Application and Comparison of Time Series Methods on Tuberculosis Incidence Data : A case study of Zimbabwe 1990-2013

by

Nothabo Dube

A Thesis

In

Statistics

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for the Degree of

MASTER OF SCIENCES

Approved

Clyde F. Martin
Co-chair of Committee

Bijoy Ghosh
Co-chair of Committee

Jingyong Su

Mark Sheridan
Dean of the Graduate School

December, 2015

ACKNOWLEDGMENTS

I would like to extend my gratitude to Dr. Clyde F. Martin for his support, to Dr. Bijoy Ghosh, my co-chair, and Dr. Su, a member of my committee. The support you have all given me is greatly appreciated. Sincere thanks also go to my family and friends for their continued support.

CONTENTS

ABSTRACT

Tuberculosis remains a major global public health problem, especially in countries that are considered as high burden countries. Zimbabwe is considered by WHO as one of the high burden countries and tuberculosis incidence continues to be very high. Therefore, there is need to continue monitoring and predicting tuberculosis incidence in an effort to make the control of tuberculosis more effective. The Box-Jenkins approach, specifically the autoregressive integrated moving average (ARIMA) model, is typically applied to predict the incidence of infectious diseases. This method takes into account changing trends, periodic changes, and random disturbances in time series. Autoregressive conditional heteroscedasticity (ARCH) models are the prevalent tools used to deal with time series heteroscedasticity. Holt Winters (HW) methods also play a significant role in time series forecasting and are especially effective for short term forecasting. In this study, based on the data of the tuberculosis incidence from 1990 -2003 in Zimbabwe, we establish the single ARIMA (2, 2, 1)model, the combined ARIMA (2, 2, 1)-ARCH (1) model, and the HW model, which can be used to predict the tuberculosis incidence successfully in Zimbabwe. Comparative analyses show that the ARIMA and ARIMA-ARCH models perform reasonably well, with the ARIMA model being the best in our case. To the best of our knowledge, this is the first study to establish the ARIMA model and ARIMA-ARCH model for prediction and monitoring the yearly incidence of tuberculosis in Zimbabwe. Based on the results of this study, the ARIMA (2, 2, 1)and ARIMA(2, 2, 1)-ARCH (1) models are suggested to give tuberculosis surveillance by providing estimates on tuberculosis incidence trends in Zimbabwe.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

Tuberculosis (TB) is a chronic respiratory infectious disease caused by the pathogen *Mycobacterium tuberculosis* and spreads through air droplets by sneezing and coughing of the infected person [33]. TB infection, if not timely treated, can be a serious health threat [40]. It is one of the biggest health challenges worldwide and it is the second major cause of mortality, particularly in poor and low income countries [13] and [14]. An estimated 9.0 million people developed TB in 2013; 1.5 million died from the disease according to the World Health Organization [31]. Many efforts to control this disease, have been put in place, but TB still remains a major public health issue with a high global health burden. The emergence of multidrug resistant (MDR), extensive drug resistant (XDR), and the recent emergence of total drug resistant (TDR) strains along with coinfection with human immunodeficiency virus (HIV), have made it particulary so, especially in developing countries [8].

Tuberculosis is a significant public health problem in Zimbabwe with high morbidity and mortality rates. According to WHO Global Tuberculosis Report 2014, the estimated TB incidence in 2013 was 552 cases per 100,000 population. In the same year, the estimated prevalence of TB (all forms) in Zimbabwe was 409 cases per 100, 000 population. The total case notification for all TB cases in 2013 was 35 278. Of all the countries that report their TB statistics to WHO, there are 22 countries that are sometimes referred to as the TB high burden countries, and they have been prioritized at a global level since 2000. These 22 countries, between them accounted for 82 percent of all estimated cases of TB worldwide in 2013 [31]. Zimbabwe has among the highest estimated TB incidence per capita (603/100,000 population) in the world [31, 25]. Sixteen percent of adults are HIV infected, and approximately three-quarters of active TB cases occur among persons with HIV [6].

In the Millennium Development Objectives agreed upon in September 2000 in the United Nations and accepted by 189 countries, the TB control program had to achieve

the objectives of reducing by 50 percent, mortality from TB in comparison with 1990. Also to stop or reduce its incidence and prevalence by 2015 and finally eliminating, that is, reducing incidence to less than one case per million population) by 2050. Accordingly, the global plan to stop TB started its activity in January 2006 with an investment of more than 67 billion dollars and presented guidelines and strategies to control or eliminate tuberculosis based on dynamics of TB infection in societies [29]. In spite of these achievements and other effective attempts, achieving the predicted objectives is very difficult due to some uncontrollable problems.

Epidemiological studies have long been used to explain TB incidence and prevalence and its mortality. Considering epidemiological transition, emerging of Multi-Drug Resistant (MDR) and Extensively-Drug Resistant (XDR) TB and spread of HIV/AIDS, they can predict new challenges and present solutions [34, 7]. A review of temporal changes and prediction of tuberculosis can play an important role in the presentation of future health problems. This includes developing and expanding controlling and intervention programs and allocating resources optimally [2]). To predict tuberculosis incidence and to study its temporal changes, different mathematical and statistical models have been used in different studies [3, 1, 38, 37, 17, 27], and based on data nature and evaluation, a certain model has been used in every study. For example, Abdullah et al., applied a univariate time series model to the TB incidence data in Malaysia in order to determine the best forecasting model [1] or Kilicman et al., showed that Holts trend corrected exponential smoothing is the best forecasting model, followed by the quadratic trend model [17]. Zhang et al., used the ARIMA model in order to forecast tuberculosis [38]. The ARIMA-ARCH model was applied by Zheng et al., to forecast the morbidity of tuberculosis in Xinjiang, China [39].
In this study, we establish the best single ARIMA model for prediction. In order to improve the accuracy of the single ARIMA model, we make an analysis of the residual of the model and we find that the residual sequence appears to exhibit heteroscedasticity. Heteroscedasticity is a critical aspect of data non-stationarity in time

series forecasting, it implies that different observations in time series have different variances. Heteroscedasticity can pose some problems, for example, in the ordinary least squares (OLS) estimate, the presence of heteroscedasticity gives a false sense of precision, and the standard errors and confidence intervals estimated by OLS will be too narrow although the regression coefficients of OLS are still unbiased [9]. Considering this reason, we further establish the autoregressive integrated moving average and autoregressive conditional heteroscedasticity (ARIMA-ARCH) combined model. The results show that our ARMA model was actually a good fit to our data, even though the ARIMA-ARCH also gives a good fit.

CHAPTER II
METHODS

2.1 Introduction

We obtained TB incidence data for Zimbabwe for 1990 to 2013 and this data will be analyzed using three time series analysis methods. These methods are ARIMA, ARIMA-ARCH and HW and the results obtained are discussed in the next chapter.

2.2 Data Source

The data of the TB incidence cases in Zimbabwe for 1990 to 2013 were obtained from World Bank Group website, last accessed on 9/26/2015. Incidence of tuberculosis is the estimated number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population. All forms of TB are included, including cases in people living with HIV. Estimates for all years are recalculated as new information becomes available and techniques are refined, so they may differ from those published previously [15].

2.3 Model Descriptions

2.3.1 ARIMA Model

The autoregressive integrated moving average (ARIMA) method is a widely used model, [12, 19, 22], etc. The ARIMA method is a reflection of the time dynamic dependency and can reveal the quantitative relationship between the research object and other objects with the development and change of time. For forecasting, the ARIMA method is more widely applied than other methods. It can take into account changing trends, periodic changes, and random disturbances in time series, and it is very useful in modeling the temporal dependence structure of a time series. The model can be written as:

$$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t \tag{2.1}$$

where Xt represents a non-stationary time series at time t, ϵ_t is a white noise (zero

mean and constant variance), d is the order of differencing, B is a backward shift operator defined by $BX_t = X_{t-1}$, $\phi(B)$, is the autoregressive operator defined as:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p, \tag{2.2}$$

$\theta(B)$ is the moving average operator defined as:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta q B^q. \tag{2.3}$$

The periodic repetition of performance norms is very common in time series analyses, a characteristic known as seasonality, and it is also a form of non-stationarity. In this case, two different components constitute the ARIMA model: a regular component, which constructs the predictions based on the previous delays in values and disturbances of the variable (with its regular auto regressive (p), moving average (q), and order of differencing (d) components), and a seasonal component, which constructs the predictions based on seasonal delays of values and disturbances of the variable (with its seasonal autoregressive(P), moving average (Q), and order of differencing (D) components). A seasonal ARIMA model with s observations per period, denoted by ARIMA (p, d, q) (P, D, Q)s is given by:

$$\Phi(B^s)\phi(B)(1-B)^d(1-B^s)^D X_t = \Theta(B^s)\theta(B)\epsilon_t \tag{2.4}$$

$$\Phi(B^s) = 1 - \phi_{s,1} B^s - \phi_{s,2} B^{2s} - ... - \phi_{s,Q} B^{Qs} \tag{2.5}$$

$$\Theta(B^s) = 1 - \theta_{s,1} B^s - \theta_{s,2} B^{2s} - ... - \theta_{s,Q} B^{Qs} \tag{2.6}$$

Generally, the standard statistical methodology to construct an ARIMA model includes four steps:

First step, to transform the non-stationary time series into stationary time series by differencing processes, d is the order of non-seasonal (regular) difference, D is the order of seasonal difference. Augmented Dickey-Fuller (ADF) test can determine whether the time series after differencing was stationary or not.

Second step, to plot the graphs of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the transformed series. According to ACF and PACF, we can determine the possible values of p, q, P and Q. This process requires both skill and experience. Generally, more than one tentative model is chosen in this step. Then, model identification and parameter estimation is carried out.

Third step, to verify the goodness of fit of the possible models by the diagnostic checking of residuals. Residuals must be equivalent to white noises (significant level $p > 0.05$) by using the Box-Jenkins Q test. Generally speaking, if the p value of Q-statistics is not bigger than 0.8, the tentative model is inadequate [23].

Fourth step, to select the best ARIMA model from possible models by the Akaike information criterion (AIC) and Schwarz criterion (SBC) [5]. The preferred model is the one with the lowest AIC and SBC values. These steps however, are cumbersome and can lead to the selection of wrong models if not carefully executed. As a result, the "'auto.arima"' function in R is utilized for the selection of a good fit ARIMA model. This is the option that was preferred for our model selection.

2.3.2 ARIMA-ARCH model

Autoregressive conditional heteroscedasticity (ARCH) models are the prevalent tools used to deal with time series heteroscedasticity [9]. The error term ϵ_t of the ARIMA is the random component and commonly assumed to be zero mean and constant variance. However, for some practical time series, the error term ϵ_t does not satisfy the homoscedastic assumption of constant variance. The time varying variance (i.e., volatility or heteroscedasticity) depends on the observations of the immediate past and is called the conditional variance. In this case, the Histogram-Normality test of the error term ϵ_t has a heavier-tailed distribution [4], as well as, the autoregressive conditional heteroscedasticity Lagrange multiplier (ARCH LM) test of the error term ϵ_t shows $p < 0.05$. ARCH model is introduced to accommodate the possibility of serial correlation in volatility. Models for volatility forecasting were first developed by Engle (1982) [9], these models known as ARCH models were developed to capture

the non-constant variance. Therefore, when the error term ϵ_t of the ARIMA has ARCH effect, we can consider a combined model, which may have higher accuracy.

The ARIMA-ARCH model is one model, in which the variance of the error term of the ARIMA model follows an ARCH process, the model can be written as [21]:

$$\Phi(B^s)\phi(B)(1-B)^d(1-B^s)^D X_t = \Theta(B^s)\theta(B)\epsilon_t, \tag{2.7}$$

$$\epsilon_t = \sqrt{v_t} z_t, \tag{2.8}$$

$$v_t = c_0 + \eta_1 \epsilon_{t-1}^2 + \eta_2 \epsilon_{t-2}^2 + ... + \eta_l \epsilon_{t-l}^2 \tag{2.9}$$

where the error term ϵ_t is said to follow an ARCH process of orders l , [9], z_t is a white noise sequence with mean 0 and variance 1. Assume that v_t is conditioned on the l previous errors, c_0 and η_i are constant coefficients.

2.3.3 Holt-Winter algorithm (HW)

Holt-Winters (HW) refers to a set of procedures that form the core of the exponential smoothing family of forecasting methods. The basic structures were provided by C.C. Holt in 1957 and Peter Winters in 1960. The HW algorithm uses a set of simple recursions that generalize the exponential smoothing recursions to generate forecasts of series containing a locally linear trend. Holt (1957) extended simple exponential smoothing to allow forecasting of data with a trend. This method involves a forecast equation and two smoothing equations (one for the level and one for the trend): Forecast equation

$$\widehat{y}_{t+h|t} = l_t + hb_t \tag{2.10}$$

Level equation

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.11}$$

Trend equation

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \tag{2.12}$$

where l_t denotes an estimate of the level of the series at time t, b_t denotes an estimate of the trend (slope) of the series at time t, α is the smoothing parameter for the level, $0 \leq \alpha \leq 1$ and β is the smoothing parameter for the trend, $0 \leq \beta \leq 1$. The level equation shows that l_t is a weighted average of observation y_t and the within-sample one-step-ahead forecast for time t, here given by $l_{t-1}+b_{t-1}$. The trend equation shows that b_t is a weighted average of the estimated trend at time t based on $l_t - l_{t-1}$ and b_{t-1}, the previous estimate of the trend.

The forecast function is no longer flat but trending. The h-step-ahead forecast is equal to the last estimated level plus h times the last estimated trend value. Hence the forecasts are a linear function of h.

The error correction form of the level and the trend equations show the adjustments in terms of the within-sample one-step forecast errors:

$$l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t \tag{2.13}$$

$$b_t = b_{t-1} + \alpha\beta\epsilon_t \tag{2.14}$$

where,

$$\epsilon_t = y_t - (l_{t-1} + b_{t-1}) = y_t - \widehat{y}_{t|t-1} \tag{2.15}$$

2.3.4 Assessing Forecast Accuracy

Three performance measures were employed in determining prediction efficiency between single ARIMA model, ARIMA-ARCH model, and HW, namely root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). These measures have been used by many researchers to compare the accuracy of their models with other known models [18, 11, 16].

The first performance measure is root mean square error (RMSE), which is used to compare the predicted value with actual value. The RMSE is computed as:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(X_t - \widehat{X_t})^2}{n}} \tag{2.16}$$

The second performance measure is mean absolute error (MAE). The MAE is defined as:

$$MAE = \frac{\sum_{t=1}^{n}|X_t - \widehat{X_t}|}{n} \tag{2.17}$$

And then, the third performance measure is mean absolute percentage error (MAPE), a measure of relative overall fitness. This performance measure is defined as:

$$MAPE = \frac{\sum_{t=1}^{n}\frac{|X_t - \widehat{X_t}|}{X_t} \times 100}{n} \tag{2.18}$$

where $\widehat{X_t}$ is the predicted value, X_t is the actual value and n is the number of observations.

CHAPTER III
RESULTS

3.1 Introduction

Three methods, ARIMA, ARIMA-ARCH and Holt Winters were used for the data analysis. The results of the analysis indicate that the ARIMA(2,2,1) fits our data better, followed by the ARIMA -ARCH and lastly the HW. Even though in most cases the ARIMA-ARCH is suppossed to be an improvement of the ARIMA and thus should give better results, this was not the case for our data, showing that the ARIMA was reasonably adequate and can be reliably used for short-term forecasts.

3.2 ARIMA

The original time series for the TB incidence includes the incidence for the years 1990-2013. An initial plot of the data did not indicate any seaonalitity, but an increasing, then decreasing trend as seen in the figure below.

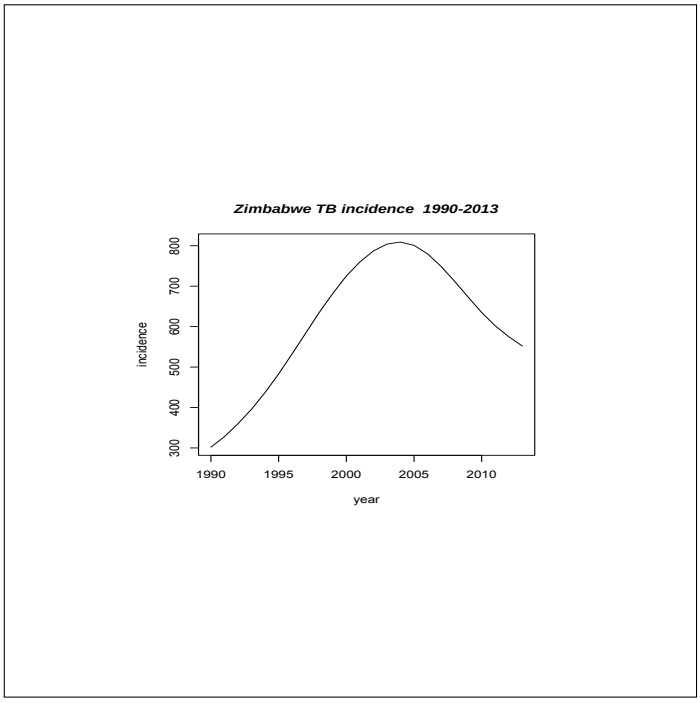

Figure 3.1: Zimbabwe TB Incidence 1990-2013.

A preliminary analysis of the data was carried out to determine the ARIMA model for the data. The ACF and PACF plots were produced. We note that the correlation plot, (ACF) cuts off at lag 3 and the partial correlation (PACF) plot at lag 1. See figures 3.2 and 3.3, respectively. In most cases, we would attempt to use these to come up with the appropriate model for our data. In our case however, the "'auto.arima"' function was utilized and this yielded an ARIMA(2,2,1) as the best model. This model is a good fit for the data as shown by the diagnostic plots of the residuals. The ACF plot cuts of at lag zero and the Box-Ljung plot shows that the residuals are uncorrelated, see figure 3.4. This is in agreement with the Box-Ljung test which yielded a p-value of 0.4528. Also the ACF and PACF of residuals also show no significant lags beyond zero, an indication that our data fits the ARIMA model reasonably well.

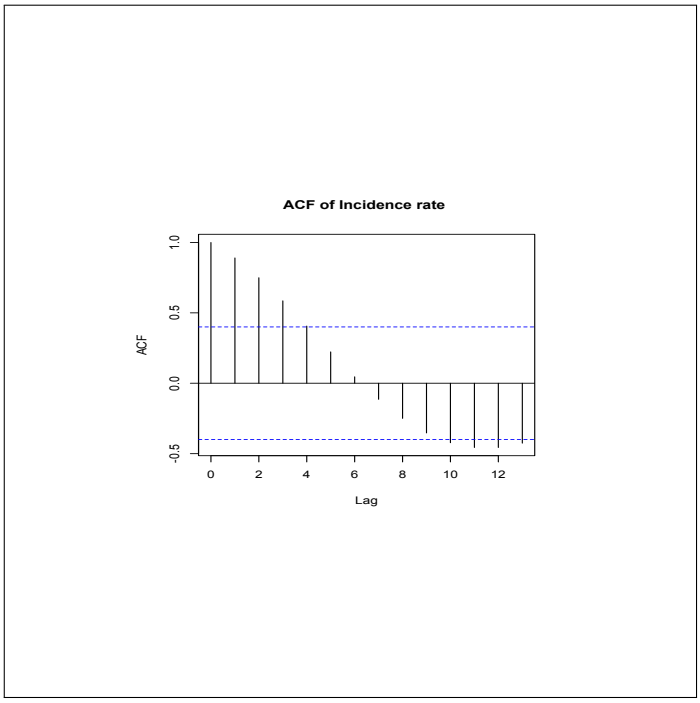

Figure 3.2: ACF Plot.

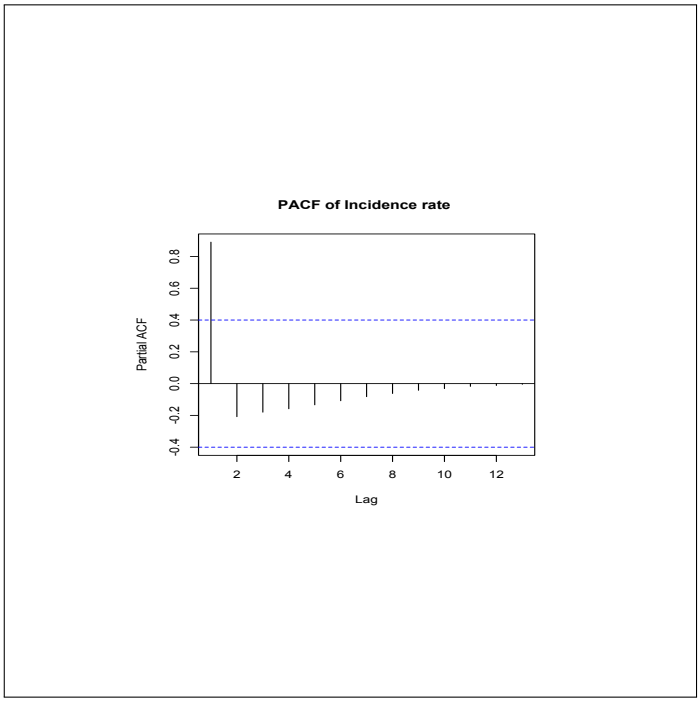

Figure 3.3: PACF Plot.

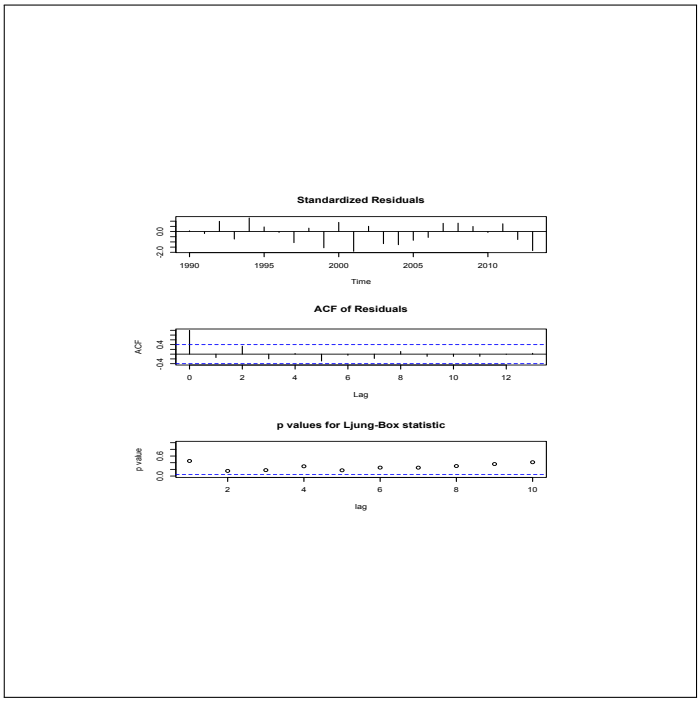

Figure 3.4: Residuals Diagnostic Plot.

Using the "'forecast"' function in R we obtain a forecast of the incidence of the next

ten years, 2014-2023, as well as 80%and 95% prediction intervals for those predictions, see table III.1. The incidence for 2013 was 552 per 100,000 thousand people (the last observed value in our time series), and the ARIMA model gives the forecasted incident rate for the following year as approximately 533 per 100,000 people. According to the forecasts the incidence continues to show a gradual decline, reaching a rate of 430 per 100,000 by the year 2023. However, the prediction intervals become wider as the years increase, an indication that the method is best for short term forecasts. Plots of the observed incidence, as well as the predicted incidences using our model are also shown in figure 3.5

Table III.1: ARIMA Model Forecasts.

Year	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	532.88	530.63	535.14	529.43	536.33
2015	517.25	509.67	524.83	505.66	528.84
2016	504.48	487.55	521.41	478.60	530.37
2017	493.82	462.87	524.76	446.49	541.15
2018	484.41	434.39	534.44	407.91	560.92
2019	475.44	401.17	549.72	361.85	589.03
2020	466.13	362.59	569.66	307.78	624.47
2021	455.81	318.36	593.27	245.60	666.03
2022	443.99	268.50	619.50	175.60	712.40
2023	430.35	213.33	647.37	98.45	762.25

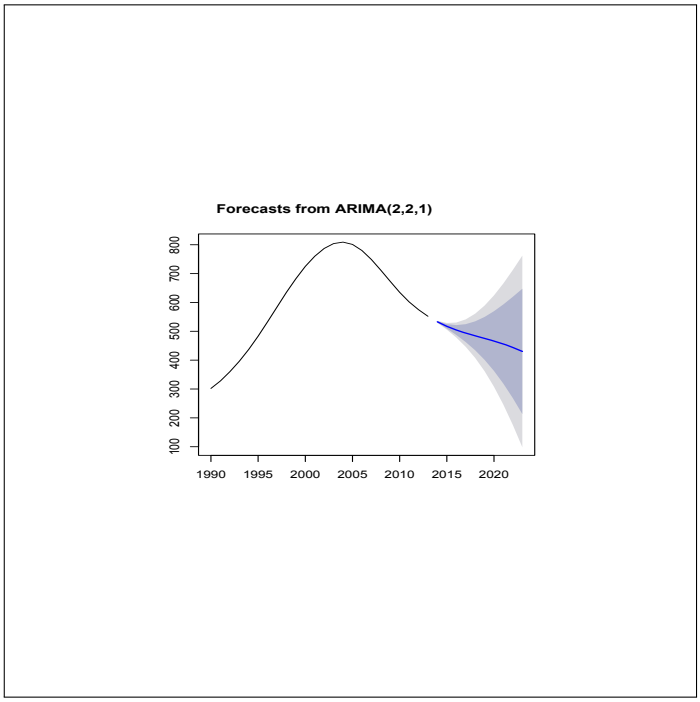

Figure 3.5: ARIMA(2,2,1) Forecasts Plot.

We already noted that the correlogram shows that none of the sample autocorrelations exceed the significance bounds, and the p-value for the Ljung-Box test was 0.4528, we can conclude that there is very little evidence for non-zero autocorrelations in the forecast errors .

To investigate whether the forecast errors are normally distributed with mean zero and constant variance, we can make a time plot and histogram (with overlaid normal curve) of the forecast errors. The time plot of the in-sample forecast errors shows that the variance of the forecast errors seems to be roughly constant over time. The histogram of the time series shows that the forecast errors are roughly normally distributed and the mean seems to be close to zero. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance.

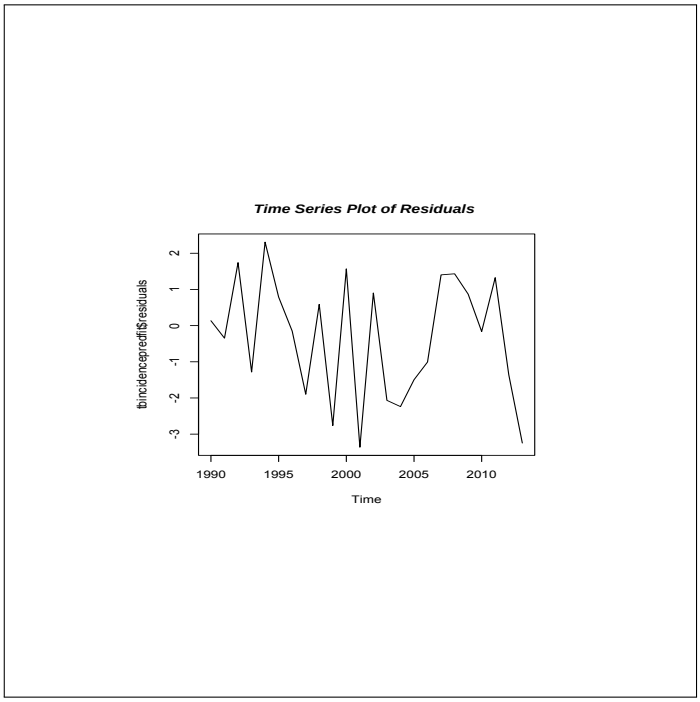

Figure 3.6: Time Series Plot of Residuals.

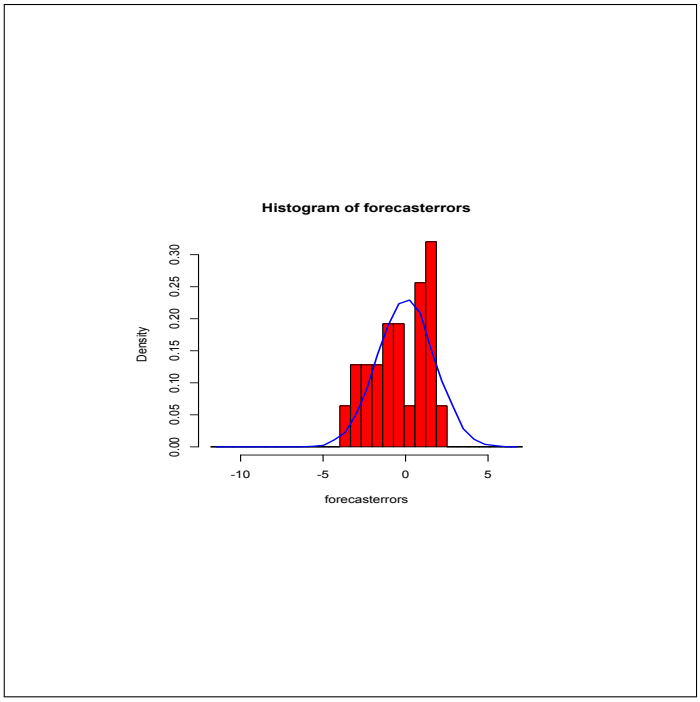

Figure 3.7: Histogram of Forecast Errors.

Since successive forecast errors do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the ARIMA(2,2,1) does seem to provide an adequate predictive model for the TB incidence for Zimbabwe for the years 1990-2013 in the short term.

3.3 ARIMA-ARCH

In an effort to improve the precision of ARIMA (2, 2, 1) model, we carry out further residual analysis. The Box-Jenkins test suggests that autocorrelation function of residual series with different lags differ from zero at some lags. A plot of the squared residuals hints at some correlation 3.8. After that, we do Histogram-Normality test,see 3.9, the result shows that heavier-tailed distribution of residual series exists. The ARCH LM test at around lag 2, though greater than 0.05 suggests a possible ARCH effect of residual series exists. The ARCH effect does not exist when lag is greater than 2. Therefore, we consider establishing ARIMA (2,2,1)-ARCH (1) model in an effort to improve the precision of prediction. We proceeded to fit this model to our data.

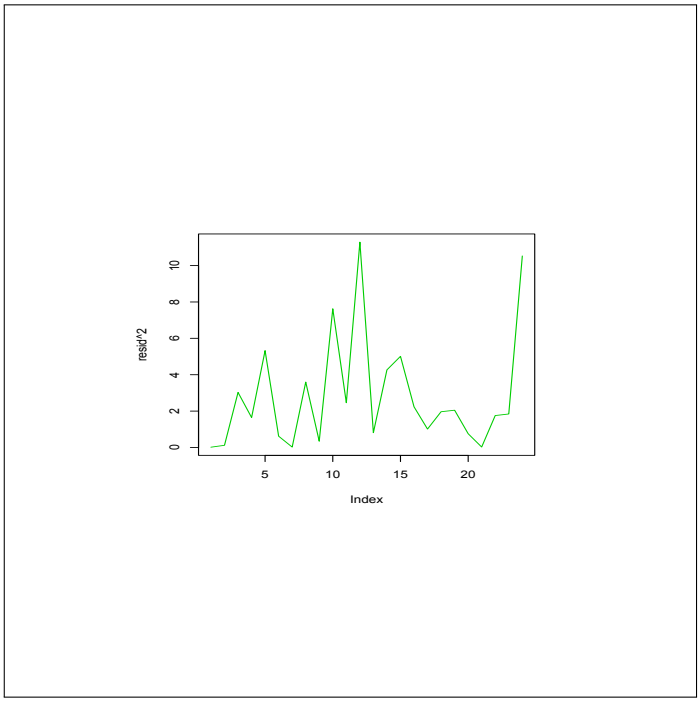

Figure 3.8: Residuals Plot.

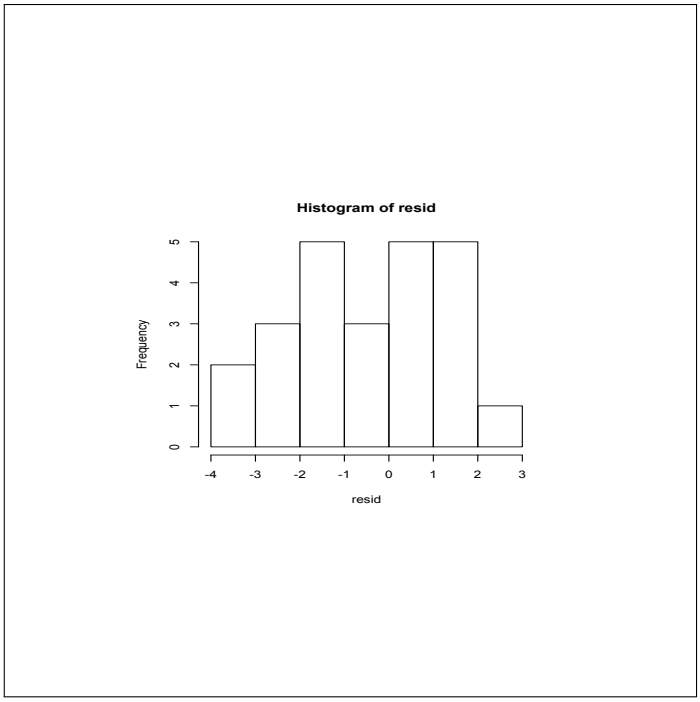

Figure 3.9: Histogram of Residuals.

The diagnostic tests of the residuals indicate normality. A histogram of the forecast errors, indicates that the errors are relatively normally distributed, as is also indicated by the box plot, see figure 3.10. The QQ-plot indicates that most of the values lie on or close to the line, with just a few values away from the line, see figure 3.11. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance.

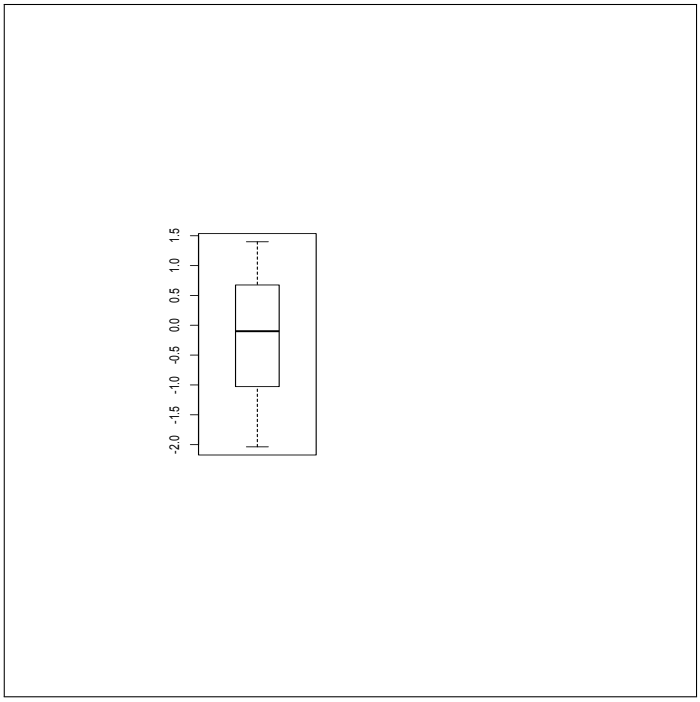

Figure 3.10: Box Plot of Residuals.

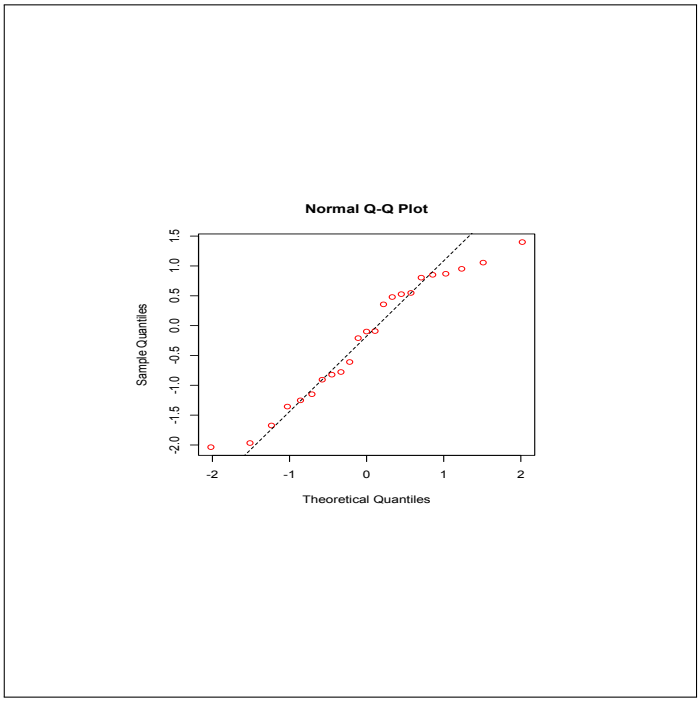

Figure 3.11: Normal Plot of Residuals.

Plots of the observed incidence, as well as the prediction intervals from our model indicate that the values lie within the intervals, an indication that the model provides a reasonably good fit, see figure 3.12

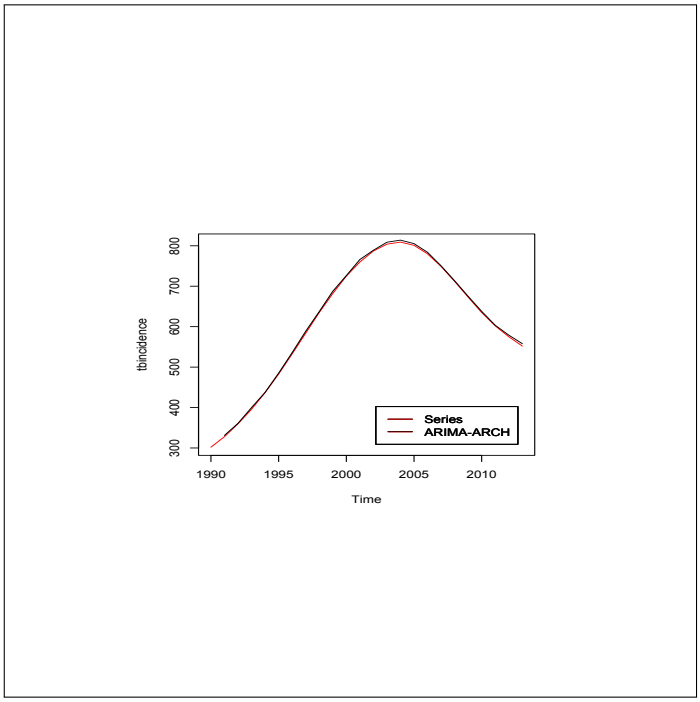

Figure 3.12: Forecasts Plot ARIMA-ARCH.

The results of the tests for normality for the forecast errors do not seem to point to any correlation, suggesting that the forecast errors seem to be normally distributed with mean zero and constant variance, the ARIMA(2,2,1) -ARCH(1) model does seem to provide an adequate predictive model in the short term for the TB incidence for Zimbabwe for the years 1990-2013.

3.4 Holt Winters

We already observed that our time series as shown in figure 3.1, indicated there was an increase in incidence from about 300 in 1990 to about 810 in 2004, and that afterwards the incidence decreased to about 550 in 2013. So our series can be described using an additive model with increasing or decreasing trend and no seasonality, and for this purpose we use Holts exponential smoothing to make short-term forecasts. We fit a predictive model using the HoltWinters() function in R,setting the parameter gamma to FALSE.

Holt Winters exponential smoothing estimates the level and slope at the current

time point. Smoothing is controlled by two parameters, alpha, for the estimate of the level at the current time point, and beta for the estimate of the slope b of the trend component at the current time point. As with simple exponential smoothing, the paramters alpha and beta have values between 0 and 1, and values that are close to 0 mean that little weight is placed on the most recent observations when making forecasts of future values.

The value of alpha obtained for our forecast was 0.96 and that of beta was estimated to be 1.00. Both values are high, and tell us that both the estimate of the current value of the level, and of the slope b of the trend component, are based mostly upon very recent observations in the time series. This makes good intuitive sense, since the level and the slope of the time series both change quite a lot over time. The value of the sum-of-squared-errors for the in-sample forecast errors is 2510.84. We can see from the figure 3.13 that the in-sample forecasts agree pretty well with the observed values, although at times they tend to lag behind the observed values a little bit.

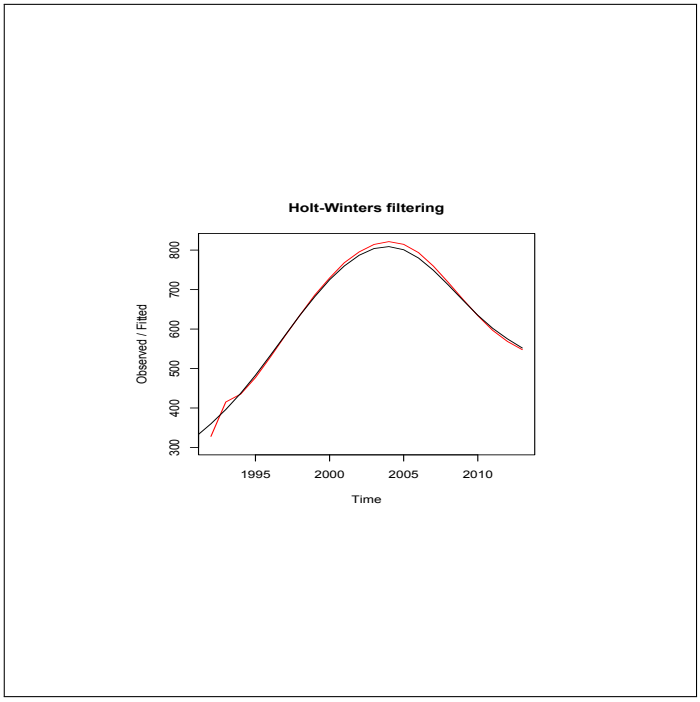

Figure 3.13: Holt Winters Forecasts Plot.

We make forecasts for future times not covered by the original time series by using the forecast function in the forecast package in R. For example, our time series data for TB incidence was for 1990-2013, so we make predictions for 2014 to 2023 (10 more data points). The results are shown in figure 3.14 below. We can also obtain a forecast of the incidence of the next ten years, 2014-2023, as well as 80% and 95% prediction intervals for those predictions, see table III.2, below. The incidence for 2013 was 552 per 100,000 thousand people (the last observed value in our time series), and the Holt Winters model gives the forecasted incident rate for the following year as approximately 529 per 100,000 people. According to the forecasts the incidence continues to show a gradual decline, reaching a rate of 322 per 100,000 by the year 2023. However, the prediction intervals become wider as the years increase, an indication that the method is best for short term forecasts. Plots of the observed incidence, as well as the predicted incidences using our model are also shown in figure 3.14

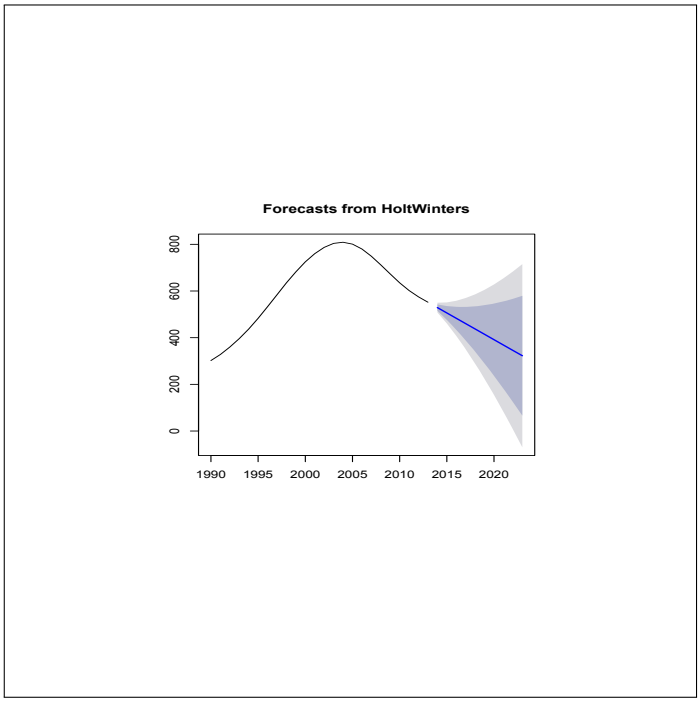

Figure 3.14: Holt Winters Predicted Incidence Plot.

Table III.2: Holt Winters Predicted Incidence Table.

Year	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2014	528.9014	515.22399	542.5788	507.98361	549.8192
2015	505.9918	476.47075	535.5128	460.84325	551.1403
2016	483.0822	433.97579	532.1886	407.98043	558.1840
2017	460.1726	388.41522	531.9300	350.42915	569.9161
2018	437.2630	340.17154	534.3545	288.77442	585.7516
2019	414.3534	289.50735	539.1995	223.41783	605.2890
2020	391.4438	236.61909	546.2685	154.65983	628.2278
2021	368.5342	181.66146	555.4070	82.73699	654.3314
2022	345.6246	124.76052	566.4887	7.84213	683.4071
2023	322.7150	66.02168	579.4084	-69.86356	715.2936

The figure 3.15 shows that the forecast errors have roughly constant variance over time. Thus, the Ljung-Box test, p value of 0.6699 shows that there is little evidence of autocorrelations in the forecast errors, while the histogram of forecast errors show that it is plausible that the forecast errors are normally distributed with mean zero and constant variance. Therefore, we can conclude that Holts exponential smoothing provides an adequate predictive model for TB incidence in Zimbabwe. In addition, it means that the assumptions that the 80% and 95% predictions intervals were based upon are probably valid.

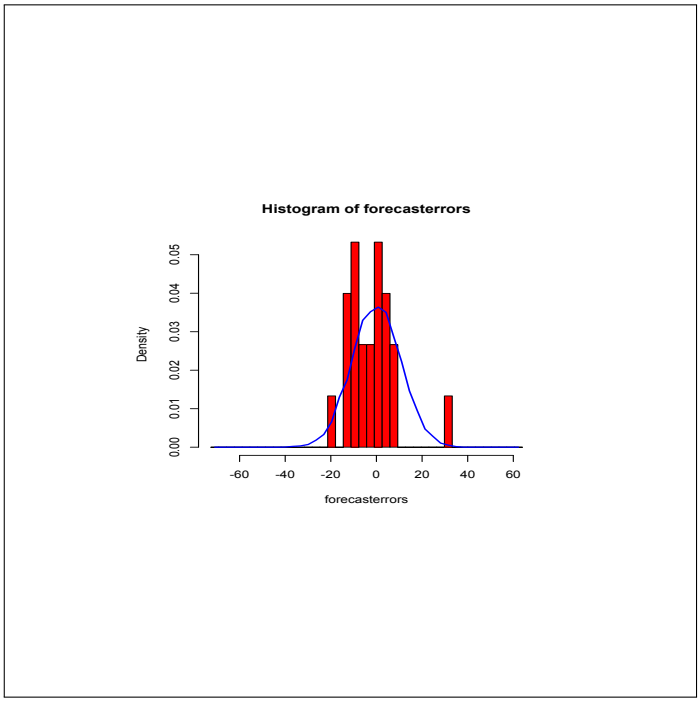

Figure 3.15: Holt Winters Histogram of Residuals.

3.5 Model Comparison

We compare the perfomance of the three methods. The figure 3.16, shows the actual incidence of TB and fitted incidence of Holt Winters, ARIMA and ARIMA-ARCH models. We also calculated the MAE, MAPE, and RMSE for each model. We note that the RMSE, MAE and MAPE, see table III.3, of the ARIMA model are smaller, followed by the ARIMA-ARCH, and finally the Holt Winters model. This indicates that the ARIMA(2,2,1) model selected for our data was adequate on its own. However, the values for the combined model also make it a likely model for our data. A plot of the observed TB incidence and confidence intervals for the ARIMA and ARIMA-ARCH indicates that both models agree at most points, see figure 3.17.

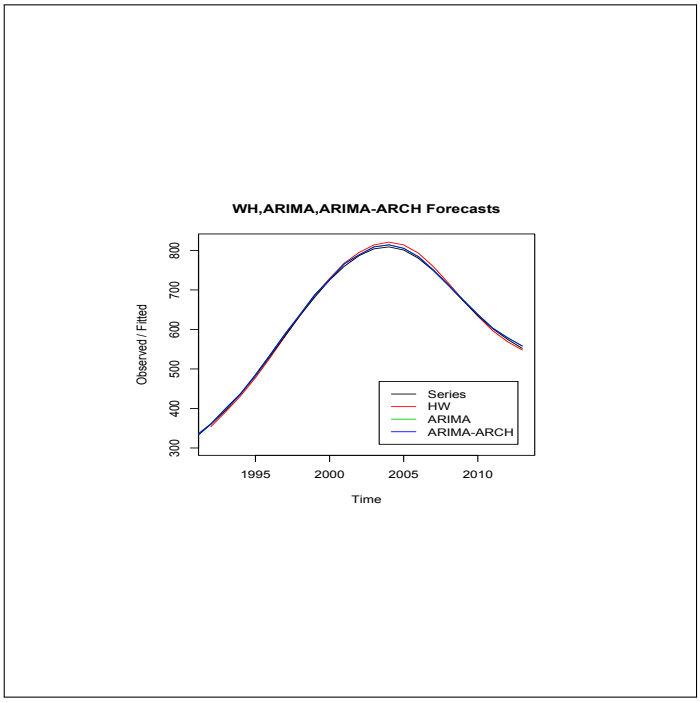

Figure 3.16: HW,ARIMA,ARIMA-ARCH Fitted.

Table III.3: Model Performance Comparison.

Method	RMSE	MAE	MAPE
ARIMA	2.8485	1.4346	0.2379
ARIMA-ARCH	11.86016	2.9600	0.4873
Holt Winters	104.6184	7.3406	1.3193

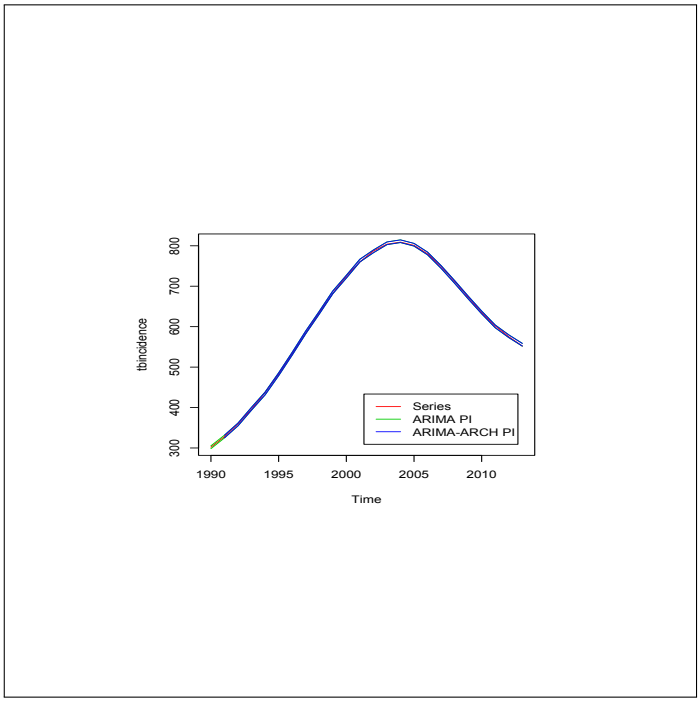

Figure 3.17: ARIMA,ARIMA-ARCH Prediction Intervals.

CHAPTER IV
DISCUSSION AND CONCLUSION

4.1 Discussion

TB remains a major global public health problem and even though a lot of progress has been made, it still is, an often fatal infectious disease in the world. In Zimbabwe, it still remains a large burden, with Zimbabwe considered as one of the world's 22 high burden countries by WHO. The increase of TB patients with HIV co-infection as well as the emergence of drug-resistant strains, has further lead to an increased difficulty in prevention and control of TB. Therefore, it is highly cost effective to detect a TB epidemic in its early stages in order to optimize disease control and intervention. Early warning based on forecasts is very important for improving vector control, community intervention and personal protection. This study aims to develop an appropriate model for predicting TB incidence in Zimbabwe.

Time series analysis of surveillance data on incidence of various infections is very helpful in developing hypotheses which can explain and anticipate the dynamics of the observed phenomena so as to establish a quality control system. ARIMA model is one of the most widely used time series forecasting techniques because of its structured modeling basis and acceptable forecasting performance [32].

ARIMA models have been successfully applied to predict the incidence of infectious disease. In this study, the yearly incidence data of TB for 1990-2003 in Zimbabwe was used. First, we use the ARIMA method to establish the single ARIMA (2,2,1) model for predicting the yearly incidence of TB in Zimbawe. After that, we analyze the residual series based on the ARIMA (2,2,1) model, and the results indicate that a possible ARCH effect exists. ARCH models are the prevalent tools used to deal with time series heteroscedasticity. In order to remove the heteroscedasticity and improve prediction accuracy, we establish the ARIMA (2,2,1)-ARCH (1) model. We develop an ARIMA model and an ARIMA-ARCH combined model to predict yearly incidence of TB in Zimbabwe. We also develop a HW model to forecast our data. When we

test the performances of the models based on our data using three indexes, namely, RMSE, MAE and MAPE. The smaller the values of these indexes , the higher the precision of the model. From table III.3, we note that the ARIMA model seems to provide the best fit to our data. We can thus conclude that even though the combined model provides a good fit, our data did not have that significant volatility to warrant that ARCH effect be taken into account. We believe that the model combining ARIMA and ARCH effect did not necessarily contain more data characteristics than the single ARIMA (2,2,1) model to render it better for forecasting yearly incidence of TB in Zimbabwe. The ARIMA model is generally used for short-term forecasts. Since the incidence of TB is not stationary, new observations series should be added continually into the sequence over time to ensure the model continues to provide a best fit. In this case it is also essential to continually evaluate the combined model as well as the behavior of the data can change as new data is continually added. If the actual data falls outside the confidence level of the forecast value, there is need that the model be updated immediately.

There are some limitations in our analyses. Our data source regularly gets updated as data becomes available, so the data used in this analysis may change over time and so with these limitations, the results and forecasts presented need to be interpreted with caution. It might also be important to consider other mechanisms that shape TB trends, such as other associated diseases e.g. HIV, that have proven critical to TB trends variability. Thus, those factors may be incorporated into the TB trends models to improve the predictability but they could have important dynamic consequences, which are worth exploring in future research. Other models such as the generalized regression models could be considered for analysis in order to provide clear recommendations for control strategies.

4.2 Conclusion

TB is a serious public health issue in Zimbabwe and continuos monitoring of this epidemic is essential for its control and intervention, which can reduce the substantial

morbidity and mortality caused by this disease. ARIMA models are an important tool for infectious disease surveillance. ARCH models are the prevalent tools used to deal with time series heteroscedasticity. We established the ARIMA (2,2,1)and ARIMA (2,2,1)-ARCH (1) models, which can be utilized to forecast the incidence of TB in Zimbabwe. Comparative analyses show that the ARIMA model has better performance, but the combined model also performed very well. Therefore, this study suggests that these models be utilized in an effort to optimize TB prevention by providing estimates on TB incidence trends in Zimbabwe. We found that our forecasts suggest that Zimbabwe will continue to be faced with high TB incidence in the coming years and so it is essential that a close watch be kept towards monitoring this disease. From our results, we anticipate that our analysis can be extended to the context of other high burden countries, where M. tuberculosis remains a substantial cause of expected morbidity and mortality.

BIBLIOGRAPHY

[1] Sapii, N., Dir, S., Mardiah, T., Abdullah, S. Aplication of univariate forecasting models of tuberculosis cases in kelantan. *ICSSBE*, 2012.

[2] Mohammad, H.G., Akhtar, S., Seasonality in pulmonary tuberculosis among migrant workers entering kuwait. *BMC Infect Dis.*

[3] Romanyukha, A.A., Avilov, K.K. Mathematical modeling of tuberculosis propa-gation and patient detection. *Automation and Remote Control*, 68(9):1604–2007. 1617,

[4] Bollerlev, T. Generalized autoregressive conditional heteroskedasticity. *Econometrics.*

[5] Hjort, N.L., Claeskens, G. *Model selection and model averaging.* Cambridge University Press, Cambridge, first edition, 2008.

[6] Bandason, T., Duong, T., Corbett, E.L., et al. Comparison of two active case-finding strategies for community-based diagnosis of symptomatic smear-positive tuberculosis and control of infectious tuberculosis in harare, zimbabwe (detectb): a cluster-randomised trial. *Lancet.*

[7] Bielefeld, R.A., Cauthen, G.M.,Daniel, T.M., Rowland, D.Y., Debanne, S.M. Multivariate markovian modeling of tuberculosis: forecast for the united states. *Emerg Infect Dis*, 6(2):148–157, 2000.

[8] Dye, C. Global epidemiology of tuberculosis. *The Lancet*, 367(9514):938–940, 2006.

[9] Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation. *Econometrica.*

[10] Fares, A. Seasonality of tuberculosis. *J Glob Infect Dis.*

[11] Faruk, D.O. A hybrid neural network and arima model for water quality time series prediction. *Eng Appl Artif Intell.*

[12] Quenel, P., Gustave, J., Cassadou, S., La Ruche, G., Girdary, L., Marrama, L., Gharbi, M. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC Infect Dis.*

[13] Floyd, K., Raviglione, M., Glaziou, P. Global burden and epidemiology of tuberculosis. *Clin. Chest Med.*, 30:621636, 2009.

[14] Sismanidis ,C., Floyd, K., Raviglione, M., Glaziou, P. Global epidemiology of tuberculosis. *Cold Spring Harb. Perspect Med.*, 5, 2014.

[15] The World Bank Group. Incidence of tuberculosis. 2015.

[16] Wang, H., Liu, Q., Yang, J., Guo, Z. A feature fusion based forecasting model for financial time series. *PLoS One*.

[17] Roslanb, U., Kilicmana, A. Tuberculosis in the terengganu region. *Forecast and data analysis. Scie Asia*.

[18] Tong, L.I., Lee, Y.S. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowl-Based Syst*.

[19] Han, Z.Y., Zhang, Y.B., Qi, S.X., Xu, Y.G., Wei, Y.M., Han, X., Liu, Y.Y., Li, Q. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. *Am J Trop Med Hyg*.

[20] Hsieh, N.H., Huang, T.L., Cheng, Y.H., Lin, Y.J., Chio, C.P., Liao, C.M., et al. Assessing trends and predictors of tuberculosis in taiwan. *BMC Public Health*.

[21] Erdem, E., Shi, J., Liu, H.P. Comprehensive evaluation of armagarch (-m) approaches for modeling the mean and volatility of wind speed. *Applied Energy*.

[22] Liu, X., Jiang, B., Yang, W., Liu, Q. Forecasting incidence of hemorrhagic fever with renal syndrome in china using arima model. *BMC Infect Dis.*

[23] Box, G.E.P., Ljung, G.M. On a measure of lack of fit in time series models. *Biometrika*, 65:297–303, 1978.

[24] Sanchez-Padilla, E., Simon-Soria, F., Eiros, J.M., Golub, J.E., Luqureo, F.J. Trend and seasonality of tuberculosis in spain, 1996-2004. *Int J Tuberc Lung Dis*.

[25] Makumbirofa, S., Makamure, B., Sandy, C., Bara, W., Mungofa, S., Hopewell, P.C., Mason, P., Metcalfe, J.Z. Drug-resistant tuberculosis in high-risk groups, Zimbabwe. *Emerg Infect Dis*.

[26] Khanjani, N., Moosazadeh, M. The existing problems in the tuberculosis control program of iran: A qualitative study. *J Qual Res Health Scie*, 1(3):189–201, 2012.

[27] Nasehi, M., Bahrampour, A., Khanjani, N., Sharafi, S., Ahmadi, S., Moosazadeh, M. Forecasting tuberculosis incidence in iran using box-jenkins models. *Iran Red Crescent Med J*, 16(5):e11779, 2014.

[28] Mirhaghghani, L., Nasehi, M., Guidelines for combat with TB, disease management center of health ministry. *Arjemand publication*, pages 6–44, 2010.

[29] Moosazadeh, M., Amiresmaeili, M., Parsaee, M., Nezammahalleh, A., Nasehi, M. The epidemiology of factors associated with screening and treatment outcomes of patients with smear positive pulmonary tuberculosis: A population-based study. *J Mazandaran Univ Med Sci.*, 21(1):9–18, 2012.

[30] Parzen, E., Newton, H. J. *Forecasting and time series model types of economic time series*. Wiley, major time series methods and their relative accuracy 1984. edition,

[31] World Health Organization. Who,global tuberculosis report 2014. 2014.

[32] Simonsen, L., Sharma, A., Pardo, S.A., Fedson, D.S., Miller, M.A., Reichert, T.A. Influenza and the winter increase in mortality in the united states, 1959-1999. *Am J Epidemiol.*

[33] Cain, K.P., Oeltmann, J.E., Kammerer, J.S., Moonan, P.K., Ricks, P.M. Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the u.s., 20052009. *PLoS One*, 6:e27405, 2011.

[34] Rieder, H. Annual risk of infection with mycobacterium tuberculosis. *Eur Respir J*, 25(1):181–185, 2005.

[35] Valenzuela, O., Rojas, F., Guillen, A., Rojas, I., Herrera L.J. et al. Soft-computing techniques and arma model for time series prediction. *Neurocomputing.*

[36] Winston, C.A., Heilig, C.M., Cain, K.P., Walter, N.D., MacKenzie , W.R., Willis, M.D. Seasonality of tuberculosis in the united states, 1993-2008. *Clin Infect Dis.*

[37] Guan, P., Guo, J.Q., Zhou, B.S., Wu, W. Comparison of gm (1, 1) gray model and arima model in forecasting the incidence of hemorrhagic fever with renal syndrome. *J China Med Univ.*

[38] Tang, G., Wang, W., Zhang, Y. Application of arima model in forecasting incidence of tuberculosis. *Modern Prevent Med.*

[39] Zhang, L.P., Zhang, X.L., Wang, K., Zheng, Y.J., Zheng, Y.L. Forecast model analysis for the morbidity of tuberculosis in Xinjiang,China. *PLoS ONE*, 10(3), 2015