European Journal of Statistics and Probability
Vol.10, No.1, pp., 21-32, 2021
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)

Application of SARIMA Model and Simple Seasonal Exponential Smoothing on Diabetes Mellitus: A Case of Enugu State Teaching Hospital, Nigeria

Christogonus Ifeanyichukwu Ugoh[a*], Anyadiegwu Chinelo Ujunwa[a], Thomas Chinwe Urama[b], Ugwu Gibson Chiazortam[b]

[a]Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria
[b]Department of Statistics, Institute of Management and Technology, Enugu, Nigeria

Citation: Christogonus Ifeanyichukwu Ugoh, Anyadiegwu Chinelo Ujunwa, Thomas Chinwe Urama, Ugwu Gibson; Chiazortam (2022) Application of SARIMA Model and Simple Seasonal Exponential Smoothing on Diabetes Mellitus: A Case of Enugu State Teaching Hospital, Nigeria, *European Journal of Statistics and Probability*, Vol.10, No.1, pp., 21-32

ABSTRACT: *This paper aims at obtaining better model between seasonal ARIMA and simple seasonal exponential smoothing that will be used for forecasting number of diabetes patients in a given hospital. Monthly dataset from January 2009 to December 2019 from Enugu State Teaching Hospital was used for this research. Seasonal ARIMA was modelled using the techniques of Box-Jenkins, and simple seasonal exponential smoothing modelled using the least squares method. Bayesian Information Criterion (BIC) was employed to obtain the best seasonal ARIMA model, while the Theil's U statistics and MAPE were used to obtain the best forecast model. ARIMA* $(1,1,2)(0,0,0)_{12}$ *was selected as the best SARIMA model with the BIC of 7.873, and simple seasonal exponential smoothing was considered the best forecast model with a Theil's U Statistic of 0.11241 and MAPE of 23.450. The fitted model was used to make out-sample forecast for the period January 2020-December 2025. The fitted model in this findings will help Enugu state government to plan efficiently, expand public sensitization, and allocate adequate resources for emergencies.*

KEYWORDS: Diabetes Mellitus, SARIMA, Simple Seasonal Exponential Smoothing, BIC, Theil's U Statistic, MAPE

INTRODUCTION

Diabetes Mellitus (DM) is one of the metabolic diseases characterized by high level of sugar in the blood and urine, resulting from the inability of the body to respond properly to the hormone insulin released by the pancreas [1-4], if not treated on time, can lead to many health complications like damage of heart, brain and kidney [4]. There are three major types of DM: Type 1, Type 2, and Gestational DM [2,5].

Type 1 DM is characterized by the autoimmune destruction of beta cells in the pancreas through T-cell mediated inflammatory response (insulitis) and humoral (B-cell) response [6], and it accounts for 80-90% of diabetes in children and adolescents [7,8]. Type 2 DM is characterized by the inefficiency of the pancreas to produce enough insulin [9]; and more than 90-95% of diabetes

patients belong to type 2 and most of the patients are adults [2]. Gestational DM is a condition whereby the hormone made by the placenta prevents the body from using insulin effectively [6].

According to [10], 537 million adults (20-79 years) worldwide are living with diabetes, and that by 2030, the number will rise to 643 million, and by 2045, 783 people are predicted to die of DM; 3.6 million adults live with DM in Nigeria. Diabetes Mellitus has been a current problem prevailing in Enugu state and the Nigeria at large. The number of diabetic patients in Enugu state has being increasing with an alarming rate thereby increasing health expenditure and causing life.

Due to the fact there was no separate records of type 1 and type 2 diabetes at Enugu State Teaching Hospital, this paper will then focus on obtaining the best model that will be used to forecast the future trend for diabetes patients as a whole, using monthly dataset from January 2009 to December 2019.

[11] carried a time series analysis of diabetes patients using monthly dataset for the period January 2006-December 2016 from Jigme Dorji Wangchuk National Referral Hospital. They selected ARIMA(0,1,1) model using the techniques of Box-Jenkins as the best model to predict for future trend of diabetes. [12] on the other hand analyzed the glucose in the artificial pancreas using stochastic seasonal models approaches, and by applying the techniques of Box-Jenkins. Their findings showed that SARIMA$(4,0,4)(1,0,1)_{33}$ was the best model to predict the glucose in the pancreas.

MATERIAL AND METHODS

Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

Seasonal ARIMA (SARIMA) is an extension of non-seasonal ARIMA that explicitly supports univariate time series data y_t with a seasonal component. Non-seasonal ARIMA is a statistical model which is used to predict future values based on past values. The 'AR' stands for Autoregressive, 'MA' stands for Moving Average, and 'I' stands for Integrated (which implies that the data values are replaced by difference between the data values and the previous values). SARIMA models are denoted by $SARIMA(p,d,q)(P,D,Q)_m$ and expressed as

$$y'_t = c + \left[\sum_{i=1}^{P} \Phi_i y'_{t-im} + \sum_{i=1}^{p} \phi_i y'_{t-i}\right] + \left[\sum_{j=1}^{Q} \Theta_j \varepsilon_{t-jm} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}\right] + \varepsilon_t \quad (1)$$

where $m =$ number of seasonal periods; $p =$ non-seasonal AR order; $q =$ non-seasonal MA order; $d =$ number of non-seasonal differencing; $P =$ seasonal AR order; $Q =$ seasonal MA order; $D =$ number of seasonal differencing; Φ_i and Θ_j are coefficients of seasonal AR and MA respectively; ϕ_i and θ_j are coefficients of AR and MA respectively

Seasonal AR and MA

Seasonal AR (SAR) and MA (SMA) terms predict time series data y_t using data values and errors at times with lags that are multiples of m, for example, seasonal first order AR will use y_{t-m} to predict y_t, and seasonal second order AR will use y_{t-m} and y_{t-2m} to predict y_t, and again for example, seasonal first order MA will use ε_{t-m}, and seasonal second order MA will use ε_{t-m} and ε_{t-2m} as predictors.

Pure Seasonal AR (SAR) model is expressed as

$$y_t = c + \sum_{i=1}^{P} \Phi_i y'_{t-im} + \sum_{i=1}^{p} \phi_i y'_{t-i} + \varepsilon_t \quad (4)$$

Pure Seasonal MA (SMA) model is expressed as

$$y_t = c + \sum_{j=1}^{Q} \Theta_j \varepsilon_{t-jm} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (5)$$

Seasonal Autoregressive Moving Average (SARMA) Model

When the time series data with seasonal component do not require any differencing before becoming stationary, and it is a combination of both seasonal and non-seasonal ARMA, the resultant model is SARMA. SARMA is denoted by $ARMA(p,q)(P,Q)_m$ and expressed as

$$y_t = c + \left[\sum_{i=1}^{P} \Phi_i y_{t-im} + \sum_{i=1}^{p} \phi_i y_{t-i}\right] + \left[\sum_{j=1}^{Q} \Theta_j \varepsilon_{t-jm} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}\right] + \varepsilon_t \quad (6)$$

SARIMA Model Fitting

SARIMA model is fitted to the time series data using the Box-Jenkins methodology. This study adopts four (4) steps, which are: Model Identification using ACF and PACF Plots, Estimation of Parameters, Model Diagnostic (Adequacy Check) and Forecast.

Model Identification using ACF and PACF Plots

Time series data is stationary if it does not change with respect to time. Stationarity of a time series data is usually determined through the following [13]:

a. If the lags in the ACF Plot fall rapidly as the number of lags increases, then the time series data is stationary
b. If the lags of the ACF Plot do not fall rapidly as the number of lags increases, then the time series data is non-stationary

Seasonal Differencing of Time Series Data

This is the process of making a non-stationary time series stationary, and it is the series of changes from one season to the next season. It stabilizes the mean of time series by removing the changes in the series and eliminating or reducing trend and seasonality.

First Order Seasonal Differenced series denoted as y_t' is expressed as

$$y_t' = (1 - B^m)(1 - B)y_t = (y_t - y_{t-1}) - \left(y_{t-m} - y_{t-(m+1)}\right) \tag{7}$$

In general, the seasonal differencing of any seasonal order can be obtained using equation (8)

$$y_t' = (1 - B^m)^D (1 - B)^d y_t \tag{8}$$

where y_t' stands for any seasonal order

Model Adequacy Check

When there are many models of the same type obtained from the time series data, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two such methods that can detect which of the models is the optimal model to use in forecasting. The AIC is written as

$$AIC = nlog(\hat{\sigma}^2) + 2k \tag{9}$$

Bayesian Information Criterion (BIC) is written as

$$BIC = nlog(\hat{\sigma}^2) + klog(n) \tag{10}$$

where k is the number of model parameters; $\hat{\sigma}^2$ is the residual sum of squares, and n is the sample size

The SARIMA model with the lowest AIC or BIC is considered the best Seasonal ARIMA model among others to fit to the data series.

Simple Seasonal Exponential Smoothing

Simple seasonal exponential smoothing model is used mostly when the time series data do not have trend but has a seasonal effect that is constant over time. Its smoothing parameters are level and season.
The Level is estimated as

$$L_t = \alpha(y_t - S_{t-m}) + (1-\alpha)L_{t-1} \quad (11)$$
$$L_t = L_{t-1} + \alpha[y_t - (S_{t-m} - L_{t-1})] \quad (12)$$

$$L_t = L_{t-1} + \alpha[y_t - F_{t-1}] = L_{t-1} + \alpha\varepsilon_t \quad (13)$$

where L_t is the time series level a time t; y_t is the observed value at time t; S_{t-m} is the seasonal effect for season t-m; and L_{t-1} is the level forecast for time t made at time t-1

The smoothing form of the equation for seasonal indexes is expressed as

$$S_t = \delta(y_t - L_t) + (1-\delta)S_{t-m} \quad (14)$$

and the smoothed forecast value is given as

$$F_t = L_t + L_{t-p+m} \quad (15)$$

where δ is the seasonal smoothing parameter; α is the smoothing coefficient for trend, and $F_t$is smoothed forecast value

Performance Measures

The measures of forecast accuracy adopted in this study is Theil's U Forecast Accuracy and Mean Absolute Percentage Error (MAPE).

Theil's U Forecast Accuracy

The Theil's U shows how the forecast conforms to the values of the future periods. It is written as

$$U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} y_t^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n} \hat{y}_t^2}} \quad (16)$$

where Y_t is the actual value of a point for a given time period t, $\hat{Y}_t$ is the forecast value, n is the number of the data points.

For $0 \leq U < 1$, the model is a good fit,
For $U = 0$, the model is a perfect fit,
For $U \geq 1$, the model is not a good fit [13]

Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is used to measure the error of both methods (SARIMA and Simple Seasonal Exponential Smoothing). The model with the smallest MAPE is considered the appropriate model. It is defined as

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \quad (17)$$

RESULTS

Figure 1 shows the Timeplot of the number of Diabetes Mellitus patients in Enugu State of Nigeria from January 2009 to December 2019.

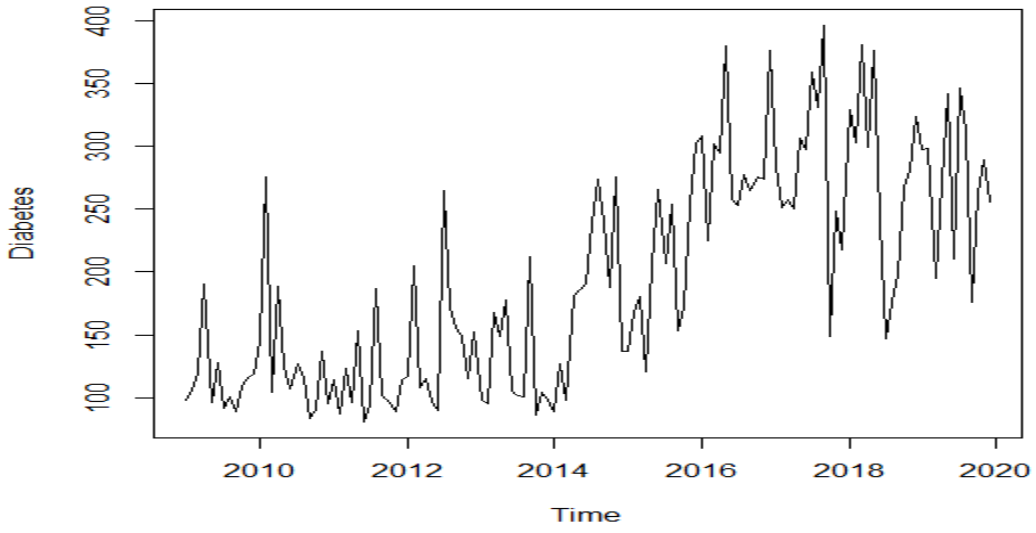

Figure 1. Timeplot of the Number of Diabetes Patients Recorded at Enugu State Teaching Hospital

Table 1. Descriptive Statistics for the Data Series

Descriptive Statistics						
	N	Minimum	Maximum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Number of Diabetes Patients	132	81	396	193.81	7.555	86.804

Table 1 shows the descriptive statistics for the data series (number of diabetes patients recorded at Enugu State Teaching Hospital). The average number of diabetes patients recorded is 193.81, approximately 194 patients monthly; the standard deviation of the number of diabetes patients is given as 86.804, approximately 87 patients.

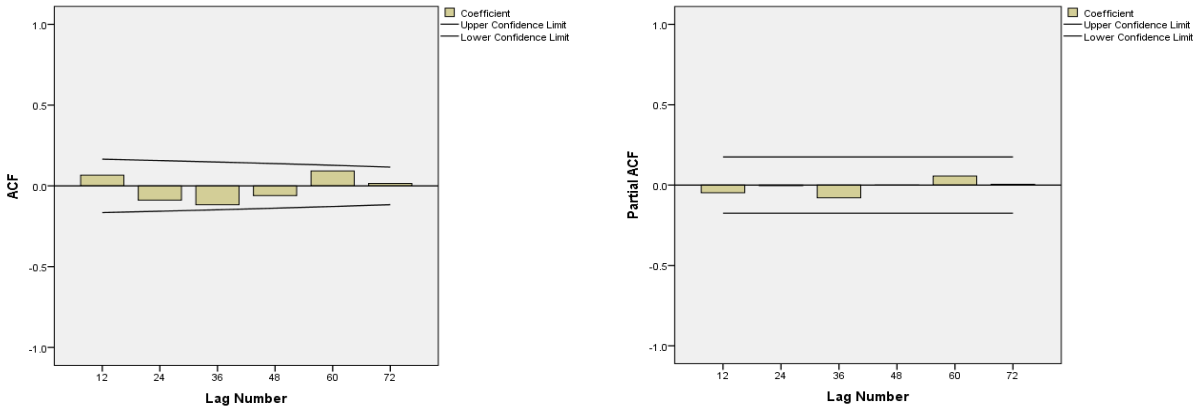

Figure 2. (a) ACF Plot at Periodic Lags (b) PACF Plot at Periodic Lags for the Data Series (Number of Diabetes Mellitus Patients)

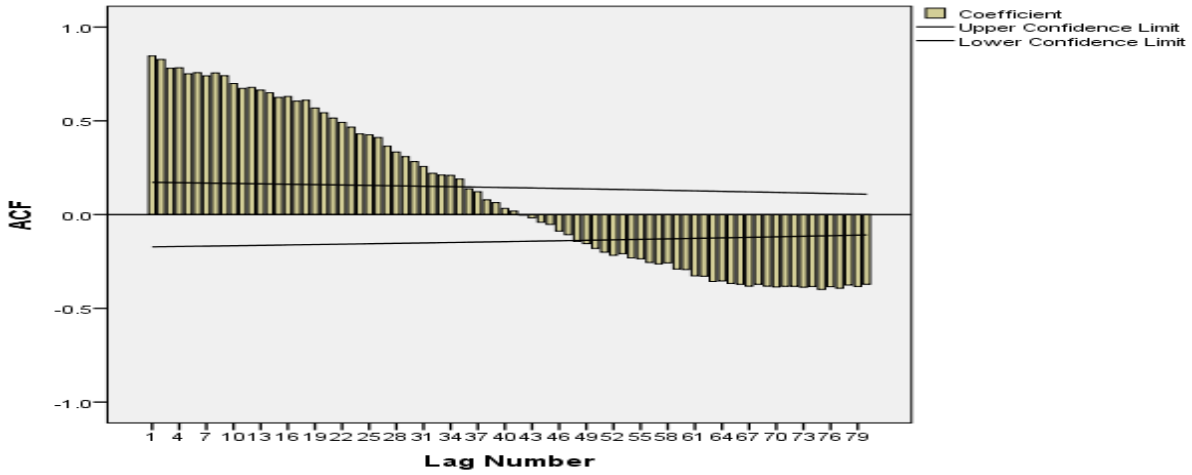

Figure 3. ACF Plot for the Data Series (Number of Diabetes Mellitus Patients)

In Figure 2a, there is a rapid fall of the periodic lags as the lag number increases at the multiple of 12. This therefore indicates that there is no need for seasonal differencing. Furthermore, no significant lag is found in Figure 2a and 2b, and this however implies that there is no evidence of the inclusion of seasonal AR term and seasonal MA term in the model.
Figure 3 shows an exponential decay of the lags as the lag number increases. This is an evidence that there is need for non-seasonal differencing.

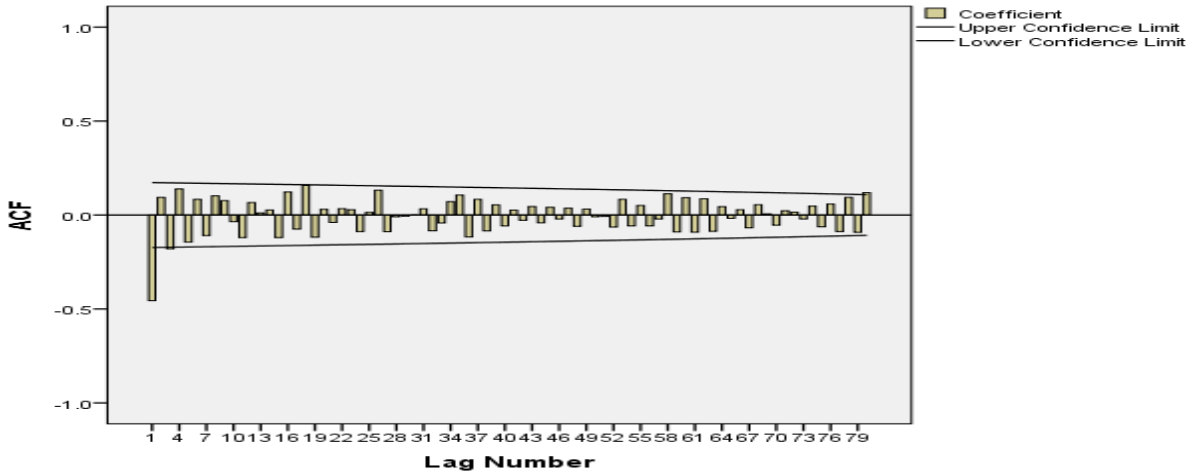

Figure 4. First Difference ACF Plot for the Data Series (Number of Diabetes Mellitus Patients)

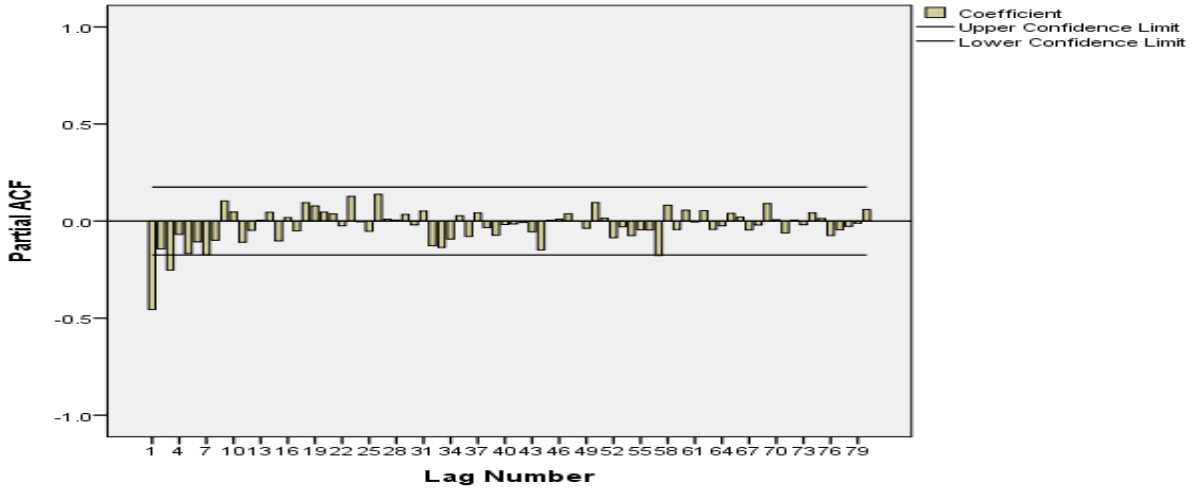

Figure 5. First Difference PACF Plot for the Data Series (Number of Diabetes Mellitus Patients)

The quick drop at lag 1 in the first differenced ACF plot in Figure 4 is an evidence of stationarity. Furthermore, negative spikes are observed at the low lags (at lag 1 and lag 3), which is an indication that non-seasonal AR term will be a useful part of the model. Moreover, negative spikes are observed at the low lags of the first differenced PACF plot in Figure 5 (at lag 1, lag 3, lag 5 and lag 7), implying that the non-seasonal MA term will also be a useful part of the model.

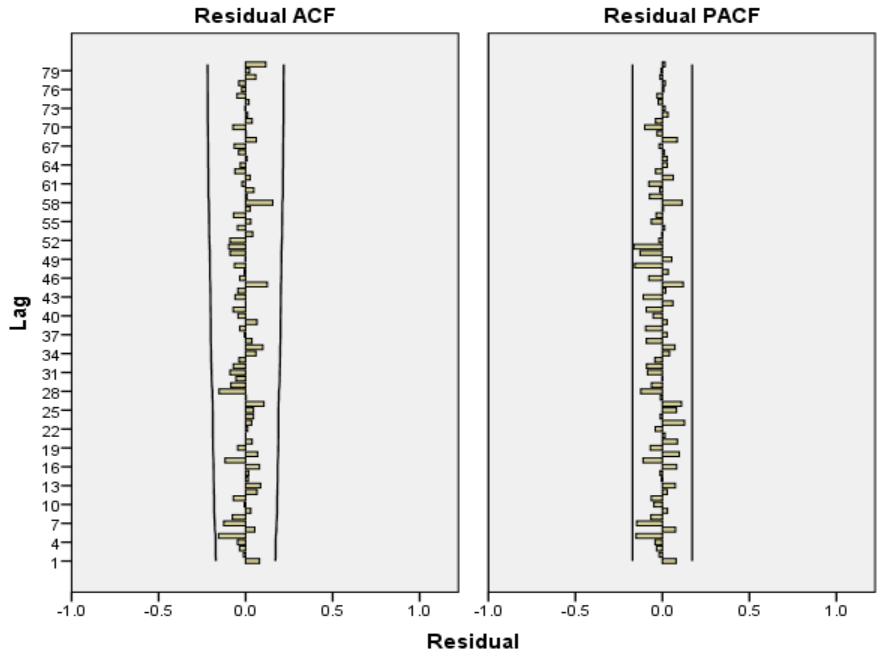

Figure 6. Residual Plot (a) ACF Plot (b) PACF Plot for the Data Series

The lags in the ACF Plot in Figure 6a and in PACF Plot in Figure 6b all fall within the upper and lower bound, thereby indicating that there is no serial correlation existing in the data (Number of Diabetes Mellitus Patients)

Table 2. Model Comparison using Bayesian Information Criterion (BIC)

$ARIMA(p,d,q)(P,D,Q)_{12}$	BIC
$ARIMA(1,1,1)(0,0,0)_{12}$	7.886
$ARIMA(1,1,2)(0,0,0)_{12}$	7.873
$ARIMA(1,1,3)(0,0,0)_{12}$	7.944
$ARIMA(2,1,1)(0,0,0)_{12}$	7.921
$ARIMA(3,1,1)(0,0,0)_{12}$	7.953

ARIMA$(1,1,2)(0,0,0)_{12}$ model in Table 2 has the lowest BIC of 7.873, and it is considered the best among the other models. Table 2 shows the estimates of the ARIMA$(1,1,2)(0,0,0)_{12}$ model parameters.

Table 3. Estimated ARIMA$(1,1,2)(0,0,0)_{12}$ Model Parameters

ARIMA Model Parameters						
			Estimate	SE	T	Sig.
Number of Patients	AR	Lag 1	-.789	.066	-11.951	.000
	Difference		1			
	MA	Lag 2	.468	.096	4.893	.000

Using the estimates in Table 3, the ARIMA$(1,1,2)(0,0,0)$ model is written as

$$y_t' = -0.789y_{t-1}' + 0.468\varepsilon_{t-2} + \varepsilon_t \quad (18)$$

Table 4. Simple Seasonal Exponential Smoothing Model Parameter Estimates

Exponential Smoothing Model Parameters			
Model		Estimate	Sig.
Simple Seasonal Exponential Smoothing	Alpha (Level)	.300	.000
	Delta (Season)	4.528E-5	.999

Using the estimates in Table 4, the smoothed forecast value is given as

$$F_t = 0.3y_t - 0.3S_{t-12} + 0.7L_{t-1} + L_{t-p-12} \quad (19)$$

Table 5. Measures of Forecast Accuracy using Theil's U Statistic and MAPE

Models	MAPE	Theil's U Statistic
ARIMA$(1,1,2)(0,0,0)_{12}$	23.450	0.11718
Seasonal Simple Exponential Smoothing	21.837	0.11241

The Seasonal Simple Exponential Smoothing has the lowest MAPE of 21.837 and lowest Theil's U Statistic of 0.11241 in Table 5. This indicates that the best forecast model for obtaining the number of Diabetes Mellitus patients based on this study is the Seasonal Simple Exponential Smoothing. Figure 7 shows the timeplot for the in-sample data and out-sample forecast for the number of Diabetes Mellitus for a period January 2009-December 2019 (In-sample) and January 2020-December 2025 (out-of-sample), using simple seasonal exponential smoothing.

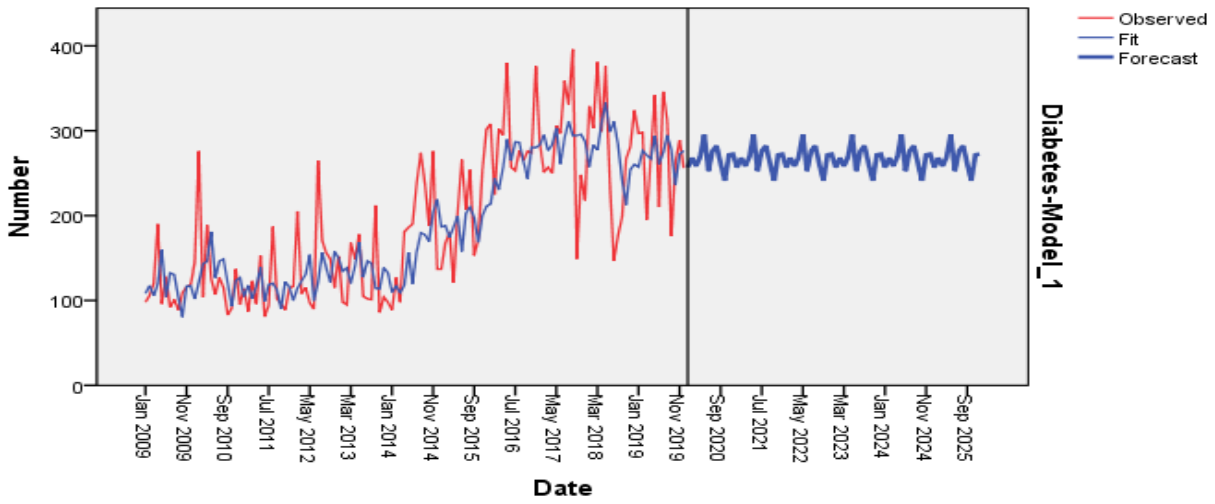

Figure 7. The In-Sample Data and Out-of-Sample Forecast for Number of Diabetes Mellitus Patients

CONCLUSION

The aim of this paper was to obtain the best model that will be used to forecast the number of diabetes patients in Enugu State Teaching Hospital, using monthly dataset from January 2009 to December 2019. ARIMA and simple seasonal exponential smoothing were modelled on the dataset. List of ARIMA models were obtained using the techniques of Box-Jenkins, and after testing the adequacy of the models using the Bayesian Information Criterion (BIC), ARIMA $(1,1,2)(0,0,0)_{12}$ was selected as the best SARIMA model; comparing the obtained SARIMA model

with the simple seasonal exponential smoothing, using the Theil's U Statistic and MAPE, simple seasonal exponential smoothing model was selected as the best forecast model. The obtained model was used to forecast the number of diabetes patients for the period January 2020 to December 2025.

References

[1] World Health Organization, 2022. Diabetes, available from https://www.who.int/Health-topics/diabetes
[2] Akran TK & Hisham MD. Diabetes Mellitus: The epidemic of the century. World Journal of Diabetes. 2015; 6(6): 850-867
[3] Moltchanova EV, Schreier N, Lammi N, Karvonen M. Seasonal variation of diagnosis of type 1 diabetes mellitus in children worldwide. Diabet Med. 2009; 26:673-678
[4] Jiahua W, Jiaxing T, Cheng T, Xinyi F, Runyu M, Haoran W, Xiuge W, Xiaolin T. The Influence of different types of diabetes on vascular complications. Journal of Diabetes Research, Vol 2022, Article ID 3448618, 12 pages, 2022. https://doi.org/10.1155/2022/3448618
[5] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Cares. 2014; 37 Suppl 1: S81-S90
[6] Devendra D, Liu E, Eisenbarth GS. Type 1 diabetes: Recent developments. BMJ. 2004; 328: 750-754
[7] Craig ME, Hattersley A, Donaghue KC. Definition, epidemiology and classification of diabetes in children and adolescents. Pediatr Diabetes. 2009; 10 Suppl 12: 3-12
[8] Dabelea D, Mayer-Davis EJ, Saydah S, Imperatore G, Linder B, Divers J, Bell R, Badaru A, Talton JW, Crume T, et al. Prevalence of type 1 and type 2 diabetes among children and adolescents from 2001 to 2009. JAMA. 2014; 311(17): 1778-1786
[9] Mayo Clinic. Type 2 diabetes, available from https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193
[10] International Diabetes Federation. Diabetes around the world 2021, available from https://diabetesatlas.org/
[11] Thukten Singye & Suntaree Unhapipat. Time series of diabetes patients: A case study of Jigme Dorji Wangchuk National Referral Hospital in Bhutan. Journal of Physics: Conference Series. 2018. 1039 012033, doi: 10.1088/1742-6596/1039/1/012033
[12] Montaser E, Diez J-L, & Bondia J. Stochastic seasonal models for glucose prediction in the artificial pancreas. Journal of Diabetes Sciences and Technology, 2017; 11(6): 1124-1131
[13] Ugoh CI, Amaeze OG, Henry NC, Nwabueze NC, Ifeanyi AC & Godson MI. Application of Autoregressive Integrated Moving Average Model and Weighted Markov Chains on Forecasting Under-Five Mortality Rates in Nigeria. Asian Journal of Probability and Statistics. 2022; 16(1): 30-43