

Original Research Article

ARIMA MODEL TO PREDICT THE PREVALENCE OF DIABETES TYPE 1 AND TYPE 2 PATIENTS: A CASE STUDY OF JOS UNIVERSITY TEACHING HOSPITAL

Comment [u1]: Over all it is good but minor revision was needed

ABSTRACT

Diabetes Mellitus is a huge burden for human health, increasing number of patient is likely to result in rising demand for the medical emergencies. Due to having limited number of hospital with sophisticated laboratory test to differentiate between type 1 and type 2 diabetes it is important to forecast the future incidences and prepare with proper resource planning. The monthly number of Diabetes patients obtained from Jos University Teaching Hospital is fitted by autoregressive integrated moving average (ARIMA) model. Dataset starting from January, 2010 to December,2020. Using ARIMA, several models were evaluated based on the Bayesian Information Criterion (BIC) and Ljung-Box Q statistics. ARIMA(3, 1, 1) is better to describe and predict the future trends of Diabetes type 1 an ARIMA(1,1,1) is the better model to predict the future prevalence of diabetes type 2. Therefore, the proposed model will help to plan appropriately and allocate resources for emergencies.

Keywords: Diabetes; Mellitus, Hospital, ARIMA; Statistics, Forecasting; BIC.

Comment [u2]: It is too short , please describe in detail methods, results, conclusion

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 β -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 [1]. Diabetes is preceded by impaired fasting glucose (IFG) resulting in a pre-diabetic state which can exist undetected for many years, [2], causing irreversible damage to vital organs. Pre-diabetes is a practical term referring to Impaired Fasting Glucose (IFG), impaired glucose tolerance [3] or a glycosylated hemoglobin (A1c) of 6.0 to 6.4%, each of which places individuals at high risk of developing diabetes and its 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 [4].

Type1 diabetes occur as a results of auto-immune beta-cell destruction in the Pancreas characterized by a total absence of insulin production. Type 1 diabetes is responsible for 5% to 10% of all cases of diabetes. Associated risk factors include autoimmune, genetic, and environmental factors, [5].

Type2 diabetes can be linked to be accounting for around 90 per cent of all cases, it is a chronic metabolic disorder, in which the body is unable utilize glucose from food because of the inability of the pancreas to produce insulin or produces insufficient insulin, or the insulin itself is inactive [6].

Gestational diabetes could be described as a form of glucose intolerance that affects some women during pregnancy. This kind of diabetes is triggered during pregnancy. Most gestational diabetes is resolved naturally after delivery, but 5-10 percent of women affected during pregnancy are later found to have diabetes, especially Type 2, after pregnancy.

Pre-diabetes is described as a precursor condition to diabetes in which a person is experiencing elevated blood glucose levels but not yet up to diagnostic criteria for diabetes. People with pre-diabetes normally suffer from impaired fasting glucose or impaired glucose tolerance, or both. Prevention of diabetes mellitus disease requires.

There is an increasing prevalence of diabetes and pre-diabetes worldwide [7]. Over 5 million people suffer from the disease in Africa and the number is expected to skyrocket to 15 million by 2025, [7]. In Nigeria the prevalence varies from 0.65% in rural Mangu village to 11.0% in urban Lagos [8]. With the incidence of diabetes in Africa, diabetic complications are also expected to rise proportionately ([9]; [10]). 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 [9]. In Nigeria the National prevalence of diabetes was 2.2% [8]. In South Eastern Nigeria the overall prevalence of diabetes was 10.51% [11], whereas in South Western Nigeria the prevalence of diabetes ranges from 4.76% in Ile-Ife, Osun State to 11.0% in Lagos ([8]; [7]). Also 0.8% of diabetes mellitus, and 2.2% of Impaired Glucose Intolerance in Ibadan [12]. This was comparable to WHO reported a prevalence of 2.8% in Ibadan [13], and 6.8% in Port Harcourt, Nigeria [14]. In 2004, a survey in Jos ([14]) reported a prevalence of 10.3%. [14] reported a prevalence of 2.2% in Port Harcourt in 2003. A prevalence of 4.7% was reported by ([15]) which was higher than the national prevalence of 2.2% reported by ([7]). A review of studies on the prevalence of diabetes in adults in Africa ([16]) 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, [17]. This analysis deals with time dependent variables with an advantage of being not necessary to consider the influence of intricate factors([18]; [19]).

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 [20]. For example, decomposition methods were used to forecast nine notifiable infectious diseases in China [21]. Autoregressive integrated moving average models (ARIMA) are widely applied in infection time series modelling including tuberculosis ([22]), typhoid fever ([20]), gonorrhoea ([23]) and hepatitis ([24]). Autoregressive conditional heteroscedasticity and generalized autoregressive conditional heteroscedasticity models have been used to investigate the risk factors associated with syphilis, [25]. the seasonal autoregressive integrated moving average (SARIMA) model has been increasingly favoured and successfully used in the prediction of communicable diseases, such as dengue ([26]), tuberculosis ([27]), mumps ([19]) and others ([27]; [28]; [29]).

Different models have been used in the analysis of diabetes but there are few of such in the study area and none has considered the types of diabetes. Hence, this research is aimed at developing a

model that will forecast the prevalence of diabetes type 1 and type 2 using secondary data by selecting the best fit models to predict the trends of diabetes type1 and type2 diabetes and forecasting diabetes type 1 and type 2 in the next 5 years using the selected models

MATERIALS AND METHOD

Autoregressive (AR) Model:An Autoregressive (AR) model of order P satisfied the following

$$x_t = \mu' + \sum_{k=1}^p \phi_{kk} x_{t-k} + \varepsilon_t$$

$$x_t = \mu + \phi_{11}x_{t-1} + \phi_{22}x_{t-2} + \dots + \phi_{pp}x_{t-p} + \varepsilon_t \dots \dots \dots (3.1)$$

For t=0 where $\varepsilon_t = 0$

For n> 0 where the error term $\varepsilon_t > 0$, is a series of independently, identically distributed (i.i.d) random variables and assumed to be normally distributed and μ is some constant. The p denotes the order of autoregressive model, defining how many previous values the current value is related to. The model is called autoregressive because the series is regressed on to past values of itself. The error term ε_t in equation (3.1) refers to the noise in the time series. Above, the errors were said to be independently identically distributed. Commonly, they are also assumed to have a normal distribution with mean zero and variance σ^2 . For the model in equation (3.1) to be of use in practice, the estimator must be able to estimate the value of ϕ_k

Moving Average (Ma) Model: An Autoregressive moving Average of order q is given as

$$x_t = \mu' + \sum_{j=1}^q \theta_{jj} x_{t-j} + \varepsilon_t$$

$$x_t = \mu + \theta_{11}x_{t-1} + \theta_{22}x_{t-2} + \dots + \theta_{pp}x_{t-q} + \varepsilon_t \dots \dots \dots 3.2$$

For t=0 where $\varepsilon_t = 0$

The error (or noise) term in this equation is the one step ahead forecasting error which can be expressed as a function of previous forecasting errors. It shows that MA (q) models make forecast based on the error made in the past, and so one can learn from the error made in the past to improve current forecast.

Autoregressive Moving Average (ARMA) Models:This is Autoregressive Component of order p and Moving Average Component of order q. In a situation where AR (q) and MA(p) were not able to solve the case at hand, a combination of the two produces another interesting model known as Autoregressive moving average (ARMA)(p, q) model where p is the number autoregressive component and q is the number of moving average component. This model is as presented as follows

The Form ARMA (p, q) model is given by the equation

$$x_t - \phi_{11}x_{t-1} - \phi_{22}x_{t-2} - \phi_{33}x_{t-3} - \dots - \phi_p x_{t-p} = \varepsilon_t - \phi_{11}x_{t-1} - \phi_{22}x_{t-2} - \phi_{33}x_{t-3} - \dots - \phi_p x_{t-p} \dots \dots \dots 3.3$$

Where ϕ 's (phis) are the autoregressive parameters to be estimated, the θ 's (thetas) are the moving average parameters to be estimated the x 's are the original time series values and the ε 's are residuals or errors which are assumed to follow the normal probability distribution.

In order to make the writing of model easier, Box-Jenkins use the backshift operator. The backshift

operator B can change the time period t to time period $t - 1$. For example, $BY_t = Y_{t-1}$, $B^2Y_t = Y_{t-2}$ and so on.

So the above model can be rewritten as:

$$(1 - \phi_1 B - \dots - \phi_p B^p)X_t = (1 - \theta_1 B - \dots - \theta_q B^q)\varepsilon_t \dots \dots \dots 3.4$$

Further can be abbreviated as:

$$\phi_p(B)X_t = \theta_q(B)\varepsilon_t$$

where $\phi_p(1 - \phi_1 B - \dots - \phi_p B^p)$ and $\theta_q(1 - \theta_1 B - \dots - \theta_q B^q)$

Autoregressive Integrated Moving Average (ARIMA) Model: In statistics, ARIMA(pdq) models, sometimes called the Box-Jenkins models after the iterative Box-Jenkins methodology usually used to estimate them, are typically applied to time series data for forecasting. Given a time series of data $X_{t-1}, X_{t-2}, X_2, X_1$, the ARIMA model is a tool for understanding and, perhaps, predicting future values in the series. The model consists of three parts: an Autoregressive (AR) part, a moving Average (MA) part and the differencing part. The model is usually then referred to as the ARIMA (p, d, p) model where p is the order of the Autoregressive part, d is the order of differencing and q is the order of the moving Average part, [30].

However, it is impossible to have the data always stationary. Therefore, to make the process we have to do differencing in non-stationary series so the differenced series $(1 - B)^d X_t$ must be added in the process ARMA (p, q). Where the a_t is a sequence of identically distributed uncorrelated deviates, referred to as "white noise. In many situations where differencing is employed, a non-zero constant term will not be required. For brevity, the equation is generally written as:

$$\phi_p(B)(1 - B)^d X_t = \theta_0 + \theta_q(B)\varepsilon_t \dots \dots \dots 3.6$$

Where

$$X_t = a_t + \sum_{k=1}^p \phi_k X_{t-k} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

and $(1 - B)^d X_t$ is a differenced series and ε_t 's is a sequence of identically distributed uncorrelated deviates, referred to as "white noise".

FORECASTING USING ARIMA MODEL

After describing various time series models, the next issue to our concern is how to select an appropriate model that can produce accurate forecast based on a description of historical pattern

in the data and how to determine the optimal model orders. Statisticians George Box and Gwilym Jenkins developed a practical approach to build ARIMA model, which best fit to a given time series and also satisfy the parsimony principle. Their concept has fundamental importance on the area of time series analysis and forecasting. In this section we will consider the Box Jenkins model building techniques, these consist of the following four steps:

Preliminary Transformation: if the data display characteristics violating the stationarity assumption, then it may be necessary to make a transformation so as to produce a series compatible with the assumption of stationarity. After appropriate transformation, if the sample autocorrelation function appears to be non-stationary, differencing may be carried out

Model Identification: if $\{yt\}$ is the stationary series obtained in step 1, the problem at the identification stage is to find the most satisfactory ARIMA (p,q) model to represent $\{yt\}$. [31]determined the integer parameters (p,q) that governs the underlying process $\{yt\}$ by examining the autocorrelations function (ACF) and partial autocorrelations (PACF) of the stationary series.

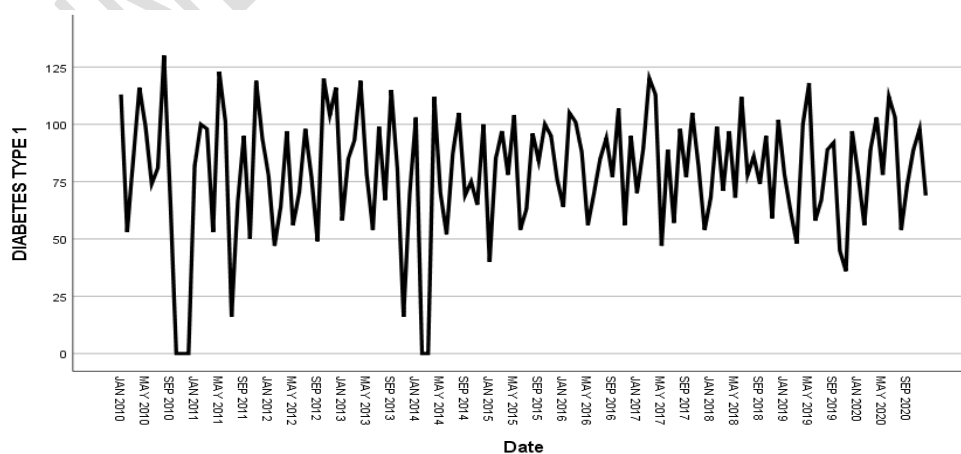
Estimation of the model: This deal with estimation of the tentative ARIMA model identified in step 2. In this study, we use the Statistical PackageSocial Sciences (SPSS 25) software to estimate the coefficient..

Diagnostic checking: Having chosen a particular ARIMA model, and having estimated its parameters, the adequacy of the model is checked by analyzing the residuals. If the residuals are white noise; we accept the model, else we go to preliminary transformation stage again and start over.

RESULTS AND DISCUSSION

MODEL FOR TYPE 1 DIABETES

Model Identification: The monthly Diabetes type 1 recorded in Jos University Teaching Hospital from 2010-2020 is plotted in the figure 1.The data display trends, which indicate that the mean is not stationary. However, it is necessary to check variance stationarity first.



Comment [u3]: Thediscussion was too short???

Figure 1: Monthly number of diabetes type 1 patients from Jos university teaching hospital (January 2010- December 2020)

Test for stationarity of the variance

It is necessary to check the stationarity of the variance. We can use Shapiro-Wilk and Kolmogorov-Smirnov for the normality test. According to the result providing in Table1, we can conclude that the variance was not constant (p -value < 0.05). This problem can be corrected by a first-degree order of differentiation.

Further, to make the mean stationary, first order differencing is used. The monthly diabetes type 1 data with first order differencing is plotted in the Figure 2. Notice that the differenced series is mean stationary and constant variance

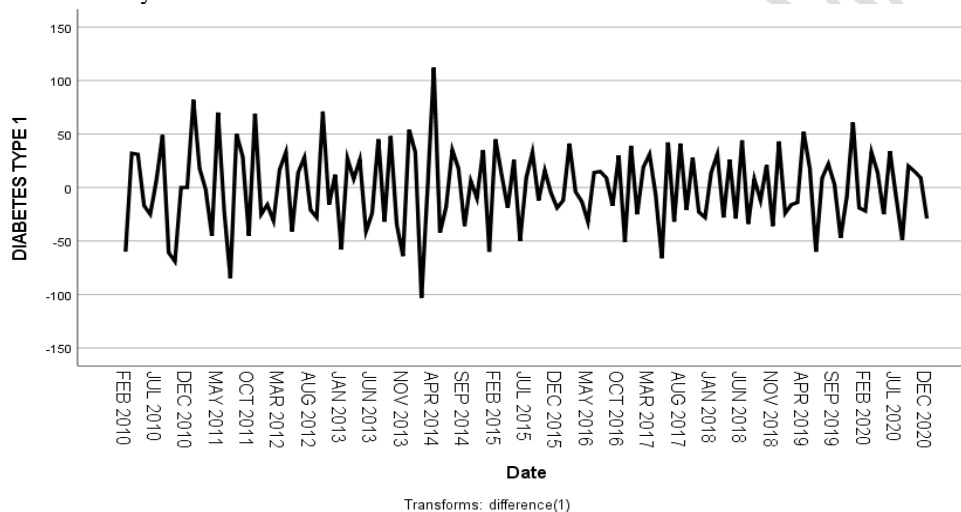


Figure 2: Time series plot of Diabetes type 1 patients after the first differencing.

Graphical representation of autocorrelations

From the ACF plots below (Figure 3), the value of its function is less than one. The parameter of any given function is less than one; it means that there is stationarity. Further the plot decays to zero on the both sides of the mean. From the Figure 3 it can be concluded that an appropriate model is moving average process of order 1.

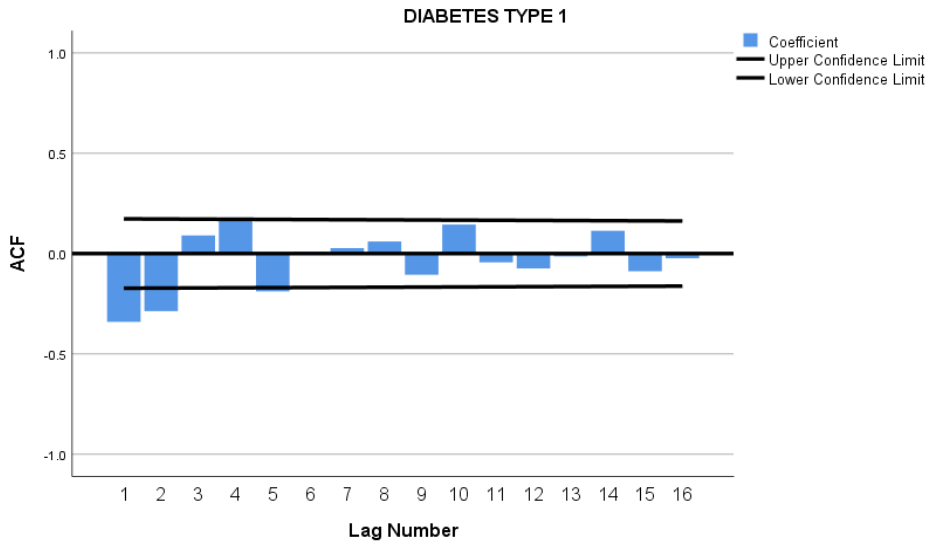


Figure 3: Sample ACF plots of monthly transformed Diabetes type1 data after first differencing

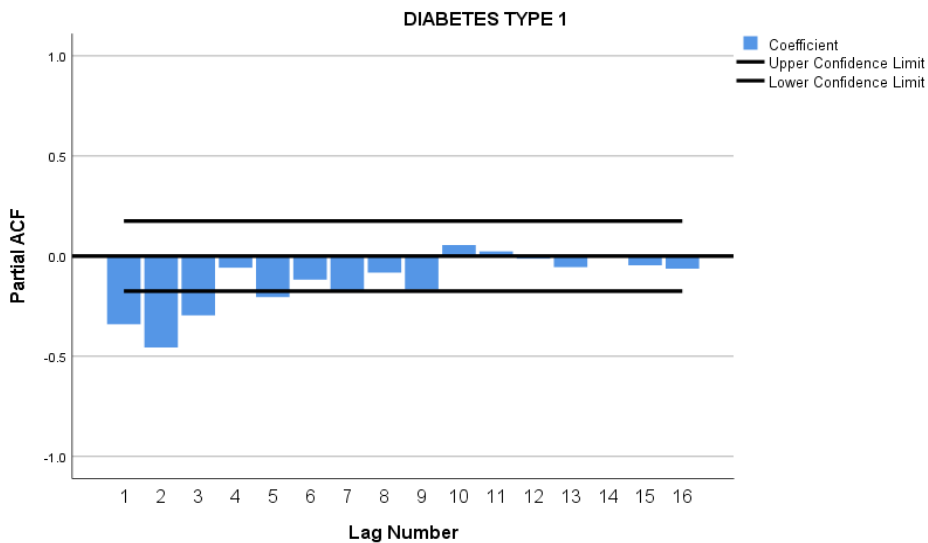


Figure 4: Sample PACF plot of monthly transformed Diabetes type1 data after first differencing.

Observing the plots of ACF and PACF plot above, it could be noticed that the ACF displays the sharper cuts off compared to the PACF. According to the PACF plot in Figure 4, we use the autoregressive process of order 3, i.e., AR(3).

Parameter estimation

Four models are chosen and tested to find a model with good fit. These models include ARIMA(1,1,0), ARIMA(2,1,0), ARIMA(3,1,0), and ARIMA(3,1,1)

Table 1, shows the model summary of ARIMA (1,1,0)

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.116	.	.116	.116	.116	.116	.116	.116	.116	.116	.116
R-squared	-.635	.	-.635	-.635	-.635	-.635	-.635	-.635	-.635	-.635	-.635
RMSE	35.012	.	35.012	35.012	35.012	35.012	35.012	35.012	35.012	35.012	35.012
MAPE	42.199	.	42.199	42.199	42.199	42.199	42.199	42.199	42.199	42.199	42.199
MaxAPE	577.529	.	577.529	577.529	577.529	577.529	577.529	577.529	577.529	577.529	577.529
MAE	27.658	.	27.658	27.658	27.658	27.658	27.658	27.658	27.658	27.658	27.658
MaxAE	112.218	.	112.218	112.218	112.218	112.218	112.218	112.218	112.218	112.218	112.218
Normalized BIC	7.186	.	7.186	7.186	7.186	7.186	7.186	7.186	7.186	7.186	7.186

Table 2 ARIMA Model Parameters of ARIMA(1,1,0)

				Estimate	SE	t	Sig.
DIABETIC PATIENTS-Model_1	DIABETIC PATIENTS	No Transformation	Constant	-.162	2.273	-.071	.943
			AR Lag 1	-.346	.083	-4.188	.000
			Difference	1			

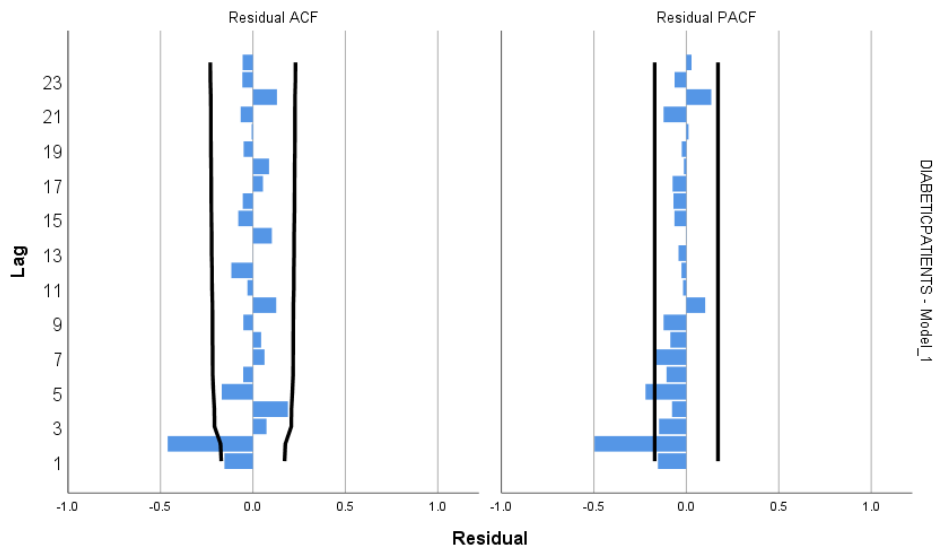


Figure 5: Residual plots for ACF and PACF after estimating ARIMA (1,1,0)

Table 3: shows the model summary of ARIMA(2,1,0)

Model Fit

Fit Statistic	Mean	S E	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.301	.	.301	.301	.3	.301	.301	.301	.301	.301	.301
R-squared	-.294	.	-.294	-.294	-.294	-.294	-.294	-.294	-.294	-.294	-.294
RMSE	31.271	.	31.271	31.271	31.271	31.271	31.271	31.271	31.271	31.271	31.271
MAPE	36.591	.	36.591	36.591	36.591	36.591	36.591	36.591	36.591	36.591	36.591
MaxAPE	397.829	.	397.829	397.829	397.829	397.829	397.829	397.829	397.829	397.829	397.829
MAE	25.216	.	25.216	25.216	25.216	25.216	25.216	25.216	25.216	25.216	25.216

MaxAE	82.151	.	82.151	82.151	82.151	82.151	82.151	82.151	82.151	82.151	82.151
Normalized BIC	6.997	.	6.997	6.997	6.997	6.997	6.997	6.997	6.997	6.997	6.997

UNDER PEER REVIEW

Table 4: ARIMA Model Parameters ARIMA(2,1,0)				Estimate	SE	t	Sig.	
DIABETIC PATIENTS-Model_1	DIABETIC PATIENTS	No Transformation	Constant	-.077	1.393	-.055	.956	
			AR	Lag 1	-.503	.079	-6.371	.000
				Lag 2	-.461	.078	-5.882	.000
			Difference	1				

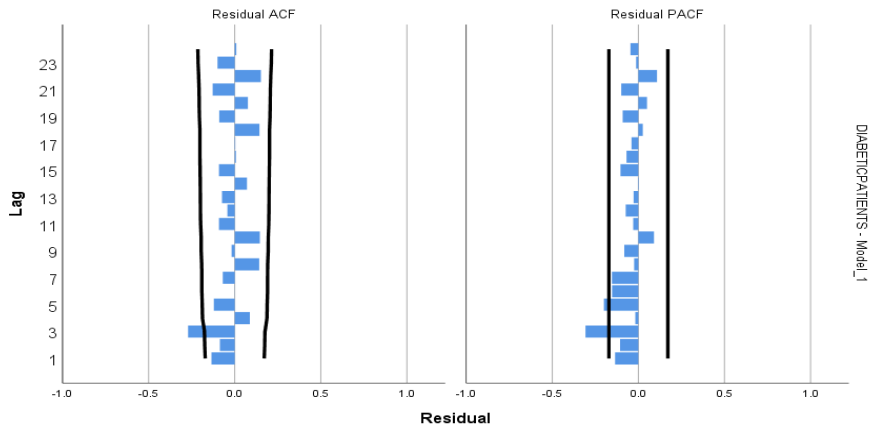


Figure 6: The residual plots of ACF and PACF after estimating ARIMA(2,1,0)

Table5: shows the model summary of ARIMA(3,1,0)

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.362	.	.362	.362	.362	.362	.362	.362	.362	.362	.362
R-squared	-.181	.	-.181	-.181	-.181	-.181	-.181	-.181	-.181	-.181	-.181
RMSE	29.991	.	29.991	29.991	29.991	29.991	29.991	29.991	29.991	29.991	29.991
MAPE	33.699	.	33.699	33.699	33.699	33.699	33.699	33.699	33.699	33.699	33.699
MaxAPE	434.673	.	434.673	434.673	434.673	434.673	434.673	434.673	434.673	434.673	434.673
MAE	23.023	.	23.023	23.023	23.023	23.023	23.023	23.023	23.023	23.023	23.023
MaxAE	75.782	.	75.782	75.782	75.782	75.782	75.782	75.782	75.782	75.782	75.782
Normalized BIC	6.951	.	6.951	6.951	6.951	6.951	6.951	6.951	6.951	6.951	6.951

UNDER PELL

Table 6 ARIMA Model Parameters of ARIMA(3,1,0)

			Estimate	SE	t	Sig.	
DIABETIC PATIENTS-Model_1	DIABETIC PATIENTS	No Transformation	Constant	-.077	1.031	-.075	.940
			AR Lag 1	-.639	.085	-7.504	.000
			AR Lag 2	-.610	.086	-7.085	.000
			AR Lag 3	-.298	.085	-3.517	.001
			Difference	1			

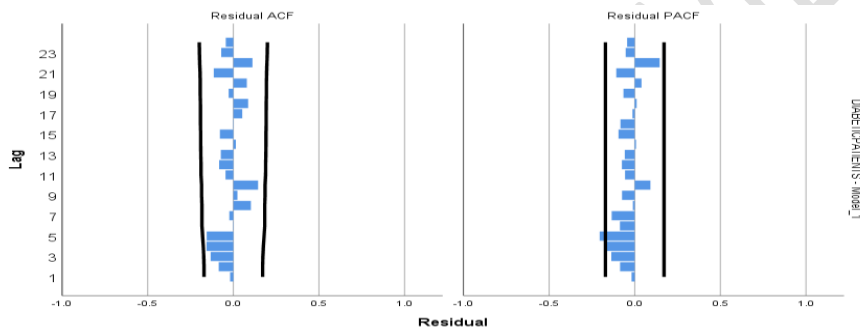


Figure 7: Shows the residual plots of ACF and PACF after estimating ARIMA(3,1,0)

Table 7: Model Summary of ARIMA (3,1,1)

Model Fit

Fit Statistic	Mean	S.E.	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.455	.	.455	.455	.455	.455	.455	.455	.455	.455	.455
R-squared	-.009	.	-.009	-.009	-.009	-.009	-.009	-.009	-.009	-.009	-.009
RMSE	27.832	.	27.832	27.832	27.832	27.832	27.832	27.832	27.832	27.832	27.832
MAPE	29.653	.	29.653	29.653	29.653	29.653	29.653	29.653	29.653	29.653	29.653
MaxAPE	349.071	.	349.071	349.071	349.071	349.071	349.071	349.071	349.071	349.071	349.071
MAE	21.432	.	21.432	21.432	21.432	21.432	21.432	21.432	21.432	21.432	21.432
MaxAE	80.625	.	80.625	80.625	80.625	80.625	80.625	80.625	80.625	80.625	80.625

Normalized BIC	6.838	.	6.838	6.838	6.838	6.838	6.838	6.838	6.838	6.838	6.838
-------------------	-------	---	-------	-------	-------	-------	-------	-------	-------	-------	-------

UNDER PEER REVIEW

Table 8: ARIMA Model Parameters of ARIMA(3,1,1)

				Estimate	SE	t	Sig.
DIABETIC PATIENTS- Model_1	DIABETIC PATIENTS	No Transformation	Constant	.056	.067	.835	.405
			AR Lag 1	.134	.093	1.443	.152
			Lag 2	-.207	.089	-2.335	.021
			Lag 3	.102	.092	1.109	.269
			Difference	1			
			MA Lag 1	.998	.570	1.751	.082

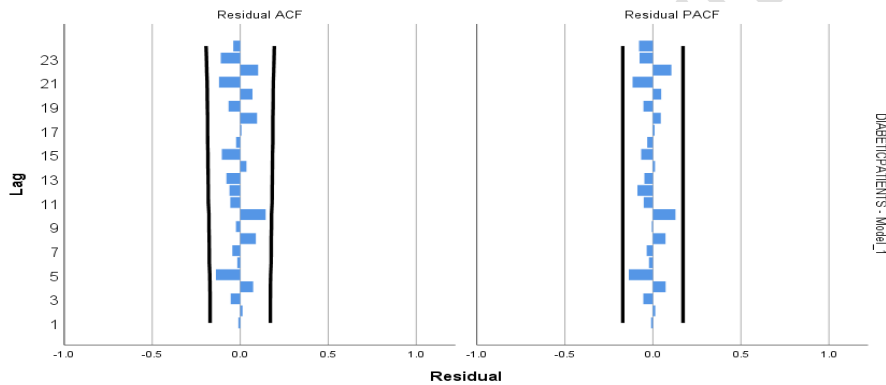


Figure 8: Residual plots of ACF and PACF after estimating ARIMA (3,1,1)

Based on the Ljung-Box Q statistics, the models with significant p-values are selected and shown in Table 9 below. The resultant BIC providing in Table 10 shows that the most appropriate model to fit diabetes type 1 patients' data is ARIMA (3, 1, 1).

Table 9: Competitive models for the monthly diabetes patients based on the Ljung-Box Q statistics test

Models	Q-statistics	P-value
ARIMA(1,1,0)	52.091	.000
ARIMA(2,1,0)	30.689	.015
ARIMA(3,1,0)	19.379	.197
ARIMA(3,1,1)	13.597	.480

Table10: Competitive models for the monthly diabetes patients

Fit statistic		ARIMA(1,1,0)	ARIMA(2,1,0)	ARIMA(3,1,0)	ARIMA(3,1,1)
Stationary square	R-	.116	.301	.362	.455
R-square		-.635	-.294	-.181	-.009
RMSE		35.012	31.271	29.991	27.832
MAPE		42.199	36.591	33.699	29.653
MaxAPE		577.529	397.829	434.673	349.071
MAE		27.658	25.216	23.023	21.432
Normalized BIC		7.186	6.997	6.951	6.838

Diagnostic checking and parameter estimation

Diagnostic checking is also known as the verification. It is to do with the testing of the goodness of fit test statistics of a model. We study the ACF and PACF of the residual plot to see it is white noise. If all the autocorrelation and partial correlation are small then the model is working fine but if some of the autocorrelations are large, the values of p or q are adjusted and the model is re-estimated. The checking of residuals and adjusting the values of p and q continues until the resulting residuals contain no additional Structure.

Form Figure 8 the sample ACF and PACF of the model shows that the autocorrelations of the residual are all close to zero which mean they are uncorrelated, hence the residual assume mean of zero and constant variance. Finally, the p-value (0.480) for the Ljung-Box statistic clearly exceeds 5% for all lag orders. Thus, the selected model ARIMA (3, 1, 1) satisfies all the model assumptions.

Forecasting using ARIMA (3, 1, 1)

The forecast values with 95 percent forecast limit of the ARIMA (3, 1, 1) of model for monthly diabetesPatients are shown in Table 16 with standard error, lower and upper limit and its actual forecasted values.

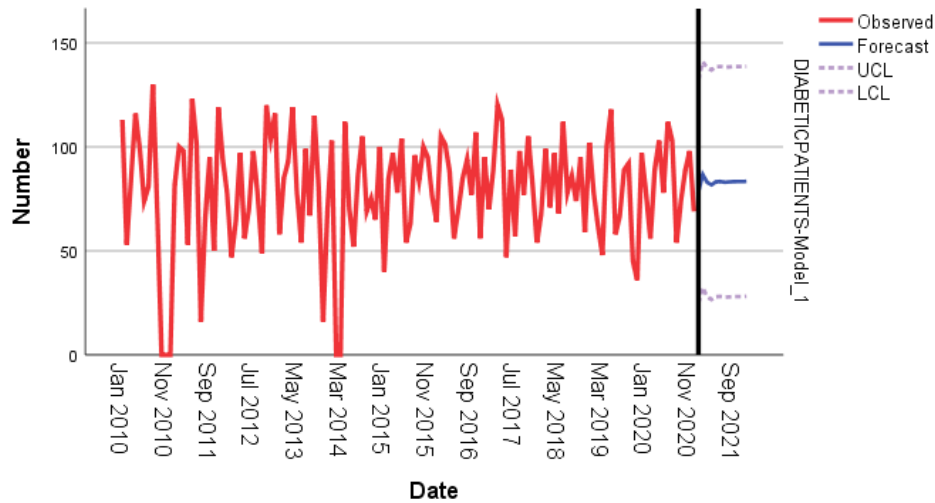


Figure 9: Forecast chat of ARIMA (3,1,1)

MODEL FOR TYPE 2 DIABETES

Model identification

The monthly Diabetes type 2 recorded in Jos University Teaching Hospital from 2010-2020 is plotted in the figure13. The data plot display trends, which indicates that the mean is not stationary. This can be corrected by first order differentiation.

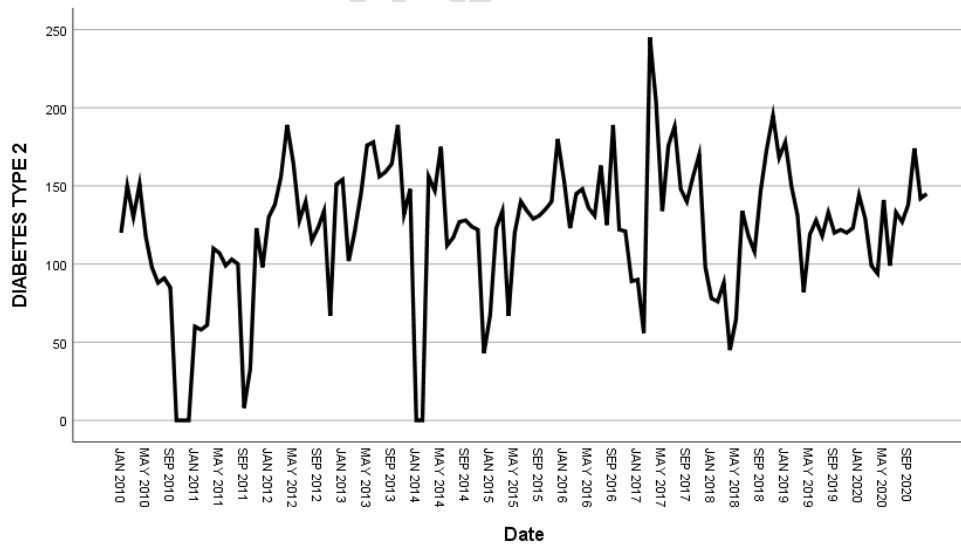
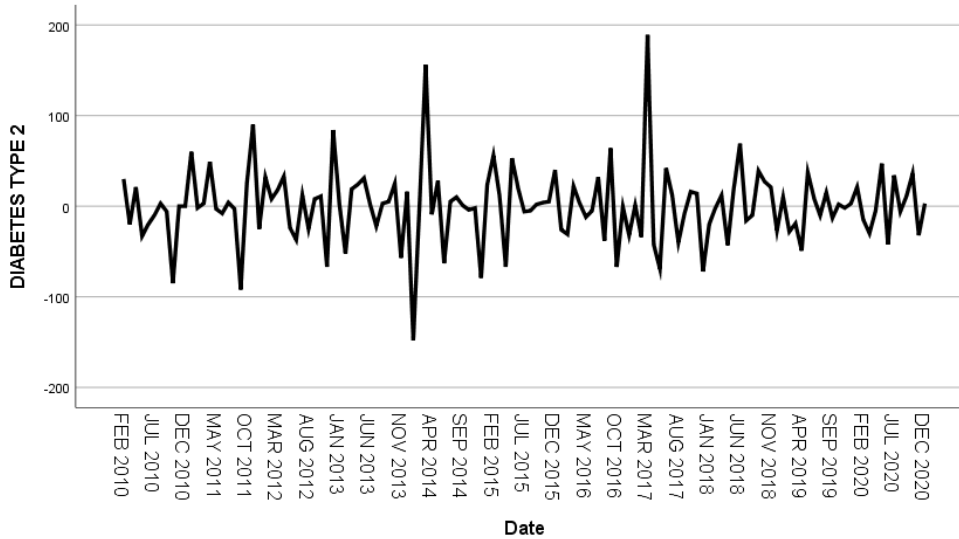


Figure 10: Monthly number of diabetes type 2 patients from Jos University Teaching Hospital (January 2010- December 2020)



Transforms: difference(1)

Figure11: Time series plot of Diabetes patients after the first differencing

Graphical representation of autocorrelations

From the ACF plots below (Figure 15), the value of its function is less than one. The parameter of any given function is less than one; it means that there is stationarity.

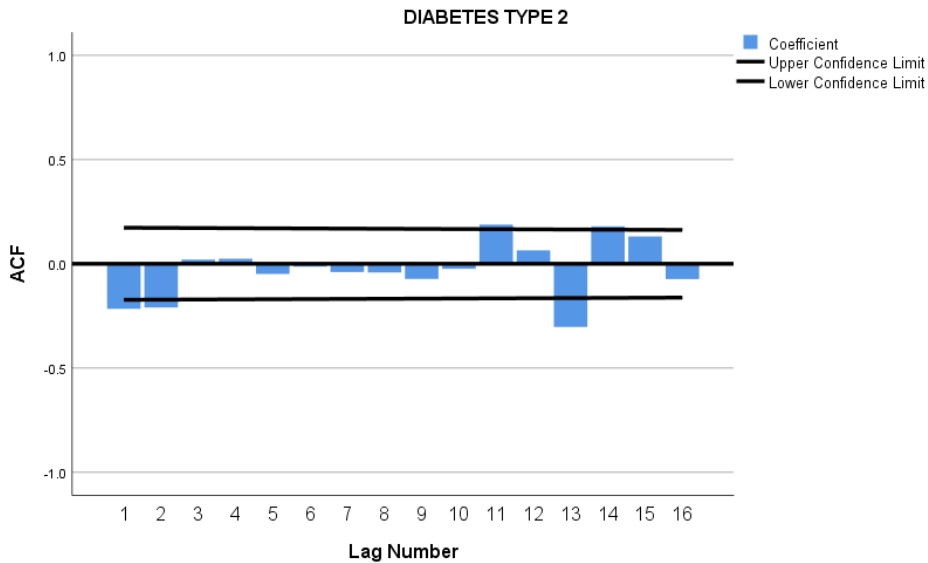


Figure12: Sample ACFplot of monthly Diabetes type2 Data after first differencing.

The plot shows a slight negative spike at lag 13 and there is no other spikes showing that it is over differenced. Further the plot decays to zero on the both sides of the mean

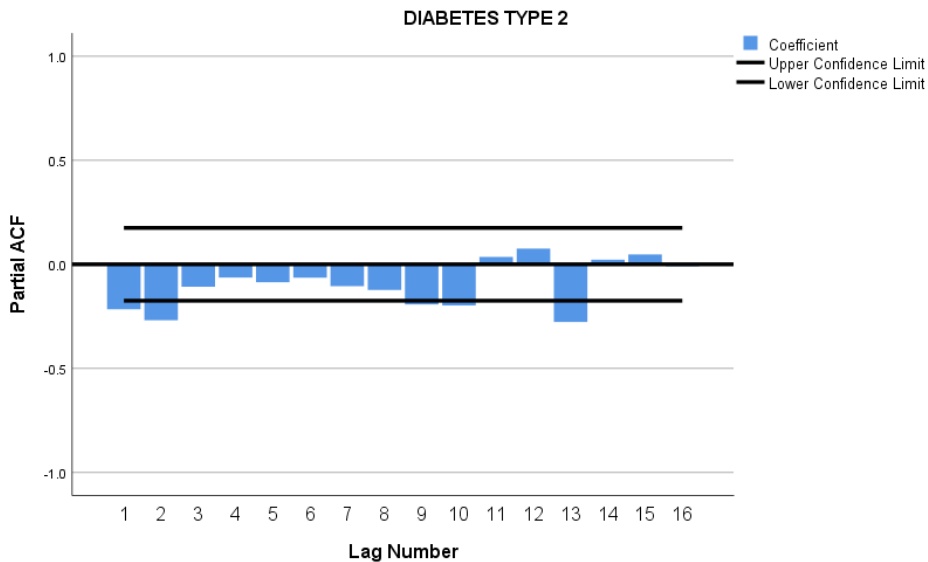


Figure 13: Sample PACF plot of monthly Diabetes type 2 Data after first differencing

Parameter Estimation

Four models are chosen and tested to find a model with good fit. These models include ARIMA(1,1,1), ARIMA(1,1,0), ARIMA(0,1,1), and ARIMA(0,1,0)

Table 11: ARIMA Model Parameters

			Estimate	SE	t	Sig.	
DIABETE	DIABETES	No	Constant	.290	.207	1.401	.164
S	TYPE	TYPE 2	AR Lag 1	.553	.084	6.568	.000
2-		Transformation	Difference	1			
Model_1			MA Lag 1	.999	.644	1.551	.123

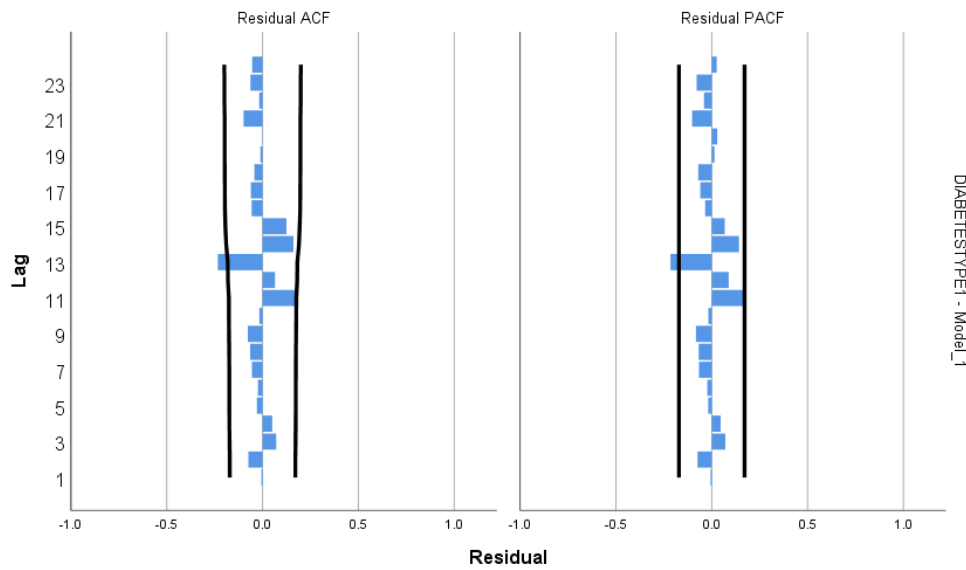


Figure 14, residual plots of ACF and PACF after estimating ARIMA (1,1,1)

Table 12, Model Fit of ARIMA(0,1,1)

Fit Statistic	Mean	SE	Percentile								
			Minimum	Maximum	5	10	25	50	75	90	95
Stationary R-squared	.092	.	.092	.092	.092	.092	.092	.092	.092	.092	.092
R-squared	.197	.	.197	.197	.197	.197	.197	.197	.197	.197	.197
RMSE	39.755	.	39.755	39.755	39.755	39.755	39.755	39.755	39.755	39.755	39.755
MAPE	31.745	.	31.745	31.745	31.745	31.745	31.745	31.745	31.745	31.745	31.745
MaxAPE	1159.469	.	1159.469	1159.469	1159.469	1159.469	1159.469	1159.469	1159.469	1159.469	1159.469
MAE	27.847	.	27.847	27.847	27.847	27.847	27.847	27.847	27.847	27.847	27.847
MaxAE	171.586	.	171.586	171.586	171.586	171.586	171.586	171.586	171.586	171.586	171.586
Normalized BIC	7.440	.	7.440	7.440	7.440	7.440	7.440	7.440	7.440	7.440	7.440

Table 13 Model Parameters of ARIMA(0,1,1)

				Estimate	SE	t	Sig.
DIABETE	DIABETE	No	Constant	.132	2.019	.065	.948
S	TYPE	2	Difference	1			
2-		on	MA Lag 1	.422	.080	5.269	.000
Model_1							

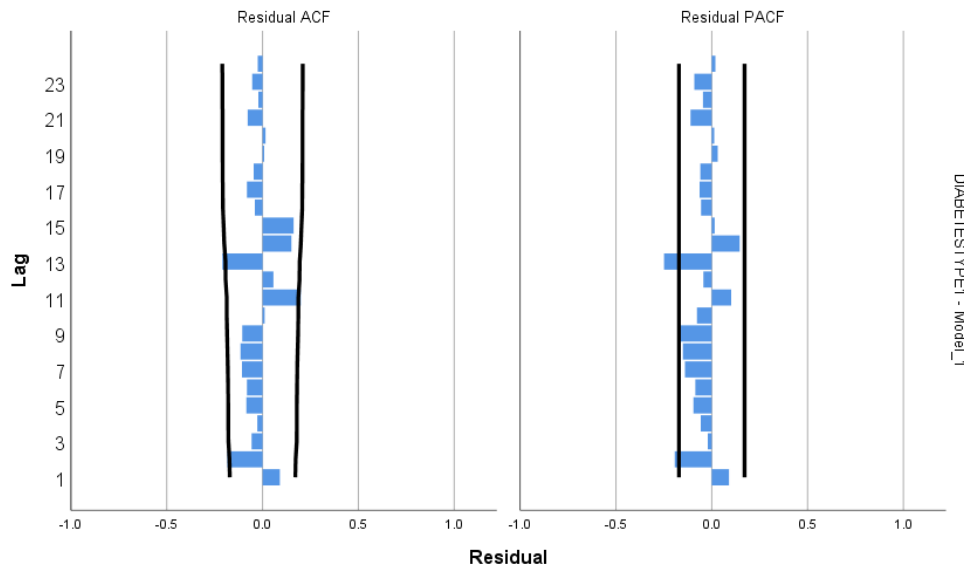


Figure 15, residual plots of ACF and PACF after estimating ARIMA (0,1,1)

Table 14: Model Fit ARIMA(1,1,0)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.047	.	.047	.047	.047	.047	.047	.047	.047	.047	.047
R-squared	.157	.	.157	.157	.157	.157	.157	.157	.157	.157	.157
RMSE	40.741	.	40.741	40.741	40.741	40.741	40.741	40.741	40.741	40.741	40.741
MAPE	31.491	.	31.491	31.491	31.491	31.491	31.491	31.491	31.491	31.491	31.491
MaxAPE	1160.286	.	1160.286	1160.286	1160.286	1160.286	1160.286	1160.286	1160.286	1160.286	1160.286
MAE	27.415	.	27.415	27.415	27.415	27.415	27.415	27.415	27.415	27.415	27.415
MaxAE	181.55	.	181.55	181.55	181.55	181.55	181.55	181.55	181.55	181.55	181.55
Normalized BIC	7.489	.	7.489	7.489	7.489	7.489	7.489	7.489	7.489	7.489	7.489

Source: spss

Table 15: ARIMA Model Parameters of ARIMA(1,1,0)

				Estimate	SE	t	Sig.
DIABETES TYPE 2- Model_1	DIABETES TYPE 2	No Transformation	Constant	.147	2.934	.050	.960
			AR Lag 1	-.215	.086	-2.500	.014
			Difference	1			

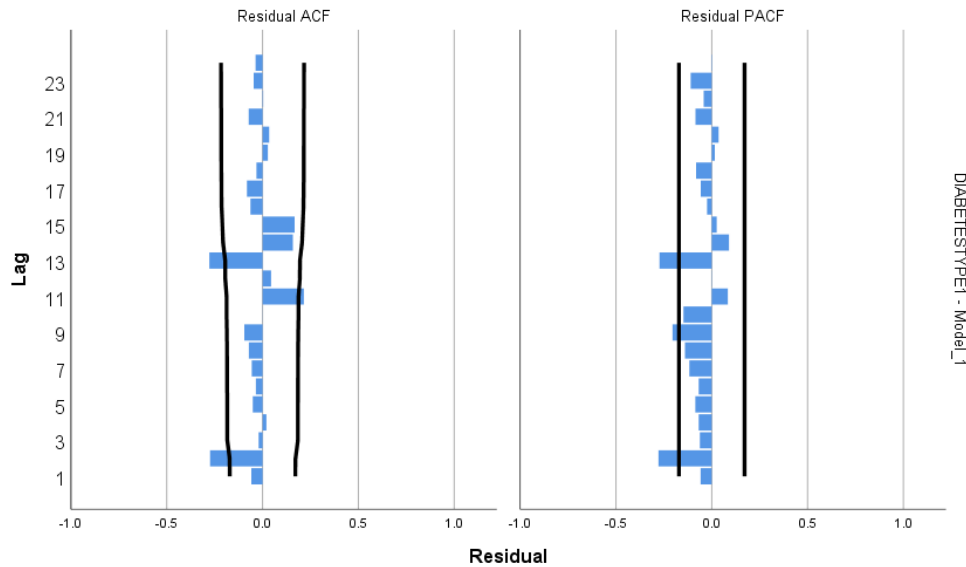


Figure 16: Residual plots of ACF and PACF after estimating ARIMA (1,1,0)

Table 16: Model Fit of ARIMA(0,1,0)

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.00	.	.000	.000	.000	.000	.000	.000	.000	.000	.000
R-squared	.11	.	.116	.116	.116	.116	.116	.116	.116	.116	.116
RMSE	41.562	.	41.562	41.562	41.562	41.562	41.562	41.562	41.562	41.562	41.562
MAPE	32.512	.	32.512	32.512	32.512	32.512	32.512	32.512	32.512	32.512	32.512
MaxAPE	1152.385	.	1152.385	1152.385	1152.385	1152.385	1152.385	1152.385	1152.385	1152.385	1152.385
MAE	27.970	.	27.970	27.970	27.970	27.970	27.970	27.970	27.970	27.970	27.970
MaxAE	188.809	.	188.809	188.809	188.809	188.809	188.809	188.809	188.809	188.809	188.809

Normalized BIC	7.492	7.492	7.492	7.492	7.492	7.492	7.492	7.492	7.492	7.492	7.492
----------------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 17: Model Parameters of ARIMA(0,1,0)

				Estimate	SE	t	Sig.
DIABETES TYPE Model_1	DIABETES TYPE 2	No Transformation	Constant Difference	.191	3.631	.053	.958
				1			

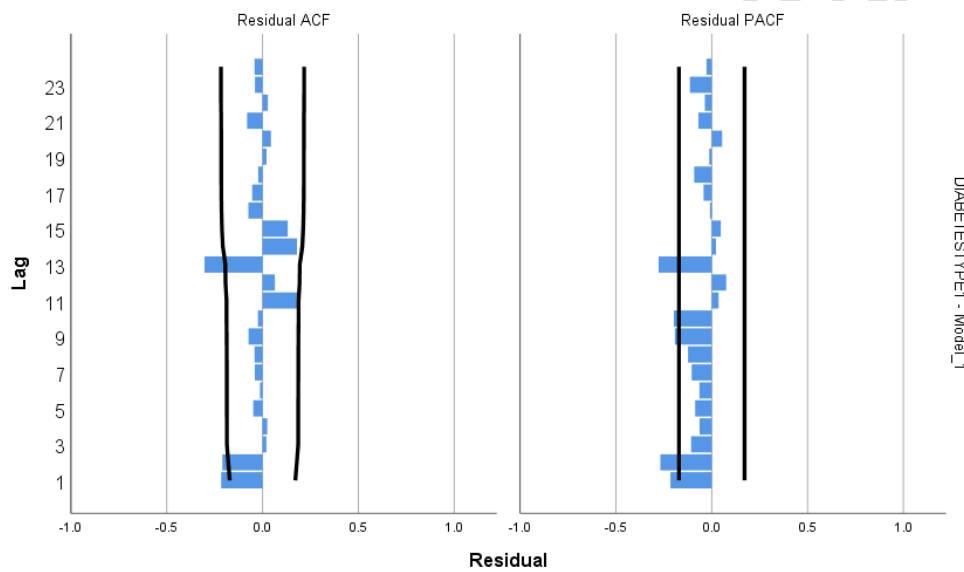


Figure 17: Residual plots of ACF and PACF after estimating ARIMA (0,1,0)

Based on the Ljung-Box Q statistics, the models with significant p-values are selected and shown in Table 18 below. The resultant BIC providing in Table 19 shows that the most appropriate model to fit diabetes type 2 patients' data is ARIMA (1, 1, 1).

Table 18: Competitive models for the monthly diabetes type 2 patients based on the Ljung-Box Q statistics test

Models	Q-statistics	P-value
ARIMA(1,1,1)	24.311	.083
ARIMA(0,1,1)	34.899	.006

ARIMA(1,1,0)	41.814	.001
ARIMA(0,1,0)	41.991	.001

Table 19: Competitive models for the monthly diabetes type 2 patients

Fit statistic		ARIMA(1,1,1)	ARIMA(0,1,1)	ARIMA(1,1,0)	ARIMA(0,1,0)
Stationary R-square	R-	.215	.092	.047	.000
R-square		.306	.197	.157	.116
RMSE		37.114	39.755	40.741	41.562
MAPE		30.204	31.745	31.491	32.512
MaxAPE		1107.781	1159.469	1160.286	1152.385
MAE		26.553	27.847	27.415	27.970
Normalized BIC		7.340	7.440	7.489	7.492

Diagnostic checking and parameter estimation

Form Figure 14 the sample ACF and PACF of the model shows that the autocorrelations of the residual are all close to zero which mean they are uncorrelated, hence the residual assume mean of zero and constant variance. Finally, the p-value (0.83) for the Ljung-Box statistic clearly exceeds 5% for all lag orders. Thus, the selected model ARIMA (1, 1, 1) satisfies all the model assumptions.

Forecasting using ARIMA (1, 1, 1)

The forecast values with 95 percent forecast limit of the ARIMA (1, 1, 1) of model for monthly diabetes type 2 Patients are shown in figure 18 with lower and upper limit and its actual forecast.

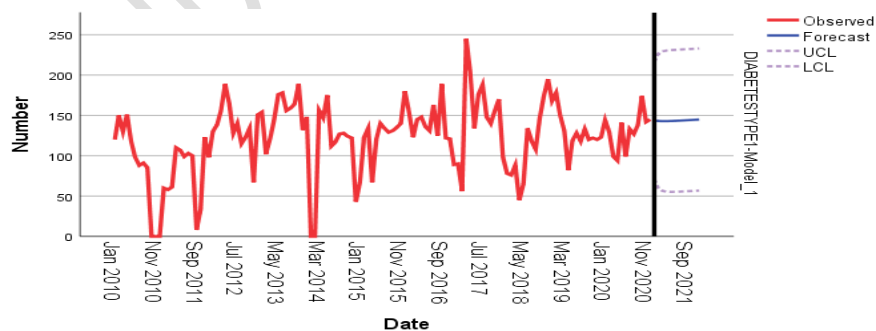


Figure 18: forecast chat of ARIMA(1,1,1)

CONCLUSION

In this research, Diabetes type 1 and type 2 patients were studied using the ARIMA modeling and forecasting method. Box-Jenkins was used to study the monthly Diabetes type 1 and type 2 patients records in Jos University Teaching Hospital for the period of January 2010 to December 2020. The purpose of this research work is to forecast the monthly diabetes type 1 and type 2 cases for the coming period of January 2020 to December 2025. The results also show the forecast for next five years. List of tentative ARIMA models are developed based on the Ljung-Box Q statistic test, BIC and after testing the significant of the estimated parameters, ARIMA(3, 1, 1) model was seen best fit for forecasting the Diabetes type 1 patient and ARIMA(1,1,1) was best fit for forecasting diabetes type 2 patients. The study came out with two better fit models for predicting the prevalence of diabetes type 1 and type 2 recorded in Jos Plateau state, these models are ARIMA(3,1,1) for type 1 and ARIMA(1,1,1) for type 2. The forecast of these models shows that both diabetes type 1 and type 2 increase gradually over time, this will give researchers and health workers information on the prevalence of diabetes type 1 and type 2. Therefore, the proposed model will help to plan appropriately and allocate resources for emergency.

Comment [u4]: Add strength and limitation of study and recommendation

REFERENCE

- [1] Diabetes Care. (2006). 29, S43–S48. Retrieved from Carediabetesjournals.org
- [2] 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.
- [3] 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.
- [4] World Health Organization, WHO (2019)
- [5] Deshpande, A. D.; Harris-Hayes, M. and Schootman, M. (2008) Epidemiology of diabetes and Diabetes-Related complications. *Diabetes Special Issue. Physical Therapy*. Vol.88, No.11, pg.1255.
- [6] Naemiratch, B. & Manderson, L. 2007. Lay Explanations of Type 2 Diabetes in Bangkok, Thailand. *Anthropology & Medicine*. Vol. 14, No. 1, 83.
- [7] International Diabetes Federation, IDF (2007).
- [8] 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.
- [9] 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.
- [10] Zimmet, P. (2003). The burden of type 2 diabetes: Are we doing enough? *Diabetes & Metabolism*, 29, 659–681.
- [11] 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.
- [12] 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.
- [13] 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.

Comment [u5]: Add references

- [14] Nyenwe, E. A., Odia, O. J., Ihekweba, 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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 of Public Health* 28, 336–346.
- [19] 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.
- [20] 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
- [21] 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
- [22] 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.
- [23] Dowell D, Tian LH, Stover JA, Donnelly JA, Martins S, Erbeling EJ, et al (2011). Changes in Fluoroquinolone Use for Gonorrhoea Following Publication of Revised Treatment Guidelines. *Am J Public Health*. 102(1):148–55. doi: 10.2105/ajph.2011.300283 PMID: 22095341.
- [24] 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.
- [25] Williams K. Investigating Risk Factors Associated with Syphilis Rate in the United States Based on ARIMA and ARCH/GARCH Time Series Models.
- [26] 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.
- [27] Zheng YL et al. (2015). Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. *PLoS One* 10, e116832.
- [28] 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.
- [29] Song X et al. (2016). Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011. *Medicine (Baltimore)* 95, e3929.
- [30] Uzuke, C. A., H. O. Obiora-Ilouno, Eze F. C and J. Daniel (2016) Time series Analysis of All Shares Index of Nigerian Stock Exchange: A Box-Jenkin Approach, vol.6, pg.25-26