Journal of the Nigerian Association of Mathematical Physics
Volume 25, No. 2 **(November, 2013), pp 95 – 106**

ARIMA Modeling of Students Enrolment in University of Lagos (1965 – 2011)

ADEWARA, Johnson Ademola[1]. MBATA, Ugochukwu A.[2] and KESHINRO. A. O[3]

[1,3]Distance Learning Institute,
University of Lagos, Akoka-Lagos, Nigeria.
[2]Department of Mathematics,
University of Lagos, Akoka-Lagos, Nigeria.

Abstract

Time series ARIMA model is applied to students' enrolment data of the University of Lagos from 1962 to 2012 to check the stability and detect relatively small shift in the admission process mean. The descriptive statistics of the data was presented. The data was modeled using ARIMA (0, 1, 1) and (2, 1, 2). The R square for ARIMA (0, 1, 1) is 37.24% and ARIMA (2, 1, 2) is 52.84% for the undergraduate. The postgraduate enrolment ARIMA (0, 1, 1) is 57.73% and ARIMA (2, 1, 2) is 62.73% respectively. The result shows that the undergraduate and postgraduate were out of control due to the common cause such as incremental rate in the admission process. Also, the special-cause plays an important role which is decay in the facilities, inadequate staff and equipments for teaching and research. The model diagnostic check indicates that ARIMA (2, 1, 2) was more fitted than ARIMA (0, 1, 1).

Keywords: Processes, Control charts, Control limits, Assignable causes, Corrective actions.

1.0 Introduction

Control charts are used to detect anomalies in processes. They are most often used to monitor production-related processes. Samples taken from such processes are assumed to be independent and identically distributed. Many business-related processes, for instance sales volumes or product prices, behave very differently. They typically contain a trend, local or global, and serial correlation. The control chart helps in keeping processes in control by focusing or monitoring the key quality characteristics or variables. There are two phases of control charting in SPC, Phase I and Phase II. In Phase I analysis, historical observations are analyzed to determine whether the process is in control, to understand the sources of variation in the process, and to estimate the in-control parameters of the process. In contrast, Phase II control charting aims at on-line monitoring of future observations by using the control limits, constructed based on the estimated in-control process parameters from Phase I, to determine if the process continues to be in control. The objective of Phase II analysis is to quickly detect process changes. Obviously, a successful Phase II process monitoring depends heavily on a successful Phase I analysis. When the process is changed by some assignable causes, an effective control chart should be able to detect the changes quickly and signal requests for investigation. If assignable causes are found, then subsequent corrective actions should be taken to eliminate them. Substantial work has been done over the years on various charts such as Shewhart-type charts based on the generalized variance [1 – 8] and multivariate exponentially weighted average (MEWMA) control charts [9 – 13] on Time-Series. This paper focused on ARIMA (p, d, q) modeled to know whether the process is stable or to detect relatively small shift in the admission process mean of the University over the years. The remainder of the paper is organized as follows: *Section 2 presents Autoregressive integrated moving average (ARIMA) model, section 3 presents The Data and Method, section 4.0 is Discussion and section 5.0 presents the Summary and conclusion.*

2.0 Autoregressive integrated moving average (ARIMA) model

Autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models i.e. ARIMA and ARMA are fitted to time series data either to understand the data or to forecast points in the series. They are applied in some cases where the data exhibit non-stationarity, where an initial differencing step

Corresponding author: ***ADEWARA, Johnson Ademola,*** E-mail: -, Tel.: +2348023875722

(corresponding to the "integrated" part of the model) can be applied to remove the non-stationarity. ARIMA models are generally represented by ARIMA (*p, d, q*) where parameters *p, d,* and *q* are non-negative integers that refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. ARIMA models form an important part of the Box-Jenkins approach to time-series modeling [14]. Autoregressive Integrated Moving-Average (ARIMA) models consist of unit-root non-stationary time series which can be rendered stationary by the differencing operator. They are useful models in real applications. The well-known exponential smoothing model for forecasting is just a special case of ARIMA (0, 1, 1) models.

2.1. ARIMA (0,1,1) Model

This is perhaps the most commonly used model in forecasting. It is the exponential smoothing model. The general form of the model is

$$Z_t - Z_{t-1} = c + a_t - \theta a_{t-1}, \text{ with } |\theta| < 1. \quad (1)$$

Following the common practice, we shall assume c = 0. Since the model is invertible, the Π-weights are $\Pi_i = \theta^{i-1}(1-\theta)$ for i ≥1. Thus,

$$Z_t = (1-\theta)Z_{t-1} + \theta(1-\theta)Z_{t-2} + \theta^2(1-\theta)Z_{t-3} + \cdots + at = \sum_{i=1}^{\infty} \theta^{i-1}(1-\theta)Z_{t-i} + a_t \quad (2)$$

The current value Z_t is, therefore, a weighted average of the past values plus an innovation. The weights decay exponentially at the rate θ. As will be seen later in forecasting, such a weighted average makes common sense as the most recent past values are more relevant than the remove past in determining the current value. If $0 < \theta < 1$, the weights remain positive. This is the exponential smoothing model in the forecasting literature. The only difference is that in the forecasting literature the rate θ is determined usually by minimizing sum of squares of one-step ahead forecast errors or by pre-specification. In time series analysis, we often estimate θ by the (exact) maximum likelihood method. The φ -weights of the model are $\varphi_i = (1-\theta)$ for all $_i > 0$. Thus,

$$Z_t = a_t + (1-\theta)\, a_{t-1} + (1-\theta)a_{t-2} + \cdots \quad (3)$$

It is seen that the effect of past innovations on Z_t is persistent, but at a discounted rate $(1-\theta)$ when $\theta > 0$.

This integrated moving average model can also be viewed as follows: The "true" time series is a random walk

$$Y_t - Y_{t-1} = b_t \quad (4)$$

where $\{b_t\}$ is a Gaussian white noise process with mean zero and variance σ_b^2 .However, we cannot observe Y_t directly. Instead, what we observe is contaminated by an independent Gaussian white noise $\{e_t\}$ with mean zero and variance σ_ε^2, i.e.

$$Z_t = Y_t + e_t. \quad (5)$$

where e_t is a measurement error such as rounding error. By applying $(1 - B)$ to Z_t, we have

$$(1 - B)Z_t = (1 - B)(Y_t + e_t) = b_t + (1 - B)e_t = b_t + e_t - e_{t-1}. \quad (6)$$

Let $w_t = b_t + e_t - e_{t-1}$. It is easily seen that

$$E(w_t) = 0, \text{ Var}(w_t) = \sigma_b^2 + \sigma_\varepsilon^2, \quad (7)$$

and

$$\text{Cov}(w_t, w_{t-l}) = \begin{cases} -\sigma^2 & for\ l = 1 \\ 0 & for\ l > 1 \end{cases} \quad (8)$$

Thus, w_t is an MA (1) process and can be written as $w_t = a_t - \theta_{at-1}$ where {at} is a Gaussian white noise series with parameters θ and σ^2 satisfying

$$(1 + \theta^2)\sigma_a^2 = \sigma_b^2 + \sigma_\varepsilon^2 \text{ and } \theta\sigma_a^2 = \sigma_\varepsilon^2 \quad (9)$$

Consider an AR (p) time series, say

$$\emptyset(B)Y_t = b_t. \quad (10)$$

Very often we cannot observe Y_t precisely. Instead, only a contaminated version of it is observed, say (5) above. Here, among many possible sources, e_t can simply represent a measurement error which, for simplicity, is serially uncorrelated. Then, we have

$$\emptyset(B)Z_t = \emptyset(B)(Y_t + e_t) = b_t + \emptyset(B)_{et}. \quad (11)$$

It is easy to check that the right hand side of the prior equation is an MA (p) process. Therefore, Z_t follows an ARMA (p, q) model. This is one of the reasons for using MA models. This point, however, appears not to be fully appreciated in the economic literature. Of course, MA parameters are usually hard to interpret and the AR models are "relatively" easier to estimate.

2.2 ARIMA (0, 2, 2) Model

Another model often used in forecasting is the ARIMA (0, 2, 2) model

$$(1 - B)^2 Z_t = (1 - \theta 1B - \theta_2 B^2)_{at} \quad (12)$$

where the zeros of $\theta(B)$ are outside the unit circle. Write the model as

$$(1 - B)(1 - B)Z_t = (1 - \lambda_1)(1 - \lambda_2)_{at} \quad (13)$$

and define a new time series

W_t by $(1 - B)W_t = (1 - \lambda_2)_{at}$. (14)

Clearly, this newly defined series is an exponential smoothing process. That is, W_t follows an ARIMA (0,1,1) model. It is also easy to see that $(1 - B)\ Z_t = (1 - \lambda_1)\ W_t$, saying that Z_t is an exponential smoothing model with an exponential innovational series. Thus, Z_t is referred to as a double exponential smoothing model. It is informative to calculate the π-weights and ψ-weights of this ARIMA (0, 2, 2) model.

In statistical process control, [15] opined that ARIMA models are used for process improvement by identifying the common causes in auto correlated behavior. Statistical process control (SP) is known as a random process-that is, a process generating independent and identically distributed (iid) random variables. Once a state of statistical control is attained, departures from statistical control may occur. These departures are usually seen in extreme individual observations (outliers) or deviant sequences of observations (runs above and below a level or runs up and down). Also, according to [15], departures from a state of statistical control are discovered by plotting and viewing data on a variety of control charts, such as Shewhart, cumulative sum (CUSUM), exponentially weighted moving average (EWMA), and moving-average charts. These departures could be special causes - introduced by [16] or common causes - suggested by[17].

3.0 The Data and Method

The data for this research work was collected from the Academic planning unit of the University of Lagos on the students admitted into both undergraduate and postgraduate programmes of the Institution from 1965 to 2011. There were two faculties at the start of the University and as a result of growth over fifty years; the Institution has been able to have twelve faculties altogether including College of Medicine. The University has eighty-three (83) departments all together. The School of Postgraduate of the University of Lagos has eighty-three (83) programs which cut across all the departments in the University. The data was modeled using ARIMA (p,d,q) to detect changes and stability in the admission process of the University of Lagos over the years. The control chart for the process was developed and prediction was made.

4.0 Results and Discussion

The model for ARIMA (0, 1, 1) is

$$X_t = X_{t-1} + 0.400\varepsilon_t \quad (15)$$

and the ARIMA (2,1, 2) is of the form

$$X_t = 187.214 + 0.0009X_{t-1} - 0.77999X_{t-2} + 0.395\varepsilon_{t-1} - \varepsilon_{t-2} \quad (16)$$

The Table1 shows the result of ARIMA (0, 1, 1) that the t-statistic of parameter MA1 and their associated p values is not statistically significant. This implies that when the time series model is correctly specified the Autocorrelation function (ACF) and the Partial Autocorrelation function (PACF) of error series should be significantly different from zero. Table 2 result shows that ARIMA (2, 1, 2) has a t-statistic of parameters AR2, MA1, MA2 and their associated p values are statistically significant. This implies that ACF and the PACF of error series is not different from zero. Furthermore, the results of descriptive statistics has a sample mean of 3615.61 with a confidence interval of (2129.11, 5102.10) for ARIMA (0, 1, 1) while ARIMA (2, 1, 2) has a sample mean of 3625.78 with a confidence interval of (1612.14, 5639.43).This implies that ARIMA (2, 1, 2) has a wider length than the ARIMA (0, 1, 1) when looking at the confidence interval. Also, the mean of ARIMA (2, 1,2) is greater than that of ARIMA (0,1,1)

Table 1: ARIMA (0, 1, 1) Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|
MA1	1	0.40078793	0.1973708	2.03	0.0516
Intercept	0	163.854085	140.38989	1.17	0.2527

Table 2: ARIMA (2, 1, 2) Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1	1	0.00090602	0.0328767	0.03	0.9782
AR2	2	-0.7998832	0.122621	-6.52	<.0001
MA1	1	0.39464846	0.1202438	3.28	0.0029
MA2	2	-1	0.1837452	-5.44	<.0001
Intercept	0	187.214829	167.41238	1.12	0.2737
Constant Estimate		336.795198			

Tables 3&4 are the summary for undergraduate and the post graduate enrolments from 1965 to 2011. The time series ARIMA models for (0, 1, 1) and (2, 1, 2) was applied to the data. The result for undergraduate enrolment shows that the R square for ARIMA (0, 1, 1) is 37.24% and ARIMA (2, 1, 2) is 52.84%. Also, for the postgraduate enrolment for ARIMA (0,1,1) is 57.73% and ARIMA (2,1,2) is 62.73%.The implies that ARIMA (2,1,2) was more fitted than ARIMA (0,1,1). The model was both stable and invertible.

Table 3: Model Summary for Undergraduate Enrolment Data

Model	ARIMA (0, 1, 1)	ARIMA (2, 1, 2)
DF	29	26
Sum of Squared Errors	50024068	34040738.5
Variance Estimate	1724967.86	1309259.17
Standard Deviation	1313.38032	1144.22864
Akaike's 'A' Information Criterion	449.182288	446.634133
Schwarz's Bayesian Criterion	452.050263	453.804069
R Square	0.37247883	0.52842026
R Square Adj	0.35084017	0.45586953
-2LogLikelihood	443.289958	435.394084
Stable	Yes	Yes
Invertible	Yes	Yes

Table 4: Model Summary for Postgraduate Enrolment Data

Model	ARIMA (0, 1, 1)	ARIMA (2, 1, 2)
DF	29	26
Sum of Squared Errors	13283010.4	11137459.6
Variance Estimate	458034.843	428363.83
Standard Deviation	676.782714	654.495095
Akaike's 'A' Information Criterion	408.075717	411.999574
Schwarz's Bayesian Criterion	410.943691	419.16951
R Square	0.57729887	0.62725528
R Square Adj	0.56272296	0.56990993
-2LogLikelihood	403.064432	399.64229
Stable	Yes	Yes
Invertible	Yes	Yes

The Figures1&2 is based on the assumption that there is no correlation between successive observations. However, since the assumption of independence of observations is questionable in admission process in the University, existence of autocorrelation should be investigated in the first place for both undergraduate and postgraduate data collected. The individual control chart plot of observed undergraduate enrolment showed that six points were out of control. The plot reveals that the first three years of the University enrolment were out of control. The enrolment was also shown out of control in 1998/1999, 2001/2002 and 2002/2003 sessions. The individual observed control chart for postgraduate enrolment also reveals that the first four years of enrolment were out of control. The enrolment was statistically out of control in 1992/1993, 2001/2002, 2007/2008 and 2010/2011 sessions. Further, chart of common and special causes are carried out to detect in the statistical process (SP) if any inherent cause regarding the lack of control in the process.

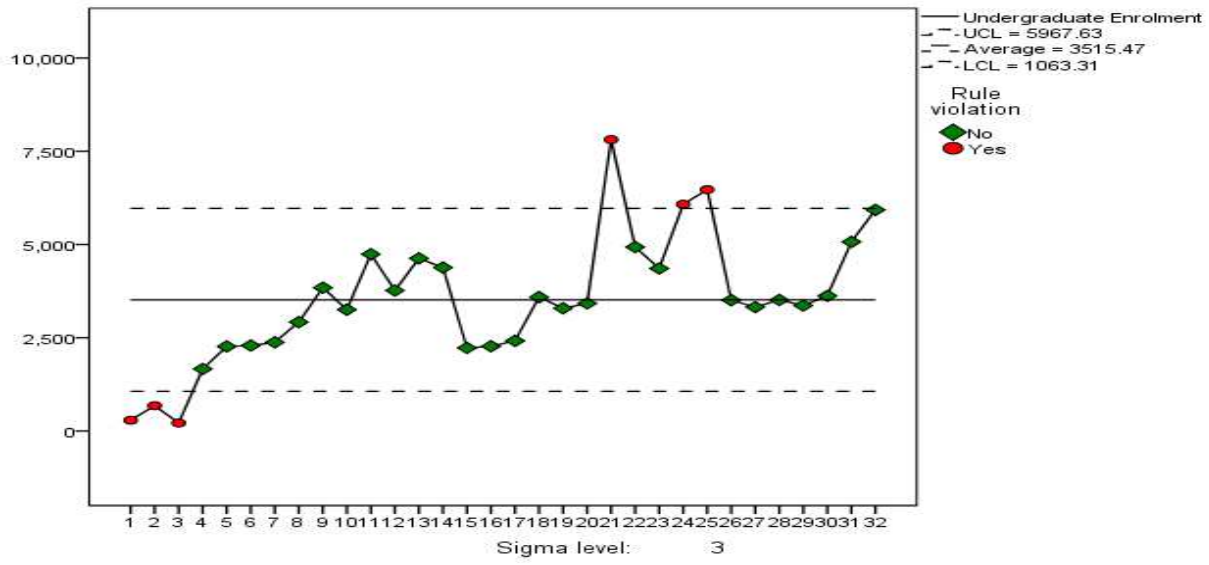

Fig1: Control chart for individual undergraduate

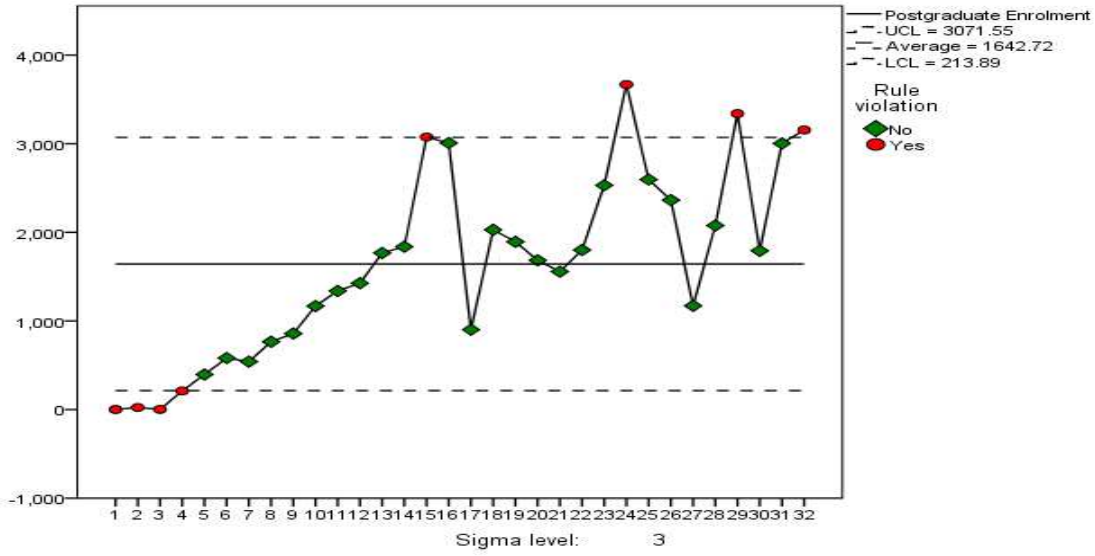

Fig 2: Control chart for individual Postgraduate

The autocorrelation function and the partial autocorrelations function in the figures 3, 4, 5 and 6 above show that the plots indicate positive correlation in the observations. Since there is a serious amount of autocorrelations in the data, an ARIMA model should be fitted in to them. Also, before the ARIMA model is fitted the trend analysis was carried out since identification process for Autoregressive (AR) and Moving Average (MA) requires stationary series. The result of the Autocorrelation function and the partial autocorrelation function reveals that there is no indication of the presence of the trends since ACF and PACF decay fast to zero after the first few lag. Therefore ARIMA (0, 1, 1) and ARIMA (2, 1, 2) model should be fitted and the series is stationary.

Lag	AutoCorr
0	1.0000
1	0.5940
2	0.3678
3	0.3125
4	0.1651
5	-0.0366
6	-0.0599
7	0.0748
8	-0.0023
9	-0.0207
10	0.1067
11	0.1673
12	0.0877
13	0.0919
14	0.0815

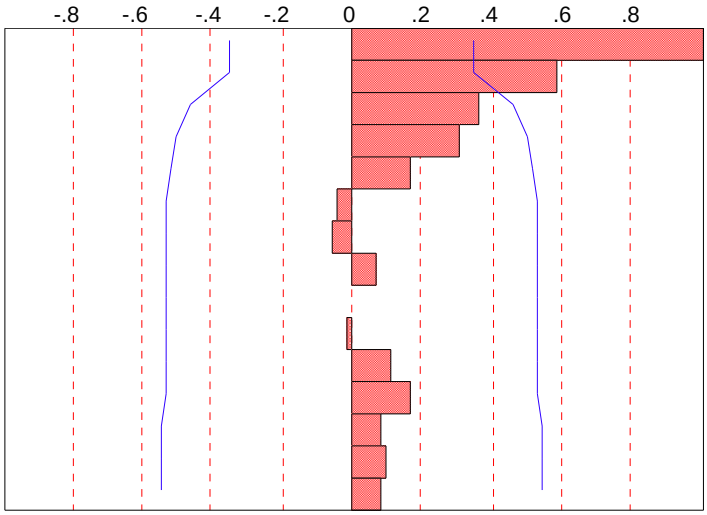

Fig3:Autocorrelation for undergraduate

Lag	AutoCorr
0	1.0000
1	0.6562
2	0.4779
3	0.4517
4	0.3207
5	0.3200
6	0.2540
7	0.1962
8	0.2255
9	0.1116
10	0.0383
11	-0.0843
12	-0.1741
13	-0.0635
14	-0.1217

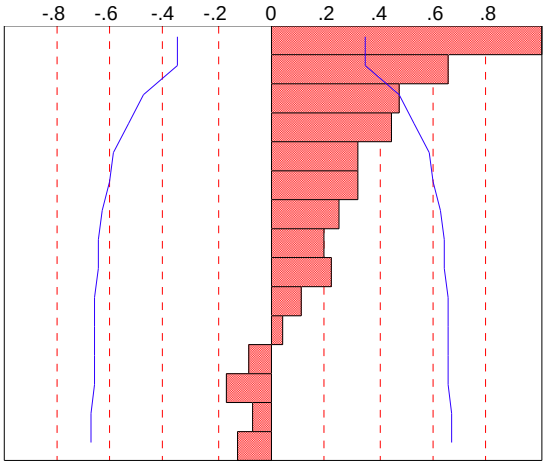

Fig 4:Autocorrelation for postgraduate

Lag	Partial
0	1.0000
1	0.7238
2	0.0900
3	0.2773
4	-0.1825
5	-0.3081
6	0.0259
7	0.4919
8	0.2580
9	0.2526
10	0.2326
11	-0.1747
12	11.9751
13	0.0000
14	0.0000

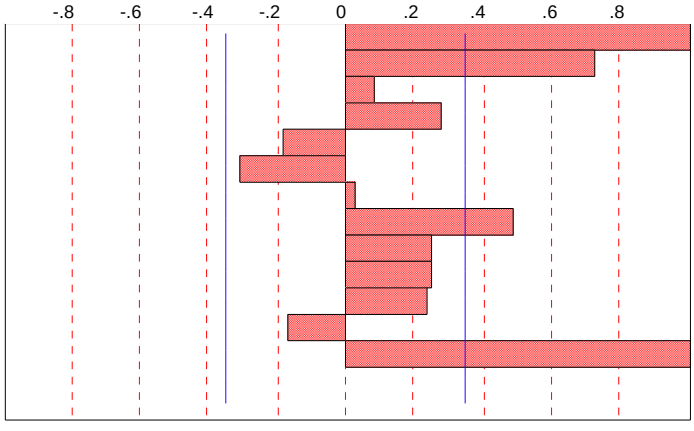

Fig 5: PartialAutocorrelation for undergraduate

Lag	Partial
0	1.0000
1	0.7679
2	0.1948
3	0.3027
4	0.0519
5	0.3598
6	-0.1531
7	0.1944
8	0.4405
9	0.0832
10	0.0807
11	-0.9106
12	-0.9605
13	0.3605
14	-0.4673

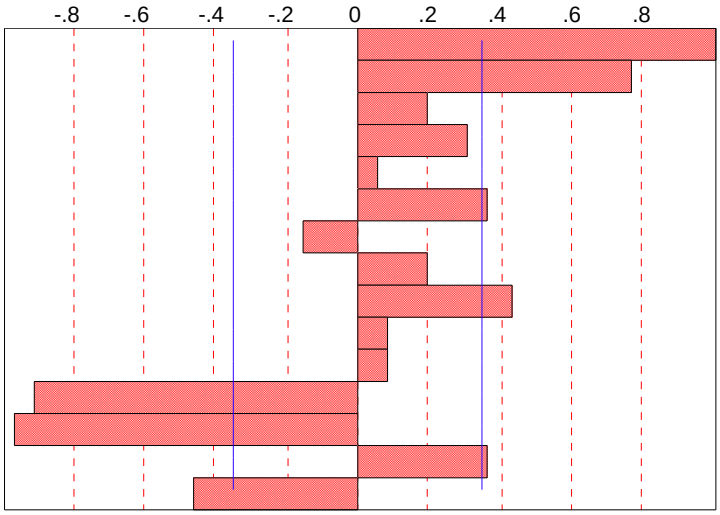

Fig 6: Partial Autocorrelation for postgraduate

The charts for ARIMA (0, 1, 1) and ARIMA (2, 1, 2) were plotted in the figures 7&8. The result shows that the statistical control chart based on the fitted values of ARIMA (2, 1, 2) showed evidence of a better understanding of the common-cause of the University enrolment system than the ARIMA (0, 1, 1) model. Nine (9) data were out of control in ARIMA (0, 1, 1) model and eight (8) data were out of control in ARIMA (2, 1, 2). The confidence intervals implied that all points are within these limits. Generally, from visual examination, it is observed that the series is obviously out of control, with strong evidence of positively auto-correlated behavior (see also Box and Jenkins (1976). This implies that corrective actions are needed to control undergraduate

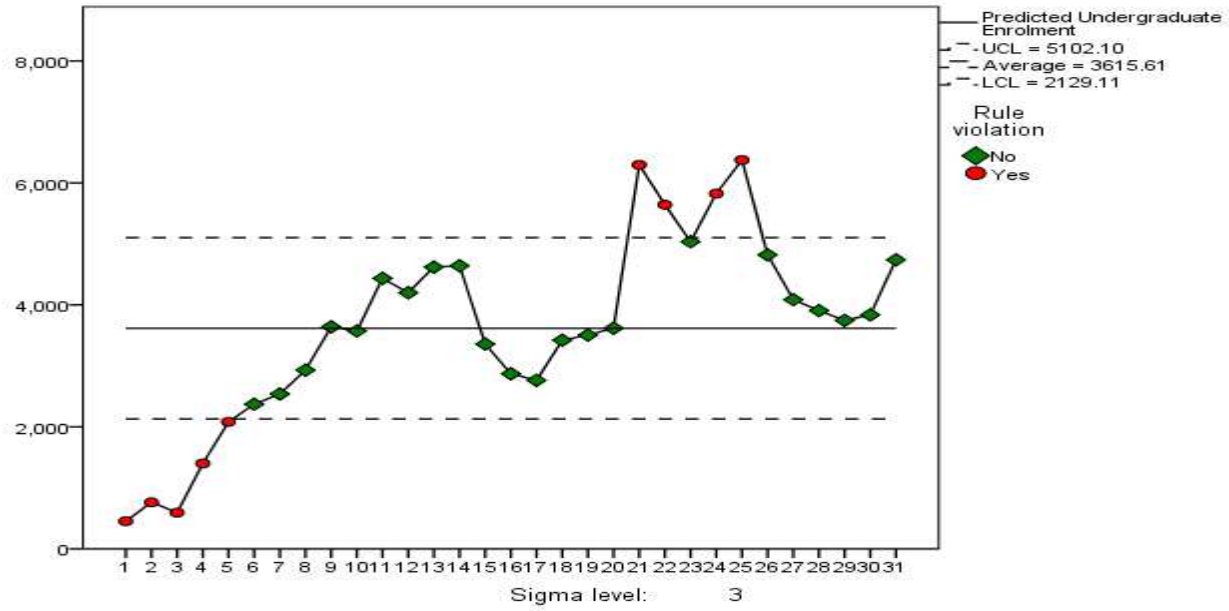

Fig 7: Control chart based on ARIMA (0, 1, 1)

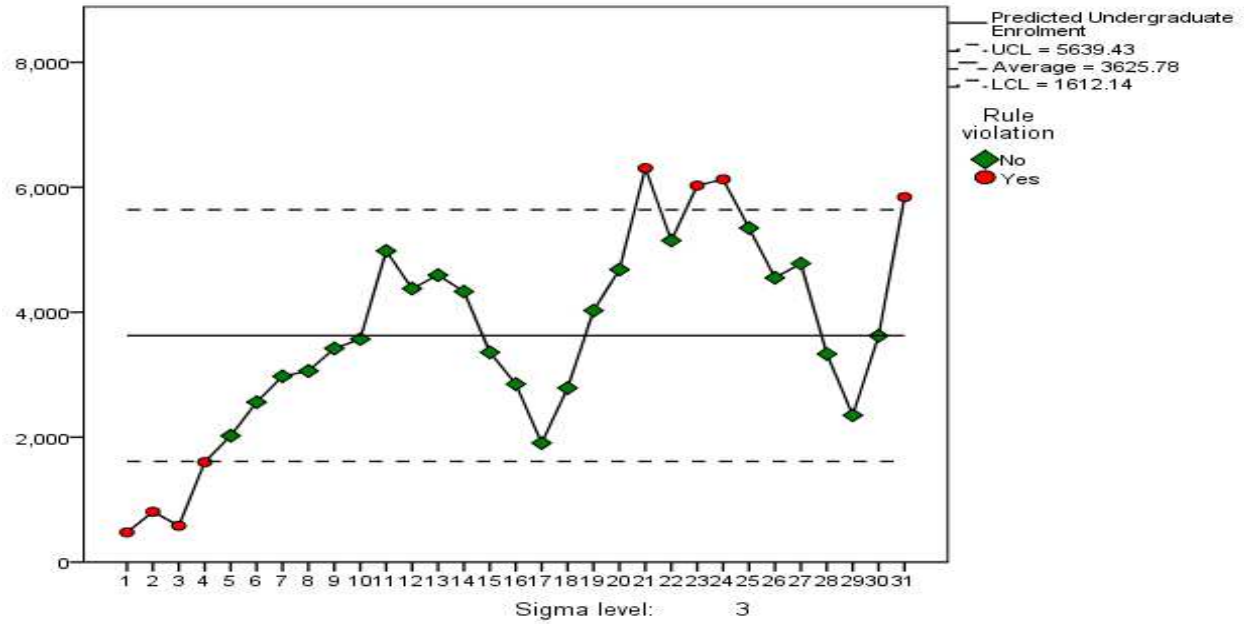

Fig 8: Control chart based on ARIMA (2, 1, 2)

Figures 9&10 showed Special-Cause Chart, which is the sequence Plot of Residuals, based on ARIMA (0, 1, 1) and ARIMA (2, 1, 2). The chart depicts that one individual points-observation 21 was out of control in ARIMA (0, 1, 1) and ARIMA (2, 1, 2) had no points out of control Since the residual has been estimated with no violation of control limits, it has taken care of case of assignable causes. Also in term of stability when we compare Fig. 9 and Fig. 10 we can see that ARIMA (2, 1, 2) is less stable than ARIMA (0,1,1). This further implied that there is no inherent special-cause in the undergraduate enrolment process system in University of Lagos.

Figures 11&12 of postgraduate enrolment process showed the sequence plot of residuals, based on ARIMA (0, 1, 1) and ARIMA (2, 1, 2). The chart depicts that one individual points-observation 27 was out control in ARIMA (0, 1, 1) and ARIMA (2, 1, 2) has no points out of control.

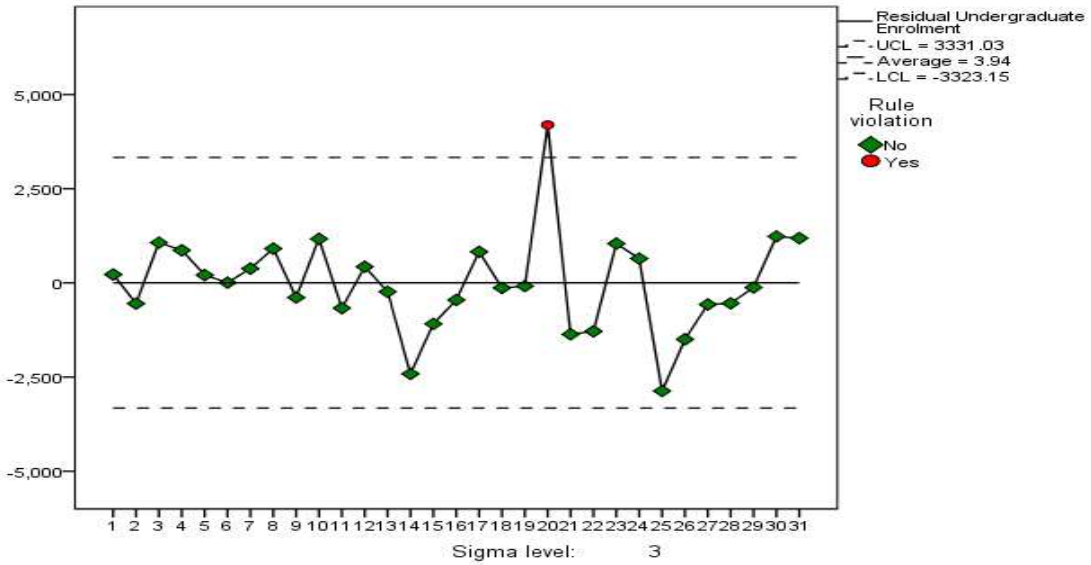

Fig 9: Undergraduates Control chart of residual based ARIMA (0, 1, 1)

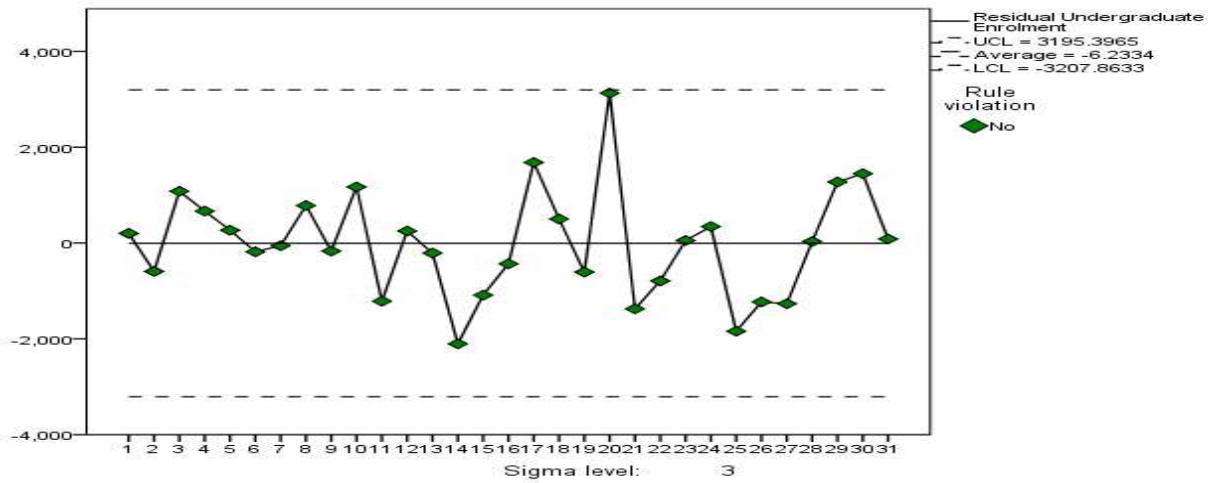

Fig 10: Undergraduate Control chart of residual based ARIMA (2, 1, 2)

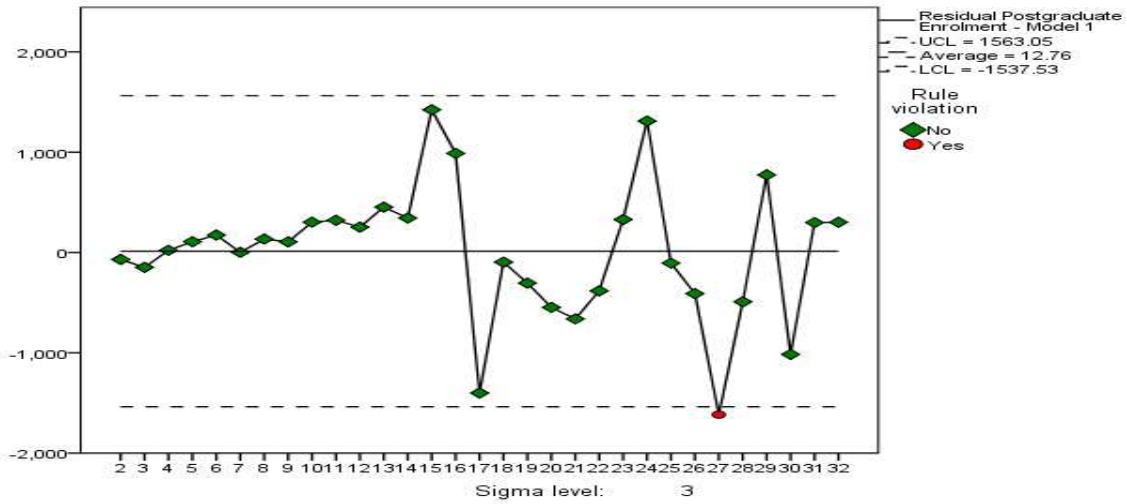

Fig 11: Postgraduates Control chart of residual based ARIMA (0, 1, 1)

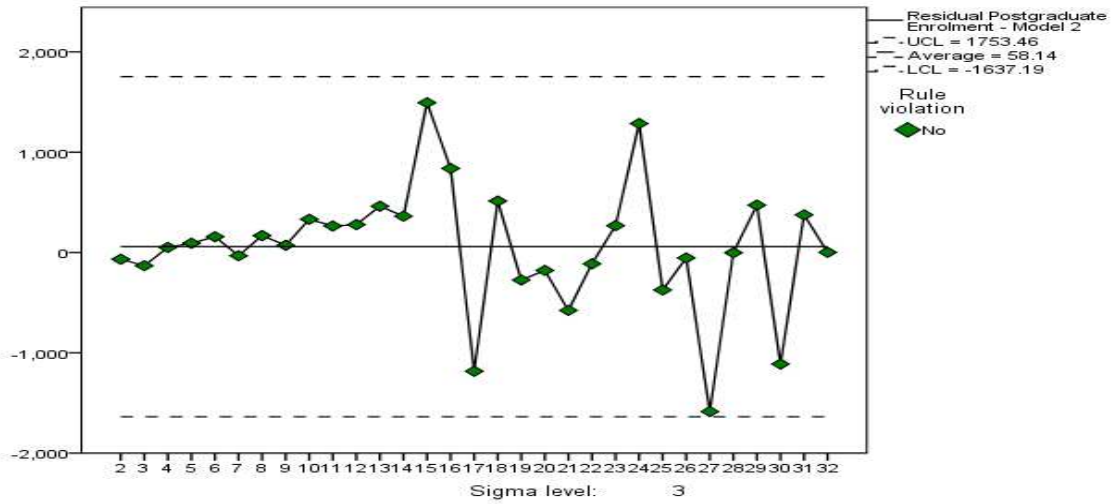

Fig 12: Postgraduate Control chart of residual based ARIMA (2, 1, 2)

The Figure13&14 of ARIMA (0, 1, 1) and (2, 1, 2) shows that the undergraduate control chart was out of control due to the common cause such as incremental rate in the admission process. The issue of the special causes also plays an important role in the process such as decay in infrastructural facilities such as laboratories, lecture theaters and Libraries. The control charts in Figure15&16 show that ARIMA (2, 1, 2) gives a better understanding of the common-cause of the postgraduate enrolment system than the ARIMA (0, 1, 1) model. Furthermore, the results of descriptive statistics for ARIMA (0, 1, 1) gave sample mean of 3615.61 with a confidence interval of (2129.11, 5102.10) and ARIMA (2, 1, 2) gave a sample mean of 3625.78 with a confidence interval of (1612.14, 5639.43).

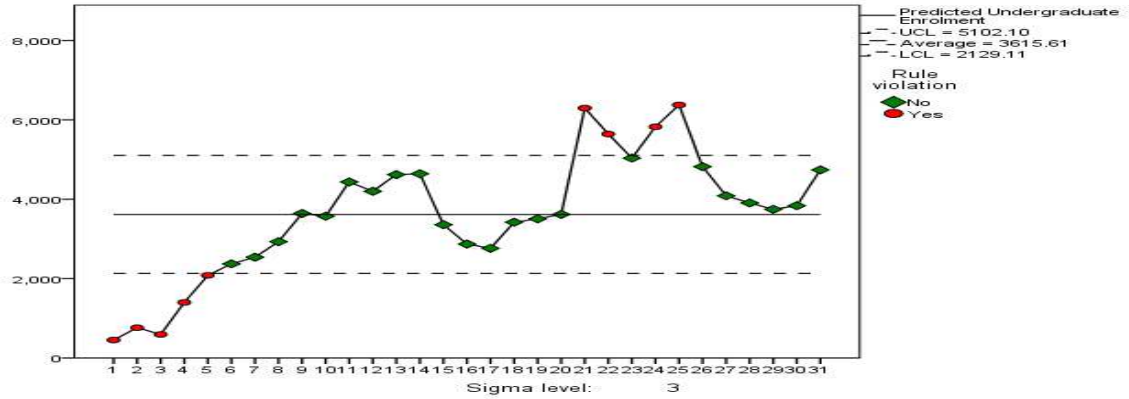

Figure 13: Common Cause for undergraduate on ARIMA (0, 1, 1)

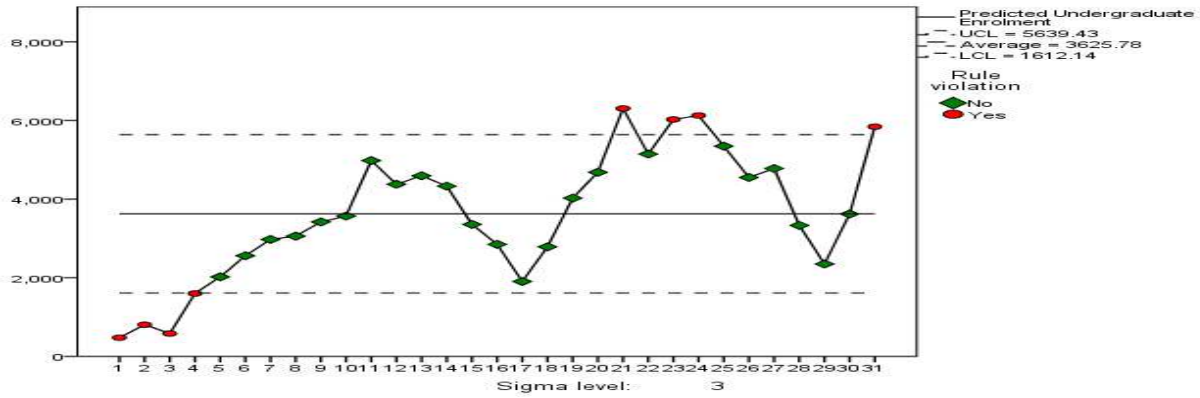

Figure 14: Common Cause Chart for undergraduate based on ARIMA (2, 1, 2)

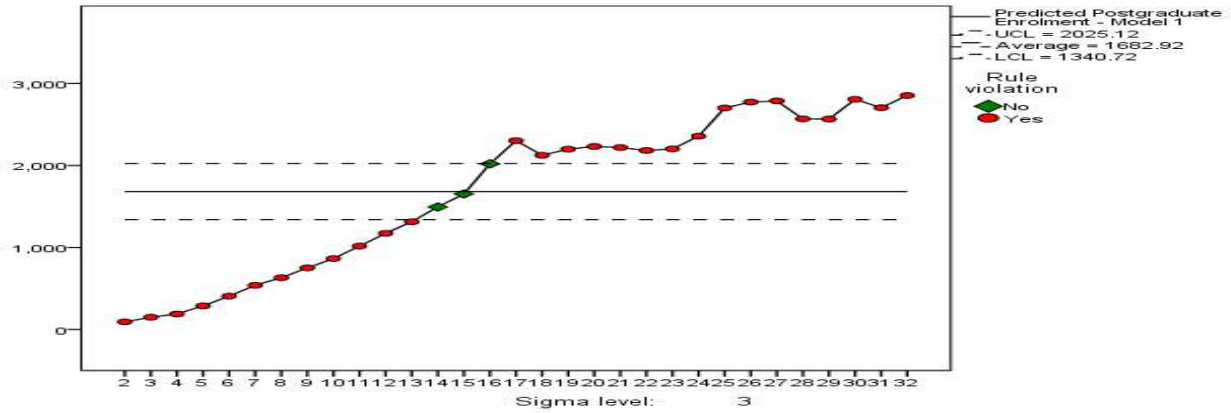

Fig15: Common Cause Chart based on ARIMA (0, 1, 1)

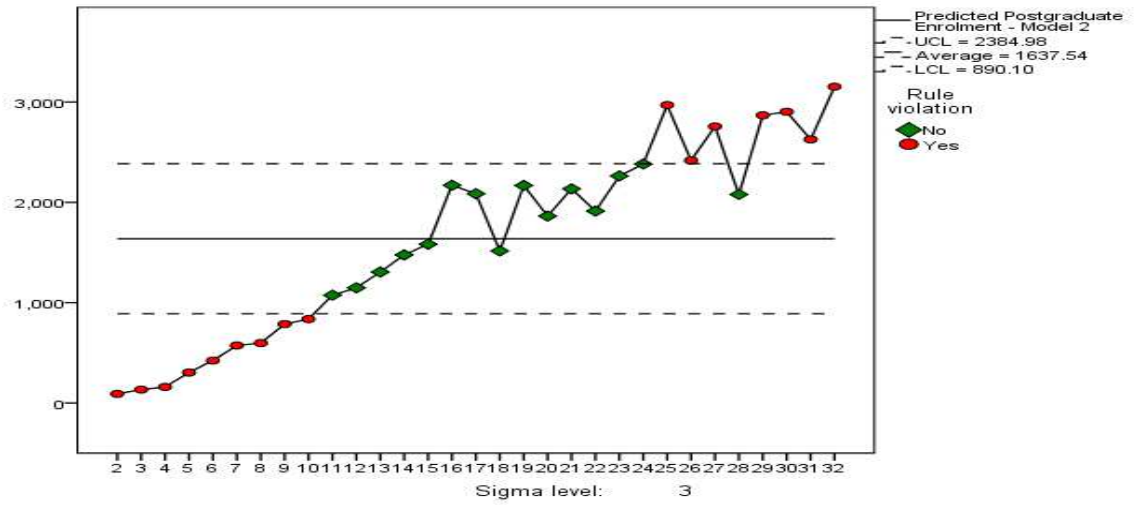

Fig 16: Common Cause Chart based on ARIMA (2, 1, 2)

5.0. Conclusion

The t-statistic of parameters AR2, MA1, MA2 and their associated p-values is statistically significant. This implies that when time series model is correctly specified the ACF and the PACF of error series should not be significantly different from zero. Also, the results of descriptive statistics has a sample mean of 3615.61 with a confidence interval of (2129.11, 5102.10) for ARIMA (0, 1, 1) and ARIMA (2, 1, 2) has a sample mean of 3625.78 with a confidence interval of (1612.14, 5639.43). The autocorrelation function and the partial autocorrelations functions show that the plots indicate positive correlation in the observations. The result of the Autocorrelation function and the partial autocorrelation function reveals that there is no indication of the presence of the trends since ACF and PACF which decayed fast to zero after the first few lags. The result of the ARIMA (0, 1, 1) and ARIMA (2, 1, 2) model show that specific corrective actions are needed to control enrolments. ARIMA (2, 1, 2) gives a better stability than ARIMA (0, 1, 1).

6.0 Reference

[1]. Alt, F.B., 1985. Multivariate quality control. Encyclopedia Stat. Sci. 6, 113-122.

[2]. Alt, F.B. & Bedewi, G.E., 1986. SPC of dispersion for multivariate data. In *Annual Quality Congress Transactions*. 40th Annual Quality Congress. Anaheim, CA, pp. 248-254.

[3]. Alt, F.B and Smith N.D (1998) Multivariate Process Control. Hand Book of Statistics, P.R. Krishnainah and C.R. Rao (eds). North-Holland Elsevier Science PublisherB.V.7 333-351

[4]. Djauhari, M.A. (2005). A measure of multivariate data concentration. *Journal of Applied Probability and Statistics*, 2, 139-155.

[5]. Sakata, T. (1987). Likelihood ratio test for one-sided hypothesis of covariance matrices of two normal populations. *Communications Statistics: Theory and Methods*, 16, 3157-3168.

[6]. Calvin, J. A. (1994). One-sided test of covariance matrix with a known null value. *Communication in Statistics: Theory and Methods*, 23, 3121-3140

[7]. Levinson, W., Holmes, D. S. and Mergen, A. E. (2002). Variation charts for multivariate processes. *Quality Engineering*, 14, 539-545.

[8]. Vargas, N. J. A. and Lagos, C. J. (2007). Comparison of multivariate control charts for process dispersion. *Quality Engineering*, 19, 191-196.

[9]. Yeh, A.B., Huwang, L., Wu, Y.-F., 2004. A likelihood-ratio-based EWMA control chart for monitoring variability of multivariate normal processes. IIE Trans. 36, 865-879

[10]. Reynolds, R. M. and Cho, G. Y. (2006). Multivariate control charts for monitoring the mean vector and covariance matrix. *Journal of Quality Technology*, 38, 230-253.

[11]. Reynolds, M. R., Jr., and Stoumbos, Z. G. (2006). "Comparisons of Some EWMA Control Charts for Monitoring the Process Mean and Variance" *Techno metrics*, 48, 550-567

[12]. Huwang, L., Yeh, A. B. and Wu, C. W. (2007). Monitoring multivariate process variability for individual observations. *Journal of Quality Technology*, 39, 258-278

[13]. Maboudou-Tchao, E. M. and Hawkins, D. M. (2011). Self-starting multivariate control charts for location and scale. *Journal of Quality Technology*, 43, 2, 113-126.
[14]. Box G.E.P. and Jenkin G,M (1976) Time Series Analysis Forecasting and Control (2nd ed.) San Francisco: Hoden-Day
[15] Alwan, L C. and Roberts, H. V. (1988)**.**Time-Series Modeling for Statistical Process Control Journal of Business & Economic Statistics, Vol. 6,(1), 87-95.
[16]. Shewhart, W. A. (1931), Economic Control of Quality of Manufactured Product, New York: Van Nostrand. (Republished in 1981, with a dedication by W. Edwards Deming, by the American Society for Quality Control, Milwaukee, WI.)
[17]. Deming, W. E. (1982), Quality, Productivity and Competitive Position, Cambridge, MA: MIT Center for Advanced Engineering Study
[18]. Box, G.E.P., Jenkins, G.M., Reinsel, G.L., (1994), *Time Series Analysis. Forecasting and Control*, Prentice Hall, Englewood Cliffs, New Jersey