Regression And Time Series Analysis Of Loan Default At Minescho Cooperative Credit Union, Tarkwa

Otoo, H., Takyi Appiah, S., Wiah, E. N.

Abstract: Lending in the form of loans is a principal business activity for banks, credit unions and other financial institutions. This forms a substantial amount of the bank's assets. However, when these loans are defaulted, it tends to have serious effects on the financial institutions. This study sought to determine the trend and forecast loan default at Minescho CreditUnion, Tarkwa. A secondary data from the Credit Union was analyzed using Regression Analysis and the Box-Jenkins method of Time Series. From the Regression Analysis, there was a moderately strong relationship between the amount of loan default and time. Also the amount of loan default had an increasing trend. The two years forecast of the amount of loan default oscillated initially and remained constant from 2016 onwards.

Keywords: Loan, forecasting, Stationary, differencing, partial autocorrelation, cyclic

1 Introduction

Loan portfolio is typically the largest asset and the most predominant source of income for any financial institution (Aballey, 2009). In spite of the huge income generated from their loan portfolio, available literature shows that huge portions of financial institution loans usually ends in default and therefore affect the financial performance of these institutions (Aballey, 2009). Loan default is the inability to repay the loan by either failing to complete the loan as per the loan agreement or neglect to service the loan. In finance, default to occurs when a debtor has not me this or her legal obligations according to the debt contract (Murray, 2001).According to Pearson and Greeff(2006),loan defaults a riskthres hold that describes the point in the borrower's repayment this tory where he or she miss edatleast three (3)instalments within a twenty four(24)month period. This represents apointintimeandan indicator of behaviour, where in here is ademonstrable increase in the risk that the borrower eventually will truly default, by ceasing all repayments. Many factors have been identified as major determinant so loan defaults. The nature, time of disbursement, supervision and profitability of enterprises which benefits from the small holder loan scheme contributes other repayment ability and consequently high default rates (Okorie, 1986). Moreover, GorterandBloem(2002) observed that non-performing loan sure mainly caused by an inevitable number of wrong economic decisions by individuals and plain bad luck(inclement weather, unexpected price changes for certain products etc.).Also high interest rates and incorrect information address of guarantors are also possible factors of loan default.

- *Otoo Henry–-Department of Mathematics, University of Mines and Technology, Tarkwa, Ghana,E-mail: hotoo@umat.edu.gh*
- *Appiah Takyi Sampson –Department of Mathematics, University of Mines and Technology, Tarkwa, Ghana, e-mail: stappiah@umat.edu.gh*
- *Wiah Neebo Eric- Department of Mathematics University of Mines and Technology, Tarkwa, Ghana. enwiah@gmail.com*

2 Study Areas

The Minescho Cooperative Credit Union or Minescho Credit Union (Minescho) is one of the vibrant Co-operative Credit Unions in the Tarkwa-Nsuaem Municipal Assembly. It has over the years engaged in the important role of credit creation by way of advancing loans to its customers. Since its inception, the credit union has given credit to wide varieties of customers. Currently, Minescho Credit Union is a proud registered member of the Ghana Co-operative Credit Union Association. Its management includes Board of Directors, the loan committee and the manager/accountant. The objectives of this Credit Union are:

- To promote thrift among its members by providing a means of savings.
- To provide loans to its members for provident or productive purposes.
- To provide quality financial services to its members.

The vision is to become a leading co-operative financial institution in Ghana, providing financial and technical product and services to members, and build a strong and committed staff to create an enviable image in the country. Therefore, since Loan default is inevitable in most financial institution, this research seeks to determine and forecast the trend of loan default at Minescho Credit Union.

3 Methods Used

3.1 Data Collection

Data collection deals with gathering data/information for statistical analysis. The data may be primary data or secondary data. For purposes of this research, secondary data were obtained from the Minescho Co-operative Credit Union, Tarkwa.

3.2 Regression Analysis

Regression analysis is used to predict the value of one variable on the basis of other variables. The technique involves developing a mathematical equation that describes there lationship between the variable to before

casted, which is called the dependent variable, and variables that the statistician believes are related to the dependent variables which are mostly denoted by $x_1, x_2, ..., x_k$ (where k is the number of independent variables) (Bluman, 2009).

Simple Linear Regression

The line a regression model is a statistical technique in which the expected value of a dependent variable is expressed as a linear combination of asset of explanatory variables. When the linear regression model involves only one independent variable it is termed as a simple linear regression. Linear regression is denoted by

$$y_t = \alpha + \beta x_t + \varepsilon_t \ (1)$$

Where $i = 1, 2, 3, .., n$

y_t = Dependent variable

α = Intercept parameter

β = Coefficient of regression orthe slope parameter

ε_t = Error term

x_t = Independent variable

Hypothesis Testing

Hypothesis testing is a method of statistical inference used for testing a statistical supposition (Montgomery and Runger, 2003). In carrying out any hypothesis test, two hypotheses are considered, namely: the null hypothesis H_0 and the alternate hypothesis H_1. In hypothesis testing, sample data are used together with knowledge of statistical theories to decide whether the sample data support the null hypothesis. On the contrary, if the sample data do not support the null hypothesis, it is rejected as being an unreasonable supposition. The null hypothesis H_0, is a statistical hypothesis that states that there is no differences between a parameter and a specific value, or that there is no difference between two parameters (Bluman, 2009). The alternate hypothesis, denoted by H_1 is the hypothesis that states the existence of a difference between a parameter and a specific value, or states there is a difference between two parameters (Bluman, 2009).

3.3 Time Series

Times series is an ordered sequence of values of a variable at equally spaced time intervals (Otoo *et al*., 2015). It is mathematically defined as a set of vectors $X(t) = 0, 1, 2, 3, ...$ where t represents the time elapsed. The variable $X(t)$ is treated as a random variable. Time series can be continuous ordiscrete (Brockwell and Davies, 2001).The mathematical expression describing the probability structure of a time series is termed as a stochastic process. Thus the sequence of observations of the series is actually a sample realization of stochastic process that produced it (Hipel and McLeod, 1994). A time series in general is affected by four main components, which can be separated from the observed data. These components are: Trend, Cyclical, Seasonal and Irregular components (Chatfield, 1996). Thus, considering the effects of these four components, two different types of models are generally used for a time series. These are Multiplicative and Additive models.

Multiplicative model

This model assumes that the components interact with each other and do not move independently. It is expressed mathematically as

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t) \ (2)$$

Additive model

The additive model of time series is generally expressed as

$$Y(t) = T(t) + S(t) + C(t) + I(t) \ (3)$$

Where $Y(t)$ is the observation and $T(t), S(t), C(t)$, and $I(t)$ are respectively the trend, seasonal, cyclical and irregular variation at time t.One major weakness of the additive model is that of its unrealistic assumption that the components are independent of each other (Brockwell and Davies, 2001).

3.4 Time Series Models

Autoregressive model AR(P)

The notation AR (p) refers to the autoregressive model of order p. The AR (p) model is denoted by

$$X_t = c + \sum_{i=1}^{p} \theta_i x_{t-i} + \varepsilon_t \ (4)$$

where $\theta_i, ..., \theta_p \ (\theta_p \neq 0)$ are the parameters, c is a constant, and the random variable ε_t is white noise. Some constraints are necessary on the values of the parameters so that the model remains stationary. Processes in $AR(1)$ model with $|\varphi_1| \geq 1$ are not stationary (Yavuz and Gazanfer, 2011).

Moving-Average model

The notation $MA(q)$ refers to the movingaverage model of order q:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i} \ (5)$$

where $\varphi_i,...,\varphi_q\left(\varphi_q \neq 0\right)$ are the parameters ofthe model, μ is the expectation of X_t, and the $\varepsilon_t,\varepsilon_{t-i},...,\varepsilon_{t-q}$,are white noise error terms.

Autoregressive, Moving Average Model (ARMA)

The notation $ARMA\left(p,q\right)$ refers to the model with p autoregressive terms and q moving-average terms. This model contains the $AR\left(p\right)$ and $MA\left(q\right)$ models,

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p}\theta_i x_{t-i} + \sum_{i=1}^{q}\theta_i\varepsilon_{t-i} \quad (6)$$

Auto Regressive Integrated Moving Average Model (ARIMA).

In Time Series Analysis, an Autoregressive Integrated Moving Average (ARIMA) model is a generalization of an Autoregressive Moving Average (ARMA)model (Otoo *et al*, 2015; Adhikari, 2013). These models are fitted to time series data either to better understand thedata or to predict future points in the series. They are applied in some cases where data show evidence of non-stationarity, where an initial differencingstep can be applied to remove then on-stationarity. The model is generally referred to as $ARIMA\left(p,d,q\right)$ model where parameters p,d,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 modelling.

3.5 The Box- Jenkins Methodology

GeorgeBox and Gwilym Jenkins developed a practical approach to build Autoregressive Integrated Moving Average (ARIMA) model, which best fit to a given time series and a lsosatisfy the parsimony principle.

The approach methodology does not assume any particular pattern in the historical data of the series to before casted but rather uses the following basic Iterative approach of

- Differencing the series to achieve stationarity
- model identification
- parameter estimation and
- diagnostic checking
- using models for forecasting

To determine the best parsimonious model from a general class of ARIMA models. The Box-Jenkins ARMA model is a combination of the Autoregressive (AR) and Moving Average (MA) models where the terms in the equation have the same meaning as given for the AR and MA model.

$$X_t = \delta + \theta_1 X_{t-1} + \theta_2 X_{t-2} + ... + \theta_n X_n + A_t - \varphi_1 A_{t-1} + \varphi_2 A_{t-2} + ... + \varphi_r A_{t-r} \quad (7)$$

Where X_t is the time series, A_t is the white noise and $\delta = \left(1 - \sum_{i=1}^{n}\theta_i\right)\mu$ with μ denoting the process mean. The Box-Jenkins model assumes that the time series is stationary. Non-stationary series are difference done or more times to achieve stationarity. The differencing of the non-stationary series results in an ARIMA model, with the"I" standing for "Integrated". Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero (Brockwell and Davis, 2001).

Table 1 *Summary of Loan Default Data from Minescho*

Month	Year						
	Amount of Loan Defaulted						
	2008	2009	2010	2011	2012	2013	2014
Jan	3867.57	1935.47	22287.59	42729.25	40661.78	38292.01	73506.7
Feb	5112.38	11780.25	3286.85	8330.55	27035	93983	125684
Mar	999	4458.35	23293.74	33844.6	48673.57	55116.64	23126.18
Apr	1904.23	5521.67	24218.5	11158	11549.5	39445.16	39260
May	1524	10706.78	29042.08	4293	47508	50974.12	40714
Jun	4328	23243	19574.98	21385	65232	65615.2	32286.61
Jul	1157.9	11491.37	62502.5	50906.54	44730	115942.23	758576
Aug	13544.6	7555.42	29542	26786	17486	92261.3	29818
Sept	58926.09	10039.6	5239.34	7669	24418.8	66632.33	5288

Oct	192803.48	10046	15761	44227.13	33494	84383.7	24887.52
Nov	8827.21	8072.94	3770	87465	50484.53	88197.8	23228.52
Dec	4537	18666	21656.2	8679	45032	165946.5	2920
Total	**297531.46**	**123516.85**	**260174.78**	**347473.07**	**456305.18**	**956789.99**	**1179295.53**

Source: Minescho Credit Union

4 Data Analysis, Results and Discussions

4.1 Regression Analysis of Loan Default

The Trend of Loan default is analysed using data from the year 2008 to 2014 as displayed in table 1. Regression analysis is employed to fit a model for the data in Table 4.1 to establish the relationship between the dependent variable (Amount in Default) and the independent variable (Months/Time). A scatter plot of the Loan Default versus time (months) with the trend line (in blue) is shown in Fig.1 below.

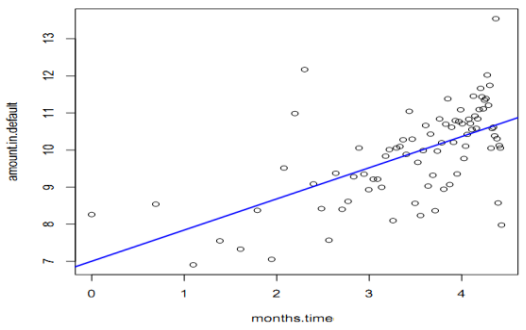

Fig.1 *Scatter Plot of Amount in Default Versus Time showing Trend Line.*

From fig.1, the estimated regression model expressed as

$$X_t = -0.9746 + 0.4480t + \varepsilon_t \qquad (8)$$

Where $\alpha = -0.9746$ and $\beta = 0.4480$ .Thus, the expected number of Loan default when time, *t* = 0 is -0.9746. The slope of the trend line is 0.4480. This means the slope line is positive and slopes upward from left to right. Hypothesis testing and analysis of variance (ANOVA) are performed on the parameters to know whether they are statistically significant.

Hypothesis Testing

The parameter β was tested at 95% confidence level to determine whether is statistically significant or not.
Hypothesis. The formulated hypothesis to be tested is:

The Null Hypothesis, $H_0 : \beta = 0$

The Alternate Hypothesis, $H_1 : \beta \neq 0$

Analysis of Variance

The results of the analysis of variance are tabulated below.

Table 2 *Analysis of Variance (ANOVA)*

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F-value	Pr(>F)
Time	1	26.409	26.4086	49.464	5.534e-10
Residuals	82	43.779	0.5339		
Total	83				

From the F-distribution table, the critical value, F_{tab} is 3.957 on 1 and 82 degrees of freedom. The calculated F-value, F_{cal} is 49.464 on 1 and 82 degrees of freedom from table 2. The estimated p-value is 5.534e-10. Since the $F_{cal} > F_{tab}$, the null hypothesis, H_0 is rejected and the alternate hypothesis, H_1 is accepted. Thus, $\beta \neq 0$. The p-value being very small also suggests that, the H_0 is to be rejected.

The Coefficient of Determination, R^2 and the Correlation Coefficient, R

The estimated R^2 is 0.3763. This means that, the time explains 37.6% of the variations in the number of loan default. The estimated Pearson's Product-Moment correlation (correlation coefficient), R is 0.6132. This shows a moderately strong correlation between the time and the amount of Loan default at Minescho Credit Union.

Diagnostic Check: Evaluation of Residuals

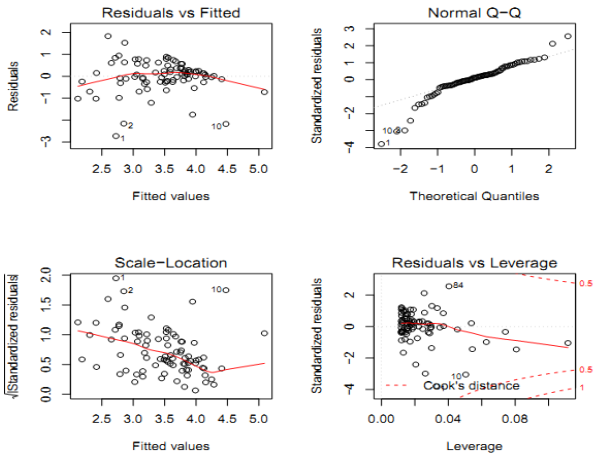

Fig. 2 *Diagnostic Plot of the Regression Model*

Analysis and Discussion

The plot in the upper left of fig. 2 shows the residual errors plotted versus their fitted values. The residuals are randomly distributed around the line representing a residual error of zero; that there is no distinct trend in the distribution points. The identified outliers are data points 1, 2 and 10. The plot in the upper right of fig. 2 depicts the standard Q-Q norm plot, which suggests that the errors are normally distributed. Thus most of the plots were found on the 45° line. The plot on the lower left of fig. 2 which is the scale-location plot shows the square root of the standardized residual as a function of the fitted values which shows no obvious trend. The identified potential outliers are data points 1, 2 and 10. The fourth plot (bottom right in fig.2) shows the plotted contour lines for the Cook's distance. Since the distances are small, it implies that the observations have little effect on the regression results.

Shapiro-Wilk Normality Test

The Shapiro-Wilk Normality Test is another approach of testing for the normality of residue. With regards to this test, when the p-value is 0.5 or more then it is normally distributed. The Shapiro-Wilk test gave a p-value of 0.7942 which indicates the errors are normally distributed. On this account, the model is valid.

Addison Darling Normality Test

Finally the Addison Normality Test was used to confirm the Shapiro-Wilk test. It gave a p-value of 0.6778. Since it has a p-value greater than 0.5 then the errors are normally distributed.

Analysis and Trend Determination

The estimated trend equation in equation (6) gives a trend line with a positive slope. From Fig.1, the trend identified from the scatter plot of fig.1 is an upward trend. Since the coefficient of time, $\beta = 0.4480$ is statistically significant at the 95% significant level, it is concluded that, the Loan Default time series has an upward trend.

Explanation

The upward trend shown by model means that the monthly record of Loan Default at Minescho Credit Union is increasing in number. This can be reduced if appropriate measures are put in place when giving out loans to customers.

4. 2 Box-Jenkins Analysis

First a Time Series Plot of the data in Table 1 is made to test the stationarity of the data and is shown on the graph in fig. 3

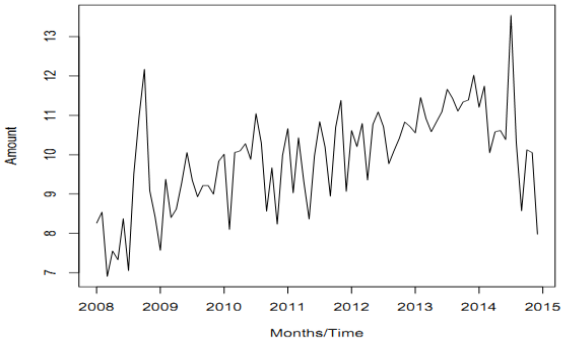

Fig. 3 *Time Series plot of Loan Default*

From fig. 3 the highest loan default was recorded in July 2014 and lowest was recorded in March 2008. It can be observed that the data is not stationary since it shows irregularly increase and decrease on the graph.

Decomposition

Decomposition is performed to identify the various components present in the Loan Default Time Series. The decomposed components in the series are displayed in fig. 4 below. Components like random, trend and seasonal are depicted in the graph.

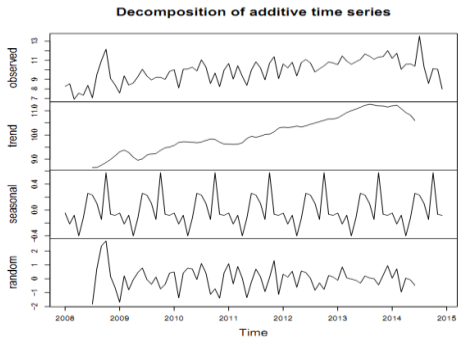

Fig. 4 *Components of the Loan Default Time Series*

Differencing Data to Achieve Stationarity

Stationarity of the Loan Default Time Series is achieved by differencing the data. The graphs of the first and second differencing are shown below in fig. 5 and 6

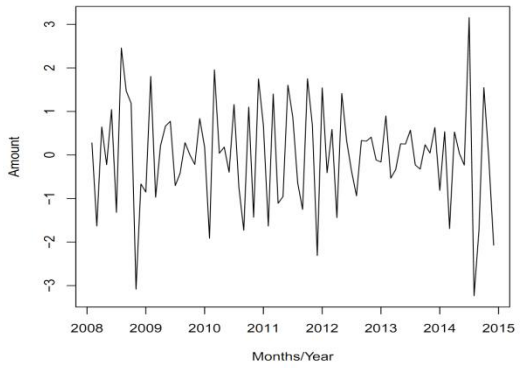

Fig. 5 *Time Series Plot of First Differenced Data*

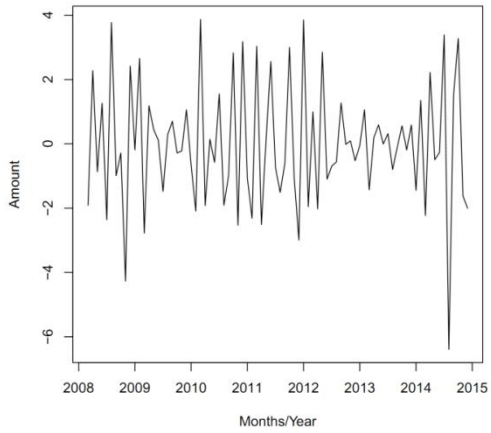

Fig. 6 *Time Series Plot of Second Differenced Data*

From fig 5 and 6, the means of both the first and second differences are around zero and with constant variance. Thus both cases show the attainment of stationarity in data, with the second difference being better skewed toward zero.

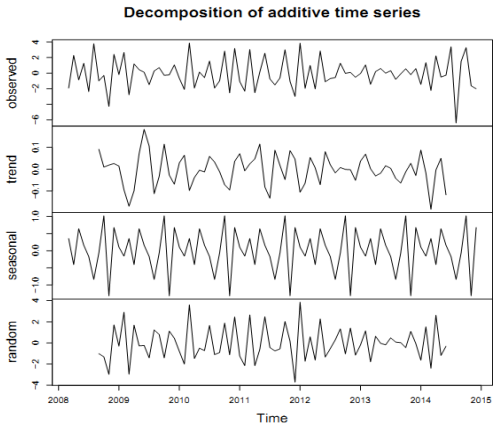

Fig.7 *Components of the Second Differenced Time Series Data*

The trend component has been removed as depicted in fig. 7. The seasonal variations have decreased and only the random component is left. This indicates that the time series of loan default is now stationary. The suggested model for the time series of the amount of loan default is ARIMA (p,2,q). This is as a result of the second differencing.

Model Identification

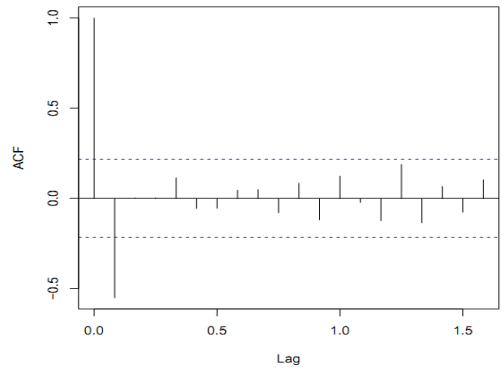

Fig.8 *Correlogram of ACF of Second Differenced Data*

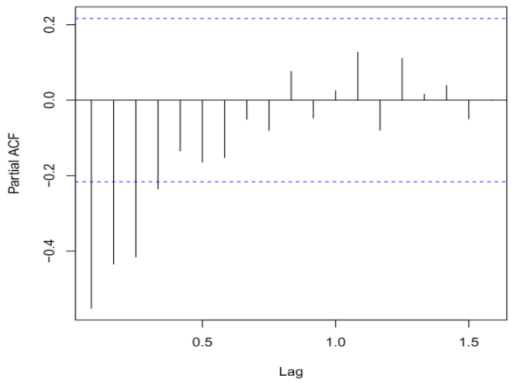

Fig. 9 *Correlogram of PACF of Second Differenced Data*

The correlogram of both ACF and PACF respectively of the second differenced data is shown in fig 8 and 9 respectively.The ACF plot has two significant spikes at lag 1 and lag 2. These lags ensure the existence of q in the ARIMA model which represents the Moving Average. It can also be observed from fig.9 that the PACF has about 3 significant spikes which are all negatives. Therefore, the possible models identified for the Loan Default time series are ARIMA(1,2,1),AR1MA(1,2,2), AR1MA(2,2,1), AR1MA(2,2,2), ARIMA(3,2,1) and ARIMA(3,2,2)

Model Estimation

The estimated ARIMA models with their model statistics are summarized in the table below.

Table 3 *The Estimated ARIMA Models*Model

	MAE	MAPE	BIC	AIC	RMSE
ARIMA(1,2,1)	1.229089	270.8152	327.10	319.92	1.612642
ARIMA(1,2,2)	0.8574073	200.5054	282.92	273.34	1.143721
ARIMA(2,2,1)	1.074815	342.3341	313.19	303.78	1.429285
ARIMA(2,2,2)	0.7999818	151.834036	279.63	267.66	1.081911
ARIMA(3,2,1)	0.9382942	183.4036	300.64	288.81	1.277621
ARIMA(3,2,2)	0.7782332	120.285	280.64	266.28	1.054089

Analysis and Discussions

Among the six estimated models, ARIMA (3,2,2) has the smallest AIC value of 266.28, MAPE value of 120.285 and MAE value of 0.7782332 as indicated in table 3. It also has the lowest RMSE value of 1.054089. Since the ARIMA (3,2,2) has the smallest AIC values then it is the best model among the six estimated models. The estimated parameter for the best model is tabulated in table 4.

Table4: *ARIMA (3,2,2) Parameters*

Model ARIMA (3,2,2)					
	ar1	ar2	ar3	ma1	ma2
Coefficient	-0.4856	-0.4039	-0.2085	-1.9834	1.0000
Standard error	0.1102	0.1159	0.1113	0.0594	0.0591

Mathematically, the ARIMA (3,2,2) model is written as

$$X_t = -0.4856X_{t-1} - 0.4039X_{t-2} - 0.2085X_{t-3} + 1.9834\varepsilon_{t-1} - 1.0000\varepsilon_{t-2} + \varepsilon_t \quad ()$$

Diagnostic Check of the Model

In order to ensure that the ARIMA (3,2,2) model is the best model fit for the incident time series, a diagnostic check is done. The mean of the residual plot is approximately zero. Also the Ljung-Box statistic was 13.954 with a p-value 0.646. This indicates the model is significant. The standard residual plot together with its ACF and the Ljung Box statistic is shown in the fig. 10 below

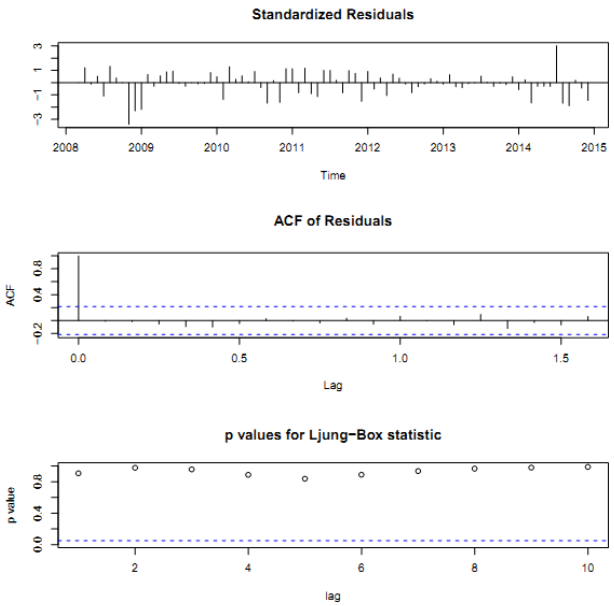

Fig.10 *Diagnostic Check ARIMA(3,2,2) Model*

Therefore the best fit model for Loan Default from January 2008 to December 2014 is

$$X_t = -0.4856X_{t-1} - 0.4039X_{t-2} - 0.2085X_{t-3} + 1.9834\varepsilon_{t-1} - 1.0000\varepsilon_{t-2} + \varepsilon_t$$

Forecasting with the model

The ARIMA (3,2,2) model was used to forecast the number incidents for a 24-month period. Thus, January 2015 to December 2016. The time series plot of the forecast is shown below.

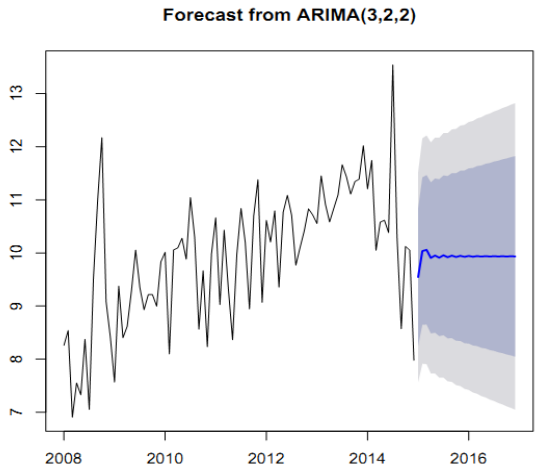

Fig.11 *A Time Series Forecast from ARIMA (3,2,2) model*

From Figure 4.11, the forecast from ARIMA (3,2,2) model shows an oscillatory trend for some time and remain constant from January 2015 to December 2016.

5 Conclusions

The Loan default at Minescho Credit Union follows an upward trend from the scatter plot as shown in fig. 4.1. Equation (8) is a good predictive model for the trend of loan default. The forecast for future loan default oscillates and remains constant from 2016 onwards. The Minescho Credit Union should make thorough background checks of customers before giving out loans.

REFERENCES

[1] Aballey, F.B. (2009), "Bad Loan Portfolio: A Case Study of Agricultural Development Bank", Unpublished Dissertation, KNUST.

[2] Adhikari, R. and Agrawal, R.K. (2013), "An Introductory Study on Time Series Modelling and Forecasting, Pearson Education Press, Delhi. pp. 12-14.

[3] Bluman, G.A. (2009), Elementary Statistics: A Step by Step Approach, Von Hoffman Press, St Louis, USA. 7th Edition. 401pp

[4] Brockwell, P. J. and Davis, R. A. (2001),

[5] Introduction to Time Series and Forecasting, Springer-Verlag, New York, 2nd Ed. 449 pp.

[6] Chatfield, C. (1996), "The Analysis of Time Series", Chapman and Hall, New York, USA. 5th Edition, 248pp.

[7] Gorter, N. and Bloem, M. (2002). "The Macroeconomic Statistical Treatment of Non-Performing Loans", Publication of The Organisation for EconomicCorporation and Development,[http://www.dbj.go.jp/english/IC/active/hot/adfiap/pdf /nagarajan.pdf], (accessed 18 March, 2015)

[8] Hipel, K.W and Mcleod, A.I, (1994), Time Series Modelling of Water Resources and Environmental Systems, Elsevier Science B.V, 1012pp.

[9] Montgomery, D.C. and Runger, G.C. (2003), Applied Statistics and Probability for Engineers, John Wiley and Sons Inc., New York, USA. pp. 279-280.

[10] Murray, J. (2011). "Default on loan", United States Business and Taxes Guide. 84pp

[11] Okorie, A. (1986). "Major Determinants of Agricultural Loan Repayments, Savings and Development, X(1)

[12] Otoo, H., Wiah, E.N., Nabubie, I. B, Ahialekedzi, I. K., (2015) "Determining and Forecasting the Trend of Incidents at a Mining Company in Tarkwa", International Journal of Statistics, Vol. 39, No. 1, pp. 1121-1130

[13] Pearson, R. and Greef, M. (2006). "Causes of Default Among Housing Micro Loan Clients, FinMark Trust Rural Housing Loan Fund", National Housing Finance Corporation and Development Bank of Southern Africa, South Africa.

[14] Yavuz Yildirim and Gazanfer Ünal, (2011), "Modelling ISE100 with continuous AR (1) Model", International Journal of Business and Management Studies, Vol. 3, No. 1 , pp. 1309-8047