TIME SERIES ANALYSIS AND FORCASTING OF SEXUAL TRANSMITTED INFECTIONS USING THE BOX JENKINS APPROACH

Abstract

Sexually Transmitted Diseases (STDs) pose a significant public health challenge globally. Understanding the temporal patterns and trends in STD incidence is crucial for effective prevention and intervention strategies. The study aims to analyze the pattern of sexual transmitted diseases in the Kasena Nankana Municipal. The sample data was the monthly entries from the year 2009 to 2018 from the "War Memorial Hospital". Through the use of the Box-Jenkins Methodology, various SARIMA models were estimated and the best among them was selected. The model building was based on three stages, the first stage was the model identification where in the series was already stationary so did not require any form of differencing based on the results that was provided by the KPSS and ADF test. An ACF and PACF plot of the data was obtained and from which the tentative models, SARIMA (2,0,4) $(1,01)_{[12]}$ was the selected best model. After forecasting 2years ahead with the model, it was observed that the incidence of STDs will continue to fluctuate in the same pattern as the original series.

1.0 Introduction

According to Nordqvist(2018), “Sexually transmitted diseases (STDs) are infections that are passed from one person to another through sexual contact. They are also known as sexual transmitted infections (STIs) or venereal diseases (VD). Some STDs can be spread through the use of unsterilized drug needles, from mother to infant during childbirth or breast-feeding, and blood transfusions”. The Centers for Disease Control also reported that, the causes of STDs are bacteria, virus, yeast and parasites. The most common types of STDs are syphilis, gonorrhea, HIV/AIDS, genital herpes, chlamydia, and others CDC (2015).

“Sexually transmitted infections (STI’s) have become increasingly prevalent as a growing global health condition. In the United States, human papillomavirus (HPV) is the most common newly acquired STI, with an estimated annual incidence of 6.2 million cases. Infection with high-risk oncogenic HPV types is a necessary cause of cervical cancer worldwide” Jane et al (2019).

2.0 Literature Review

World Health Organization WHO (2002) have defined sexual transmitted diseases as diseases that are transmitted predominantly by sexual contact, including vaginal, anal and oral sex. STDs can be transmitted through non-sexual means such as blood products.

The National Health Service (NHS) – referring to the four freely subsidized social insurance frameworks in the United Kingdom – determines that the transmission is through unprotected sex (sex without a condom) and furthermore through genital contact NHS (2013).

In a baseline survey on STDs conducted in the Northern Region and Upper East Region of Ghana, 5.2% admitted that they have ever been treated of STDs ranging between 1-2 times during their lifetime. In the year before the study, 2.9% of the respondents were treated o STIs. The most common STIs that respondents contracted were gonorrhea (67.7%), followed by chlamydia (19.4%) and then syphilis (12.9%), Kannae and Anafi (1999).

The CDCs center for Health Statistics Report (2004) indicates that chlamydia, gonorrhea, HIV/AIDS, syphilis, and hepatitis B are the five most common STDs. Furthermore, STDs disproportionately affect women, who suffer more frequent and more serious STDs complications than men. They established difference in STD prevalence rates on gender but did not examine the association between blood type and the transmission of STDs.

3.0 Methodology

In order to gather relevant data and information for this research work, the secondary source of data collection was adopted. The people of Kasena Nankana Municipal in the Northern Region of Ghana are the target population, specifically the patients of 'War Memorial Hospital'.

3.1 THE BOX-JENKINS METHODOLOGY

The Box-Jenkins approach also known as the Autoregressive Integrated Moving Average (ARIMA) modeling, is the most widely used methodology for analyzing time series data, this is

because it can handle any time series data that is stationary or not with or without seasonal elements. The following are the steps involved in Box-Jenkins approach:

- Plot the time series data
- Test for stationarity
- Difference the data if it is not stationary
- Identify a tentative model
- Estimate the model
- Diagnostic checking

4.0 ANALYSIS AND DISCUSSION OF RESULTS

4.1 DESCRIPTVE ANALYSIS

Table 4.1 clearly indicates that, the average incidence of STDs per month for the recorded incidence of STDs from (2009 to 2018) is 18.00 with the median also being 18.00. Also the maximum incidence of STDs recorded was 27 with the minimum value being 10.

Table 4.1 Descriptive statistics of STDs

Variable	Mean	Median	Minimum	Maximum	Standard Dev	Variance	Sample(N)
STDs	**18.00**	**18.00**	**10.00**	**27.00**	**3.21**	**1.79**	**120**

The plot below shows irregular features as a result of random changes in the incidence of sexual transmitted diseases within the given time period. It could be noted that the incidence of STDs fluctuates regularly resulting into a zigzag shape. The lowest incident of STDs was recorded between the periods of 2010 to 2012 and the highest incidence was recorded within 2016 to 2018.The visual representation of the time series data did not show any form of trend (increasing or decreasing) since the values do not go up or down over a long term and hence given us an idea that the time series data already is stationary. But then to confirm this, a number of statistical stationarity test was performed.

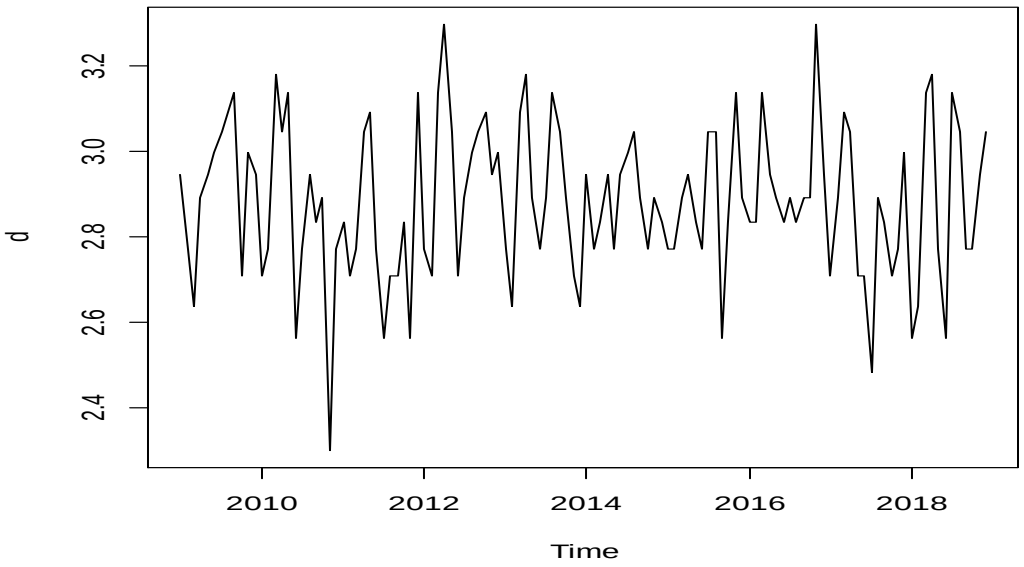

Fig 4.1 Time Series Plot of the Data

4.2 TEST FOR STATIONARITY

Table 4.2 Stationarity test for the data

Test Type	Test Statistic	Lag order	P-value
ADF	-4.3166	**4**	**0.01**
KPSS	**0.051286**	**4**	**0.1**

Table 4.2 above presents the Augmented Dickey-Fuller (ADF) Test for the null hypothesis of a unit root against the alternative of a stationary series together with the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for the null hypothesis of a level of stationary against an alternative of a unit root.

The ADF test statistic of -4.3166 with a p-value of 0.01 which is less than 0.05, hence we reject the null hypothesis and conclude that the data is stationary. The KPSS test also presents a test statistic of 0.051286 with a p-value of 0.1 which is greater than 0.05, hence we fail to reject the null hypothesis and conclude that the data is stationary.

In conclusion, it is clear from the time series plot and the statistical tests performed that the data is stationary and hence do not require any form of differencing to become stationary.

4.3 DEOMPOSITION OF ADDITIVE TIME SERIES

Figure 4.3 below clearly indicates the estimated components of the time series data, that the trend, seasonal, random and observed components. It could be observed from the trend component that, there was a sharp decrease in the incidence of STDs within the period of 2010 to 2012 and fluctuates throughout the remaining years. The seasonal component in the time series clearly indicates some form of seasonality though it was not really clear in the original plot of the series and must be captured in our model in order obtain a good forecast.

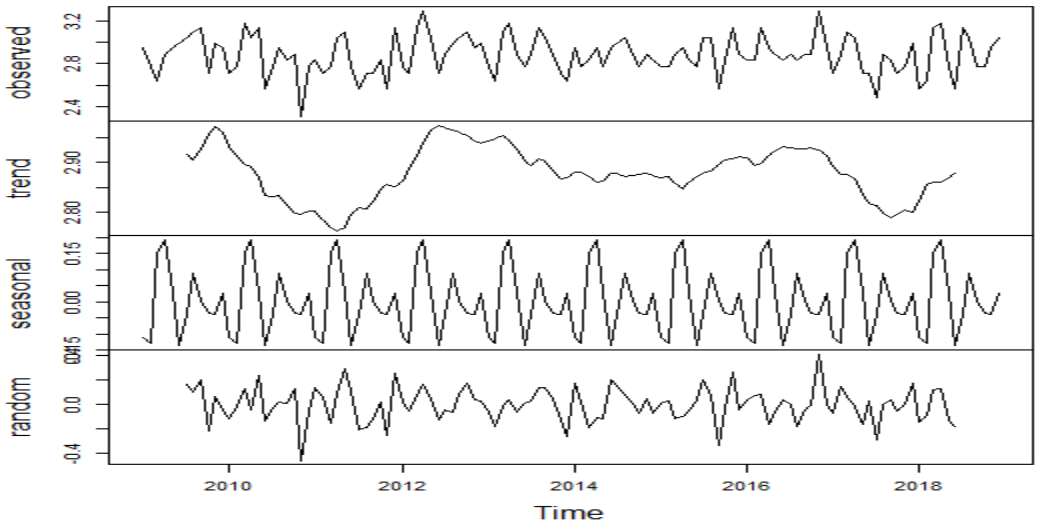

Fig 4.2 Decomposition of Additive Series

Fig 4.4 shows the ACF and PACF plot of the data. The top part is the plot of the autocorrelation function and the down part is the plot of the partial autocorrelation function. From the ACF plot, the autocorrelation is significant at lag 1, 2 and 4 indicating MA (1), MA (2) and MA (4) process respectively. Also, the PACF plot is significant at lag 1 and 2 suggesting AR (1) and AR (2) process. Observing the seasonal lags, we notice that there is a significant spike at lag 12 in the ACF (Q=1) and PACF (P=1) indicating that we must capture the seasonal component in our model. Due to fact that the seasonality is not that visible in the original time series plot, we set

D=0 in the seasonality component. Based on the above, the following tentative models were obtained;

- ARIMA (1, 0, 1) $(0, 0, 1)_{[12]}$
- ARIMA (2, 0, 4) $(1, 0, 0)_{[12]}$
- ARIMA (1, 0, 2) $(0, 0, 1)_{[12]}$
- ARIMA (2, 0, 0) $(0, 0, 1)_{[12]}$
- ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$
- ARIMA (2, 0 ,4) $(0, 0,1)_{[12]}$

To select the best model, the AIC, BIC and the Log likelihood values for the models were computed and then compared to obtain the model with the least of the values.

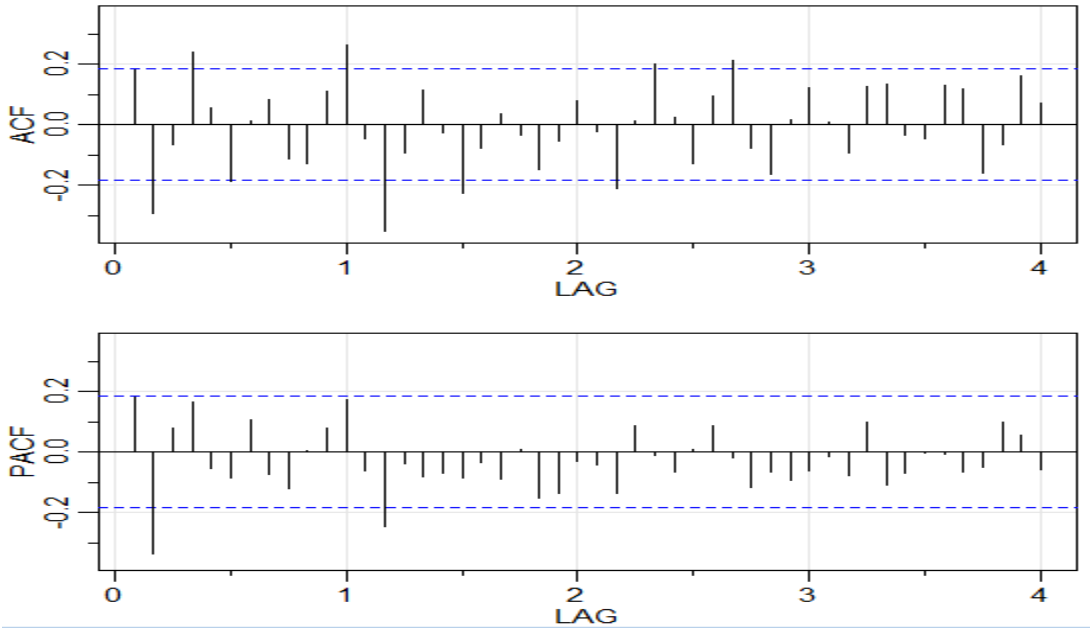

Fig 4.3 ACF and PACF

Table 4.3 ARIMA model selection of STDs

Model	AIC	BIC	Log likelihood
ARIMA (1, 0, 1) $(0, 0, 1)_{[12]}$	610.89	624.82	-300.44
ARIMA (2, 0, 4) $(1, 0, 0)_{[12}$	609.89	634.98	-295.95
ARIMA (1, 0, 2) $(0, 0, 1)_{[12]}$	611.23	627.95	-299.61
ARIMA (2, 0, 0) $(0, 0, 1)_{[12]}$	607.26	621.2	-298.63
ARIMA (2, 0, 4) (1, 0, 1)$_{[12]}$	**602.53***	**630.4***	**-291.26***
ARIMA (2, 0, 4) $(1, 0, 0)_{[12}$	610.08	635.16	-296.04

From Table 4.3, we can conclude that ARIMA (2, 0, 4) $(0, 0, 1)_{[12]}$ is our best model since it has the least AIC, BIC and Log likelihood value.

4.5 ESTIMATION OF THE PARAMETERS OF THE MODEL SELECTED

Parameter estimates determine the coefficients of the time series equation that is generated from the data and the diagnostics test is used to check the correlation and significance of the residuals.

4.5.1 PARAMETER ESTIMATES AND DIAGNOSTICS OF ARIMA (2, 0, 4) MODEL

Table 4.4 Parameter estimates

Type	ar1	ar2	ma1	ma2	ma3	ma4	sar1	sma2	mean
Coefficients	-0.0286	-0.9990	0.2819	0.9688	0.2290	-0.0429	-0.5561	0.7202	18.1317
s.e.	0.0110	0.0027	0.0937	0.1034	0.0995	0.0952	0.2879	0.2581	0.321

Table 4.4 clearly indicates the estimates of the parameters of ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$ with their corresponding standard errors and mean.

4.5.1.2 DIAGNOSTIC CHECKING OF ARIMA (2, 0, 4) (1, 0, 1)[12]

In order to check for the adequacy of the model, an L-jung Box test together with the time plot of the standard errors, acf plot of the residuals and a plot of the p-values of the residuals was performed.

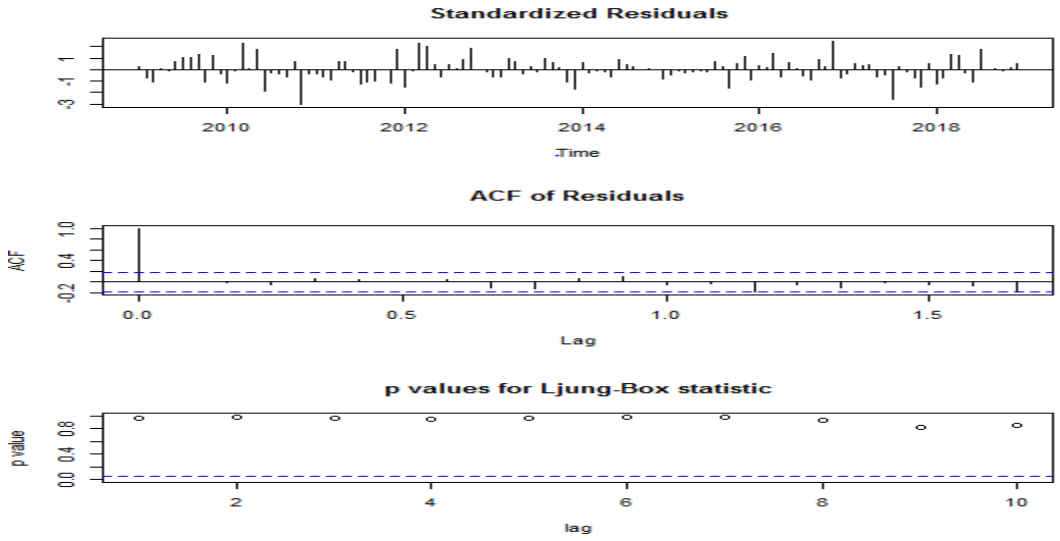

Fig 4.4 Residual Plot of ARIMA (2, 0, 4) (1, 0, 1)[12]

Figure 4.4 above represents the diagnostic residual plots of the model. The top box is the time plot of the standardized residual, the middle is the ACF plot of the residuals and the last box is the probability plot of the residuals. The time plot of the residuals clearly shows that the residuals appear to be randomly scattered about zero, no evidence exists that the error terms are correlated with one another. The residuals or errors are therefore conceived as an independently and identically distributed (i.i.d) sequence with a constant variance and a zero mean. The ACF plot of the residuals shows no significant spikes at any particular lag. Finally, the probability plot of the residuals also indicates that the individual probabilities of the residual are greater than or above 0.05 (blue line). This shows that the residuals of ARIMA (2, 0, 4) (1, 0, 1)$_{[12]}$ is a white noise process.

The distribution of the errors by ARIMA (2, 0, 4) (1, 0, 1)$_{[12]}$ model is as shown in figure…..below. From the diagram it can be observed that the residuals appear to be normally

distributed since most of the data points are on the normal line except for some few residuals deviating from the normality.

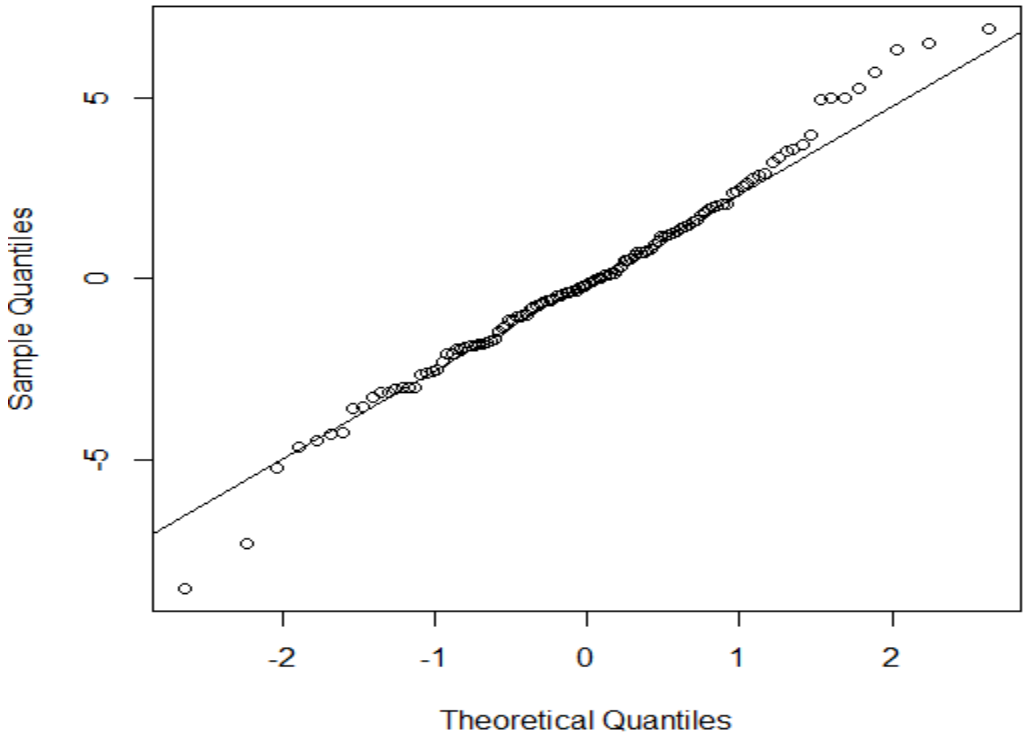

Fig. 4.5 Normal Q-Q Plot of the Residuals

In order to affirm the normality of the residuals, the Shapiro Wilk normality test was performed. The table below is summary of the results obtained from the test.

Table 4.5 Summary of Shapiro test

Test Type	Test Type	P-value
Shapiro-Wilk	0.98635	0.2704

The test provides a test statistic of 0.98635with a p-value of 0.2704which is greater than 0.05, hence we fail to reject the null hypothesis and conclude that the residuals are normally distributed which goes in with the Q-Q plot.

4.5.1.3 L-jung Box Test for ARIMA (2, 0, 4) (1, 0, 1)[12]

Table 4.6 Summary of Ljung-Box test

Test Type	Lag	Test Statistic	P-value
Ljung-Box	1	0.04	0.8417206
	2	0.30	0.8607990
	3	0.72	0.8683403
	4	0.91	0.9225114
	5	0.91	0.9691644
	6	1.30	0.9718580
	7	3.01	0.8840641
	8	5.07	0.7499122
	9	5.55	0.7843957
	10	6.75	0.7489348

The test was performed from lag 1 through to lag 10 with all the p-values greater than 0.05, hence we fail to reject the null hypothesis and conclude that the residuals are serially uncorrelated.

4.4.1.4 Arch-LM Test for ARIMA (2, 0, 4) (1, 0, 1)[12]

Table 4.7 Summary of Arch-LM test

Test Type	Lag	Test Statistic	P-value
	1	0.047486	0.8275
	2	0.31523	0.8542
	3	0.30366	0.9593
	4	0.34761	0.9865
	5	0.34516	0.9967
	6	1.6282	0.9505
	7	2.1467	0.9513
	8	**15.693**	**0.0470**
	9	16.295	0.0609
	10	16.261	0.09241

From table 4.7, it can be observed that after performing the Arch-LM test from lag 1 to lag 12, all the p-values were more than 0.05 except at lag 8. Since it is only one lag that exhibits conditional heteroskedasticity, we can say the model is good for forecasting.

4.4.1.5 McLeod.Li Test

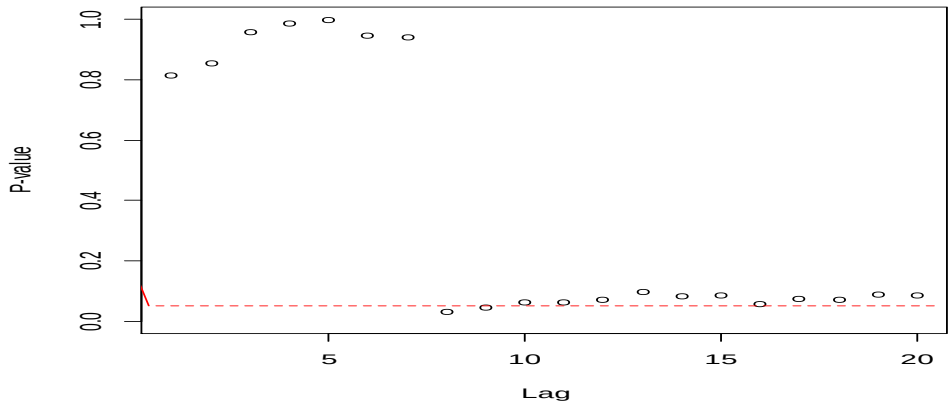

Fig 4.6 McLeod.Li. Test for Conditional Heteroskedasticity

From the McLeod.Li Test for conditional heteroskedasticity it can be observed that one point was totally below the red line indicating the presence the presence of heteroskedasticity at that lag particular probably lag 8 as suggested by Arch-LM test. Since almost all the lags indicates the absence of heteroskedasticity we conclude that the model is good for forecasting.

4.5 FORECASTING

Using the ARIMA model obtained, we forecast for the next two years.

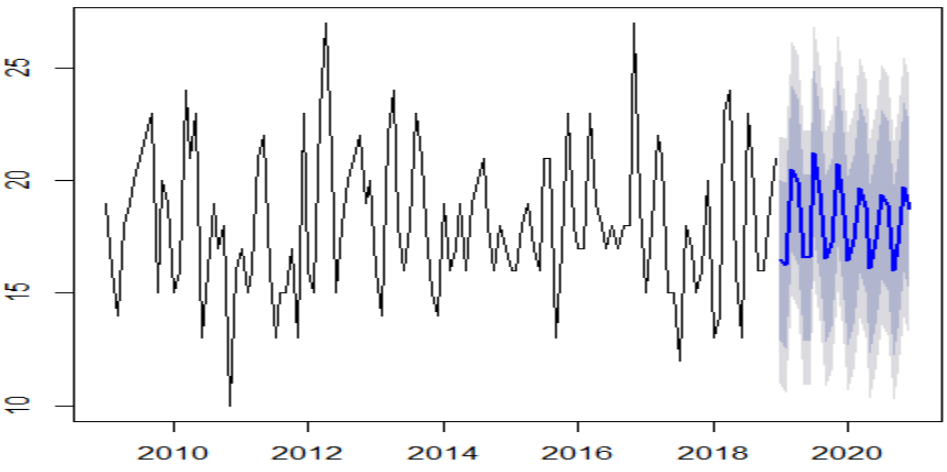

Figure 4.7 Forecast from ARIMA (2,0,4)(1,0,1)[12]

Table 4.8 Forecasted Values.

Forecast Period	Forecasted STDs	95% Lower Limit	95% Upper Limit
January 2019	16.48083	12.89932	11.00338
February 2019	16.23724	12.54297	10.58734
March 2019	20.51958	16.82092	14.86297
April 2019	19.98584	16.28605	14.32750
May 2019	16.57507	12.87551	10.91708
June 2019	16.57507	12.86962	10.91061
July 2019	21.21373	17.51327	15.55437
August 2019	19.46948	15.76796	13.80849
September 2019	16.53489	12.83353	10.87415
October 2019	17.21606	13.51368	11.55376
November 2019	20.75645	17.05419	15.09433
December 2019	19.45891	15.42984	13.44844
January 2020	16.43608	17.05419	10.52147
February 2020	17.34616	13.58150	11.58861
March 2020	19.66835	15.90333	13.91025
April 2020	18.81526	15.04900	13.05526
May 2020	16.08658	12.32037	10.91708
June 2020	17.37052	13.60313	11.60879
July 2020	19.37230	15.60491	13.61057
August 2020	18.93101	15.16250	13.16756
September 2020	16.02394	12.25537	10.26040

October 2020	17.18505	13.41541	11.41989
November 2020	19.70653	15.93678	13.94119
December 2020	18.76110	14.9903	12.99425

4.6 Discussion of Results

From the analysis in Table 4.1, the average cases of STDs within a month in the Kasena Nankana Municipal was found to 18 with the median also being 18. It was also noted the minimum and maximum incidence of STDs that was recorded during the period of study is 10 and 27 respectively.

Fig 4.1 and 4.2 shows the time series plot of the data and it's decomposition into the various components which is the trend, seasonal, irregualar and observed. Also, Fig 4.3 shows the ACF and PACF plot of the data which indicated a significant spike at lag 1, 2 and 4 in the ACF and lag 1 and 2 in the PACF suggesting MA(1), MA(2) ,MA(4) and AR(1), AR(2) respectively. Observing the seasonal lags in the ACF and PACF also gives significant spikes at lag 12 giving us an idea that there might be a seasonality in the data. After decomposing the series in Fig 4.2, it was now clear that there is seasonality in the data. T

Stationarity test was conducted using ADF and KPSS test as shown in Table 4.2. Through this test we realized the data was stationary hence did require any form of non-seasonal differencing. The seasonality component was captured as (1, 0, 1) since the seasonality is not that strong in the original plot of the series.

Also Table 4.3 displays the various models identified and of which ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$ was chosen as the best fitted model due to it having a minimum value of AIC, BIC and Log likelihood. Also Figure 4.4 was also showing the diagnostic plot for ARIMA (2 0, 4) $(1, 0, 1)_{[12]}$. From the diagnostic plot it can be seen that the ACF plot is not significant at any lag. Finally, the probability plot of the residuals also indicates that the individual probabilities of the residual are greater than or above 0.05 (blue line). This shows that the residuals of ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$ is a white noise process.

Moreover, the Ljung- Box and ARCH-LM test in Table 4.6 and 4.7 was indicating that all the test were adequate, taking into account their p-value which were all greater than the alpha value.

This adequacy shown by the tests is also an indication telling as that forecasting on the STDs can be made. Finally, Table 4.8 and Figure 4.7 is displaying the forecasted values for the best model ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$. This forecasting was done on a 2 years monthly basis from January (2019) to December (2020). From the forecasted values it can be seen that, the pattern of STDs will keep fluctuating in an increasing and decreasing manner.

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 FINDINGS

- The best model for the monthly sexual transmitted diseases in Kasena Nankana Municipal is SARIMA (2, 0,4) $(1, 0, 1)_{[12]}$
- From the seasonal values, the month of April experience high cases of sexual transmitted diseases whereas the month of June record the lowest.
- From the forecasted values, the incidence of sexually transmitted diseases will continue fluctuating from increasing and decreasing values.

5.2 CONCLUSION

This study aims to identify the best and accurate model among various ARIMA models which is adequate for forecasting the incidence of STDs in the Kasena Nankana Municipal with monthly data available from the year 2009 to 2018 with the use of the Box-Jenkins Methodology. The model building was based on three stages, the first stage was the model identification where in the series was already stationary so did not require any form of differencing based on the results that was provided by the KPSS and ADF test. An ACF and PACF plot of the data was obtained and from which the tentative models were gotten. Based on the AIC, BIC and the Log likelihood of the tentative models, ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$ was selected as the best model.

The next stage was the estimation of the parameters and diagnostic checking of the best model where the residuals of ARIMA (2, 0, 4) $(1, 0, 1)_{[12]}$ was normally distributed, random (white noise) and had no form of serial correlation and heteroskedasticity. Finally, an out sample forecast for the period of 10 years was made. It was concluded that the model for the incidence of the STDs based on the 12 steps forecasts in the future will follow the same pattern as the actual time plot of the original STDs cases data.

5.3 RECOMMENDATION.

- The research shows high peak in April and June recorded the lowest. I recommend that the health directorate should intensify STDs campaign sensitization in the month April.
- Enough education should also be given to people to avoid excessive use of alcohol and the use of drugs which may help prevent the transmission of disease because these activities may lead to risky sexual behavior.
- The government and NGOs should map up rigorous strategies such as periodic vaccination programs to minimize the spread of the disease in the municipal.

REFERENCES

CDC. (2004). National Health Report. Available at:

https://www.cdc.gov/healthreport/index.htm.

CDC.(2015). Sexually Transmited Diseases. Retrieved from NIAID:

https://.niaid.nih.gov/diseases-conditions/sexually-transmitted-diseases

Dickey, D,A & Fuller, W.A (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. Econometrics, 49(4) 1057- 1072.

Holmes KK, Levine R, Weaver M. Effectiveness of condoms in preventing sexually transmitted

infections. Bull World Health Organ. 2004; 82(6): 451-61

Hamiltonian, J.D (1981). Time Series Analysis. Princeton Univ. Press, Princeton New Jersey.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P & Shin, Y. (1992):

Testing the Null Hypothesis of stationary against the Alternative of a Unit Root. Journal of Econometrics, 54: 159-178.

NHS (2013). Sexually transmitted infections (STIs). Accessible at:

http://www.nhs.uk/conditions/Sexually-transmittedinfections/Pages/Introduction.aspx

Nordqvist, C (2018). Retrieved from Medical News Today:

https://www.medicalnewstoday.com/articles/246491.php