\alpha(L) = 0$ is unity, then we have an ARIMA (p, 1, q) model and the general solution of the homogeneous equation will include a constant term, which represents the product of the unit root with a coefficient, which is determined by the starting values. Hence, the forecast will tend to a non-zero constant. If two of the roots are unity, then the general solution will embody a linear time trend, which is the asymptote to which the forecasts will tend. In general, if d of the roots is unity, then the general solution will comprise a polynomial in t of order $d - 1$.

The data on the University of Lagos undergraduates' admission were collected for the forty-two year period (1962–2016) with 1966 and 2004 excluded due to the onset at that time of the Nigerian Civil War and incessant strikes by the Academic Staff Union of Universities (ASUU), respectively. To illustrate the above ARIMA model as explicated, the Box–Jenkins methodology of step-by-step approach was used to fit a stochastic model to the undergraduates' enrolment in the University of Lagos, Nigeria.

Figure 1 details a plot of the undergraduate students' enrolment at the University of Lagos for the sampling period of 1962 to 2016. The data plotted displays a non-stationary pattern with an upward trending behaviour. To select a suitable stochastic model, we followed the three iterative steps of identification, estimation and diagnostic checking, recommended by Box and Jenkins (1976). However, at the identification stage, we utilised the autocorrelation functions (ACF) and partial autocorrelation function (PACF).

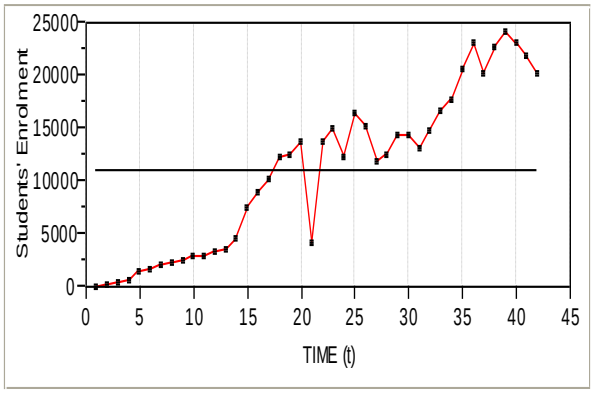

Figure 1: Undergraduates' enrolment in the University of Lagos

"Autocorrelations" are numerical values that indicate how a data series is related to itself over time. More precisely, it measures how strongly data values at a specified number of periods apart are correlated to each other over time. The number of periods apart is usually called the "lag". The sample ACF and the PACF for the original series are presented in Figures 2 and 3 to check whether the enrolment series is stationary. The plots show that both the ACF and PACF are decreasing slowly indicating that the series is non-stationary. The white noise hypothesis is also rejected very strongly, which is expected since the series is non-stationary. The p value for the test of the first six autocorrelations is printed as < 0.0001, which means the p value is less than 0.0001.

Since the series is non-stationary, the next step is to transform it to a stationary series by differencing. That is, instead of modelling the undergraduates' enrolment series itself, we model the change in undergraduates' enrolment from one period to the next. The functions have been plotted for $(1-B)y_t^{½}$, that is, the series with square root transformation and the degree of differencing $d = 1$. (see Box and Cox, 1964; Dickey and Pantula, 1987). The graph in Figure 4 shows a series that moves around a constant mean with approximately constant variance.

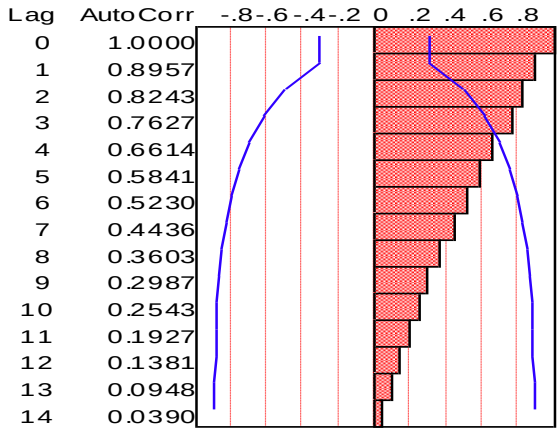

Figure 2: ACF of undergraduates' enrolment data

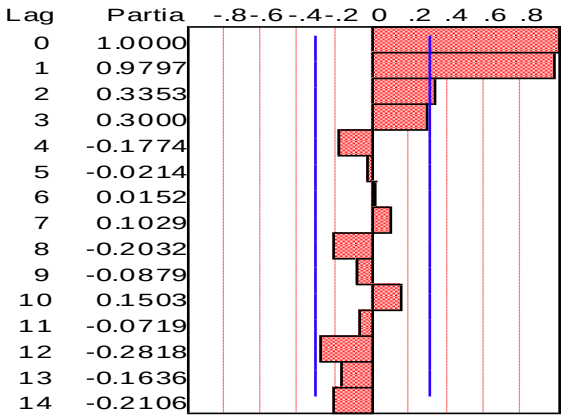

Figure 3: PACF of undergraduates' enrolment data

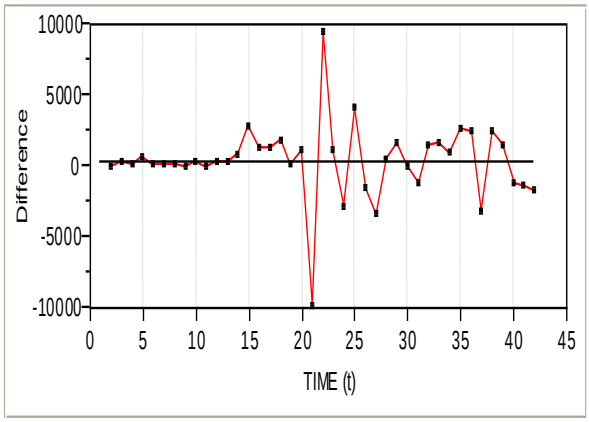

Figure 4: First difference (1–B)^1 of undergraduates' enrolment in University of Lagos, Nigeria

The ACF and PACF of differenced series are shown in Figures 5 and 6. Figure 5 shows that the autocorrelations decrease rapidly, indicating that the change in undergraduates' enrolment is a stationary time series. The sample ACF shows that the population ACF (r_1) is significantly different from zero. Hence, the model is the first order moving average type or MA(1). The partial and inverse autocorrelation function plots are also useful aids in identifying appropriate ARIMA models for the series. Looking at the sample PACF (r_1), it shows that it is significantly different from zero. The first differenced series is also of the first order autoregressive type or AR(1). But given that the first coefficients show some decreasing structure and $\hat{\phi}_{66}$ is statistically significant, perhaps an ARIMA (1, 1) model should be tried as well. In the Box–Jenkins approach to ARIMA modelling, the sample autocorrelation function, inverse autocorrelation function and partial autocorrelation function are compared with the theoretical correlation functions expected from different kinds of ARIMA models. This matching of theoretical autocorrelation functions of different ARIMA models to the sample autocorrelation functions computed from the response series is the heart of the identification stage of the Box–Jenkins modelling.

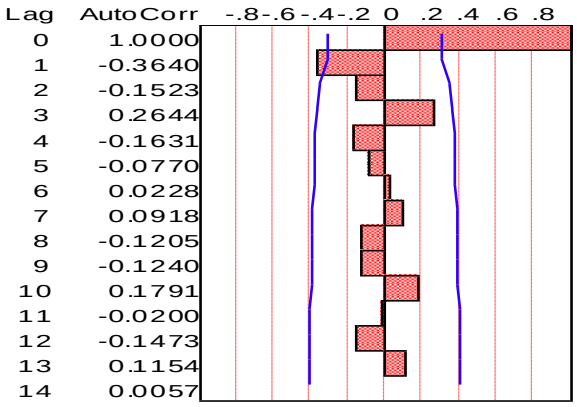

Figure 5: ACF of differenced data with square root transfomation

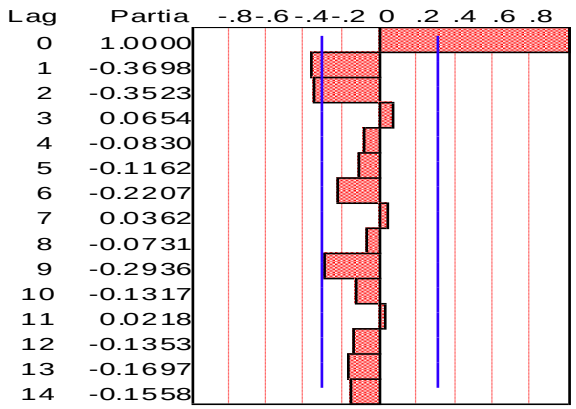

Figure 6: ACF and PACF of differenced data with square root transformation

The estimation was done using JMPin computer software and the results were presented in Table 1.

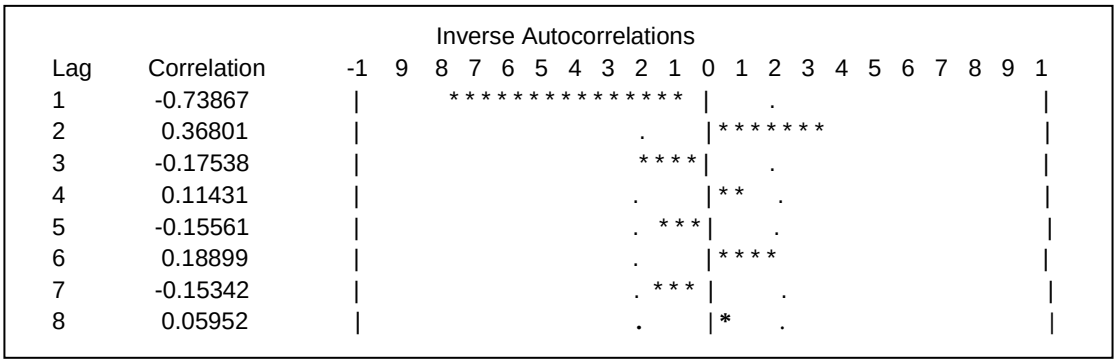

Figure 7: Inverse Autocorrelation Function plot for change in undergraduates' enrolment

Table 1: Model Comparison and Validation

Model	DF	AIC	SBC	AME	RMSE	MAPE	RSquare
ARIMA (1, 1, 0)	49	646.29109	649.71823	0.000963	0.000305	0.000028	0.892
ARIMA (0, 1, 1)	49	651.49256	654.52314	0.001191	0.003592	0.003472	0.880
ARIMA (1, 0, 1)	48	649.33765	653.47821	0.001041	0.003864	0.000329	0.905
ARIMA (1, 1, 1)	48	638.30428	643.44499	0.001472	0.004655	0.000457	0.914

Table 2: Parameter Estimates

Term	Lag	Estimate	Std Error	t value	Prob>|t|
ARI	1	-0.3610442	0.1441658	-2.50	0.0166
Intercept	0	510.915058	284.02189	1.80	0.0798
IMA	1	-0.3204135	0.1493262	-2.25	0.0169
Intercept	0	629.831074	294.101347	1.67	0.0856
ARMA	2	0.1792053	0.1453790	-1.87	0.0187
Intercept	0	533.279061	289.135742	0.99	0.6521
ARIMA	2	-0.1698421	0.1423195	-1.79	0.0155
Intercept	0	506.230165	283.095317	1.23	0.0684

The standard goodness of fit criterion in Statistics is the coefficient of determination:

$$R^2 = 1 - \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_{\tilde{\varepsilon}}^2} \text{ where } \hat{\sigma}_\varepsilon^2 = \sum \hat{\varepsilon}_t^2 / N.$$

Therefore, maximising R^2 is equivalent to minimising the sum of squared residuals. This measure presents some problems as a criterion for model selection. Firstly, the R^2 cannot decrease when more variables are added to a model and typically it will fall continuously. Besides, economic time series usually presents strong trends and/or seasonalities and any model that captures this fact to some extent will have a very large R^2. Modifications were proposed to this coefficient by Harvey (1989) to solve this problem.

Due to the limitations of the R^2 coefficient, a number of criteria have been proposed in the Literature to evaluate the fit of the model versus the number of parameters (see Postcher and Srinivasan, 1994). These criteria were developed for pure AR models but have been extended for ARIMA models. The more applied model selection criteria are the Akaike Information Criterion (AIC) (Akaike, 1974) and the Schwarz Information Criterion (SIC) (Schwarz, 1978) given by

$$AIC = \ln(\hat{\sigma}_{\tilde{\varepsilon}}^2) + \frac{2k}{N}, \qquad SIC = \ln(\hat{\sigma}_{\tilde{\varepsilon}}^2) + \frac{k}{N}\ln(N)$$

where k is the number of the estimated ARIMA parameters (p+q) and N is the number of observations used for estimation.

These criteria come down to minimise (in-sample) one-step-ahead forecast errors, with a penalty for over fitting (Qu *et al.*, 2006). Both criteria are based on the estimated variance $\hat{\sigma}_\varepsilon^2$ plus a penalty adjustment depending on the number of estimated parameters but it is in the extent of this penalty that these criteria differ. The penalty proposed by SIC is larger than AIC's since $\ln(N) > 2$ for $N \geq 8$. Therefore, the difference between both criteria can be very large if N is large; SIC tends to select simpler models than those chosen by AIC. In practical work, both criteria are usually examined. If they do not select the same model, many authors tend to recommend the use of the more parsimonious model selected by SIC (Moral and González, 2003). The parameter estimates for the models and key statistics for diagnostic testing are summarised in Table 2.

The criteria used for testing the validity of the models by comparing the three periods: estimation, validation and total periods are (i) absolute mean error (AME), (ii) root mean square error (RMSE) and (iii) mean absolute percent error (MAPE). The mean of the absolute deviation of the predicted and observed values is called absolute mean error and is defined as

$$\text{AME} = \sum_{t=1}^{\text{T}} \frac{\left|y_{obs} - y_{pred}\right|}{\text{T}}$$

The square root of the sum of square of the deviation of the predicted values from the observed values divided by their number of observations is known as the root mean square error, defined as

$$RMSE = \sqrt{\frac{1}{\text{T}} \sum_{t=1}^{\text{T}} \left(y_{obs} - y_{pred}\right)^2}$$

The mean of the sum of absolute deviation of the predicted and observed values divided by the observed value is called the mean absolute error.
For comparison, we have multiplied by 100; which is called mean absolute percent error and defined as

$$\text{MAPE} = \frac{1}{\text{T}} \sum_{t=1}^{\text{T}} \frac{\left|y_{obs} - y_{pred}\right|}{y_{obs}} \times 100$$

Our study suggests that the smaller the error, the better the forecasting performance of the observed variables and if the model variable performs well so will the whole model.

The ARIMA (1, 1, 0) model is given as

$$(1 - \phi L)\Delta \cup A_t = \mu + \varepsilon_t$$

Substituting the estimated values, we have

$$(1 - (-0.3610442\text{B}))(\text{Y}_t - \text{Y}_{t-1}) = 510.9151 + a_t$$

$$(-2.50) \qquad (1.80) \quad \text{(t statistics)}$$

$$\text{Y}_t = 510.915 + 0.639\text{Y}_{t-1} + 0.3610\text{Y}_{t-2} + a_t$$

An ARIMA (1, 1, 0) model predicts the change in undergraduates' enrolment as an average change plus some fraction of the previous change plus a random error plus some fraction of the random error in the preceeding period.
An ARIMA (1, 1, 0) model for the change in undergraduates' enrolment is the same as an ARIMA (1, 1, 1) model for the level of undergraduates' enrolment. The estimated ARIMA (1, 1, 1) model is

$$(1 - \text{B})\text{Y}_t = \mu + \left(\frac{1 - \theta_1 \text{B}}{1 - \phi_1 \text{B}}\right) a_t$$

That is, $(1 - \text{B})\text{Y}_t = 506.230651 + \left(\frac{1 + 0.3204135}{1 + 0.3610442}\right) a_t$

Several checks need to be made on the adequacy of the models. The problem arises in the time series analysis because the disturbances, which are a summary of a large number of theoretically irrelevant (and supposedly random) factors that enter into the relationship under study, are likely to be carried over into subsequent time periods.

To substantiate the assertion, Lee and Ho (2011) declared that "virtually all works in time series analysis assumes that a first-order autoregressive process is generating the disturbances". As shown under parameter estimates, all two models appear to have statistical strengths in terms of large t-values. The models pass the residual diagnostics with very similar results: the zero mean hypotheses for the residuals are not rejected and the correlograms indicate that the residuals behave as white noise processes. However, the parameters of the ARIMA (1, 1, 1) model are not statistically significant.

Given the fact that including an MA term does not seem to improve the results (see the AIC, SIC and R^2 values under model comparison in Table 1), we submit that the Autoregressive Integrated (ARIMA 1, 1, 0) model is ideal for modelling undergraduates' university admission in the University of Lagos and, by extension, other universities in Nigeria. Hence, it is used for our forecasting.

Since the model diagnostic tests show that all the parameter estimates are significant and that the residual series is white noise, the estimation and diagnostic checking stage is complete. We can now proceed to forecasting the undergraduates' admission enrolment series with the ARIMA (1, 1, 0) model for the period 2017–2040, with the assumptions of normally distributed errors, a 95% prediction interval for $y_{t+\ell}$, the future value of the series at time $t + \ell$ is

$$y_{t+\ell}^{t} + 1.96\sqrt{\hat{\sigma}_{\varepsilon}^2 \sum_{i=0}^{\ell-1} \psi_i^2}$$

Table 3 presents the forecast for this period. Estimates of students' enrolment from year to year are close to one another. Confidence intervals for forecast values have widths of 0.10 or 0.16 in all the years, showing the remarkable precision of the forecast.

For a stationary series and model, the forecasts of future values will eventually converge to the mean and then stay there. For the purpose of this study, the applied ARIMA models remain the most suitable statistical tool since the data they are applied to are not volatile as obtainable with high frequency data, such as financial data. These stylised volatile data are

captured with autoregressive conditional hetero-skedasticity (ARCH) models, proposed by Engle (1982). The ARIMA models (being a crucial forecasting tool) are equally adopted in identifying parameter orders in the generalised autoregressive conditional heteroskedasticity (GARCH) models (*p*, *q*) used in empirical applications of financial data. See Bollerslev (1986) and Onyeka-Ubaka *et al.* (2014).

Table 3: Forecast of Enrolment (2017–2040) using ARIMA (1, 1, 0)

Year	t	Approximate Forecast Value	95% C. I. Lower Limit	95% C. I. Upper Limit
2017	53	26260	19570	28951
2018	54	26823	19846	29801
2019	55	27386	20138	30636
2020	56	27950	20443	31457
2021	57	28513	20759	32269
2022	58	28076	21086	33067
2023	59	28640	21422	33858
2024	60	28203	21766	34641
2025	61	28766	22117	35416
2026	62	29330	22475	36185
2027	63	29893	22838	36948
2028	64	30456	23208	37706
2029	65	31020	23582	38458
2030	66	31583	23961	39206
2031	67	32146	24344	39949
2032	68	32710	24731	40688
2033	69	33273	25123	41424
2034	70	33836	25517	42156
2035	71	34400	25915	42884
2036	72	34963	26316	43610
2037	73	35203	26657	44205
2038	74	35675	26982	45176
2039	75	36304	27405	45478
2040	76	36567	27861	45892

Conclusion

The results show that the Kalman filter collapses, after the processing of an initial stretch of the data, to computing the exact moving average errors,

$$\varepsilon_t = \frac{\phi(\mathrm{B})}{\theta(\mathrm{B})} y_t .$$

Collapsing is analogous to the augmented Kalman filtering reducing to the ordinary Kalman filtering in the nonstationary case and emphasises that uncollapsed forms deal with the tedium of exact initialisation. The findings also show the remarkable degree of robustness of the ARIMA (1, 1, 0) model for forecasting future observations.

This study endeavoured to develop the best ARIMA model to efficiently forecast the undergraduates' admission into the University of Lagos because if it is possible to provide a better model for admission modalities, which can enable the University of Lagos, and by extension other Nigerian universities, to predict the number of students in advance, it would help the university management as well as the stability of the university's environment. Our empirical results show that the population of students keeps increasing annually without a corresponding increase in the provision of infrastructure.

Consequently, the management of the University of Lagos and the government of Nigeria are expected to provide adequate infrastructure, such as classroom blocks, hostels, health facilities, vehicular parking spaces and water distribution points to satisfy the needs of the ever-growing student population. They are also expected to stabilise power supply and ensure the maximum utilisation of facilities for effective and efficient teaching, learning and research.

References

Anderson, B. D. O. and Moore, J. B. (1979). Optimal Filtering. Englewood Cliffs, New Jersey: Prentice-Hall, p. 353–354.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, p. 716–723.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *J. Econometrics*, **31**: 307–327.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**: 211–252.

Box, G. E. P. and Tiao, G. C. (1968). A bayesian approach to some outlier problems. *Biometrika*, **55**: 119–129.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, p. 185–236.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis, Forecasting and Control*, 3rd edn., Englewood Cliffs, New Jersey: Prentice-Hall, p. 85–155.

De Jong, P. and Penzer, J. (2000). The ARIMA model in State-space form, Department of Statistics, London School of Economics Houghton Street, London, WC2A 2AE, UK, p. 1–10.

Dickey, D. A. and Pantula, S. G. (1987). Determining the order of differencing in autoregressive processes, *Journal of Business and Economics Statistics*, **5**: 455–461.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of United Kingdom inflation. *Econometrica*, **50**: 987–1007.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*, NY: John Wiley & Sons, Inc., 3–6.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press, p. 373–399.

Harvey, A. C. (1989). *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, p. 301.

Hoff, J. C. (1983). *A Practical Guide to Box–Jenkins Forecasting*, Belmont, C. A: Lifetime Learning Publications, p. 316.

Javier, C., Rosario, E., Francisco, J. N. and Antonio, J. C. (2003). ARIMA models to predict next electricity price, *IEEE Transactions on Power Systems*, **18**(3): 1014–1020.

Khashei, M., Bijari, M. and Ardal, G. A. R. (2012). Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks, *Computers and Industrial Engineering*, **63**(1): 37–45.

Klein, L. R. (1986). *An Essay on the Theory of Economic Prediction*, Markham, Chicago, p. 1–30.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika*, **80**: 117–126.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). Distribution of the estimators for autoregressive time series with a unit root, *J. Econometrics*, **54**: 159–178.

Lee, C. and Ho, C. (2011). Short-term load forecasting using lifting scheme and ARIMA model, *Expert System with Applications*, **38**(5): 5902–5911.

Moral, P. and González, P. (2003). *Univariate Time Series Modelling*, Dec. 10, p. 53–147.

Onyeka-Ubaka J. N. and Abass, O. (2013). Central Bank of Nigeria (CBN) intervention and the future of stocks in the banking sector. *American Journal of Mathematics and Statistics*, **3**(6): 407–416.

Onyeka-Ubaka J. N., Abass, O. and Okafor, R. O. (2014). Conditional variance parameters in symmetric models. *International Journal of Probability and Statistics*, **3**(1): 1–7.

Qu, N., Dark, J. and Zhang, X. (2006). *Influence Diagnostics in a Bivariate GARCH process.* Monash University, Australia, p. 278–291.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*, New York: John Wiley & Sons, Inc, p. 167–201.

Pearlman, J. G. (1980). An algorithm for the exact likelihood of a high-order autoregressive-moving average process, *Biometrika*, **67**(1): 232–233.

Postcher, B. and Srinivasan, S. (1994). A comparison of order determination procedures for ARIMA models, *Statistica Sinica*, **4**: 29–50.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**: 461–464.

Shim, J. K., Siegel, J. G. and Liew, C. J. (1994). *Strategic Business Forecasting*, Probus Publishing Company, Chicago, England, p. 152–243.

Whittle, P. (1984). *Prediction and Regulation by Linear Least Square Methods* 2nd edn., Oxford: Blackwell, 187pp.

Yao, Y. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. *The Annals of Statistics*, **12**(4): 1434–1447.