“EPIDEMILOGICAL FACTORS OF TUBERCULOSIS IN SOUTH ASIAN COUNTRIES: TIME SERIES ANALYSIS”

By

SAHANA

Reg. No.: MSCB05

Dissertation submitted to

YENEPOYA (DEEMED TO BE UNIVERSITY)

Manglore, Karnataka

In Partial fulfilment for the award of degree of

Master of Science

In

Biostatistics

Under the guidance of

Dr Sudhakar Prasad

YENEPOYA
(DEEMED TO BE UNIVERSITY)
Recognized under Sec 3(A) of the UGC Act 1956
Accredited by NAAC with 'A' Grade

Department of Statistics

Yenepoya (Deemed to be University)

Mudipu, Mangalore-574153

2020-2022

Certificate

This is to certify that Miss Sahana, a student of M.Sc. Biostatistics from the Department of Statistics, Yenepoya (Deemed to be University), Mudipu has completed her semester project under my guidance from April 2022 to October 2022. The project work entitled **"EPIDEMILOGICAL FACTORS OF TUBERCULOSIS IN SOUTH ASIAN COUNTRIES: TIME SERIES ANALYSIS"** Embodies the novel work done by her.

Name and signature of HOD **Name and Signature of Guide**

Declaration

I hereby declare that the project work entitled **"EPIDEMILOGICAL FACTORS OF TUBERCULOSIS IN SOUTH ASIAN COUNTRIES: TIME SERIES ANALYSIS"** submitted to the Department of Statistics, Yenepoya (Deemed to be University), Mudipu, is a record of the work done by me under the guidance of Dr Sudhakar Prasad This project report is submitted in partial fulfillment of the requirements for the award of degree of Master of Science in Biostatistics.

The result embodied in this thesis has not been submitted to any other Institute or University for the award of any degree/certificate.

Date:

Place: Name and signature

Acknowledgment

This project report is the end of my journey toward obtaining my Master's Degree. At this moment of accomplishment, I would like to thank all people who contributed in many ways in completion of this project work.

I would like to express my sincere gratitude to my guide Dr Sudhakar Prasad for the continuous support of my study, for their motivation, encouragement, and many ways.

I would like to offer my regards to all the faculties, and office staff of the Department of Statistics for all their valuable assistance through the tenure of the course. Finally, yet importantly, I express my heartfelt thanks to the Almighty and Parents, I am very grateful to them for their blessings and encouragement.

Abstract

BACKGROUND: Worldwide Tuberculosis is the major burden to the population. To see the pattern in annual incidence of tuberculosis statistical model is created.

OBJECTIVES: 1) To study the pattern in annual TB incidence number and annual TB and HIV incidence number of south Asian countries.2) To study the pattern in annual TB mortality number and with HIV confounding mortality number of south Asian countries.3) To predict the annual incidence rate of TB.4) To predict the annual mortality rate of TB.

METIRIALS AND METHODOLOGY: The present study is based on the secondary data gathered under the program of open data platform and stored in information system. The data is from 2000 to 2020.South Asian countries are considered in this study. Incidence rate, Mortality rate, and confounding with HIV is the main variables, are predicted using Time series analysis. ARIMA model is used for prediction.

CONCLUSION: The annual incidence number of TB, Incidence of TB and HIV, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number of TB of South Asian countries were predicted for 5 years from 2021 to 2025.The incidence rate of TB cases in Afghanistan, Bangladesh, Pakistan, and Sri Lanka is increasing. There is a decrease in the trend in India and Nepal. There is variation in the incidence rate of Bhutan. The incidence rate and HIV cases in Bangladesh and Pakistan show an increasing rate, and Afghanistan, Nepal, and Sri Lanka show that there is variation in incidence and HIV rate. It is rapidly declining in India and Bhutan.

Contents

List of tables

List of Figures

Chapter 1

1.Introduction

1.1 Motivation

Health science statistics are a form of evidence, or facts that can support a conclusion. Evidence-informed policy-making, "an approach to policy decisions that is intended to ensure that decision making is well-informed by the best available research evidence" and evidence-based medicine (EBM), or "the conscientious, explicit, judicious and reasonable use of modern, best evidence in making decisions about the care of individual patients" are essential to informing how best to provide health care and promote population health.

Not all evidence is, or should be, equally convincing in the support of a conclusion. Evidence varies in quality and whether it is applicable to a given situation. It is therefore essential that health researchers and policy makers understand how to assess evidence in a systematic way, including how to access transparent, high quality health statistics and information.

Health statistics measure four types of information. The types are commonly referred to as the four

- Correlates: See how to measure the risk factors and protective factors that impact our health.
- Conditions: Learn to assess how often and how badly diseases impact a community.
- Care: Dig into how health care is delivered to the communities that need it, to treat disease and illness.
- Costs: Get more information on what health care costs, and why.

This study is mainly about tuberculosis. When it comes to health statistics modules, it comes under the second point called "conditions". To assess how often and how badly TB impacts the country using the evidence of the statistics supported by the statistical concept called "Time series". Using the ARIMA model of time series, forecasted the five-year pattern of the TB trend and interpreted it in this study.

1.2 Tuberculosis

TB is a bacterial infection is contagious that often affects the lungs. The kidneys, spine, and brain are just a few of the body organs that it can harm. There have two different forms of TB.

• Infection with latent TB. Although there are tuberculosis pathogens in the body, they do not cause illness.

• Mycobacterium tuberculosis is the culprit behind tuberculosis sickness (active tuberculosis). In most cases, antibiotics can treat tuberculosis. It can, however, be lethal if not adequately treated.

When an untreated individual coughs, sneezes, laughs, or sings, Tb is primarily disseminated through the air. In order for the disease to be transferred to another person, an infected individual needs to be in close proximity to that person. Sharing glasses or utensils might spread the disease TB. Even if you come into personal contact with someone, it won't spread.

It will not spread even if you are in physical contact with someone. There’s no age restriction for TB; anyone can get TB. TB infection will cause a higher rate of TB infection as it compromises your immune system. Likewise, other immunosuppressive diseases will cause a higher chance of getting TB. HIV infection reduces immunity strength, which may result in a latent TB becoming active.

Certain people are more susceptible to developing TB disease after contracting the infection. They consist of those who:

- own HIV or acquired tuberculosis infection within the previous two years
- If you have conditions that make it difficult for your body to fight off TB bacteria, such as diabetes, etc
- If you have a problem with drinking or injecting drugs.
- If you were not properly treated for tuberculosis in the past were less than 5 were an older adult.

Although there is a good chance that those with latent TB will never get active TB, those with latent TB infection are more likely to do so than other persons. Those who are more likely to have active TB include:

- infected with HIV
- individuals with active TB within the last two years
- youngsters and new-borns
- those who take illicit drugs
- individuals with immunosuppressive diseases
- Older individuals
- those who in the past received inadequate TB treatment.

1.3 Epidemiology

Epidemiology is a study of distribution determinant of health-related states or event in a specified population and application of the study to control the health problems. In a simple way it can be stated that it is a study about disease and its betterment.

Basic mechanism of investigation:

- observational capacity
- Analytical mind
- Making inference

Specific objectives of epidemiology:

- ❖ To identify the etiology or cause of a disease and the relevant risk factors that is , factors that increase a person's risk for a disease .
- ❖ To determine the extent of disease found in the community
- ❖ To study the natural history and prognosis of disease

Uses of Epidemiology

- ❖ It helps to predict disease trend
- ❖ Identify the health needs of community
- ❖ To study historically rise and fall of the disease in the population
- ❖ By relating disease to inter population differences and other attributes of the population tries to identify the cause of the disease.
- ❖ Epidemiology is concerned with entire spectrum of disease in a population

1.4 Terminology

Incidence rate and mortality rates of tuberculosis are the important terminologies in this study.

They are defined as follows:

$$\text{Incidence rate} = \frac{number\ of\ new\ cases\ of\ specific\ disease\ during\ the\ given\ time\ period}{population\ at\ risk\ during\ time\ period} \times 100000$$

$$\text{Mortality rate} = \frac{number\ of\ death\ cases\ of\ specific\ disease\ during\ the\ given\ time\ period}{population\ at\ risk\ during\ time\ period} \times 100000$$

In this study the incidence number is new cases of tuberculosis occurring in a defined population during specified time. And the mortality number is death cases of tuberculosis occurring in a defined population during specified time.

1.5 Literature review

In this section a brief review of literature on epidemiological factors of tuberculosis prediction is presented. The review gives an exposure on the statistical tools available for analysis of data.

Aryee G et al (2018) used to obtain a time series model to estimate the incidence of TB cases at the chest clinic of the Korle-Bu Teaching hospital. According to this journal Various models were stated and compared and the best was found to be based on the Akaike Information Criterion and Bayesian Information Criterion.

Zheng Y et al (2021) We used a single Box-Jenkins method and a Box-Jenkins and Elman neural network (Elman N) hybrid method to do prediction analysis of TB incidence in Kashgar. Root means square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) were used to measure the prediction accuracy.

Liu K et al it describes Tuberculosis, a severe infectious disease caused by the Mycobacterium tuberculosis, arouses huge concerns globally. The notified TB incidences demonstrated a continuously declining trend and More seriously, based on the WHO's

release, only 64% of cases were diagnosed and recorded formally, an estimated 1.3 million TB cases resulted in death among HIV-positive people.

1.6 Aim

To develop appropriate statistical models for Tuberculosis prediction.

1.7 Objectives

- To study the pattern in annual TB incidence number and annual TB and HIV incidence number of south Asian countries.
- To study the pattern in annual TB mortality number and with HIV confounding mortality number of south Asian countries.
- To predict the annual incidence rate of TB.
- To predict the annual mortality rate of TB.

1.8 Data source

The secondary data is available at the official website of World Health Organization (WHO). and can be accessed through the link https://www.who.int/teams/global-tuberculosis-programme/data .

1.9 Software used

The R studio (version 4.1.1) is used for model building, MS excel and SPSS is used for exploratory analysis.

1.10 Overview

Tuberculosis is the major burden to worldwide. Prediction of tuberculosis may provide useful insights for policy making. This project work is carried out by using the secondary data streams which are available to general public. The details of the work done and the conclusions derived are provided in the report. This project report is divided into four chapters. The need and motivation for the study is mentioned in the first chapter along with the details of Tuberculosis and Epidemiological factors.

The second chapter is devoted to the materials and methods used in the study. It also consists of assumptions of the study which allow us to choose the appropriate statistical techniques.

The report of data analysis is presented in third chapter. The data analysis includes the exploratory analysis, the time series analysis.

The last chapter is devoted to the summary, conclusions and useful recommendations based on the study results.

Chapter 2

2.Materials and Methodology

The present study is based on the secondary data gathered under the program of open data platform and stored in information system. This chapter is concerned with the important materials and the statistical techniques used to draw reasonable conclusions. This chapter is organised as follows: The first section describes the exploratory analysis. The second section is deals with the Time series analysis. The last section gives the brief summary of the statistical methods used in the study.

2.1 Exploratory data analysis:

The principal objective of the exploratory data analysis is to summarise the main characteristics of the data before using any actual statistical models. It is simply the task of exploring the data using simple statistical and graphical tools. It was prompted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypothesis that could lead to new data collection and experiments. The graphical tool such as boxplots are used in this technique.

2.2 Mean:

The arithmetic mean of a set of observations x1, x2, x3……,xn is defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The calculation of mean incorporates all the values in the data according with selected countries.

2.3 Variance

Variance is the most commonly used measure of dispersion in statistical analysis. It is a measure that takes into account all the values in a set of observations. It is given by:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

The wider the dispersion of the values around their mean, the greater will be the variance. If there is no dispersion then all the values are equal to the mean which means that variance is zero.

The positive square root of the variance is called the **standard deviation** and is denoted by S.

$$S=\sqrt{\sigma^2}$$

The variance is expressed in units that are the square of the unit of measure of the variable under the study. However, the standard deviation is expressed in the original unit of measure of the variable.

2.4 Median

Median is the middle most number in the data when it is in the ascending or descending order. If there is total of odd amount of number, the median value is the number that is in the middle, with the same amount of numbers below and above. If there is total of even amount of number in the list, the middle pair must be determined, added together, and divided by two to find the median value.

2.5 Quartiles:

Quartiles are the set of values which has three points dividing the data set into four identical parts, first, second and third quartile is represented by Q_1, Q_2 and Q_3, respectively. Q_2 is nothing but the median. Inter Quartile ranges are calculated by the formula

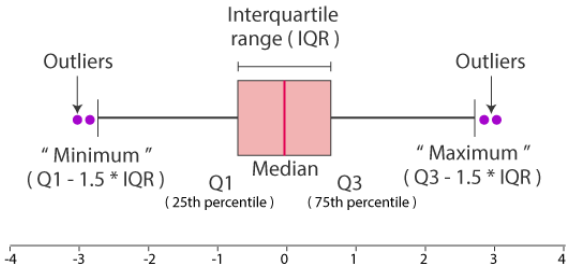
Q_1 = [(n+1)/4]th item

Q_2 = [(n+1)/2]th item

Q_3 = [3(n+1)/4]th item

2.6 Box plots

Boxplots are used to show spread and centres of the data set, even interquartile range and mean of the data set.



2.7 Description time series analysis

A Time Series is a sequence of observations ordered in equally spaced, discrete time intervals. In this kind of analysis, we make use of different statistical models to analyse the time series data. A basic assumption in any time series analysis/modelling is that some aspects of past pattern will continue to remain the same in the future. Suitable forecasting time series model can be developed which gives minimum forecasting error. At least 50 observations are necessary for performing Time Series analysis, as propounded by Box Jenkins who were pioneers in Time Series modelling. We start by plotting the time series data and look for nonstationary components. We then eliminate these components using various methods, in order to get a stationary data. After identifying a suitable probability model for the time series, these 10 models can be used for prediction. We start by plotting the time series data and look for nonstationary components. We then eliminate these components using various methods, in order to get a stationary data. After identifying a suitable probability model for the time series, this model can be used for prediction.

2.8 The concepts of time series:

An important step in analysing Time Series data is to consider the types of data patterns, so that the models which are most appropriate to those patterns can be utilized. An observed time series can be decomposed into four components.

Time series X_t can be represented in terms of these four components as follows:

$$X_t = M_t + S_t + C_t + I_t$$

1. Trend (M_t) - when there is long term increase or decrease in the data.

2. Seasonal (S_t) -when a series is influenced by seasonal factor and recurs on a regular periodic basis.
3. Cyclical (C_t) -when the data exhibits rise and falls that are not a fixed period.
4. Irregular (I_t) - when data values fluctuate around a constant value.

Whenever the time series contains these components, it is said to be non-stationary, therefore before developing a stochastic model for time series we need to test the presence of these components and eliminate them from the series.

The steps in analysing the time series are:

1. Time profile: Plot the observed time series Xt versus time point t in a graph. Examine whether data contains trend, seasonality and cyclic variations.

2. Making series stationary: Remove the non-stationary components from observed series by estimation techniques or by differencing methods.

3. Model building: Build an appropriate time series model for the stationary time series making use of sample autocorrelation and partial autocorrelation function.

4. Diagnostic checking: Once we fit the appropriate time series model, we are going to check whether fitted model is good fit for the series. For this we are using residual series of fitted time series model. If all autocorrelation of residual series is insignificant then fitted model is good fit for stationary series.

5. Forecasting: It is the main objective of Time Series analysis. It will be achieved by forecasting the stationary series then inverting transformation described in step2 to arrive at forecast of the observed data.

2.9 ARIMA Model:

Let $\varepsilon_t \sim WN(0, \sigma^2)$ and $\{x_t \ t > 1\}$ and $\{x_t \ x > 1\}$is non stationary time series. Let ∇ denote the differencing operators and let Yt~ARIMA(p,q) i.e Yt has representation

$\emptyset(B)y_t = \theta(B)\ \varepsilon_t$ 1

Substituting for Yt $= \nabla^d X_t = (1-B)^d Xt$

Model 1 reduces to $\emptyset(B)\ (1-B)^d Xt = \theta(B)\varepsilon_t$

Thus, the time series $\{x_t\}$ is said to follow Auto Regressive Integrated Moving Average process of order (p d q) if it has representation

$\emptyset(B)Xt = \theta(B)\ \varepsilon_t$

Where $\emptyset(B)$ = 1-β1B – $\beta 2B^2$ -…… βpB^p

And $\theta(B)$= 1-α_1B – $\alpha_2 B^2$ -…… $\alpha_q B^p$

We denote Xt~ARIMA (p d q)

Finding Arima model:

Let the variable of interest is denoted by x(t). Make the transformation Y(t)=ln(Xt).

Let us identify appropriate ARIMA (p d q) model for Yt

Step 1: Identification of ARIMA model.

Examine the stationary using ACF. If spikes outside the 2**σ** limit then non-stationary.

Step 2: If Yt is nonstationary, let determine d by variate difference method

- Compute variance Xt the original series v(Xt)=v_0
- Obtain $\Delta x_t = x_t - x_{t-1}$ t=2,3…..n and compute v(Δx_t)= v_1

If $v_0 > v_1$ then take d=0 i.e original series is stationary and no need of differencing the series and terms is not present.

If $v_0 < v_1$ then take d=1 and the time series is stationary after differencing it one $\nabla^2 x_t$ and calculate v ($\nabla^2 x_t$)= v_2

If $v_2 < v_1$ then take d=1 and the time series is stationary after difference it once

$\{\Delta x_t\}$ is stationary.

Otherwise continue differencing the series until get a variance of the difference series in less than variance of the previous differencing.

Step 3: plot the ACF at the last steps $\{\nabla^2 y_t\}$ for the series for checking stationary

2.10 Autocorrelation function (ACF):

Autocorrelations referred to the observations in a time series are related to each other and is measured by the simple correlation between current observation (Xt) and observations from k periods before the current one $Xt-k$ i.e., for a given series Xt, autocorrelation at lag k is the correlation between the pair ($Xt\ Xt-k$).

Sample autocorrelation function is the consistent estimator of the ACF. It is defined as:

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}$$

It is useful to check whether these coefficients are statistically nonzero or more specifically, to check whether Xt is a white noise. Then a usual test of individual significance can be applied, i.e., $H0$: pk =0 against $H1$:pk ≠0 for any k=1,2.... The null hypothesis $H0$ would be rejected at the 5% level of significance if $|rk| > 1.96\sqrt{n}$

2.11 Partial Autocorrelation function (PACF):

Partial autocorrelations are used to measure the degrees of association between Xt Xt-p when X effects at other time lags 1,2,3, …. p-1 are removed. The auto covariance coefficient γk at lag k measures the covariance between two values Xt and Xt+k separated by interval of time. The partial autocorrelation function is estimated by the OLS coefficient from the expression that is known as sample PACF. Under the assumption that Xt ~ WN (0, σ2), the distribution of the sample coefficients in large samples is identical to those of the sample ACF. In consequence, the rule for rejecting the null hypothesis of individual non significance is also applied to the PACF. The bar plot of the sample PACF is called the sample partial correlogram and usually includes the two standard error bands ±2 to assess for individual significance. In general, PACF is the tool used for determining the order of auto regression process.

2.12 Akaike's Information Criteria

AIC=$\frac{e^{\frac{2k}{n}}}{n}\sum_{i=1}^{n} e_i^2$

AIC =$e^{\frac{2k}{n}}\frac{ESS}{n}$

n = numbers of observation

k = number of regressors

or ln AIC=$\frac{2k}{n}$+ln ($\frac{ESS}{n}$)

In comparing two or more models, the model with lowest value of AIC is preferred. One advantage of AIC is that it is useful not only in sample but also out of sample forecasting performance of regression model.

2.13 Accuracy Measures: In forecasting, our objective is to produce and optimum forecast that has no error or as little error as possible, which leads us to the minimum mean square error forecast. This forecast will produce an optimum future value with the minimum error in terms of the mean square error criterion.

- **Root Mean Square Error (RMSE):**

The Root Mean Square Error (RMSE) is a frequently used measure of difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are called residuals, and the RMSE serves to aggregate them into a single measure of predictive power. The RMSE of a model prediction with respect to the estimated variable X model is defined as the square root of the mean squared error. RMSE= $\sqrt{\frac{\sum_{i=1}^{h}(X_{obs,i}-X_{model,i})^2}{n}}$ Where, X_{obs} denotes observed values and X_{model} denotes modelled values at time/place i.

- **Mean absolute error (MAE):**

The mean absolute error (MAE), is calculated as the average of the forecast error values, where all of the forecast values are forced to be positive. Forcing values to be positive is called making them absolute MAE=$\frac{\sum_{i=1}^{N}|x_i-\hat{x}|}{N}$ where, x_i is the actual observations time series, $\hat{x_t}$ is the estimated time series, N is the number of non-missing data point.

Chapter 3

3 Analysis and Discussion

3.1 Descriptive of the data

From the data mean, median, standard deviation, quartiles are calculated. The variable description is given below:

E_inc_num = estimated incidence number

E_TBHIV_inc_num = estimated TB and HIV number

E_mortality_TB_ex_HIV = estimated mortality number of TB excluding HIV

E_TBHIV_mortality_num= estimated mortality number of TB and HIV number

E_mortality_num= estimated mortality number

Descriptive statistics of Afghanistan

Estimates	**Mean**	**median**	**Standard deviation**	**Quartiles**		
				25	**50**	**75**
e_inc_num	56333.33	55000.00	24430.465	47500.00	55000.00	66000.00
e_TBHIV_inc_num	12.71	13.00	4.941	7.50	13.00	17.50
e_mortality_TB_ex_HIV	12371.43	13000.00	1413.203	11000.00	13000.00	14000.00
_TBHIV_mortality_num	111.19	120.00	38.505	75.00	120.00	140.00
e_mortality_num	12523.81	13000.0	1364.516	11000.00	13000.00	14000.00

Table 3.1

Descriptive statistics of Bangladesh

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	325619.05	326000	24430.465	305500	3260	347500
e_TBHIV_inc_num	474.19	550	222.578	250	550	665
e_mortality_TB_ex_HIV	73095.24	76000	16364.305	63500	76000	87000
e_TBHIV_mortality_num	161.43	180	99.302	55.50	180	260
e_mortality_num	73285.71	76000	16248.516	63500	76000	87000

Table 3.2

Descriptive statistics of Bhutan

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	1304.76	13	132.198	1200	13	1400
e_TBHIV_inc_num	41.38	46	26.607	9.50	46	62
e_mortality_TB_ex_HIV	140.95	140	15.781	130	140	150
e_TBHIV_mortality_num	10.95	12	7.440	2	12	17.50
e_mortality_num	151.43	150	18.244	140	150	160

Table 3.3

Descriptive statistics of Pakistan

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	489809.52	495000	58298.044	437000	495000	542500
e_TBHIV_inc_num	3721.90	2700	3598.423	745	2700	5950
e_mortality_TB_ex_HIV	47952.38	47000	3680.709	45000	47000	52000
e_TBHIV_mortality_num	821.19	610	911.025	23.50	610	1300
e_mortality_num	48857.14	49000	3021.353	46000	49000	52000

Table 3.4

Descriptive statistics of India

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	2980000	3050000	197027.917	2815000	3050000	3155000
e_TBHIV_inc_num	193428.5	195000	97043.068	103500	195000	245000
e_mortality_TB_ex_HIV	512714.2	494000	61263.482	453500	494000	566500
e_TBHIV_mortality_num	51595.24	55000	29076.459	17000	55000	74500
e_mortality_num	564285.7	549000	88569.827	477500	549000	639000

Table 3.5

Descriptive statistics of Nepal

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	83238.10	84000	11126.117	71500	84000	93500
e_TBHIV_inc_num	2188.57	2600	920.572	1400	2600	2900
e_mortality_TB_ex_HIV	22095.24	22000	4815.649	17000	22000	26500
e_TBHIV_mortality_num	529.5	520	238.493	330	520	710
e_mortality_num	22619.05	22000	4873.153	18000	22000	27000

Table 3.6

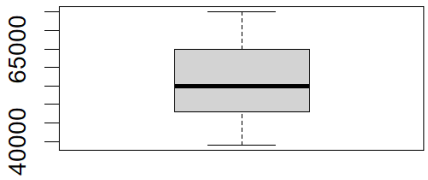
Descriptive statistics of Sri Lanka

Estimates	Mean	median	Standard deviation	Quartiles		
				25	50	75
e_inc_num	13428.57	13000	597.614	13000	13000	14000
e_TBHIV_inc_num	61.71	62	24.286	46.50	62	78
e_mortality_TB_ex_HIV	994.76	950	298.523	740	950	1200
e_TBHIV_mortality_num	19.19	19	7.068	14.50	19	23
e_mortality_num	1010.48	980	296.723	760	980	1700

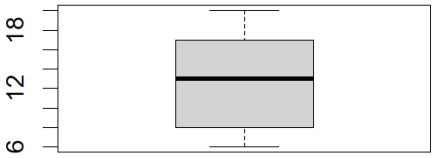
Table 3.7

BOXPLOTS

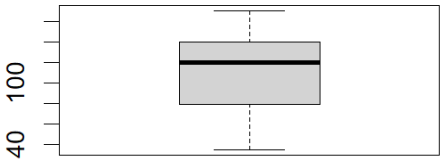
Afghanistan: Figure 3.1



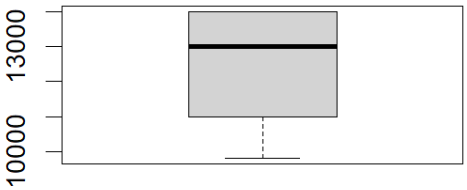
Incidence number



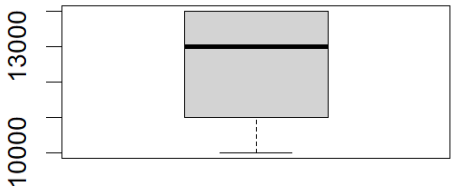
Incidence number of TB and HIV



mortality number TB and HIV

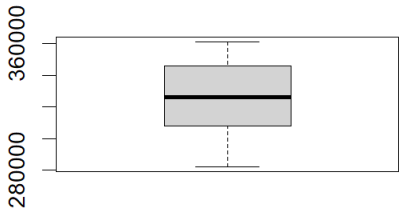


mortality number of TB excluding HIV

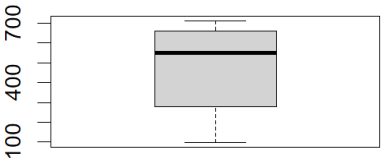


mortality number

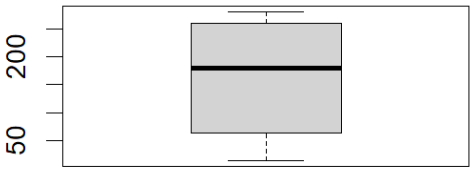
Bangladesh: Figure 3.2



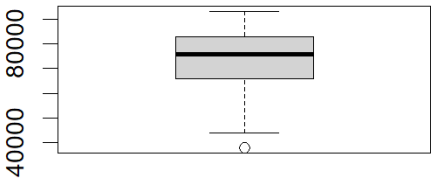
Incidence number



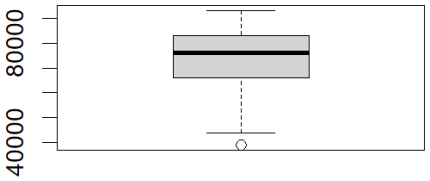
Incidence number of TB and HIV



mortality number TB and HIV

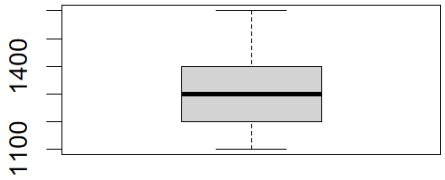


mortality number of TB excluding HIV

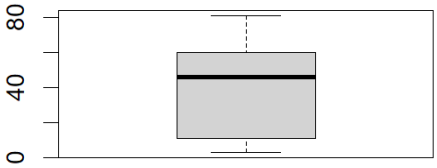


mortality number

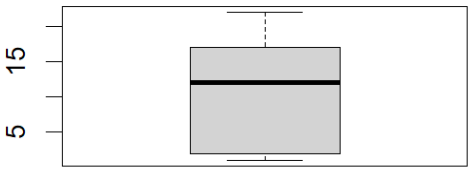
Bhutan: Figure 3.3



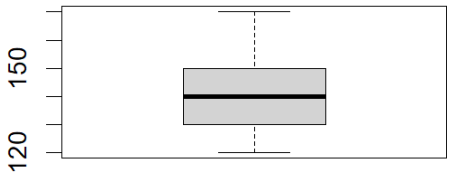
Incidence number



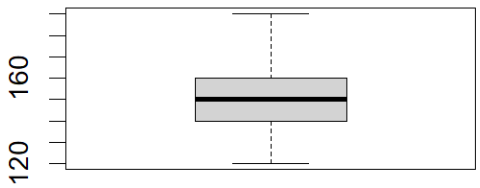
Incidence number of TB and HIV



mortality number TB and HIV

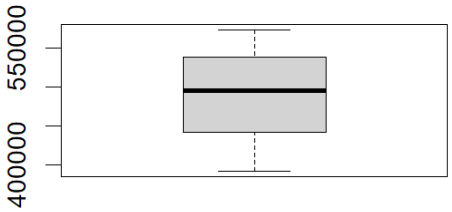


mortality number of TB excluding HIV

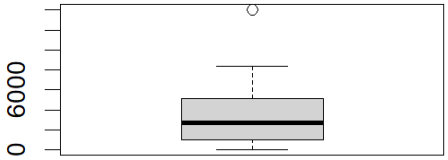


mortality number

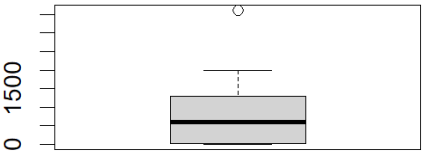
Pakistan: Figure 3.4



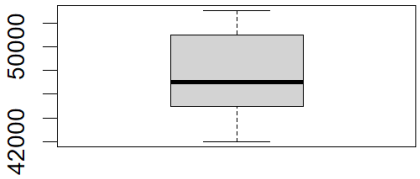
Incidence number



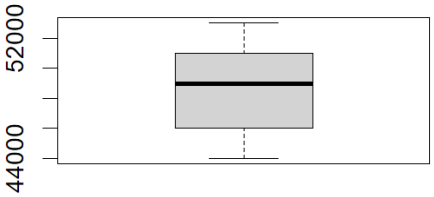
Incidence number of TB and HIV



mortality number TB and HIV

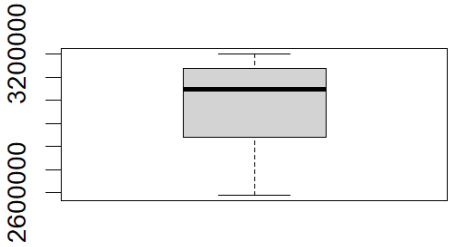


mortality number of TB excluding HIV

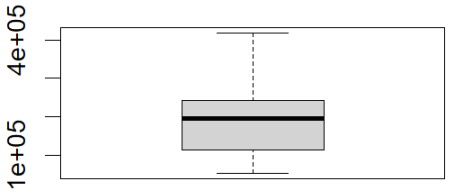


mortality number

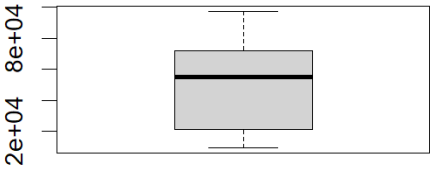
India: Figure 3.5



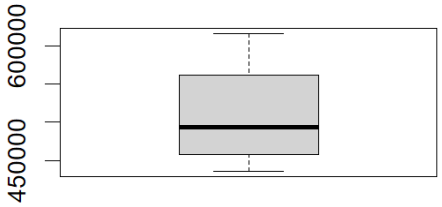
Incidence number



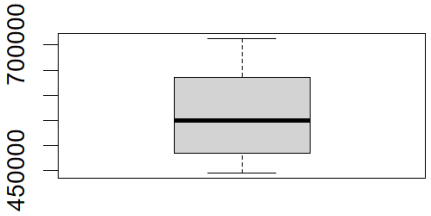
Incidence number of TB and HIV



mortality number TB and HIV

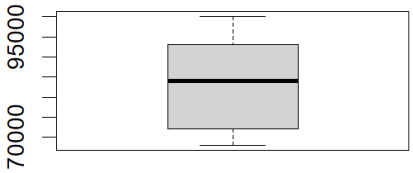


mortality number of TB excluding HIV

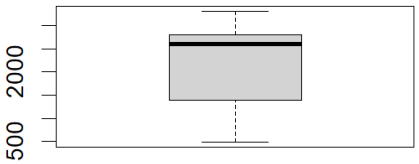


mortality number

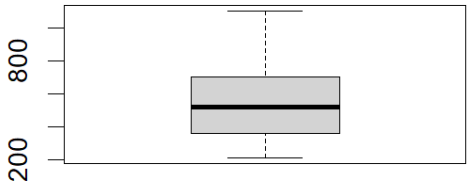
Nepal: Figure 3.6



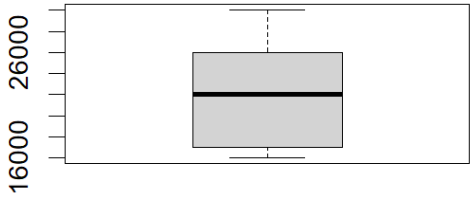
Incidence number



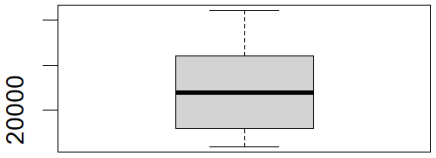
Incidence number of TB and HIV



mortality number TB and HIV

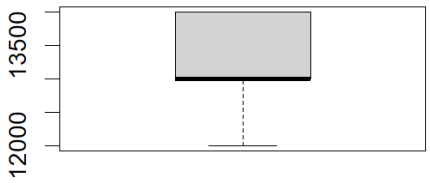


mortality number of TB excluding HIV

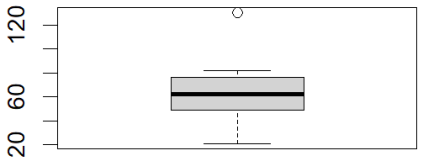


mortality number

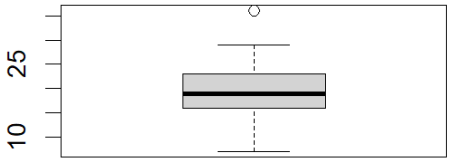
Sri Lanka: Figure 3.7



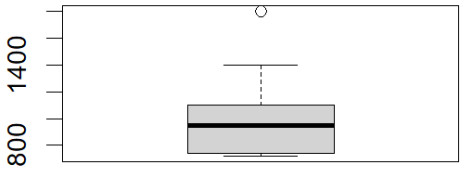
Incidence number



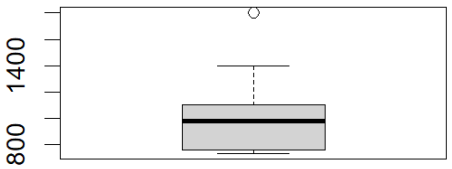
Incidence number of TB and HIV



mortality number TB and HIV



mortality number of TB excluding HIV



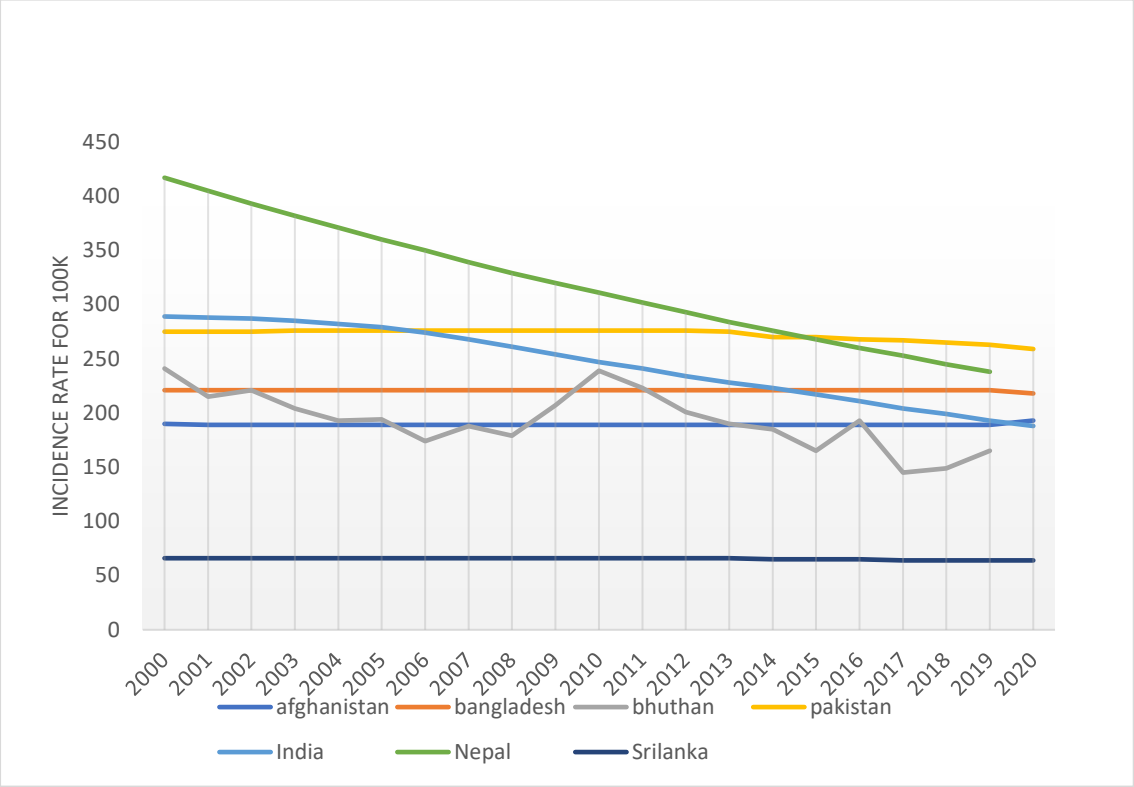
mortality number

Over all interpretation of box plots

- The median is represented by the line in the box. The median is a common measure of the centre of your data. Half the observations are less than or equal to it, and half are greater than or equal to it.
- The interquartile range box represents the middle 50% of the data. It shows the distance between the first and third quartiles (Q3-Q1).
- When data are skewed, the majority of the data are located on the high or low side of the graph.
- Outliers, which are data values that are far away from other data values, can strongly affect your results. Often, outliers are easiest to identify on a boxplot.

3.2 Time series analysis.

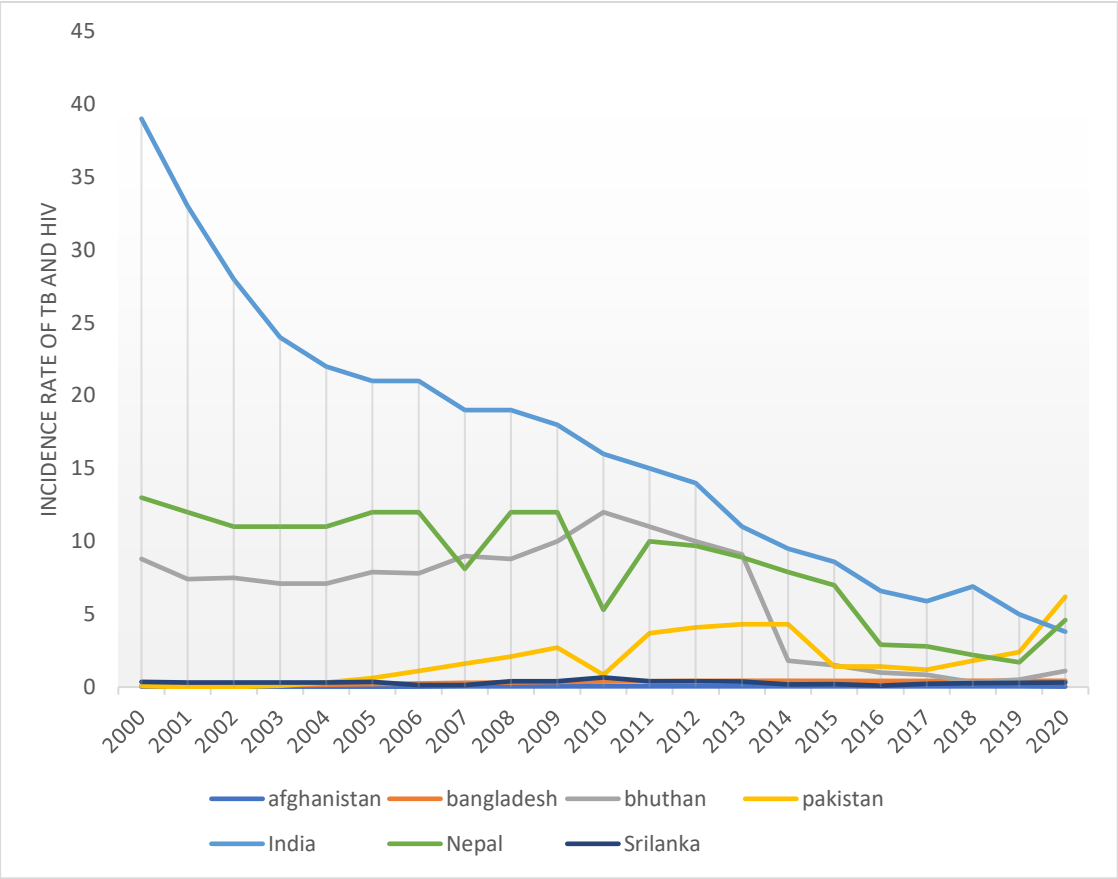
The following time profiles depict the nature of Estimated Incidence rate of TB per 100k for south Asian countries (Figure 3.8)



Country wise time profile of yearly incidence rate of TB

- From the above time profile Incidence rate of Nepal is gradually decreasing.
- From 2006 onwards Incidence rate of India is rapidly decreasing.
- There is almost constant incidence rate for Sri Lanka, Afghanistan, Bangladesh and Pakistan.
- From the time profile it is clear that there is irregular trend for incidence rate of Bhutan.

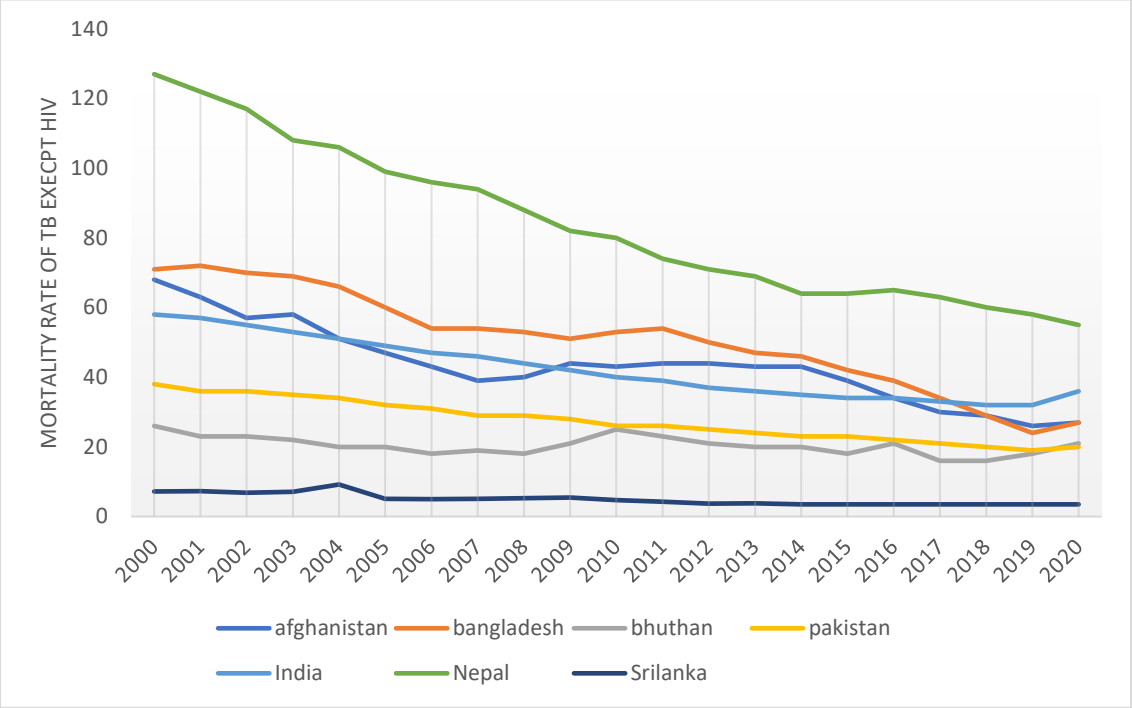
The following time profiles depict the nature of Estimated Incidence rate of TB and HIV of south Asian countries. (Figure 3.9)



Country wise time profile of yearly incidence rate of tb and HIV

- From the above time profile there is rapid decrease incidence of TB and HIV rate.
- There is an irregular variation in the incidence of TB and HIV rate of Nepal, Pakistan, Bhutan, and for Sri Lanka Incidence of TB and HIV rate is constant.

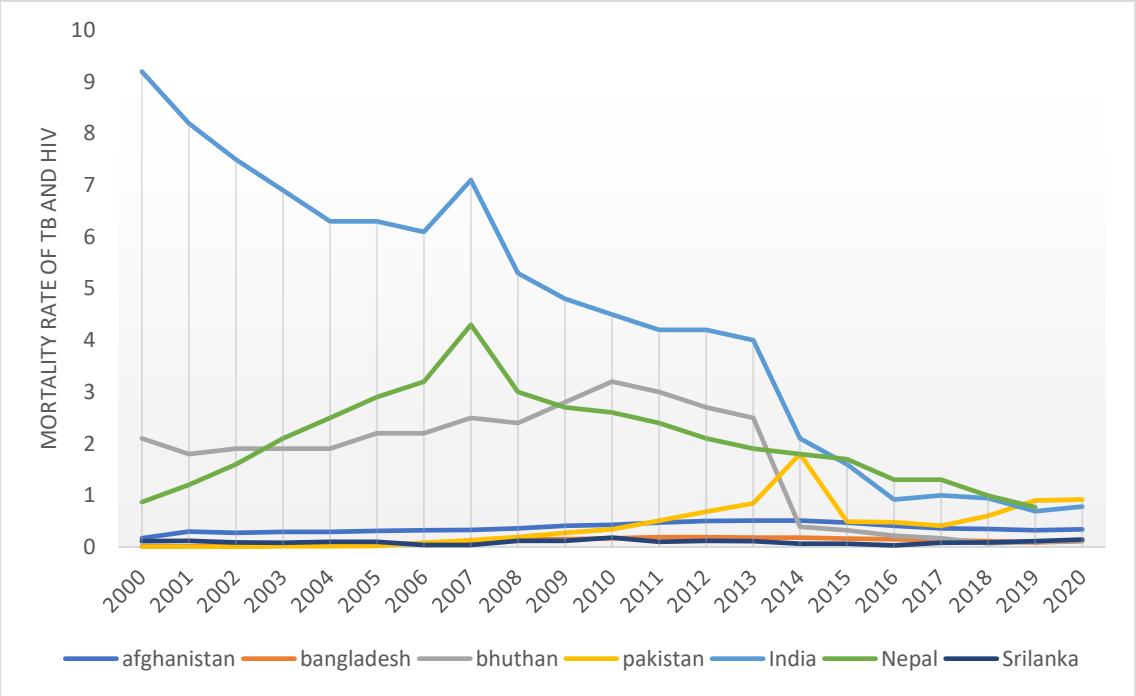
The following time profiles depict the nature of Estimated Incidence rate of TB and excluding HIV per 100k for south Asian countries. (Figure 3.10)



Country wise time profile of yearly mortality rate of TB except HIV

- The following time profile shows that there is downward trend in mortality rate of TB and excluding HIV per 100k for Nepal, Bangladesh and Pakistan.
- There is irregular variation in the time profile of mortality rate of TB and excluding HIV for Bhutan and India.
- For Sri Lanka, the mortality rate of TB and excluding HIV is almost constant.

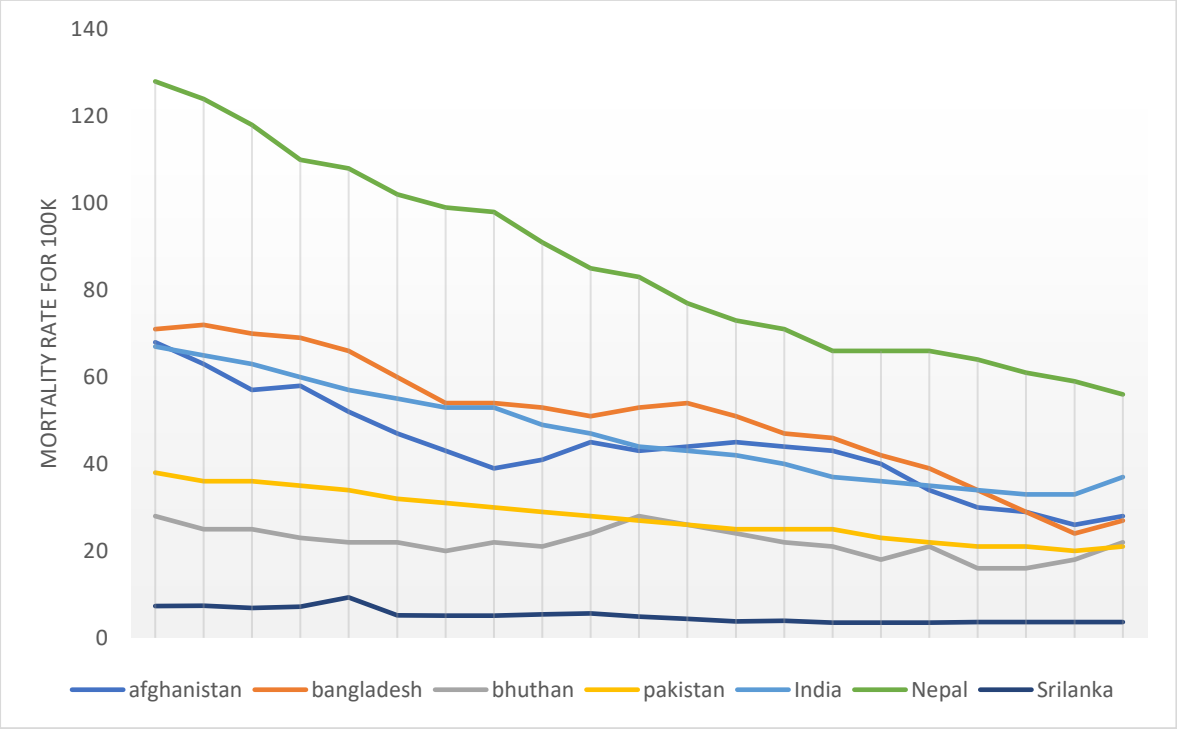
The following time profiles depict the nature of Mortality rate of TB and who have HIV for 100k positive of south Asian countries. (Figure 3.11)



Country wise time profile of yearly incidence rate of TB and HIV

- The above time profile shows that there is a rapid decrease in mortality rate of TB and HIV rate for India.
- Irregular variation in the time profile of mortality rate of TB and HIV for Nepal, Bhutan, and For Bangladesh, Sri Lanka, Afghanistan the time profile is constant.

The following time profiles depict the nature of Mortality rate of TB of south Asian countries. (Figure 3.12)



Country wise time profile of yearly mortality rate of TB

- From above time profile it is clear that there is downward trend in Nepal, Bangladesh and Pakistan and for Sri Lanka time profile shows a constant variation.
- There is irregular variation in the mortality rate of Bhutan, Afghanistan.

The time series is divided into test and train data and the autoregressive integrated moving average models are fitted for given data. The best model is selected using different criteria like AIC, ME, RMSE and MAE. The best models for the prediction of all variables are presented below along with the accuracy measures.

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Afghanistan	Estimated Incident number	0	1	0	310.13	1.7714	497.6796	382.723
	Estimated TB and HIV incident number	0	1	0	82.64	0.2860	1.7728	1.0480
	Estimated mortality number of TB cases excluding HIV	1	0	0	353.6	-81.7908	932.8482	850.1889
	Estimated mortality number of TB and HIV cases	0	1	0	159.93	0.0014	11.6451	9.0014
	Estimated mortality number	1	0	0	350.72	-79.3141	869.2394	731.1792

Table 3.8

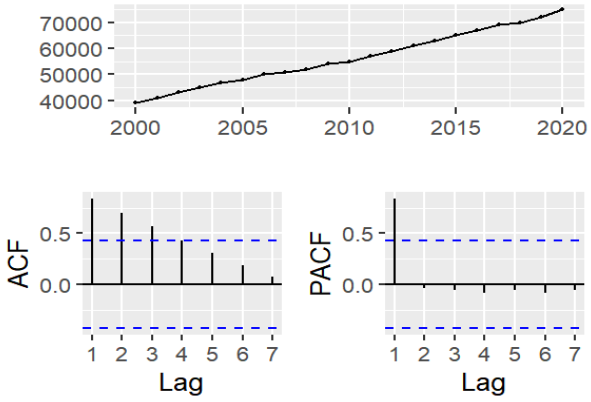
The annual incidence rate of Tb of Afghanistan's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

Prediction table

Afghanistan					
	2021	2022	2023	2024	2025
Estimated Incident number	76800	78600	80400	82200	84000
Estimated TB and HIV incident number	13	13	13	13	13
Estimated mortality number of TB cases excluding HIV	11360.10	11627.52	11826.12	11973.60	12083.12
Estimated mortality number of TB and HIV cases	134.75	139.50	144.25	149	153.75
Estimated mortality number	11354.92	11626.87	11835.25	11994.91	12117.25

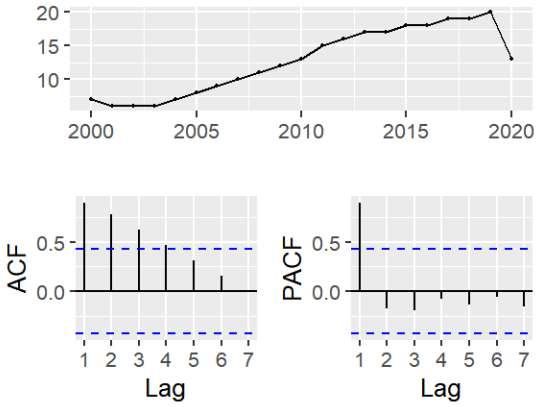
Table 3.9

.

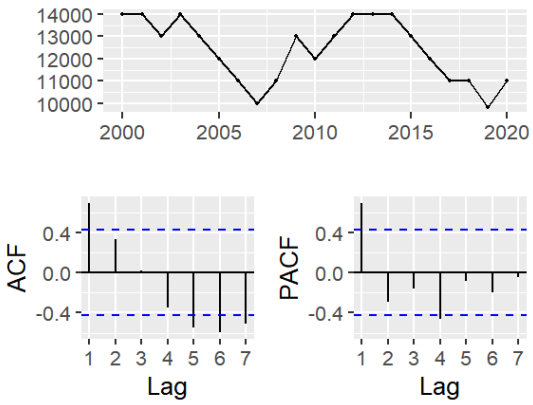


Time profile, ACF, PACF of Incidence number

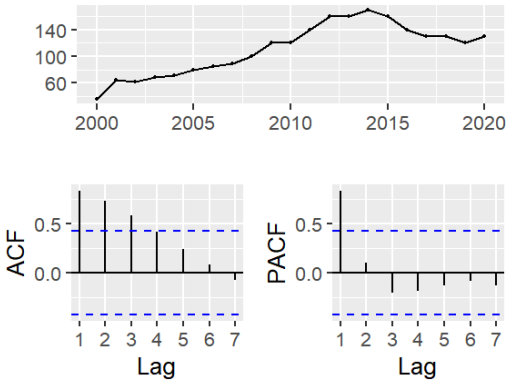
(Figure 3.13)



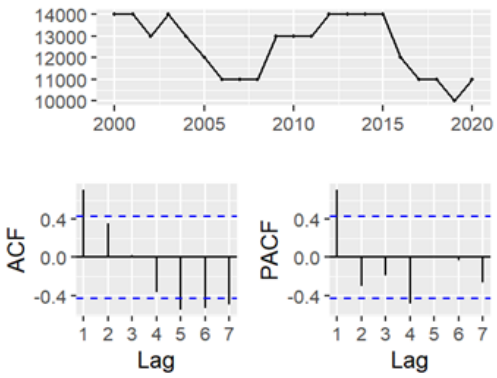
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.14)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.15)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.16)



Time profile, ACF, PACF of mortality number

(Figure 3.17)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model.
- Time profile shows the variation in the trend.
- ACF outside the threshold bound then it is considered as significant.
- PACF outside the threshold bound then it is considered as significant.

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Bangladesh	Estimated Incident number	0	2	0	328.9	-344.7053	1256.395	642.4305
	Estimated TB and HIV incident number	0	2	0	148.81	-0.0981	10.9631	10.96317
	Estimated mortality number of TB cases excluding HIV	0	1	0	394.11	4.4452	4062.363	3385.398
	Estimated mortality number of TB and HIV cases	0	1	0	198.79	7.4292	32.3463	19.8101
	Estimated mortality number	0	1	1	390.17	-1188.769	3628.134	2933.527

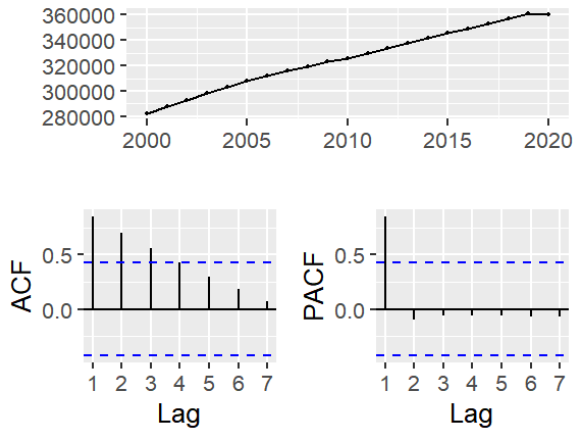
Table 3.10

The annual incidence rate of Tb of Bangladesh's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

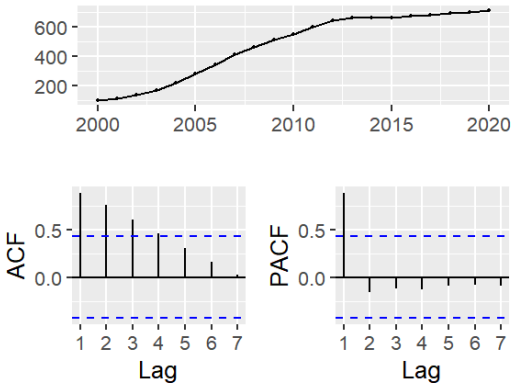
Prediction table

Bangladesh					
	2021	2022	2023	2024	2025
Estimated Incident number	359000	358000	357000	356000	355000
Estimated TB and HIV incident number	720	730	740	750	760
Estimated mortality number of TB cases excluding HIV	41650	39300	36950	34600	32250
Estimated mortality number of TB and HIV cases	170	170	170	170	170
Estimated mortality number	49462.55	49462.55	49462.55	49462.55	49462.55

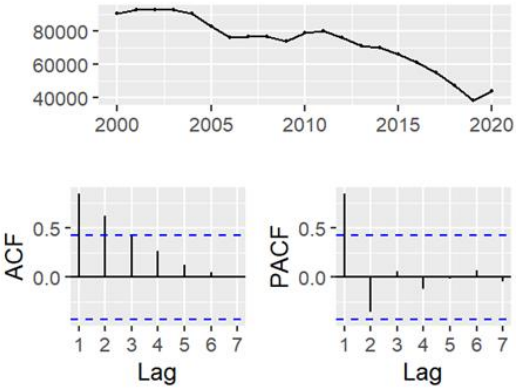
Table 3.11



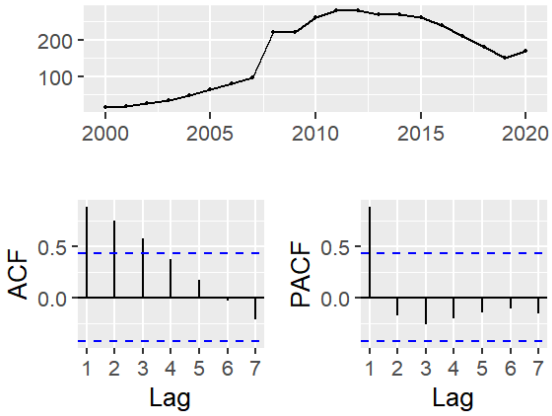
Time profile, ACF, PACF of Incidence number (Figure 3.18)



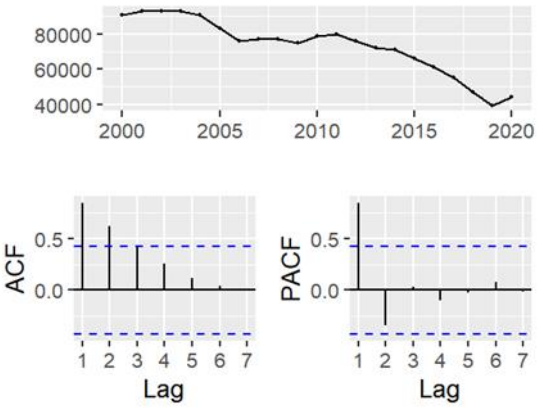
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.19)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.20)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.21)



Time profile, ACF, PACF of mortality number

(Figure 3.22)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Bhutan	Estimated Incident number	1	0	0	265.51	-2.2566	116.1938	93.6402
	Estimated TB and HIV incident number	0	1	0	159.95	-2.0928	12.24745	-23.56168
	Estimated mortality number of TB cases excluding HIV	0	0	0	178.44	0.000	15.4009	12.6077
	Estimated mortality number of TB and HIV cases	0	1	0	110.59	-0.5232	3.5657	1.6672
	Estimated mortality number	1	0	0	182.34	-0.19175	16.0337	12.78799

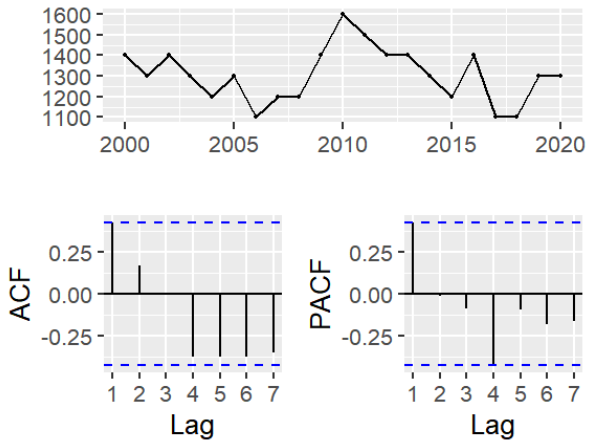
Table 3.12

The annual incidence rate of Tb of Bhutan's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

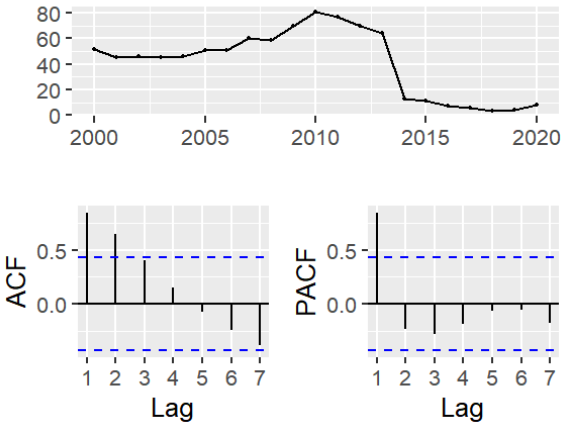
Prediction table

Bhutan					
	2021	2022	2023	2024	2025
Estimated Incident number	1304.453	1306.327	1307.115	1307.447	1307.586
Estimated TB and HIV incident number	8	8	8	8	8
Estimated mortality number of TB cases excluding HIV	140.9524	140.9524	140.9524	140.9524	140.9524
Estimated mortality number of TB and HIV cases	1	1	1	1	1
Estimated mortality number	159.9121	155.5849	153.7287	152.9325	152.5910

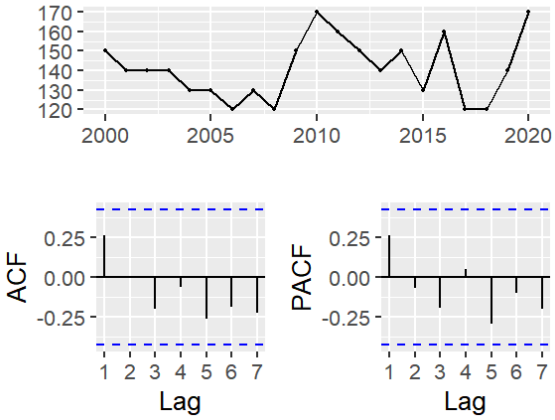
Table 3.13



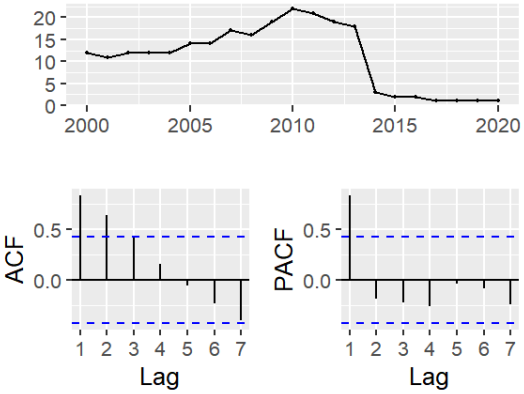
Time profile, ACF, PACF of Incidence number (Figure 3.23)



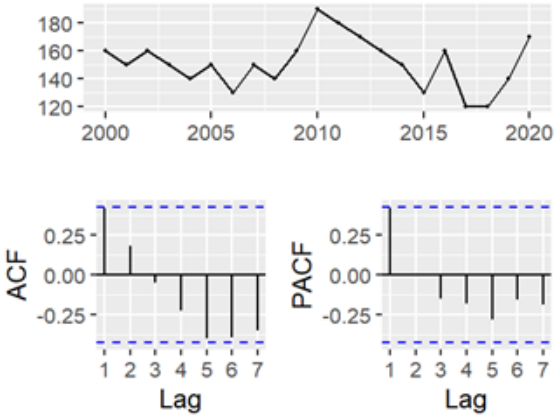
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.24)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.25)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.26)



Time profile, ACF, PACF of mortality number (Figure 3.27)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Pakistan	Estimated Incident number	0	2	1	359.1	-604.5552	2578.372	1266.636
	Estimated TB and HIV incident number	0	1	0	376.45	660.9581	2746.015	1558.101
	Estimated mortality number of TB cases excluding HIV	0	1	0	331.25	2.545237	843.8251	664.45
	Estimated mortality number of TB and HIV cases	0	1	0	324.6	94.95267	751.2351	357.4289
	Estimated mortality number	0	1	0	337.86	-330.8095	1046.6	621.5714

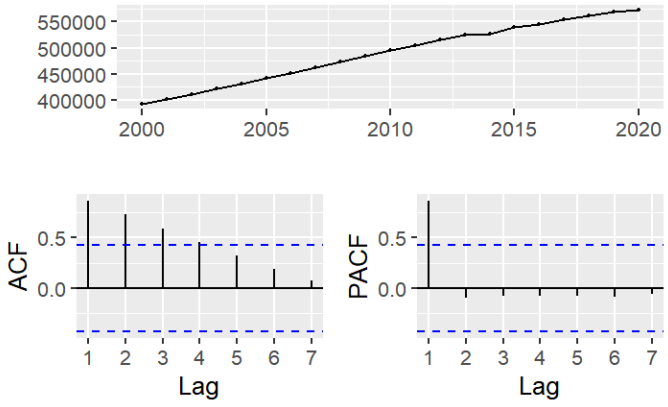
Table 3.14

The annual incidence rate of Tb of Pakistan's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

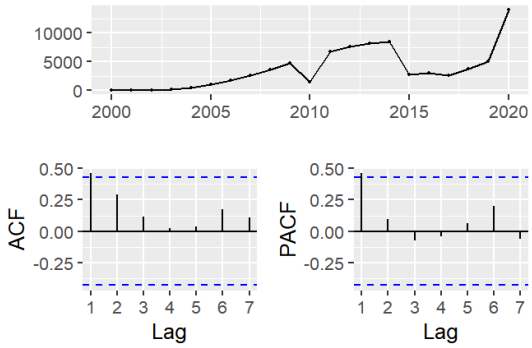
Prediction table

Pakistan					
	2021	2022	2023	2024	2025
Estimated Incident number	579926.9	586853.9	593780.8	600707.8	607634.7
Estimated TB and HIV incident number	14000	14000	14000	14000	14000
Estimated mortality number of TB cases excluding HIV	43550	43100	42650	42200	41750
Estimated mortality number of TB and HIV cases	2000	2000	2000	2000	2000
Estimated mortality number	46000	46000	46000	46000	46000

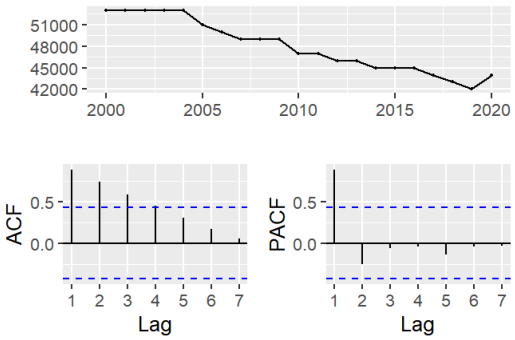
Table 3.15



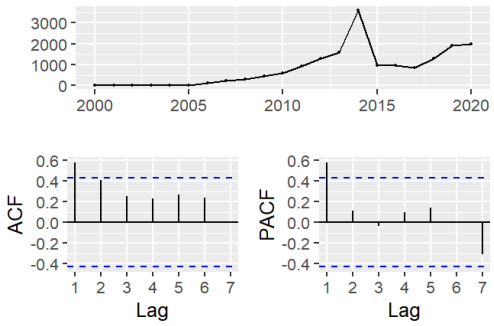
Time profile, ACF, PACF of Incidence number (Figure 3.28)



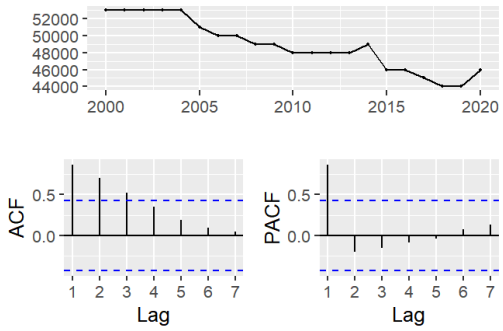
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.29)



Time profile, ACF, PACF of mortality with HIV number



Time profile, ACF, PACF of mortality without HIV number (Figure 3.31)



Time profile, ACF, PACF of mortality number (Figure 3.32)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
India	Estimated Incident number	0	2	0	408.7	-4886.485	10276.42	7873.533
	Estimated TB and HIV incident number	1	1	0	450.99	1121.476	15933.26	12346.48
	Estimated mortality number of TB cases excluding HIV	0	2	0	420.6	2925.655	14001.32	5767.206
	Estimated mortality number of TB and HIV cases	0	1	0	417.87	4.823807	7357.245	4890.538
	Estimated mortality number	0	1	0	453.58	34.39998	17649.12	9767.733

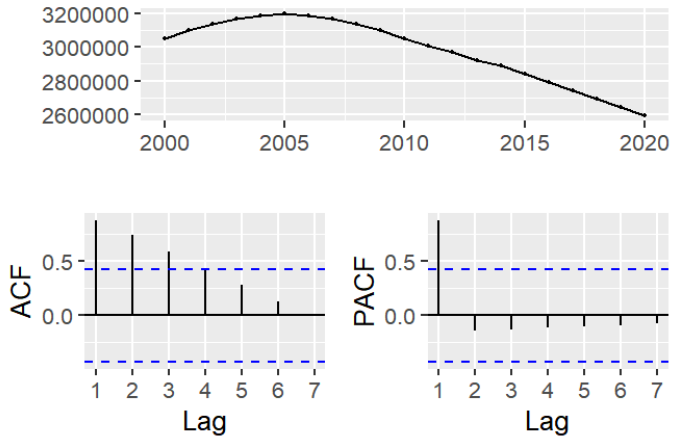
Table 3.16

The annual incidence rate of Tb of India's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

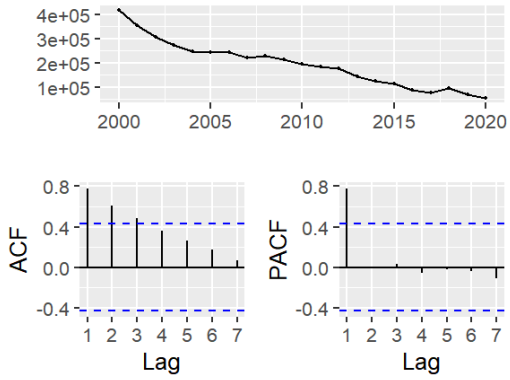
Prediction table

India					
	2021	2022	2023	2024	2025
Estimated Incident number	2540000	2490000	2440000	2390000	2340000
Estimated TB and HIV incident number	34897.218	15880.002	3534.873	23122.677	42785.683
Estimated mortality number of TB cases excluding HIV	550000	607000	664000	721000	778000
Estimated mortality number of TB and HIV cases	6700	2400	1900	6200	10500
Estimated mortality number	493600	483200	472800	462400	452000

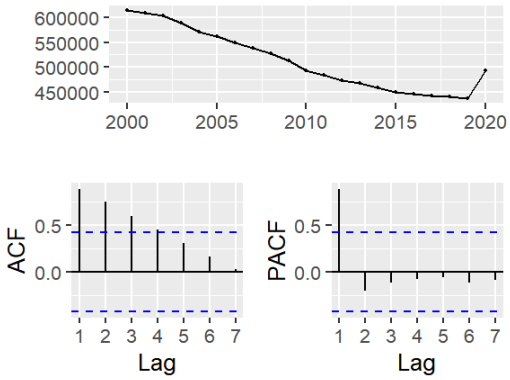
Table 3.17



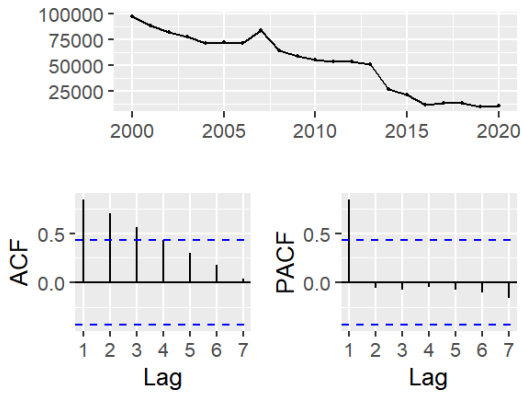
Time profile, ACF, PACF of Incidence number (Figure 3.33)



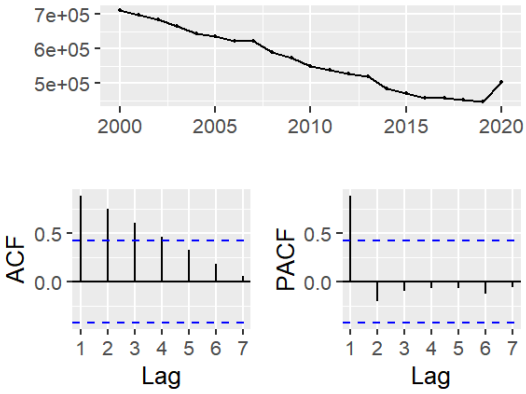
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.34)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.35)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.36)



Time profile, ACF, PACF of mortality number (Figure 3.37)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Nepal	Estimated Incident number	2	1	0	327.3	-39.4421	695.6819	547.3757
	Estimated TB and HIV incident number	0	1	1	319.05	-168.4186	618.2626	456.324
	Estimated mortality number of TB cases excluding HIV	0	1	0	330.22	1.461904	822.3353	696.7
	Estimated mortality number of TB and HIV cases	1	0	0	264.78	21.98327	113.0481	83.08083
	Estimated mortality number	0	1	0	318.11	1.511904	607.5324	501.5119

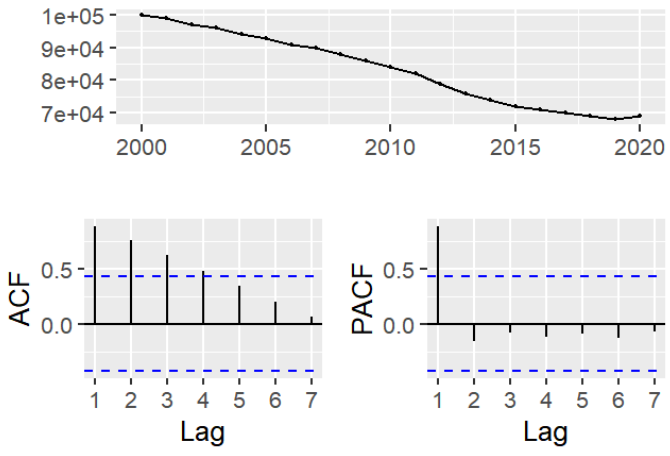
Table 3.18

The annual incidence rate of Tb of Nepal's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

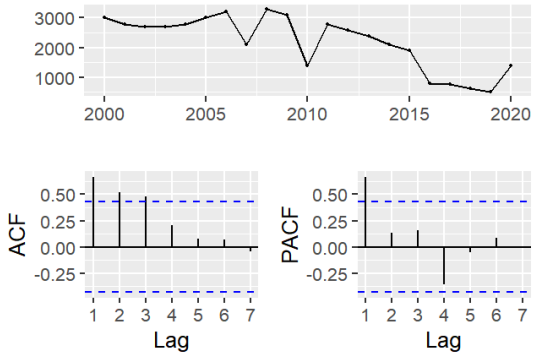
Prediction table

Nepal					
	2021	2022	2023	2024	2025
Estimated Incident number	68783.08	69275.33	69325.08	69622.43	69755.26
Estimated TB and HIV incident number	1051.281	1051.281	1051.281	1051.281	1051.281
Estimated mortality number of TB cases excluding HIV	15300	14600	13900	13200	12500
Estimated mortality number of TB and HIV cases	202.1984	194.6866	187.4539	180.4899	173.7846
Estimated mortality number	15250	14500	13750	13000	12250

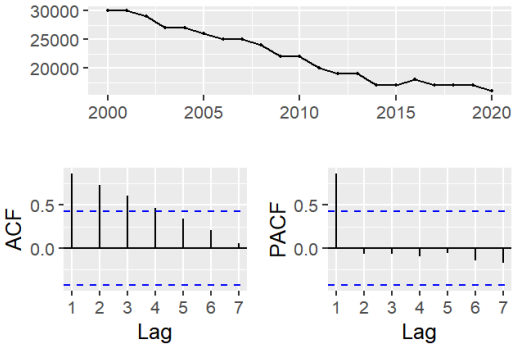
Table 3.19



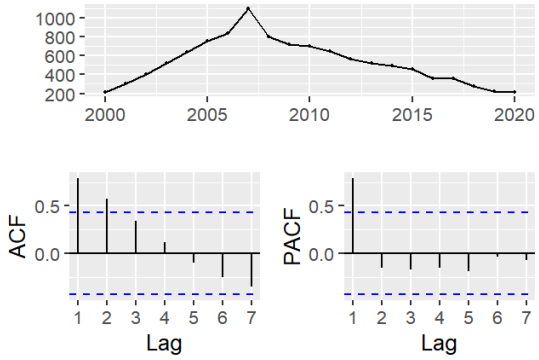
Time profile, ACF, PACF of Incidence number (Figure 3.38)



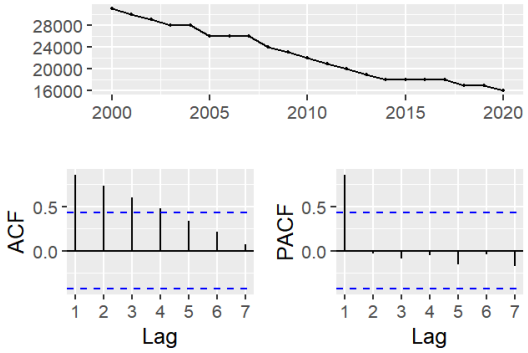
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.39)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.40)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.41)



Time profile, ACF, PACF of mortality number (Figure 3.42)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Time series model

Country	Estimates	Order			AIC	ME	RMSE	MAE
		p	d	q				
Sri Lanka	Estimated Incident number	0	1	0	289.02	95.80952	308.6178	95.80952
	Estimated TB and HIV incident number	1	0	0	193.69	-0.10230	21.00134	14.06102
	Estimated mortality number of TB cases excluding HIV	0	1	0	272.74	-26.60476	205.4266	98.15714
	Estimated mortality number of TB and HIV cases	1	0	0	143.4	-0.033150	6.350372	4.928525
	Estimated mortality number	0	1	0	272.33	-29.93333	203.3414	91.97143

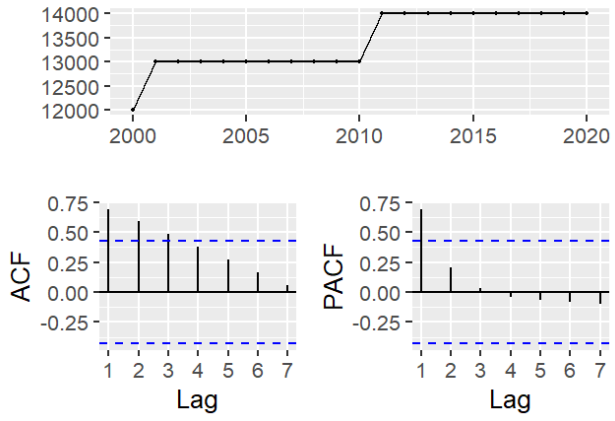
Table 3.20

The annual incidence rate of Tb of Sri Lanka's incident number, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number is predicted for 5 years from 2020 to 2021 which is given below,

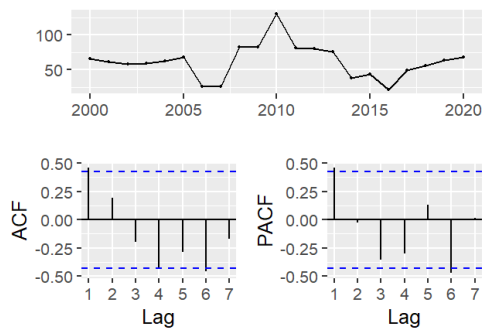
Prediction table

Sri Lanka					
	2021	2022	2023	2024	2025
Estimated Incident number	14000	14000	14000	14000	14000
Estimated TB and HIV incident number	64.71543	63.25588	62.60730	62.31910	62.19103
Estimated mortality number of TB cases excluding HIV	740	740	740	740	740
Estimated mortality number of TB and HIV cases	23.24416	20.98563	20.09940	19.75166	19.61521
Estimated mortality number	770	770	770	770	770

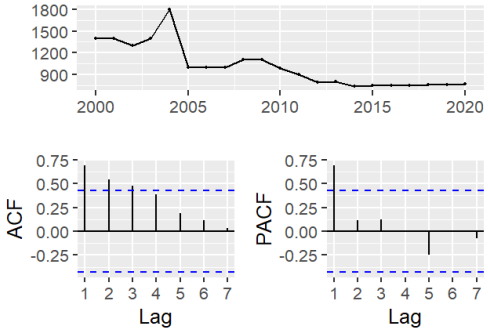
Table 3.21



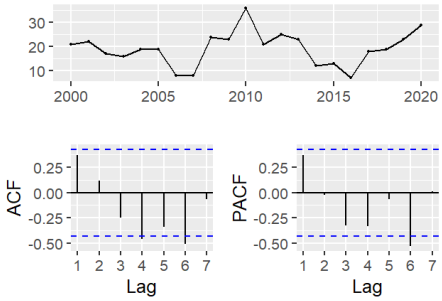
Time profile, ACF, PACF of Incidence number (Figure 3.43)



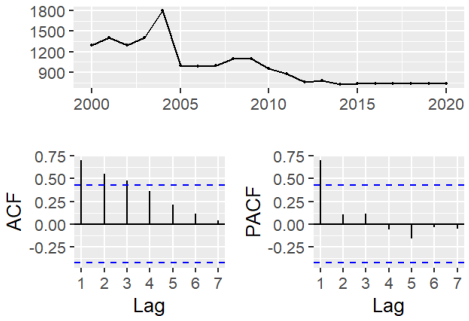
Time profile, ACF, PACF of Incidence with HIV number (Figure 3.44)



Time profile, ACF, PACF of mortality with HIV number (Figure 3.45)



Time profile, ACF, PACF of mortality without HIV number (Figure 3.46)



Time profile, ACF, PACF of mortality number (Figure 3.47)

- The goodness of fit of the model is assessed using AIC, ME, RMSE and MAE model
- Time profile shows the variation in the trend
- ACF outside the threshold bound then it is considered as significant
- PACF outside the threshold bound then it is considered as significant

Chapter

4. Summary and Conclusion

The following conclusions are drawn based on the statistical analysis.

- The time series is divided into test and train data and the autoregressive integrated moving average models are fitted for given data. The best model is selected using different criteria like AIC, ME, RMSE and MAE.
- The predicted annual incidence and mortality rates are the projected values under ideal conditions. These predictions can be compared with the actual incidence.
- The annual incidence number of TB, Incidence of TB and HIV, mortality number of TB cases excluding HIV, mortality number of TB and HIV cases, mortality number of TB of South Asian countries were predicted for 5 years from 2021 to 2025.
- The incidence rate of TB cases in Afghanistan, Bangladesh, Pakistan, and Sri Lanka is increasing. There is a decrease in the trend in India and Nepal. There is variation in the incidence rate of Bhutan.
- The incidence rate and HIV cases in Bangladesh and Pakistan show an increasing rate, and Afghanistan, Nepal, and Sri Lanka show that there is variation in incidence and HIV rate. It is rapidly declining in India and Bhutan.
- The mortality rate of Afghanistan, Bangladesh, Bhutan, and Pakistan varies. In Nepal and Sri Lanka, the mortality rate is decreasing. But only in India is the mortality rate of TB increasing.
- The mortality rate of TB and HIV varies in Afghanistan, Bangladesh, Bhutan, Pakistan, and India. And it is decreasing in Nepal and Sri Lanka.

References

1. Montgomery D.C, Jennings C.L,Kculachi M(2008). *Introduction Time Series Analysis and Forecasting*. John Wiley & Sons. New Jersey.
2. Box G.E.P, Jenkins.G.M, Reinsel G.C, Ljung G.M(2008). *Time Series Analysis Forecasting and Control*. John Wiley & Sons. NewJersey.
3. Sullivan A, Nathavitharana RR. Addressing TB-related mortality in adults living with HIV: a review of the challenges and potential solutions. Therapeutic Advances in Infectious Disease. 2022 Mar;9:20499361221084163.
4. Kober L, Solovic I, Littva V, Siska V. Prediction of the Epidemiological Situation of Tuberculosis in Slovakia by 2040-Data Update. Iranian Journal of Public Health. 2021 Nov;50(11):2229.
5. Carwile ME, Hochberg NS, Sinha P. Undernutrition is feeding the tuberculosis pandemic: a perspective. Journal of Clinical Tuberculosis and Other Mycobacterial Diseases. 2022 Mar 17:100311.
6. Aryee G, Kwarteng E, Essuman R, Nkansa Agyei A, Kudzawu S, Djagbletey R, Owusu Darkwa E, Forson A. Estimating the incidence of tuberculosis cases reported at a tertiary hospital in Ghana: a time series model approach. BMC Public Health. 2018 Dec;18(1):1-8.
7. Zheng Y, Zhang X, Wang X, Wang K, Cui Y. Predictive study of tuberculosis incidence by time series method and Elman neural network in Kashgar, China. BMJ open. 2021 Jan 1;11(1):e041040.
8. Liu K, Li T, Vongpradith A, Wang F, Peng Y, Wang W, Chai C, Chen S, Zhang Y, Zhou L, Chen X. Identification and prediction of tuberculosis in eastern China: analyses from 10-year population-based notification data in Zhejiang Province, China. Scientific reports. 2020 May 4;10(1):1-0.