TUBERCULOSIS (TB) DATA FOR HIGH BURDEN CITYS IN TAMILNADU FOR USING FORECASTING ANALYSIS

S.Poyyamozhi[1] and A.Kachi Mohideen[2]

[1]Assistant Professor and Head, Department of Statistics,
Government Arts College (Autonomous), Kumbakonam – 612 002.

[2]Assistant Professor, Department of Statistics,
Periyar EVR College (Autonomous), Trichy – 620 023.

Abstract

The Box-Jenkins approach, specifically the autoregressive integrated moving average (ARIMA) model, based on the data of the tuberculosis incidence from 2005 -2017 in Tamilnadu, 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 Tamilnadu. Comparative analyses show that the ARIMA and ARIMA-ARCH models perform reasonably well, with the ARIMA model being the best in our case. Tuberculosis (TB) remains a major global public health problem, especially for considered as high burden cities in Tamilnadu. It 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. To the best of our knowledge, this is the First study to establish the ARIMA model and ARIMA-ARCH model for pre-diction and monitoring the yearly incidence of tuberculosis (TB) in Tamilnadu. 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 Tamilnadu.

Key words: Tuberculosis, Box-Jenkins approach, the autoregressive integrated moving average (ARIMA) model, Heteroscedasticity (ARCH) model, Holt Winters (HW) method.

1. Introduction

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-stationary 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. 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.

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 is refer to Cain et.al (2011). TB infection, if not timely treated, can be a serious health threat. 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 is discussed in Floyd et.al (2009). An estimated 9.0 million people developed TB in 2017; 1.5 million died from the disease according to the World Health Organization (2017). 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.

Tuberculosis is a significant public health problem in Tamilnadu with high morbidity and mortality rates. According to WHO Global Tuberculosis Report 2017, the estimated TB incidence in 2016 was 552 cases per 100,000 populations. In the same year, the estimated prevalence of TB (all forms) in Tamilnadu was 409 cases per 100, 000 populations. 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 2016. Tamilnadu has among the highest estimated TB incidence per capita (603/100,000 population) in the world.

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. 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. To predict tuberculosis incidence and to study its temporal changes, different mathematical and statistical models have been used in different studies, and based on data nature and evaluation, a certain model has been used in every study. For example, Sapii *et.al.,* applied a univariate time series model to the TB incidence data in Malaysia in order to determine the best forecasting model (2012) showed that Holts trend corrected exponential smoothing is the best forecasting model, followed by the quadratic trend model. Zhang *et.al.,* used the ARIMA model in order to forecast tuberculosis. The ARIMA-ARCH model was applied by Zheng *et.al.,* to forecast the morbidity of tuberculosis in Xinjiang, China (2015).

2. Model Descriptions

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,

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

Where X_t 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 $B\,X_t = X_{t-1}$, $\phi\,(B)$, is the autoregressive operator defined as,

$$\phi(B) = 1 - \phi_1\,(B)\,\phi_2\,(B)^2 - \cdots - \phi_p\,(B)^p \qquad \cdots (2.2)$$

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

$$\theta(B) = 1 - \theta_1(B) - \theta_2(B)^2 - \cdots - \theta_q(B)^q \qquad \cdots (2.3)$$

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-stationary. 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,

$$\Phi(B^s)\,\phi(B)(1-B)^d\,(1-B^s)^D X_t = \Theta(B^s)\,\theta(B)\epsilon_t \quad \cdots (2.4)$$

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

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

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) of the transformed series. According to ACF, 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 t of the possible models by the diagnostic checking of residuals. Residuals must be equivalent to white noises (significant level p > .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.

Fourth step, to select the best ARIMA model from possible models by the Akaike information criterion (AIC) and Schwarz criterion (SBC). The preferred model is the one with the lowest IC 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 “autoregressive Arima” function in R is utilized for the election of a good fit ARIMA model. This is the option that was preferred for our model selection.

2.1. ARIMA-ARCH model

Autoregressive conditional Heteroscedasticity (ARCH) models are the prevalent tools used to deal with time series Heteroscedasticity. 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, 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), 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],

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

$$\epsilon_t = \sqrt{\upsilon_t z_t} \quad \cdots (2.8)$$

$$\upsilon_t = c_0 + \eta_1\epsilon_{t-1}^2 + \eta_2\epsilon_{t-2}^2 + \cdots + \eta_l\epsilon_{t-l}^2 \quad \cdots (2.9)$$

Where, the error term ϵ_t is said to follow an ARCH process of orders l, z_t is a white noise sequence with mean 0 and variance 1. Assume that υ_t is conditioned on the l previous errors, c_0 and η_i are constant co-efficient.

2.2. 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

$$\hat{y}_{t+h/t} = l_t + hb_t \quad \cdots (2.10)$$

Level equation

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

Trend equation

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

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 at 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,

$$l_t = l_{t-1} + b_{t-1} + \alpha\epsilon_t \quad \cdots (2.13)$$

$$b_t = b_{t-1} + \alpha\beta\epsilon_t \quad \cdots (2.14)$$

Where,

$$\epsilon_t = y_t - (l_{t-1} + b_{t-1}) = y_t - \hat{y}_{t\backslash t-1} \quad \cdots (2.15)$$

2.3. 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. 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,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(X_t - \hat{X}_t\right)^2}{n}}$$

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

$$MAE = \frac{\sum_{t=1}^{n}\left|X_t - \hat{X}_t\right|}{n}$$

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

$$MAPE = \frac{\frac{\sum_{t=1}^{n}|X_t - \hat{X}_t|}{X_t} \times 100}{n}$$

Where, $\hat{X}_t$ is the predicted value, X_t is the actual and n is the number of observations.

3. Result and Discussion

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 supposed 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.

The original time series for the TB incidence includes the incidence for the years 2008-2017. An initial plot of the data did not indicate any seasonality, but an increasing, then decreasing trend as seen in the figure below.

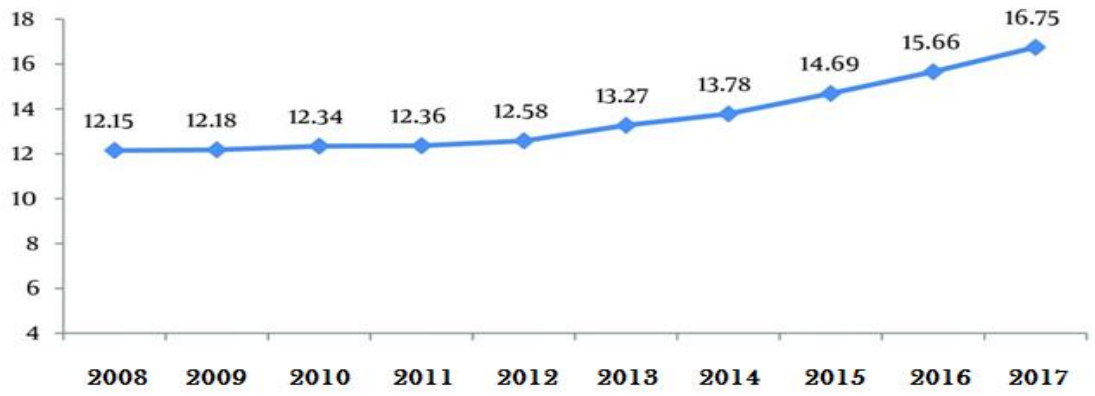

Figure -1: Observed and Predicted Number of Tuberculosis in Tamilnadu, (Per Month from 2008 until 2017)

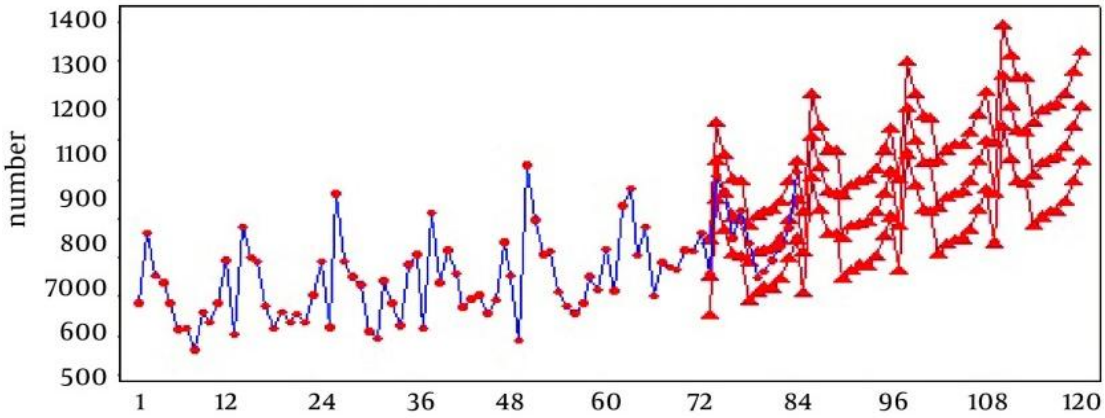

Figure-2; Time Series plot for total TB (With forecasts and their 95% confidence limits)

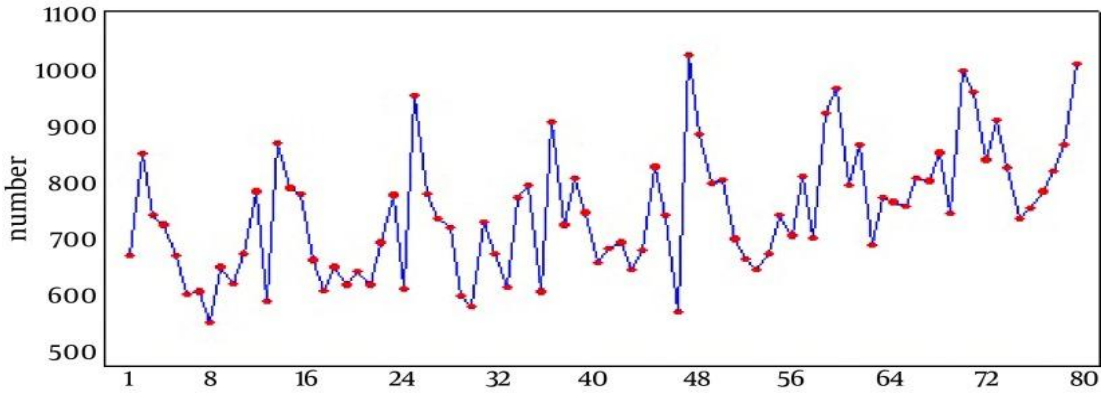

Figure-3; Time Series Plots for (Crud Data) Number of incidence in TB (Per Month from 2008 until 2017)

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 of at lag 2 See figures 5. In most cases, we would attempt to use these to come up with the appropriate model for our data. In our case however, an ARIMA (2,2,1) as the best model. This model is a good t 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 4. This is in agreement with the Box-Ljung test which yielded a p-value of 0:4528. Also the ACF of residuals also show no significant lags beyond zero, an indication that our data fits the ARIMA model reasonably well.

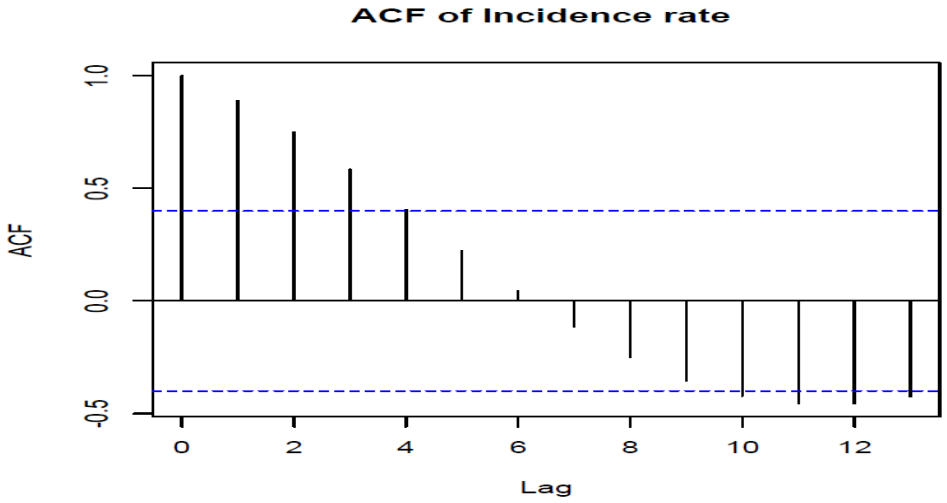

Figure-4: ACF Plot.

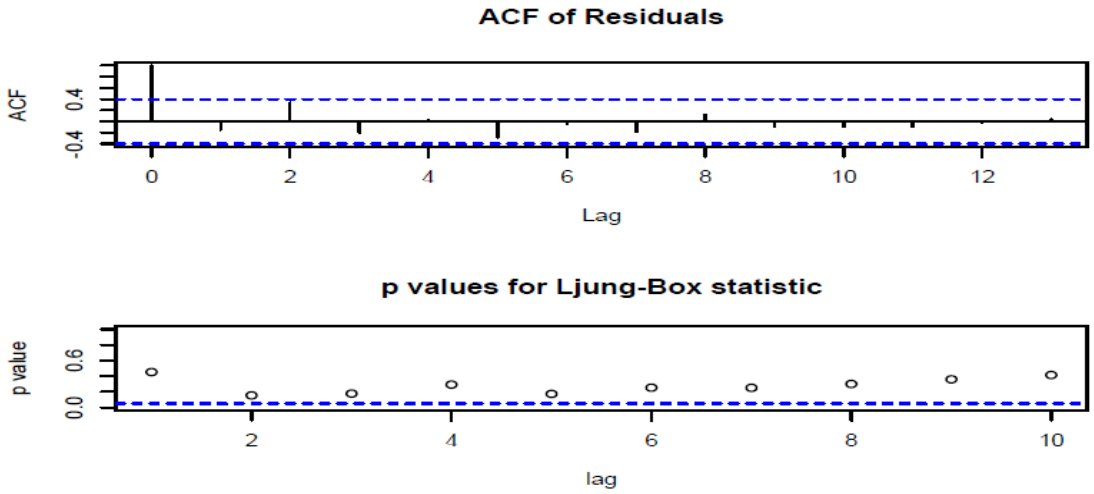

Figure-5: Residuals Diagnostic Plot.

Using the forecast function in R we obtain a forecast of the incidence of the next ten years, 2018-2027, as well as 80%and 95% prediction intervals for those predictions, see table -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 2027. 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 7

Table 1: ARIMA Model Forecasts.

Year	Forecast point	Low 80%CI	High 80%CI	Low 95%CI	High 95%CI
2018	532.88	539.57	530.65	528.34	542.87
2019	517.25	509,66	526.85	506.46	532.68
2020	504.43	488.78	520.18	479.44	534.43
2021	497.90	463.90	538.35	456.63	540.07
2022	484.76	452.72	543.90	403.68	558.41
2023	476.78	402.98	560.54	358.67	620.32
2024	466.59	376.60	575.88	305.63	654.90
2025	445.24	320.34	590.34	253.56	661.07
2026	436.60	278.38	617.56	178,44	710.28
2027	430.80	214.44	644.75	99.87	753.81

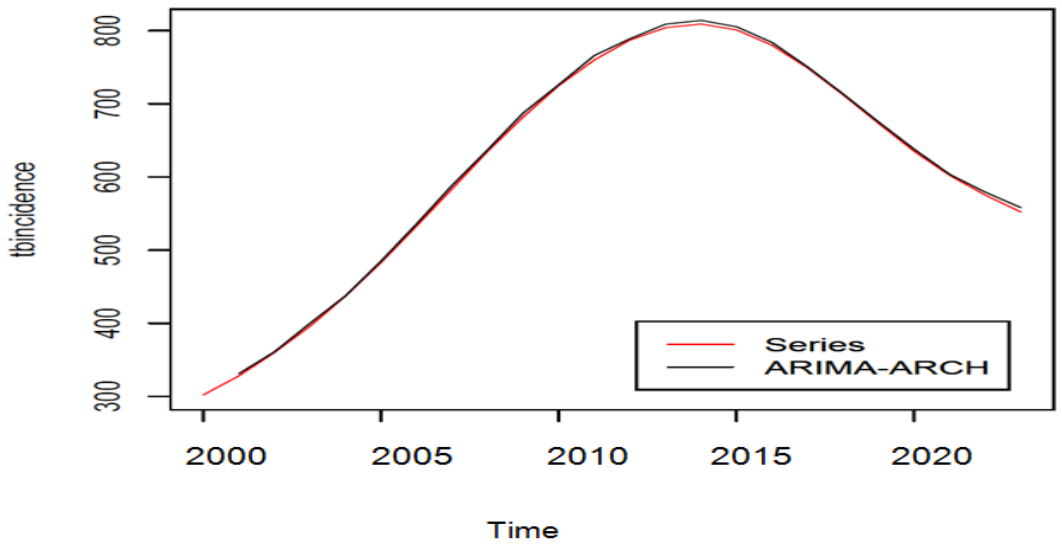

Figure- 6; Forecasts from ARIMA- Autoregressive conditional Heteroscedasticity.

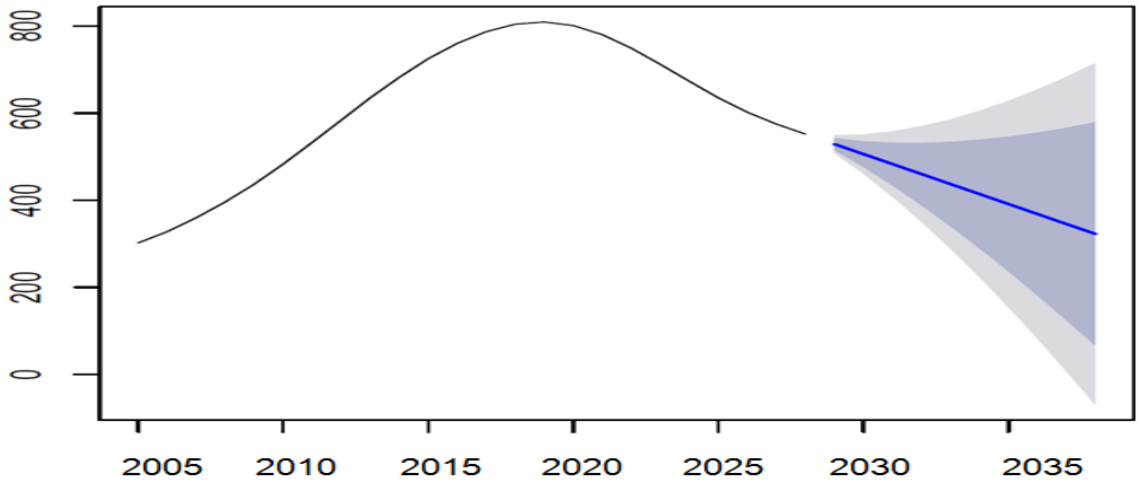

Figure- 7; Forecasts from Holt Winters Predicted Incidence of TB

Table 2: Holt Winters Predicted Incidence Table.

Year	Forecast point	Low 80%CI	High 80%CI	Low 95%CI	High 95%CI
2018	532.88	530.63	535.14	529.43	536.33
2019	517.25	509.67	524.83	505.66	528.84
2020	504.48	487.55	521.41	478.60	530.37
2021	493.82	462.87	524.76	446.49	541.15
2022	484.41	434.39	534.44	407.91	560.92
2023	475.44	401.17	549.72	361.85	589.03
2024	466.13	362.59	569.66	307.78	624.47
2025	455.81	318.36	593.27	245.60	666.03
2026	443.99	268.50	619.50	175.60	712.40
2027	430.60	224.30	647.37	98.45	762.25

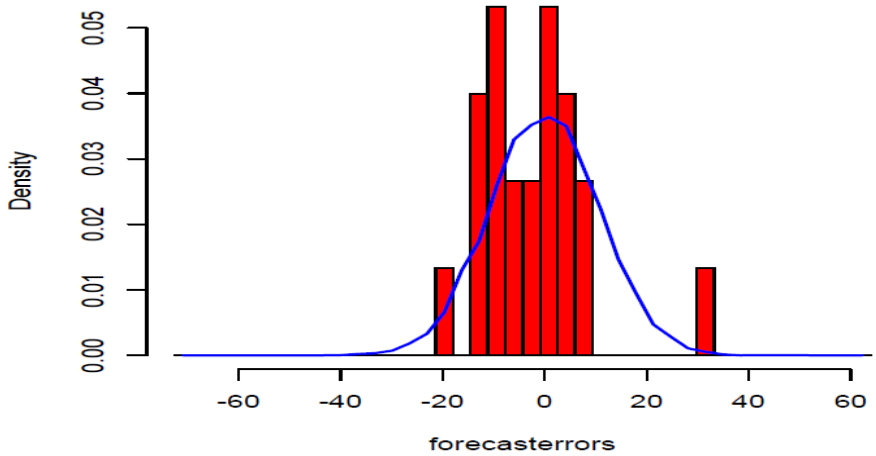

Figure- 8: Histogram of Residuals.

The figure-8 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 shows 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 Tamilnadu. In addition, it means that the assumptions that the 80% and 95% predictions intervals were based upon are probably valid.

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.

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 Tamilnadu for the years 2005-2017 in the short term. 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 figure-6 after that, we do Histogram-Normality test, see figure-8, 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 this data.

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

4. Conclusion

TB is a serious public health issue in Tamilnadu and continuous 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 Tamilnadu. Therefore, this study suggests that these models be utilized in an effort to optimize TB prevention by providing estimates on TB incidence trends in Tamilnadu. We found that our forecasts suggest that Tamilnadu 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 cities.

Reference

1. Cain, K.P., Oeltmann, J.E., Kammer , J.S., Moonan, P.K., Ricks, P.M (2011). Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the u.s., 2005-2009. *PLoS One, 6:e27405.*
2. Floyd, K., Raviglione, M., Glaziou, P. (2009) Global burden and epidemiology of tuberculosis. Clin. *Chest Med., 30:621636.*
3. World Health Organization (2017). Who, global tuberculosis report.
4. Sapii, N., Dir, S., Mardiah, T., Abdullah, S (2012). Application of univariate forecasting models of tuberculosis cases in kelantan. *ICSSBE.*
5. Roslanb, U., Kilicmana, A. Tuberculosis in the terengganu region. Forecast and data analysis. *Scie Asia.*
6. Willis MD, Winston CA, Heilig CM, Cain KP, Walter ND, Mac Kenzie WR. Seasonality of tuberculosis in the United States, 1993-2008.*Clin Infect Dis. 2 012; 54(11):1553-60.*
7. Zhang Y, Tang G, Wang W. Application of ARIMA Model in forecasting incidence of tuberculosis. Modern Prevent Med. 2008;9:6.

8. Abdullah S, Sapii N, Dir S, Mardiah T. (2012) Application of univariate forecasting models of tuberculosis cases in Kelantan, in Statistics in Science, Business, and Engineering (ICSSBE). 2012 International Conference on*: IEEE. Langkawi, Kedah, Malaysia.*

9. Engle, R.F. Autoregressive conditional Heteroscedasticity with estimates of variance offend kingdom nation. Econometric.

10. Roslanb, U., Kilicmana, A. Tuberculosis in the terengganu region. Forecast and data analysis. Scie Asia.

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

12. Liu, X., Jiang, B., Yang, W., Liu, Q. Forecasting incidence of hemorrhagic fever with renal syndrome in china using Arima model. BMC Infect Dis.

13. Tong, L.I., Lee, Y.S. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. Know-Based. Syst.

14. Central TB Division in India report Directorate General of Health Services, Ministry of Health and Family Welfare, Nirman Bhavan, New Delhi 10011(http://www.tbcindia.org)

15. Tuberculosis Control in the South-East Asia Region Annual Report 2013.

16. WHO Library Cataloguing in Publication Data Assessing tuberculosis prevalence through population-based surveys, (2007).