Science World Journal Vol. 17(No 3) 2022
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

TIME SERIES MODELLING OF DIABETES DISEASE IN TARABA STATE, NIGERIA

*[1]Pascalis Kadaro Matthew, [2]Kurutsi Nuhu Timothy, [2]Rita Ajia, [1]Solomon Antyev

[1]Department of Mathematical Sciences, Taraba State University, Jalingo – Nigeria
[2]Department of Basics Science, College of Agriculture, Jalingo – Nigeria

*Corresponding Author Email Address: pkadaro@yahoo.com

ABSTRACT
In this study, we applied an Autoregressive Integrated Moving Average (ARIMA) model to predict the spread of Diabetes disease infection in Taraba State, Nigeria. The monthly recorded cases of Diabetes between January 2010 and December 2020 in Federal Medical Centre, Jalingo was used to fit and validate the ARIMA model. A seasonal fluctuation and a slightly increasing pattern of a long-term trend were revealed in the time series of Diabetes disease. ARIMA (0,1,1) was selected as the best optimal model which has the lowest value of AIC/BIC. The root mean square error (RMSE) was used to assessed the predictive capability of the optimal model. The twenty-four (24) months forecast of Diabetes disease infection in Taraba State, Nigeria was also presented. The ARIMA model could be applied to effectively predict the short-term Diabetes disease infections in Taraba State, Nigeria and provide support for the development of interventions for disease control and prevention.

Keywords: Diabetes; ARIMA; Forecasting; AIC, BIC, RMSE.

INTRODUCTION

Diabetes mellitus is chronic non-communicable disease associated with long term complications to the brain, kidney, and the heart. There is destruction and loss of the s-cells of the pancreas causing insulin deficiency; it may also result from abnormalities arising from resistance to insulin. Symptoms of hyperglycemia include polydipsia, polyphagia polyuria, blurred vision, weight loss, generalized pruritus, neuropathy, retinopathy, etc. Life threatening consequences of uncontrolled diabetes include diabetes-ketoacidosis, lactic acidosis and hyper-osmolar non-ketotic state (Diabetes Care,2006). Diabetes is preceded by impaired fasting glucose (IFG) resulting in a pre-diabetic state which can exist undetected for many years, (Nathan, et.al, 2007) causing irreversible damage to vital organs. Pre-diabetes is a practical term referring to Impaired Fasting Glucose (IFG), impaired glucose tolerance (Ronal and Zubin, 2013) or a glycosylated hemoglobin (A1c) of 6.0 to 6.4%, each of which places individuals at high risk of developing diabetes and it complications. The World Health Organization criteria for diagnosing pre-diabetes are fasting plasma glucose level of between 6.1 and 6.9 mmol/l. A fasting plasma glucose level 7.0 mmol/l or more meets the criteria for the diagnosis of diabetes. Fasting value for venous and capillary plasma glucose are identical (WHO, 2019).

There is an increasing prevalence of diabetes and pre-diabetes worldwide (IDF,2007). Over 5 million people suffer from the disease in Africa and the number is expected to skyrocket to 15 million by 2025 (IDF,2007). In Nigeria the prevalence varies from 0.65% in rural Mangu village to 11.0% in urban Lagos (Akinkugbe, 1997). With the incidence of diabetes in Africa, diabetic complications are also expected to rise proportionately (Wild et.al, 2004; Zimmet, 2003). This will undoubtedly pose serious health and economic problems. The disease affects many people under the age of 64 years in Africa as compared to the developed world where it affects many people over the age of 64 years (Wild et.al, 2004). In Nigeria the National prevalence of diabetes was 2.2% (Akinkugbe, 1997). In South Eastern Nigeria the overall prevalence of diabetes was 10.51% (Chris et.al, 2012), whereas in South Western Nigeria the prevalence of diabetes ranges from 4.76% in Ile-Ife, Osun State to 11.0% in Lagos (Akinkugbe, 1997; IDF,2007). Also 0.8% of diabetes mellitus, and 2.2% of Impaired Glucose Intolerance in Ibadan (Olatunbosun, 1998). This was comparable to WHO reported a prevalence of 2.8% in Ibadan (Owoaje, 1997), and 6.8% in Port Harcourt, Nigeria (Nyenwe,2003). In 2004, a survey in Jos (Nyenwe, 2003) reported a prevalence of 10.3%. (Nyenwe, 2003) reported a prevalence of 2.2% in Port Harcourt in 2003. A prevalence of 4.7% was reported by (Lucia and Prisca, 2012) which was higher than the national prevalence of 2.2% reported by (IDF,2007). A review of studies on the prevalence of diabetes in adults in Africa (Unwin, Sobugwi, & Alberti, 2001) demonstrated a rising prevalence across the continent.

Time series analysis is one of the quantitative methods which can effectively predict the future incidence of communicable diseases and epidemiological trends using previously observed data and time variables (Zenq et.al,2016). This analysis deals with time dependent variables with an advantage of being not necessary to consider the influence of intricate factors (Wang et.al, 2016; Xu, 2017).

Time series methods have been widely used to analyzed infectious diseases' surveillance data in recent decades, including data for sexually transmitted diseases. Different time series models were used to forecast the epidemic behaviour in previous study (Zhang et.al 2014). For example, decomposition methods were used to forecast nine notifiable infectious diseases in China (Zhang et.al 2013). Autoregressive integrated moving average models (ARIMA) are widely applied in infection time series modelling including tuberculosis (Rios et.al, 2000), typhoid fever (Zhang et.al 2014), gonorrhoea (Dowell et.al, 2011) and hepatitis (Ture &Kurt, 2006). Autoregressive conditional heteroscedasticity and generalized autoregressive conditional heteroscedasticity models have been used to investigate the risk factors associated with

syphilis (Williams, 2014). the seasonal autoregressive integrated moving average (SARIMA) model has been increasingly favoured and successfully used in the prediction of communicable diseases, such as dengue (Martinez et.al, 2011), tuberculosis (Zheng et.al,2015), mumps (Xu, 2017) and others (Zheng et.al,2015; Peng et.al, 2017; Song et.al, 2016).

MATERIALS AND METHOD

This study utilized the monthly Diabetes data recorded at Federal Medical Centre, Jalingo, from 2010 to 2020 obtained from the records department of the Hospital. The ARMA or ARIMA model is used for forecasting. A typical ARMA model combines both the moving average (MA) model and the autoregressive (AR) model. The autoregressive moving average (ARMA) is given as

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t=1} + \cdots + \theta_q \varepsilon_{t-q} \quad 2.1$$

$$Y_t = \sum_{k=1}^{p} \phi_k Y_{t-k} + \sum_{k=1}^{q} \theta_k \varepsilon_{t=k} + \varepsilon_t \quad 2.2$$

Equation (2) can be expressed as

$$\phi(B) Y_t = \theta(B) \varepsilon_t \quad 2.3$$

Where, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$

If Z_t is a stationary series obtained after d differencing from Y_t series then we obtain

$$Z_t = \nabla^d Y_t = (1 - B)^d Y_t \quad 2.4$$

If Z_t follows an ARMA (p,) model, we can conclude that $\{Y_t\}$ is an ARIMA (p,d,q). Hence the expression of the ARIMA(p,d,q) model is given as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d Y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p) \quad 2.5$$

$$(1 - B)^d \phi_p(B) Y_t = \theta_q(B) \varepsilon_t \quad 2.6$$

The expression in (2.6) consist of an Autoregressive (AR) process which handles memory from previous values, an integrated procedure which makes the data stationary and a Moving-Average (MA) which represents terms of previous error. The Box-Jenkins' technique applies the ARMA or ARIMA models to obtain the best model that fits the time-series data. The stages involved in Box-Jenkins technique are as follows;

Model Identification: In developing a Box–Jenkins model, it starts with determining whether the time series data is stationary or not. A time series data is said to be Stationary when the mean and the variance are constant over a period of time. To make the data stationary, it is recommended to difference the data. The ACF and PACF is plotted immediately the data becomes stationary, this is carried out in order to identify the order of the ARMA model (i.e. *p* and *q*).

Parameter Estimation: The methods of Maximum likelihood estimation (MLE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are applied for the parameter estimation. The maximum likelihood estimation (MLE) method obtains values of the parameters that maximize the probability of obtaining the actual value of the parameter we are estimating. The likelihood function is as follows;

$$\log L = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t}^{T}\varepsilon_t^2 \quad 2.7$$

Where, T= time, $t = 1,..., T$, ε_t =error and σ^2 =constant variance.

The log likelihood with maximal value is selected as the best model. The Akaike Information Criterion (AIC) is useful for comparing models fitted to the same series. It is written as

$$AIC = -2\ln L + 2N \quad 2.8$$

where L= likelihood of the data and N= number of parameters in the model.

The Bayesian Information Criterion (BIC) penalizes the parameters more strongly than the AIC. It is written as

$$BIC = AIC + N\log(T) \quad 2.9$$

ARIMA model that has the lowest AIC, AICc, BIC with largest log likelihood will be the preferred model.

Diagnostic Model Assessment: The ACF, PACF, chi-square and the Ljung-Box Q statistics obtained from the residual of the ARIMA model are used for diagnostic checking. The model is considered fit if the residual satisfies the assumption of regression model and assumptions for a stationary process. The residuals should be white noise.

Forecasting: Accuracy measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Square Error (MASE) are used for measuring the performance accuracy of the ARIMA model. Any of the temporary fitted ARIMA models which has the smallest statistics is considered the best model for forecasting.

$$MASE = \frac{1}{N}\sum_{t-1}^{N}\left|\widehat{Y}_t - Y_t\right|^2 \quad 2.10$$

$$MAE = \frac{1}{N}\sum_{t-1}^{N}\left|\widehat{Y}_t - Y_t\right| \quad 2.11$$

Science World Journal Vol. 17(No 3) 2022
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

$$RMSE = \sqrt{\frac{1}{N}\sum_{t-1}^{N}\left|\widehat{Y}_t - Y_t\right|^2} \qquad 2.12$$

where, Yt= actual value of variable Y at time$t$, $Y\hat{t}$= predicted value of Y at time t, N= number of observations

RESULTS AND DISCUSSION

This section focus on the analysis and discussions of results using the methodology highlighted in the previous sections.

Visualize the Time series

Figure 1 shows the monthly time series plots of Diabetes diseases recorded at Federal Medical Centre, Jalingo, Nigeria from 2010-2020. The plot of the observations displays the series on the y-axis against the time interval on the x-axis showing the patterns in the data over time.

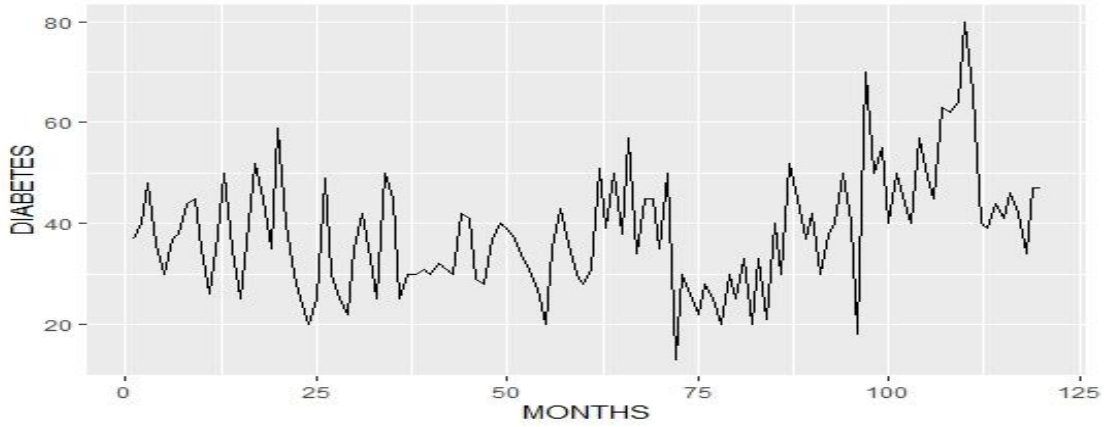

Figure 1: Monthly Time Series Plot of Diabetes cases recorded at Federal Medical Centre, Jalingo, Nigeria, from 2010-2020.

The plot of the series reveals several fluctuations across the months indicating significant variation over time, which shows that our time series is not stationary with the Augmented Dickey-Fuller Test (Dickey-Fuller = -2.9533, Lag order = 4, p-value = 0.1807). Since the p-value is greater than the significant level at 0.05 we therefore fail to reject the null hypothesis that the diabetes data is not stationary. In order to perform any successive modeling on our time series, our time series must be stationary: that is, the mean, variance, and covariance of the series should all be constant with time. We need to stationarize the time series by differencing the data.

Stationarizing the Time series

The Diabetes data attains stationarity after the first difference, with the Augmented Dickey-Fuller Test (Dickey-Fuller = -6.633, Lag order = 4, p-value = 0.01). Since the p-value is less than 0.05, we therefore reject the null hypothesis and accept the alternative hypothesis that the time series data for diabetes rate is stationary

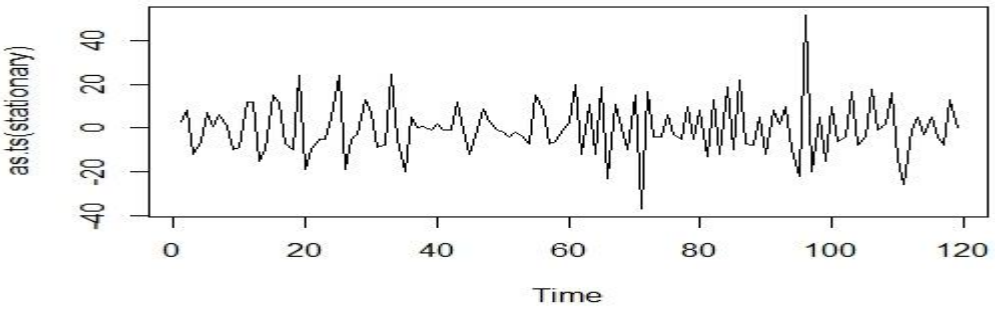

Figure 2: Time Series Plot for the differenced data

Science World Journal Vol. 17(No 3) 2022
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Model Identification

Since the diabetes data became stationary at the first lag difference, it means that the model to be considered is of this form ARIMA(p,1,q).

In Figure 3, the plots of the autocorrelation and partial autocorrelation function of the stationary time series data are examined. The ACF decay exponentially after lag 1 suggesting an MA (1) and the PACF dies off after significantly lag 1, suggesting an AR (1) and the corresponding ARIMA Model is (1,1,1).

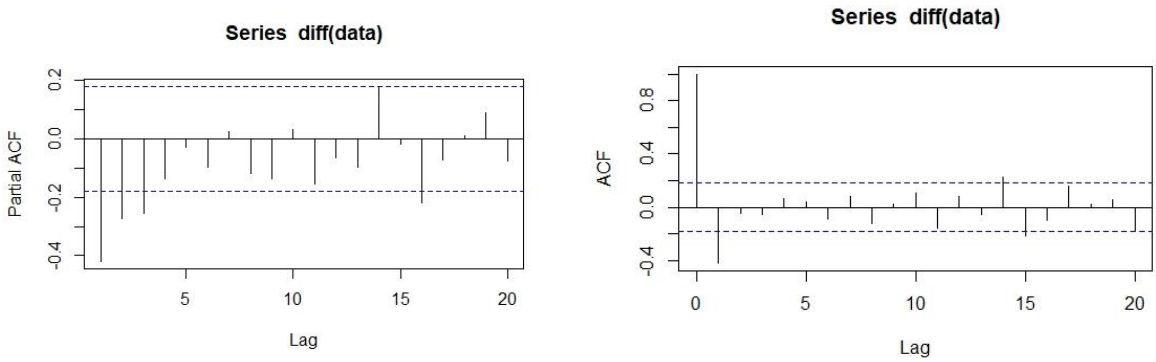

Figure 3: ACF and PACF of the first differenced of Diabetes data

Model Estimation

Our findings in the ACF/PACF section suggest that, one among the four models i.e ARIMA (1,1,1), ARIMA (1,1,2), ARIMA (2,1,1), ARIMA (2,1,2) might be the best fit. These tentative ARIMA models are assessed by choosing the least AIC, log likelihood and BIC criteria to serve as the best ARIMA model.

Table 1: Goodness of fit test and selection of optimal ARIMA model for Diabetes data

	Goodness of fit			
ARIMA Model	**RSME**	**AIC**	**BIC**	**AICc**
ARIMA (1,1,1)	11.23479	918.4	926.7394	918.61
ARIMA (1,1,2)	10.24019	901.25	912.3693	901.65
ARIMA (2,1,1)	10.79355	911.58	922.66	911.93
ARIMA (2,1,2)	10.23569	903.24	917.09	903.77

The results in Table 1 indicates that, among the four models suggested by ACF/PACF, ARIMA (1,1,2) is the best model, since it has the smallest values of AIC and BIC. However, we conducted a random experiment using an built-in function in *R software* called ''*auto.arima*'' to search for parameters, until we find the optimal parameters that yield the lowest AIC/BIC. ARIMA (0,1,1) was found to be the optimal model. Therefore, we can use ARIMA (0,1,1) for forecasting of our future values.

Science World Journal Vol. 17(No 3) 2022
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Table 2: Best ARIMA (0,1,1) Model for Diabetes data using ''**auto.arima**'' in R

	MA(1)
Coefficients	-0.7330
Standard Error	0.0777
Sigma^2	106.9
log likelihood	-446.72
AIC	897.44
AICc	897.54
BIC	903.00

Diagnostic Checking

The diagnostic checking analyzes the residuals as well as model comparisons. The Standardized residuals are independent identically distributed sequence with zero mean and variance one. The ACF and PACF of the standardized residuals show no significant lags. Also, the Ljung Box result (Q^* =4.2462, df = 7, p-value =0.751) shows that the residuals are stationary and constitute white noise. Thus the selected model is appropriate to represent the series and it is a good fit.

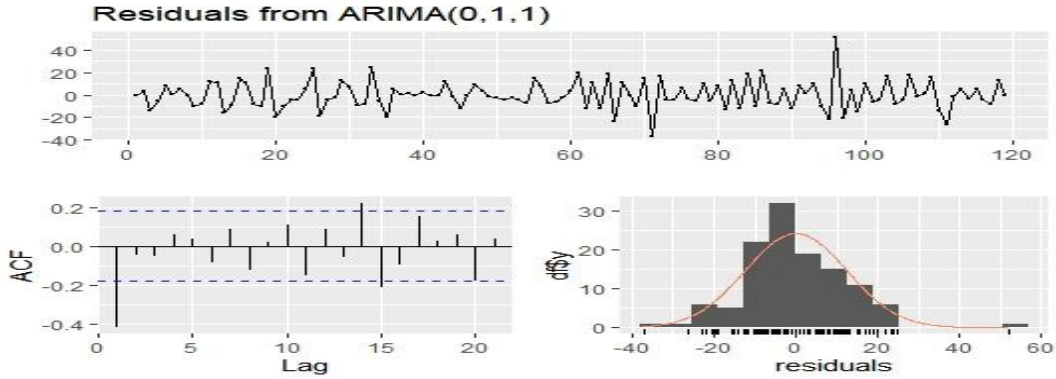

Figure 4: The standardised residual plot (a), ACF (b) and Histogram of the residuals (c) of the series after a first-order non-seasonal difference and first-order seasonal difference.

Forecasting Diabetes diseases in Taraba State, Nigeria

The parameters of ARIMA (0, 1, 1) from the result in Table 2 of the analysis is θ_1 =-0.7330.The parameters are used to modelled the Diabetes diseases. ARIMA (0,1,1) model can be obtained from equation (2.6) as follows;

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad 3.1$$

$$Y_t = \varepsilon_t - 0.7330\varepsilon_{t-1} \quad 3.2$$

Future Diabetes diseases in Taraba State, Nigeria can be forecasted using equation (3.2).

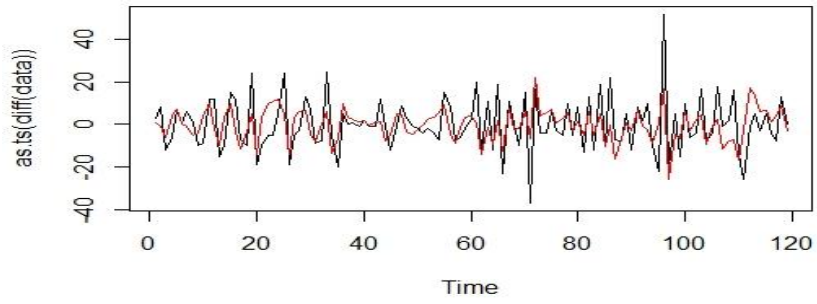

Figure 5: Timeseries plot of the actual and predicted Diabetes data

Table 4: *Predicted Diabetes diseases from January 2021 to December 2022 with the selected model*

Month	Forecast	Lower 95% Limit	Upper 95% Limit
121	40	17.7	62.3
122	38.2	15.9	60.5
123	38.2	15.9	60.5
124	38.2	15.9	60.5
125	38.2	15.9	60.5
126	38.2	15.9	60.5
127	38.2	15.9	60.5
128	38.2	15.9	60.5
129	38.2	15.9	60.5
130	38.2	15.9	60.5
131	38.2	15.9	60.5
132	38.2	15.9	60.5
133	38.2	15.9	60.5
134	38.2	15.9	60.5
135	38.2	15.9	60.5
136	38.2	15.9	60.5
137	38.2	15.9	60.5
138	38.2	15.9	60.5
139	38.2	15.9	60.5
140	38.2	15.9	60.5
141	38.2	15.9	60.5
142	38.2	15.9	60.5
143	38.2	15.9	60.5
144	38.2	15.9	60.5

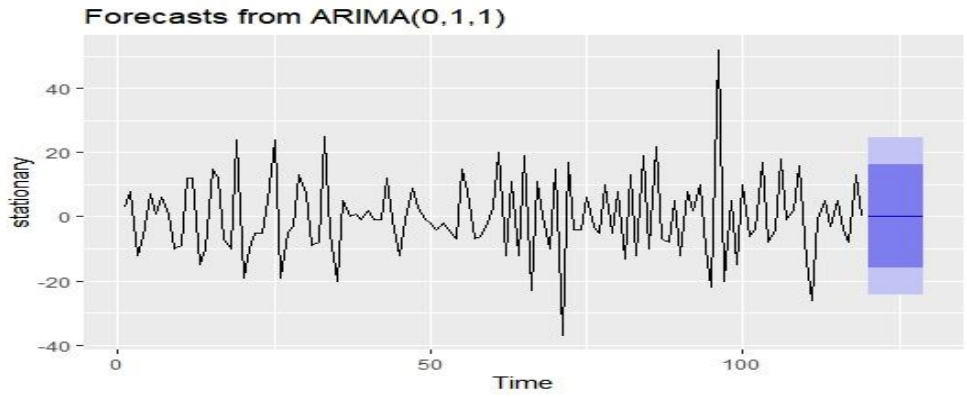

Figure 6: Time series Plot of Diabetes data in Taraba State, Nigeria and its Forecast

Conclusion

In this study, we presented the time series plot of the Diabetes diseases recorded at Federal Medical Centre, Jalingo. Nigeria. The time series plot indicated that the mean and variance were not constant, hence the need for differencing of the observed data to attain stationarity. The data attain stationarity after the first difference. ARIMA (0,1,1) was selected as the best optimal model which has the lowest value of AIC/BIC. The root mean square error (RMSE) was used to assess the predictive capability of the optimal model. The twenty-four (24) months forecast of Diabetes disease infections in Taraba State, Nigeria was also presented. The results conform with (Zhang et.al 2014), where ARIMA model was used to forecast typhoid fever incidence in China. The ARIMA model could be applied to effectively predict the Diabetes disease infection in Taraba State, Nigeria and provide support for the development of interventions for disease control and prevention.

REFERENCES

Akinkugbe, O. O. (Ed.). (1997). *Final report of national survey on non-communicable diseases in Nigeria series 1*. Lagos: Federal Ministry of Health and Social Services.

Azevedo, M., & Alla, S. (2008). Diabetes in Sub-Saharan Africa: Kenya, Mali, Mozambique, Nigeria, South Africa and Zambia. *International Journal of Diabetes in Developing Countries, 28*, 101–108. https://doi.org/10.4103/0973-3930.45268

Box GEP et al. (2015). Time Series Analysis: Forecasting and Control. Journal of the Operational Research Society 22, 199–201

Chris, E. E., Akpan, U. P., John, O. I., & Daniel, E. N. (2012). Gender and age specific prevalence and associated risk factors of Type 2 Diabetes Mellitus in Uyo metropolis, South Eastern Nigeria. *Diabetological Croatica, 41*, 17–28.

Diabetes Care. (2006). 29, S43–S48. Retrieved from Carediabetesjournals.org

Dowell D, Tian LH, Stover JA, Donnelly JA, Martins S, Erbelding EJ, et al (2011). Changes in Fluoroquinolone Use for Gonorrhea Following Publication of Revised Treatment Guidelines. Am J Public Health. 102(1):148–55. doi: 10.2105/ajph.2011.300283 PMID: 22095341

International Diabetes Federation (2007).

Lucia, Y. O., & Prisca, O. A. (2012). Type 2 diabetes mellitus and impaired fasting plasma glucose in Urban South Western Nigeria. *International Journal of Diabetes and Metabolism,21*, 9–12.

Martinez EZ, Silva EA and Fabbro AL (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of Sao Paulo. Brazil. Revista da Sociedade Brasileira de Medicina Tropical 44, 436–440.

Nathan, D. M., Davidson, M. B., DeFronzo, R. A., Heine, R. J.,Henry, R. R., Pratley, R., & Zinman, B. (2007). Impaired fasting glucose and impaired glucose tolerance: Implications for care. *Diabetes Care, 30*, 753–759.

Nyenwe, E. A., Odia, O. J., Ihekwaba, A. E., Ojule, A., & Babatunde, S. (2003). Type 2 diabetes in adult Nigerians: A study of its prevalence and risk factors in Port Harcourt, Nigeria. *Diabetes Research and Clinical Practice, 62*, 177– 185.

Olatunbosun, S. T., Ojo, P. O., Fineberg, N. S., & Bella, A. F. (1998). Prevalence of diabetes mellitus and impaired glucose tolerance in a group of urban adults in Nigeria. *Journal of the National Medical Association, 90*, 293–301.

Owoaje, E. E., Rotimi, C. N., Kaufman, J. S., Tracy, J., & Cooper, R. S. (1997). Prevalence of adult diabetes in Ibadan, Nigeria. *East African Medical Journal, 74*, 299–302.

Peng Y et al. (2017). Application of seasonal auto-regressive integrated moving average model in forecasting the incidence of hand-foot-mouth disease in Wuhan, China. Journal of Huazhong University of Science and Technology-Medical Sciences 37, 842–848.

Rios, M, Garcia, J.M, Sanchez, J.A, Perez, D (2000). A statistical analysis of the seasonality in pulmonary tuberculosis. Eur J Epidemiol. 16(5):483–8. PMID: 10997837

Ronald, G., & Zubin, P. (2013). Definition, classification and diagnosis of diabetes, pre-diabetes and metabolic syndrome. *CDA Clinical Practice Guidelines Expert Committee, 37*, S8–S11.

Song X et al. (2016). Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011. Medicine (Baltimore) 95, e3929.

Ture, M., Kurt, I., (2006). Comparison of four different time series methods to forecast hepatitis A virus infection. Expert Systems with Applications. 31(1):41–6.

Unwin, N., Sobugwi, E., & Alberti, K. G. M. M. (2001). Type 2 diabetes: The challenge of preventing a global epidemic. *Diabetes International, 11*, 4–8.

Wang K et al. (2016). The use of an autoregressive integrated moving average model for prediction of the incidence of dysentery in Jiangsu, China. Asia-Pacific Journal ofPublic Health 28, 336–346.

Williams K. Investigating Risk Factors Associated with Syphilis Rate in the United States Based on ARIMA and ARCH/GARCH Time Series Models.

Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care, 27*, 1047–1053.

World Health Organization (2019)

Wu J.B., Ye, L.X., and You, E.K., (2007). Prediction of incidence of notifiable contagious diseases by application of time series model. Journal of Mathematical Medicine 1, 90–92.

Xu Q et al. (2017). Forecasting the incidence of mumps in Zibo city based on a SARIMA model. International Journal of Environmental Research and Public Health 14, 925.

Zhang X, Zhang T, Young AA, Li X, (2014). Applications and Comparisons of Four Time Series Models in Epidemiological Surveillance Data. PLoS ONE. 9(2):e88075. doi: 10.1371/journal.pone.0088075 PMID: 24505382

Zhang X, Liu Y, Yang M, Zhang T, Young AA, Li X (2013). Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China. PLoS ONE.8(5):e63116. doi: 10.1371/journal.pone.0063116 PMID: 23650546

Zeng Q et al. (2016). Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016. *Scientific Reports* 6, 32367.

Zheng YL et al. (2015). Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. PLoS One 10, e116832.

Zimmet, P. (2003). The burden of type 2 diabetes: Are we doing enough? *Diabetes & Metabolism, 29*, 659–681.