University of Nairobi

College of Physical and Biological Sciences

School of Mathematics

Modeling and projection of HIV/AIDS epidemics in Ethiopia using ARIMA

By

Demissew Tsigemelak Gebreyohannes

I56/69903/2013

A project submitted in partial fulfilment of the requirements of the award of Master of Science Degree in Biometry at University of Nairobi

July 2015

Declaration

This project is my own original work and has not been submitted to any other university for the award of a degree.

Demissew Tsigemelak Gebreyohannes

Signature .... Date ....

This thesis has been submitted with my approval as University supervisor.

Dr. Nelson Owuor Onyango

School of Mathematics

University of Nairobi

Signature ....Date ....

Dedication

This project is dedicated to my beloved and humble families who have sacrificed all what they have for me.

Acknowledgement

I gratefully acknowledge my supervisor, Dr. Nelson O. Onyango for his prompt guidance, encouragement and support throughout my project work. He devoted his precious time to offer me practical statistical application. I am very much grateful to my beloved families; my father, Tsigemelak Gebreyohannes, who is a teacher by profession and my mother, Bizuwork Gebreyes and all my siblings especially my wonderful brothers, Getnet Tsigemelak and Dereje Tsigemelak, I owe you a big appreciation. It is because of your great sacrifices that I could make it and you have taken me this far. Moreover, my acknowledgement goes to Prof. Abiy Yenesew, who has been supporting and mentoring me in every aspect of life and Dr. Solomon Derese, Department of Chemistry, University of Nairobi, for their continuous support and advice. My acknowledgement further goes to all my compatriots, Tsegaye Deyou, Yoseph Atilaw and Fozia Ali, who are PhD students and Dr. Negera Abdissa and Ms. Sinidu Makonnen, who have showed me their consideration in every aspect. I should not cease expressing my gratefulness to all who have helped me directly or indirectly to successfully complete my study. I would also like to thank The Director, School of Mathematics, Prof. Weke Patrick Guge O. and all staff of the school for their cordiality and every consideration they have shown me. Praise to Almighty God who has helped me to overcome all the stumbling blocks even if the road was so tough.

Abstract

HIV/AIDS remains one of the lethal diseases and leading global health predicament. Ethiopia is one of the Sub-Saharan countries most affected by the HIV pandemic with a prevalence of 1.5% among adults and it is one of the top 22 countries with the highest number of pregnant women living with HIV/AIDS. This study was conducted with objective of formulating a model to determine the trend, prevalence and projecting HIV/AIDS epidemics in Ethiopia. Data were obtained from UNAIDS and Ministry of Health bulletin in Ethiopia. The data was analyzed using Autoregressive Integrated Moving Average (ARIMA) time series analysis model and the ARIMA (2,3,2) appeared to be providing the best fit for the observed data. The trend revealed that the HIV/AIDS prevalence was increasing in alarming rate from approximately mid 1990s and reached its climax in the years 2002 to 2004 and decreased onward. The prediction showed that the prevalence of HIV/AIDS would decrease in Ethiopia for the next 5 years. Both the trend and the prevalence showed that the status of HIV/AIDS in Ethiopia could be controllable. Further investigation including research on significant contributing factors and predictors of the disease will be required to perfect this study. It would also be good if this model can be compared to other models used in HIV/AIDS research.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

ACF	Autocorrelation Function
ACVF	Auto covariance Function
AIC	Akaike's Information Criterion
AIM	AIDS Impact Model
ANC	Antenatal Clinic
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
EPP	Estimating and Projecting Package
HIV/AIDS	Human Immuno Virus/Acquired Immunodeficiency Syndrome
MA	Moving Average
MOH	Ministry of Health, Ethiopia
PACF	Partial autocorrelation Function
UN	United Nation
UNAIDS	United Nations Aid and Development
WHO	World Health Organization

Chapter 1

Introduction

1.1 Background

The acquired immuno deficiency syndrome (AIDS) remains one of the lethal diseases which cause millions of deaths in a year since the first case report in 1981 [1]. Much has been done on the containing of HIV/AIDS in different parts of the world but this epidemic infectious disease remains one of the onerous health problems issue in developing countries affecting the working age cohort group of the population. This is due to mainly its intractable mode of transmission and nature of the disease. UNAIDS and WHO (2010) has reported that more than 40 million people have been infected with HIV worldwide since the beginning of the epidemic and an estimated 70% of those infected people live in Africa [2].

Young people are disproportionately affected by HIV globally of which 25% of infected persons are aged between 10 and 24 years. Those aged 15–24 years have imminently 35% probability for new infections, resulting in 900,000 new infections occurring annually [3].

Even though the prevalence of HIV/AIDS in Ethiopia is declining as a result of using antiretroviral therapy (ART) medicine by people but it is still the major public health problem with a prevalence of 2.3%. The prevalence of young women in sub Saharan Africa have almost 8 times greater of the same age man, and their annual HIV incidence is an estimated 8% [5].

The epidemiological estimates of HIV/AIDS infection and the mortality rate with this disease are crucial for planning and monitoring of trends at the national, regional, and worldwide level. Continued effort is mandatory to design a better way of improving the validity of the estimates and developing public health policy [9].

1.2 Profile of Ethiopia

Ethiopia is the second populous country in Africa and one of the largest HIV-infected population in the world [4]. According to World Bank (2005) [21], Ethiopia is defined as a low-income country with a low per capital gross national income and the country national income is the second lowest world-wide and its main economic activity depends on agriculture.

Ethiopia is one of the most afflicted sub-Saharan countries by HIV/AIDS epidemic with a prevalence of 1.5% among adults, and it is one of the top 22 countries with the highest number of pregnant women affected with HIV/AIDS. CSA (2005) [6] showed that 1.4% of the Ethiopian population is infected with HIV and it is one of the highest rates in Sub-Saharan Africa countries. The HIV/AIDS pandemic continues to present a major health challenge for sub-Saharan Africa and in Ethiopia, adult HIV prevalence in 2009 was estimated to be between 1.4% and 2.8% [7]. Report from UNAIDS (2010) [3] pointed out the number of adults and children in sub-Saharan Africa contracted with the human immunodeficiency syndrome (HIV) has reduced from 2.2 million to 1.8 million in years 2001 and 2009 respectively.

1.3 Statement of the problem

The Federal Republic of Ethiopia government has taken and put in place a lot of preventive and controlling measures to address the HIV/AIDS pandemic through creation of awareness to society, training many health extension workers to a village level to educate the society against the repercussion effect of the disease and increase the budgetary allocation to control HIV/AIDS prevalence. Despite all these measures taken by the government, the incidence and prevalence of HIV/AIDS is still high compared to other countries with minimum rates even if it goes down from time to time sporadically. There is therefore a need of considering the past and present preventive measures to come across precise and accurate information on the nature of trends of HIV/AIDS in order to develop better planning and accurate evaluation on the impact of these preventive interventions and progress in the fight against HIV/AIDS. Mathematical modeling is deemed as the only likely way of measuring the efficacy of HIV intervention in order to predict, assess the past and future events and explaining the impact of the disease.

The availability of such precise estimates and projections is thought essential in supporting decision-makers to understand the magnitude of the HIV/AIDS problem and supporting efforts to improve prevention and health-care programs. Projecting the future prevalence and its impact of HIV/AIDS demands a sound methodology for projecting the number of future HIV infections and determining the impact of those infections on the future pattern of adult and child deaths.

1.4 Objectives

The overall objective of this study is to establish a model which helps to determine the magnitude, prevalence and status of HIV/AIDS epidemic in Ethiopia which could potentially be used as a tool to monitor the status of HIV/AIDS in the country. The specific objectives of the project are:

- to recognize the trend and prevalence of HIV/AIDS for the next 5 years.
- to predict the number of people that will be infected by the disease in the country.

1.5 Justification of the study

The knowledge of the prevalence of HIV/AIDS disease helps in providing information on designing appropriate controlling and preventive integrative measures in order to bring a long term solution. The study will be helpful in order to recognize the status and predict the prevalence of HIV/AIDS so that it gives insight to take pre-cautionary measures. It also guides policy makers to make appropriate intervention on how to control and minimize the repercussion effect of HIV/AIDS.

Chapter 2

Literature Review

2.1 Theory of ARIMA

Time series analysis helps to come up with a model using a historic data and allows predicting the future. This includes naive method, moving average, trend analysis, exponential smoothing and the autoregressive integrated moving average (ARIMA). These methods are suitable in the description of the general tendencies or patterns without considering the factors of affecting the variable to be predicted [8].

In univariate time series, forecasting is based on the past values of the variables being forecast. Zhang (2003) [11] explained the autoregressive integrated moving average model (ARIMA) as the future value of a variable of interest is the linear combination of a number of previous observations and random error and the underlying process that produces time series is given by:

$$Y_t = \phi_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$$

,

Where Y_t and ε_t the current observation value and random error at time t respectively; $\beta_i (i = 1, 2, ..., p)$ and $\theta_j (j = 1, 2, ..., q)$ are the parameters of the model. p and q are the order of the autoregressive and moving average models respectively and the random errors, ε_t is assumed to be independent and identically distributed with constant mean,μ and variance of σ^2.

Time series data does not follow the normality assumption that the error and successive observation are uncorrelated each other and this effect which is autocorrelation, biases the standard error connected with regression estimates of slope parameters and the usual normal statistical tests like t-test will be invalid. The Box-Jenkins method helps to relate the time sequenced observation statistically to the others in the same series [14]. The Box-Jenkins models are particularly suitable for short term predicting as most ARIMA models gives greater importance on the recent past than the distant past.

2.2 Mathematical models for HIV/AIDS

Several mathematical and statistical models have been used to estimate and project incident HIV infection [10]. The AIDS Impact Model (AIM) is a computer software program used for projecting the trends in the impact of the AIDS epidemic [12]. This model can also be utilized for projection of future number of HIV infections, cases and deaths through HIV/AIDS taking into consideration adult HIV prevalence. Much has not been done on time series analysis of HIV/AIDS epidemics but a study in Ghana showed that time series modeling on trend analysis of past growth patterns revealed an increase in new cases of HIV infection in the Northern part of the country, with the greatest increase happening among persons aged 30 years and over. The epidemic in the southern sector appeared to be constant [10].

A time series forecasting which predicts the future values of the observed time series variables by extrapolating trends and patterns from the past values of the series was carried out in South Africa with the purpose of using the available antenatal HIV seroprevalence data to predict the future trend of the HIV epidemic. This study used quadratic model and found out that the coefficient of determination R^2 is 0.97. It also indicated that the time series forecasting exercise using the quadratic model and trend exhibited there was likelihood decreasing of HIV trend beyond the year 2010 [13].

Time series models try to predict ahead the epidemiological behaviors through modeling and considering the past surveillance data. Many researchers have been applying different time series models to forecast epidemic incidence in previous studies. Exponential smoothing and generalized regression methods by international groups in 2010 were used to forecast the epidemic infection and incidence of cryptosporidiosis respectively.
The Univariate time series modeling and Estimating and Projection Package (EPP) ap-

proaches forecasted an increase in incident HIV infection over a three-year period 2008, 2009 and 2010 in a study which was carried out in Ghana, whereas the Box-Jenkins model projected an increase in incident HIV infection among males for the three-year period and the EPP models forecasted a decline in incident of HIV infection by 2010 [10].

Chapter 3

Methodology

3.1 Data source

The data comprised annual data of HIV/AIDS infected people from 1990 to 2013 in Ethiopia. The data embraced the whole region of Ethiopia and most of the national and regional HIV/AIDS estimates made for Ethiopia were extracted from UNAIDS data source and Ministry of Health in Ethiopia.

3.2 The Box-Jenkins methodology

This method embraces three iterative procedures of model specification, model fitting and model diagnostics.

3.2.1 Model specification

The central idea behind model identification is a time series derived from ARIMA process has some sort of theoretical autocorrelation properties. Fitting the empirical autocorrelation patterns with the theoretical ones helps to identify the potential tentative model for the given time series data [11]. In this step, transformation of observed time series to stationary is inevitable. A stationary time series has constant mean, variance and covariance statistical characteristics over time. Box and Jenkins (1976) recommended the autocorrelation and partial autocorrelation function as main tool to identify the order of ARIMA model [14].

The general $ARIMA(p, d, q)$ model is explained by:

$$\beta(B) \bigtriangledown^d Y_t = \theta_0 + \theta(B)\varepsilon_t \tag{3.1}$$

Where

$$\beta(B) = 1 - \beta_1 B - \beta_2 B^2 - .... - \beta_p B^p;$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q;$$

$$\bigtriangledown Y_t = Y_t - Y_{t-1};$$

$$\bigtriangledown^d Y_t = \bigtriangledown(\bigtriangledown^{d-1} Y_t);$$

$$BY_t = Y_{t-1};$$

$$\beta^n Y_t = Y_{t-n};$$

The autocovariance function, $Cov(Y_t, Y_s)$,of a stationary time series Y_t having mean $E(Y_t = \mu$ and variance $Var(Y_t) = E(Y_t - \mu)^2 = \sigma^2$, which are constant, and the covaraince is symbolized by $\gamma(k)$ and given by:

$$\gamma(k) = Cov(Y_t, Y_{t+k}) = E(Y_t - \mu)(Y_{t+k} - \mu), \tag{3.2}$$

where k is an integer and $\gamma(k)$ is the autocovariance function (ACVF) at lag k (Wei, 2006) [15]. The covariance, $Cov(Y_t, Y_s)$, is a function of the time series difference $|t - s|$. As the size of ACVF depends on the units which Y_t is measured, it is standardized for rendering suitable interpretation and producing a function called the Autocorrelation function (ACF), given by:

$$\rho(k) = Corr(Y_t, Y_{t+k}) = \frac{\gamma(k)}{\gamma(0)} \tag{3.3}$$

$$\rho(k) = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{VarY_t}, \sqrt{VarY_{t+k}}} \tag{3.4}$$

$\rho(k)$ is the autocorrelation funciton.

The ACF of a stationary time series is a significant tool for assess its properties.

The sample autocorrelation is a good indicator of the order of the process in MA (q) models since the autocorrelation function is zero for lags beyond q. However, a different function is needed to determine the order of autoregressive models as AR (p) model does not turn into zero after a certain number of lags in autocorrelation function as the model attenuates instead of cut off. Such a function can be described as the correlation between Y_t and Y_{t-k} excluding the effect of the intervening variables, $Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-k+1}$.

This coefficient is called the partial autocorrelation at lag k and will be designated by β_{kk}.

According to Cryer and Chan (2008), the partial autocorrelation function for non-normal distribution at lag K is then defined by the correlation between the prediction errors which is given by [16]:

$$\beta_{kk} = corr((Y_t - E(Y_t Y_{t+1}, , Y_{t+k-1}), Y_{t+k} - E(Y_{t+k}|Y_{t+1}, , Y_{t+k-1})))$$

Where $E(Y_t|Y_{t+1}, , Y_{t+k-1})$ and $E(Y_{t+k}|Y_{t+1}, , Y_{t+k-1})$ are the predictions of Y_t and Y_{t+k} respectively.

Hence, β_{kk} , $K \geq 2$ is the correlation of the two residuals obtained after regressing Y_{t+k} and Y_t on the intervening observations. In other words, it is the correlation between prediction errors. Generally, the sample partial autocorrelation function is given by Levinson (1947) [20] and Durbin (1960) [19]:

$$\beta_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \beta_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \beta_{k-1,j}\rho_j} \tag{3.5}$$

Where

$$\beta_{k,j} = \beta_{k-1,_j} - \beta_{kk}\beta_{k-1,k-j} for : j = 1, 2, , k - 1$$

Therefore, the PACF can help to determine the order of an AR (p) process as the ACF helps to determine the order of an MA(q) process. For an AR(p) model, the PACF "drops off" to zero after the p^{th} lag.

Table 3.1: Typical features of a sample ACF and sample PACF for AR and MA models

Conditional Mean Model	ACF	PACF
AR(p)	Tails off gradually	Cuts off after p lags
MA(q)	Cuts off after q lags	Tails off gradually
ARMA(p, q)	Tails off gradually	Tails off gradually

3.2.2 Model fitting

After the order of the $ARIMA(p, d, q)$ for a given time series data has been specified, the parameters should be estimated. Model fitting engages estimating the parameters of the ARIMA model from the observed time series $Y_1, Y_2, ..., Y_n$ using method of moment, least square and maximum likelihood. It generally involves parameter estimates, fitted values, residuals, significance check of estimated parameters, stationary and invertibility conditions, and correlation check of estimated parameters. Here, Maximum likelihood estimation method in relation to least square estimation will be discussed as it is the most efficient method of estimation in time series data. Method of moment is less efficient even if it is relatively easy to calculate.

1. Maximum Likelihood Estimation

   Maximum likelihood estimation offers a unified approach for parameters estimation for ARMA process. It is the most efficient and preferred method of parameter estimation. It also offers a standard way to deal with models of stochastic time processes. In time series analysis because of interrelated observation, the likelihood approach using probability density function is given as follows:

   Assuming the error follows white noise which is $\varepsilon \sim N(0, \sigma^2)$, the joint probability distribution function: $f(\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_n) = f(\varepsilon_1)f(\varepsilon_2), ..., f(\varepsilon_n)$ instead of $f(Y_1, Y_2, ..., Y_n)$ as there is dependency between time series observation which will not be written as a multiplication of marginal probability density functions.

   For a general ARIMA (p, q) stationary process:

   $$\hat{Y}_t = \beta_1 \hat{Y}_{t-1} + \beta_2 \hat{Y}_{t-2} + ... + \beta_p \hat{Y}_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \tag{3.6}$$

   Where

   $$\hat{Y}_t = Y_t - \mu$$

   [15].

   The joint probability distribution function of $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is given by:

   $$f(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n | \mu, \beta, \theta, \sigma_e^2) = (2\pi\sigma_e^2)^{\frac{-n}{2}} \exp\left\{\frac{-1}{2\sigma_e^2} \sum_{t=1}^{n} \varepsilon_t^2\right\} \tag{3.7}$$

   Let $Y = (Y_1, ..., Y_n)$ and assume that the initial conditions Y_* and ε_*, The conditional log-likelihood function is given by:

$$\ln(L_*)(\mu,\beta,\theta,\sigma_e^2) = (\frac{-n}{2}\ln(2\pi\sigma_e^2)) - (\frac{S_*(\mu,\beta,\theta)}{2\sigma_e^2}) \tag{3.8}$$

Where $S_*(\mu,\beta,\theta) = \sum_{t=1}^{n} \varepsilon^2(\mu,\beta,\theta|Y,Y_*,\varepsilon_*$ is the conditional sum of square.

For the specifying the initial condition, the assumptions of Y_t stationary and $\varepsilon_t \sim N(0,\sigma_e^2)$ random variable, the unknown Y_t can be replaced by the sample mean $\bar{Y}$ and the unknown ε_t by its expected value of 0. For the model ARMA (p, q), it may be assumed $\varepsilon_p = \varepsilon_{p-1} = ... = \varepsilon_{p+q-1} = 0$, and calculate ε_t for $t \geq (p+1)$, gives

$$S_*(\beta,\mu,\theta) = \sum_{t=p+1}^{n} \varepsilon^2(\beta,\mu,\theta|Y) \tag{3.9}$$

It is this form of an equation that most computer programs also utilize in estimation.

3.2.3 Model diagnostics

This deals with the goodness of fit a model or checking the fit of the model which is an iterative process and it is also imperative if the model can be improved [16]. Model adequacy is assessed through checking whether the model assumptions are satisfied and it is carried out after estimation of parameter [15]. The principal assumption of the time series include the error ε_t is white noise, this is to say that the errors are uncorrelated random shocks having mean zero and constant variance. Hence, this indicates that the residuals are estimates of the unobserved white noise ε^s.

1. Residual analysis

   In any statistical models, residuals can be calculated as a difference between the observed (actual) and predicted value. If the residuals can nearly attain white noise properties, this reasonably would indicate that the model is appropriately specified and the parameter estimates are convincingly close to the true values. They should behave roughly like independent, identically distributed normal variables with zero mean and constant variation. Deviations from these properties can help us discover a more appropriate model.

2. Normality and Independence

   The normality assumptions can be checked by histograms and quantile-quantile (Q-Q) plot of the residuals. The hypothesis test of the normality can also be confirmed using Shapiro-Wilk test and independence is using runs test.

3. Residual autocorrelation and Partial autocorrelation Function

   The residuals of ACF and PACF should not be forecastable, that is the terms of the residual ACF and residual PACF should all approximately lie between the 95% confidence limit. If this is not the case, there are elements of residuals which can be forecastable.

4. Portmanteau Test (Ljung-Box-Pierce Q-statistic)

   This test helps to determine if there is any pattern left in the residual which can be modeled. This can be achieved by testing the significance of the autocorrelations up to a certain lag.

$$Q(k) = n(n+2)\sum_{i=1}^{k} \frac{r_j^2}{n-j} \tag{3.10}$$

   Where r_j is the j^{th} residual autocorrelation, n is the total number of data points and k is the lag.

3.2.4 Model selection criteria

In any data analysis, a given data set may sufficiently be represented by fitting model. It is sometimes easy to choose the best model but it is not always be the case. The model identification tools such as ACF and PACF are only utilized for identifying the most likely adequate models [15]. Residuals from adequate models are approximately white noise and indistinguishable in terms of these functions. For a given data set, when there are multiple adequate models, the selection criterion is normally based on summary statistics from residuals computed from a fitted model or on forecast errors calculated the out-sample forecasts.

Based on residuals, the following model selection criteria are used:

1. Akaike's Information Criteria (AIC)

   If a statistical model of k parameters is fitted to the observed data, the quality of the model fitting can be assessed using information criteria. One of the criteria is Akaike's information criterion which is given in the literature [15]

$$AIC = -2\ln(maximum \quad likelihood) + 2k \tag{3.11}$$

   Where k is the number of parameters in the model and the maximum likelihood estimates is given in equation 3.8 above. The value of AIC will be high with the number of model parameters (k).

2. Akaike's Bayesian Information Criteria (BIC)
   Akaike (1978) [22] and (1979) [23] has developed an extension of Bayesian of the minimum AIC, known as the Bayesian Information Criterion (BIC) and given by:

$$BIC = -2\ln(maximum \quad likelihood) + k\ln(n) \tag{3.12}$$

   Where n is the number of observations in the given stationary time series data and, k is the number of parameter. In similar fashion to AIC, the best model taking part in ARIMA (p,d,q) models is the one with the smallest BIC.

3.2.5 Forecasting

Forecasting in time series model involves uses of historical epoch data for the variable of interest that is going to be forecasted and it requires routine calculations to make use of a large number of events [11]. Forecasting helps to achieve one of the most important objectives in dealing with modeling exercise that able to predict the value of the random variable in the future from the currently existed one and get information in advance.

From the observed time series data, $Y_1, Y_2, ..., Y_t$, the forecasted value would be given by Y_{t+l}, where $l \geq 1$ and t is the forecast origin and l the lead time for the forecast. The value Y_{t+l} gives "l steps ahead" of the observed time series value Y_t [16]. Producing an optimum forecast with no or little error leads to minimum mean square error forecast. This forecast will generate an optimum future value having minimum error in terms of mean square error criterion [15].

1. Minimum mean square error forecasting
   The objective of minimum mean square error forecasting is producing an optimum predicts that has no error or as minimum error as possible which directs to the minimum mean square error forecast [16]. Based on this mean square error criterion, the forecast will produce an optimum future value with minimum error and the minimum mean square error forecast $\hat{Y}_t(l)$ which is given by:

$$\hat{Y}_t(l) = E(Y_{t+1}|Y_1, Y_2, ..., Y_t) \tag{3.13}$$

Equation 3.13 is derived from differentiating $E[(Y_{t+1} - \hat{Y}_t(l))^2|Y_1, Y_2, ..., Y_t]$ with respect to $\hat{Y}_t(l)$ and equating to zero. Given the $ARMA(p,q)$ time series model,

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_1 - \theta_2 \varepsilon_2 - ... - \theta_q \varepsilon_{t-q} \tag{3.14}$$

This implies that unknown Y_{t+l} is given by:

$$Y_{t+l} = \beta_1 Y_{t+l-1} + \beta_2 Y_{t+l-2} + ... + \beta_p Y_{t+l-p} + \varepsilon_t - \theta_1 \varepsilon_{t+l-1} - \theta_2 \varepsilon_{t+l-2} - ... - \theta_q \varepsilon_{t+l-q} \tag{3.15}$$

Hence, using equation 3.13 and 3.15, the minimum mean square error forecast for the Y_{t+l} is given by:

$$\begin{aligned}\hat{Y}_t(l) &= E[Y_{t+l}|Y_1, Y_2, ..., Y_t] \\ &= \beta_1 E[Y_{t+l-1}|Y_1, Y_2, ..., Y_t] + \beta_2 E[Y_{t+l-2}|Y_1, Y_2, ..., Y_t] + ... + \beta_p E[Y_{t+l-p}|Y_1, Y_2, ..., Y_t] \\ &- \theta_1 E[\varepsilon_{t+l-1}|Y_1, Y_2, ..., Y_t] - \theta_2 E[\varepsilon_{t+l-2}|Y_1, Y_2, ..., Y_t] - ... - \theta_q E[\varepsilon_{t+l-q}|Y_1, Y_2, ..., Y_t] \\ &= \beta_1 \hat{Y}_t(l-1) + \beta_2 \hat{Y}_t(l-2) + ... + \beta_p \hat{Y}_t(l-p) - \theta_1 E[\varepsilon_{t+l-1}|Y_1, Y_2, ..., Y_t] \\ &- \theta_2 E[\varepsilon_{t+l-2}|Y_1, Y_2, ..., Y_t]\end{aligned} \tag{3.16}$$

Where

$$E[\varepsilon_{t+l-j}|Y_1, Y_2, ..., Y_t] = \begin{cases} 0 & \text{if } l > j \\ \varepsilon_{t+l-j} & \text{if } l \leq j \end{cases} \tag{3.17}$$

$$\hat{Y}_t(l-j) = \begin{cases} \hat{Y}_t(l-j) & \text{if } l > j \\ Y_{t+l-j} & \text{if } l \leq j \end{cases} \tag{3.18}$$

The minimum mean square forecast error of $\hat{Y}_t(l)$ is given in the random shock model form as:

$$\begin{aligned} e_t(l) \quad &= Y_{t+1} - \hat{Y}_t(l) \\ &= \textstyle\sum_{j=0}^{l-1} \phi_j \varepsilon_{t+l-j} \end{aligned} \tag{3.19}$$

Where the $\phi_j (j = 1, 2, ...)$ weights are the functions of $ARMA(p,q)$ model parameters. The forecast error variance which is used to determine the confidence interval for Y_{t+l} under the assumption of $\varepsilon_t \sim N(0, \sigma^2)$ is given by:

$$Var(e_t(l)) = \sigma^2 \sum_{j=0}^{l-1} \phi_j^2 \tag{3.20}$$

The $\left(1-\frac{\beta}{2}\right)*100\%$ confidence interval of Y_{t+1} of $\phi \in (0,1)$ assuming the assumption holds, is given by:

$$\hat{Y}_t(l) \pm Z_{\frac{\alpha}{2}}\sqrt{Var(\varepsilon_t(l))} \tag{3.21}$$

where $Z_{\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)*100\%$ percentile of the standard normal distribution.

Model selection using forecast errors

The final choice of a model may rely on the goodness of fit like the residual mean square or information criteria. But, if the main objective of a model is to forecast future values based on the current and past values, then the criteria for model selection can be based on forecast errors [15]. If the forecast error l step ahead be,

$$e_l = Y_{n+l} - \hat{Y}_n(l) \tag{3.22}$$

where n is the forecast which is greater or equal to the length of the series. The comparison of the forecast error measures which help us to know how much we should rely on the chosen prediction method is based on the following statistics.

1. Mean percentage error (MPE), it is also called bias as it measures forecast bias. This is given by the mathematical formula:

$$MPE = \left(\frac{1}{j}\sum_{l=1}^{j}\frac{e_l}{Y_{n+l}}\right) \tag{3.23}$$

2. Mean square error (MSE)

$$MSE = \frac{1}{j}\sum_{l=1}^{j} e_l^2 \tag{3.24}$$

3. Mean absolute error (MAE)

$$MAE = \frac{1}{j}\sum_{l=1}^{j} |e_l| \tag{3.25}$$

4. Mean absolute percentage error (MAPE)

$$MAPE = \left(\frac{1}{j}\sum_{l=1}^{j}\left|\frac{e_l}{Y_{n+l}}\right|\right) \tag{3.26}$$

The model with the smallest MPE, MSE, MAE and MAPE will be selected the best model for forecasting. But, Hyndman and Koehler (2005) proposed the mean absolute scaled error become the standard measure for comparing forecast accuracy across multiple time series.

Chapter 4

Results and Discussions

4.1 Data analysis and results

A time series analysis model for HIV/AIDS was developed to examine the prevalence, trend and forecast the future in Ethiopia. Data from UNAIDS and Ministry of Health in Ethiopia was used for this study. The data was initially non-stationary and it was transformed to stationary through differencing. After making the data differencing three times, the data attained stationarity and the tentative model appeared to be $ARIMA(2,3,1)$ but the final model $ARIMA(2,3,2)$ was fitted using the information criterion. The final model was tested through different diagnostics methods and provided the best fit for the observed data. $ARIMA(2,3,2)$ was able to capture the most important features of the data. The fitted model is then used to forecast the number of people that will be infected with HIV/AIDS in Ethiopia.

4.1.1 Plotting the observed time series data against time

The first step in time series analysis is plotting the observed data against time to see whether the data has constant mean and variance. From the original graph , Figure (4.1), it can be observed that the data is non-stationary and it is imperative to change this data to stationary through differencing to deal with time series. It is not intractable to see the possible change in mean and dispersion of the data over time series. The trend of the mean may be upward and downward, so the mean is definitely varying and the series is non-stationary. This non-stationarity can also be observed from the plot of ACF and PACF. Since the original data demands transformation for stationarity, first

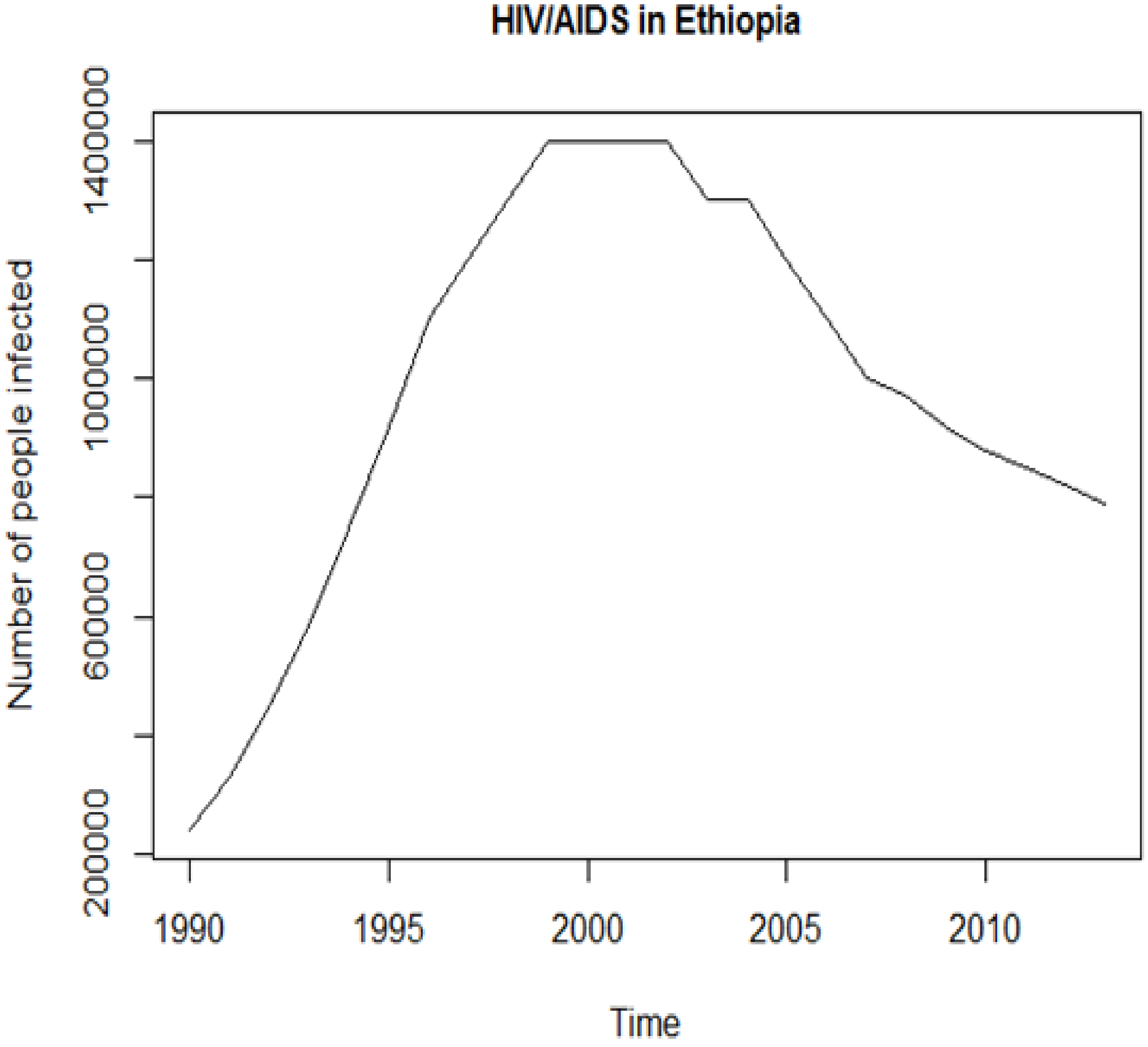

Figure 4.1: Time series plot of number of infected people with HIV/AIDS in Ethiopia

differencing was carried out and the following graphs was obtained. The resulting time series of first differencing of Figure (4.2) does not appear stationary and there is a need again to transform the data to stationary using second differencing. The time series of second differencing Figure (4.3) below does not appear to be stationary in mean and variance, as the level of the series stays differs over time, and the variance of the series appears and differencing is required further. Using the first and second differencing to make the non-stationary to stationary still does not make it stationary and it goes up to making third differencing to change this data to stationary. It can be observed from the graph in the Figure(4.4), the variation is constant over time and the data is stationary.

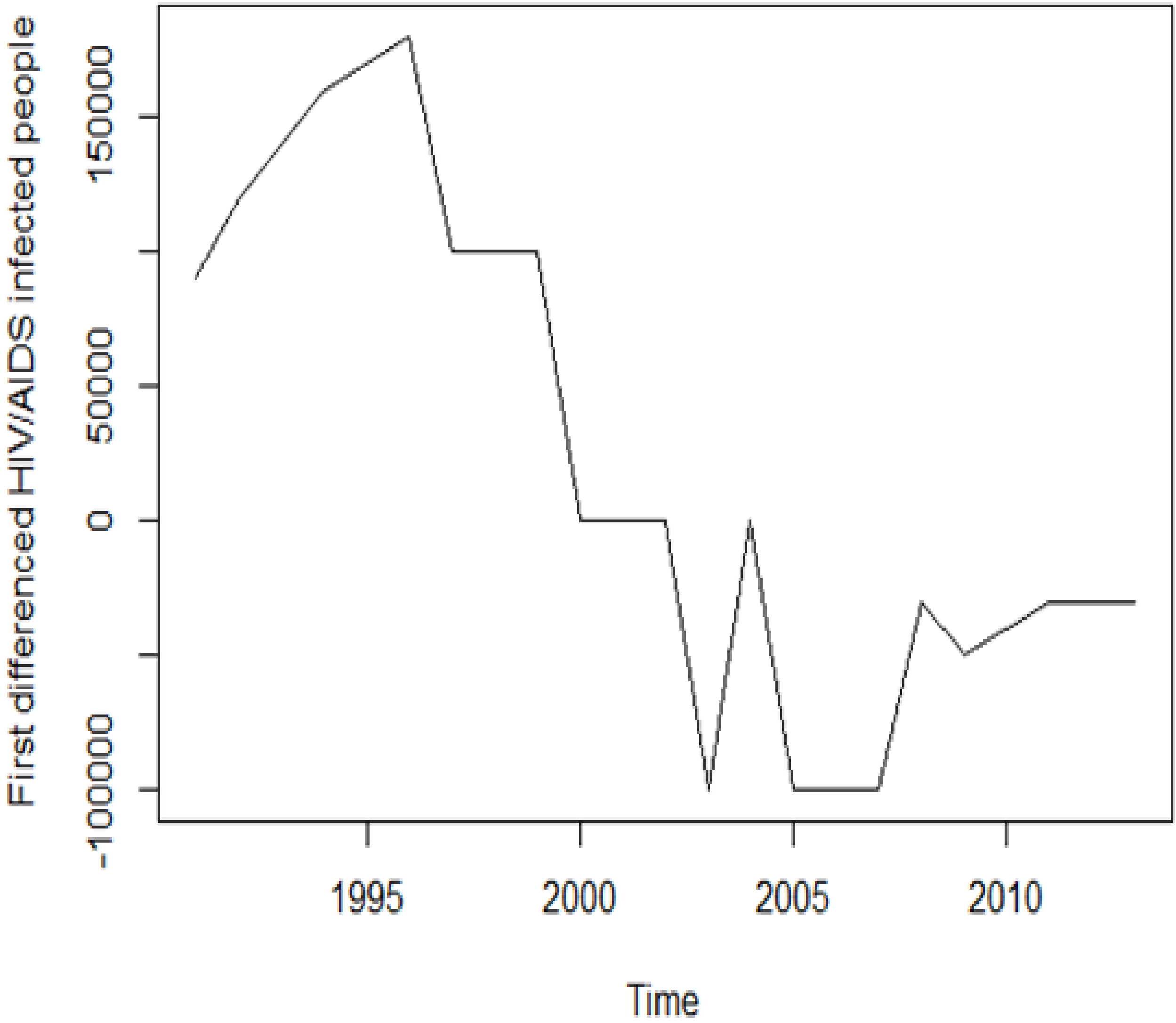

Figure 4.2: Time series plot of the HIV/AIDS infected people after first differencing

Unit root test for stationarity

The presence of unit root exhibits that the observed time series is not stationary. Unit root test is one of the methods objectively used to determine if differencing is required by the observed time series to achieve stationarity. Among the most popular method of testing stationarity, the Augmented Dickey-Fuller (ADF) test is used in this project. The ADF test is estimated by the following regression model [17].

From the given ADF test, it is confirmed the third differenced data is stationary as null hypothesis which says that the data is non stationary (random walk) is rejected ($p = 0.01$). It is actually described in the R-software that the p-value is smaller than

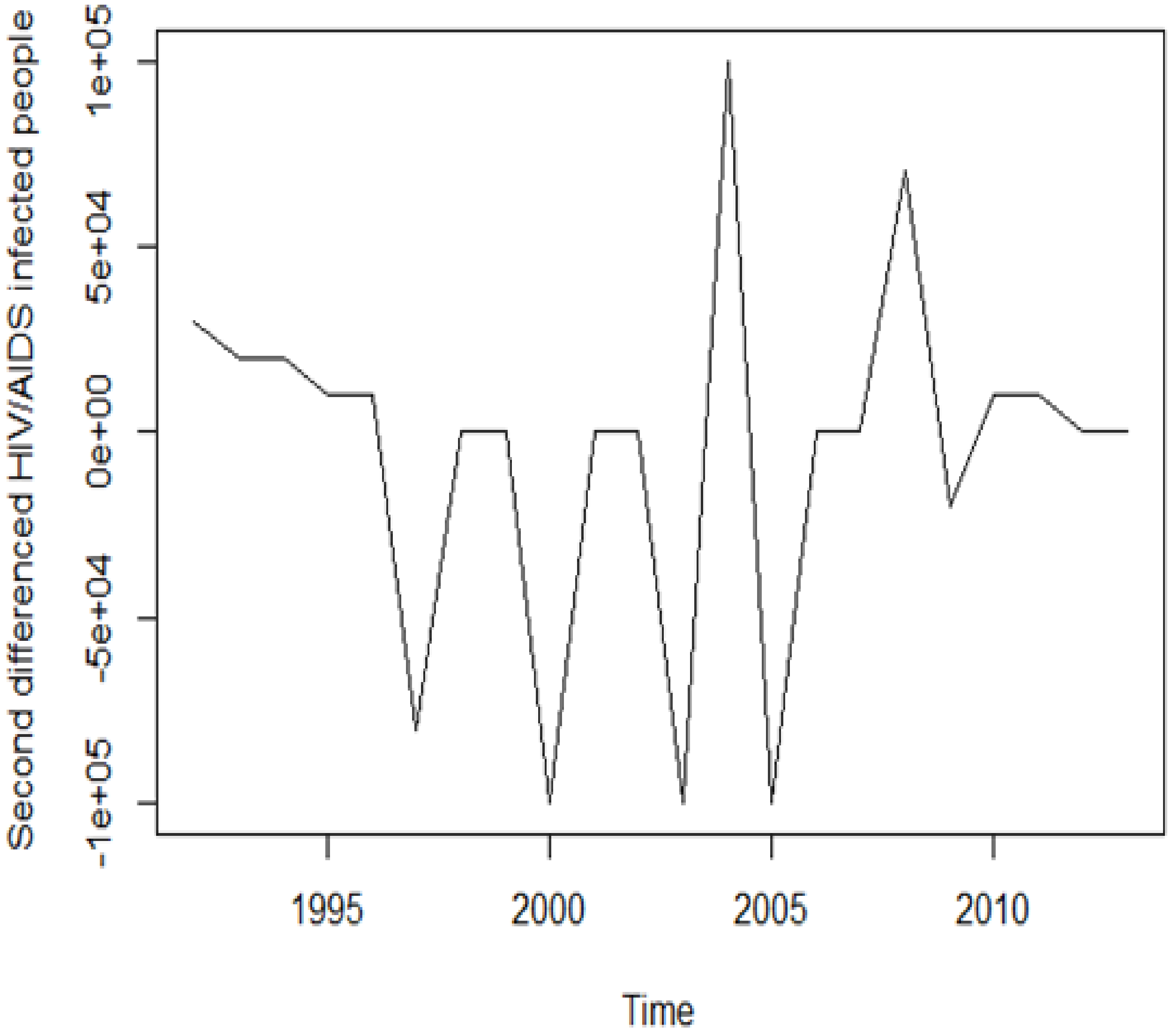

Figure 4.3: Plot of second differencing of the HIV/AIDS infected people data in Ethiopia

the given value above in the test.

4.1.2 Identifying tentative ARIMA model

Once stationarity is attained through transformation using differencing, the next step is to select the appropriate order of $ARIMA$ model, which means finding the values of most appropriate values of p and q for an $ARIMA(p, d, q)$ model. To identify tentative model, we usually need to examine the correlogram and partial correlogram of the stationary time series. It is not usually possible to recognize the values of p and q from the time plot from the given data. Therefore, it would be imperative to use the ACF

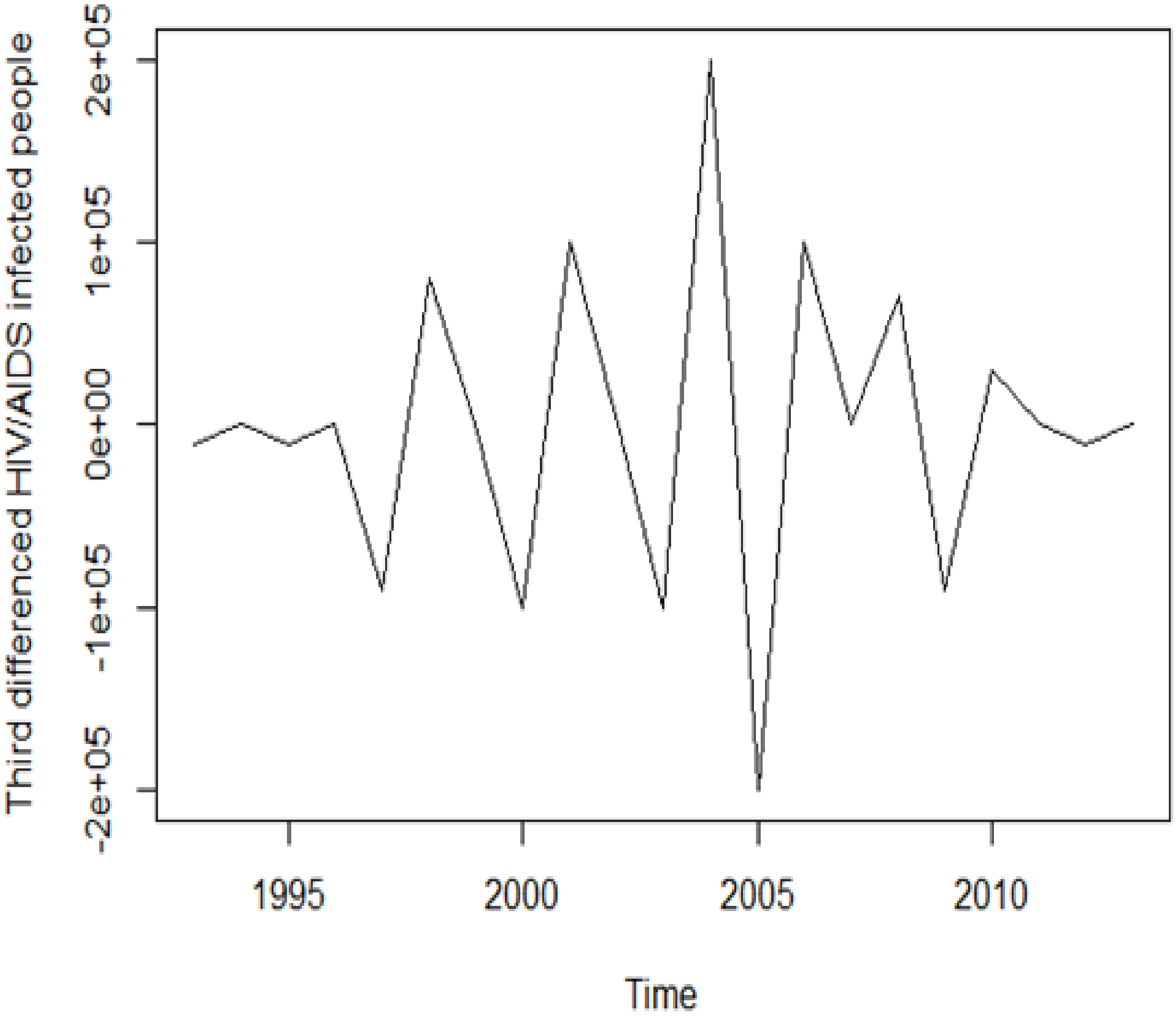

Figure 4.4: Plot of the third differencing of the HIV/AIDS data in Ethiopia

and PACF plot to determine the proper values of p and q.

If the data follow $ARIMA(p, d, 0)$, the ACF pattern is exponentially decaying and there is a significant spike at lag p of PACF and it gradually cuts off after p lags. On the other hand, if the data follow $ARIMA(0, d, q)$ model, the PACF is exponentially decaying and there is significant spike at lag q of ACF which gradually cuts off after q lags.

For this particular study p appeared to be 2, d is 3 and q is 2. It can be seen from the autocorrelation plot (correlogram), Figure(4.5), the autocorrelations at lag 1 which is about -0.715 exceeds the significance bounds and the significance shows that the q value, i.e, $q = 1$ for the order of moving average.

The partial correlogram in Figure(4.6), shows that the partial autocorrelations exceed

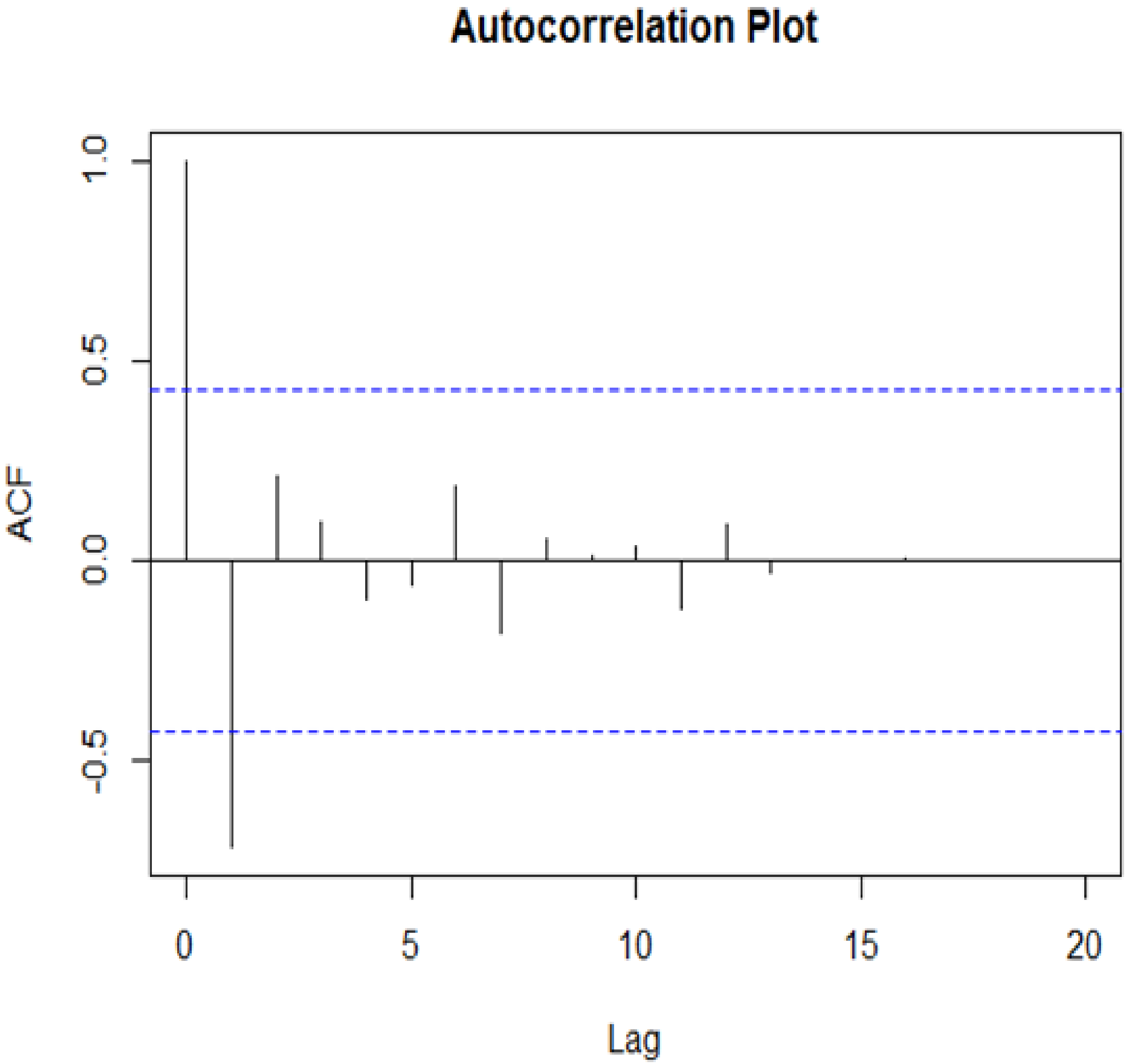

Figure 4.5: The Autocorrelation function of the stationary time series data

the non-significant bounds lag 1 and 2 which gives an idea on what should be on p, i .e, $p = 2$ for the order of the autoregressive. Pertinent to ACF and PACF plots, the model has been found out that $ARIMA(2, 3, 1)$ as the potential candidate model for the given time series data. This model should be checked with information criteria like AIC which are useful in determining the order of ARIMA model. These information criteria help to pick the one with lowest value of AIC, BIC, etc since good models are obtained by minimizing either the AIC or BIC.

The selection of ARIMA processes was conducted using Akaike's information criterion (AIC), which measures how well the model fits the series. According to Hyndman and

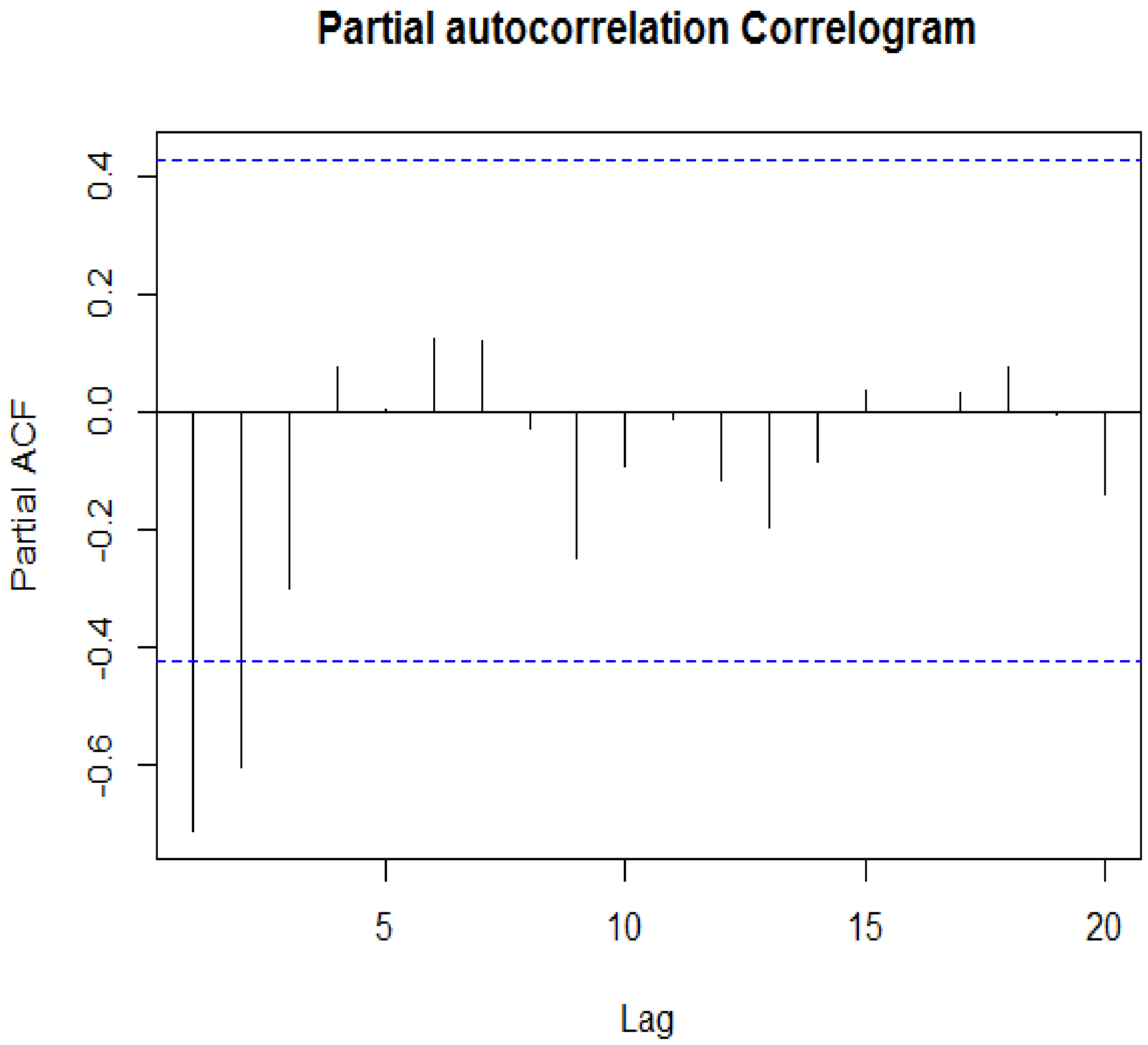

Figure 4.6: The Partial autocorrelation function of the stationary time series data

Athanasopoulos (2014) [17], the values of p and q are chosen by minimizing the AIC after differencing the data d times[17]. If $d \geq 1$, the constant in a model set to zero and the model is called ”current model.” To get the best model, vary p and/or q from the current model by ±1 and therefore, the best model with smallest AIC is selected among: $ARIMA(2, d, 2), ARIMA(0, d, 0), ARIMA(1, d, 0), ARIMA(0, d, 1)$. Therefore, based on the criteria of Hyndman and Athanapouls (2014), the best model with smallest AIC is $ARIMA(2, 3, 2)$ [17]. From the Table (4.3), it can be easily observed that ARIMA (2,3,2) model has the lowest AIC value.

Table 4.1: The coefficients of the parameters and Information Criteria of ARIMA$(2, 3, 1)$ model

	ar1	ar2	ma1	AIC	AICc	BIC	σ^2
Coefficients	-1.3249	-0.660	-1.000	489.46	492.54	493.02	1.622e+10
Standard error	0.1511	0.1456	0.152				
log likelihood -240.73							

Table 4.2: The coefficients of the parameters and Information Criteria of ARIMA$(2, 3, 2)$ model

	ar1	ar2	ma1	ma2	AIC	AICc	BIC	σ^2
Coefficients	-1.2232	-0.6180	-1.9292	0.997	479.71	484.71	484.16	6.136e+09
Standard error	0.1627	0.1567	0.2288	0.2203				
log likelihood -234.85								

Table 4.3: The value of some tested ARIMA models using AIC criteria

Model	AIC value
ARIMA (0, 3, 0)	532.13
ARIMA (0, 3, 1)	513.99
ARIMA (1, 3, 0)	514.24
ARIMA (2, 3, 0)	502.22
ARIMA (2, 3, 1)	489.46
ARIMA (2, 3, 2)	479.91

4.2 Model Diagnostic

4.2.1 Examine the residuals of ACF and PACF

Since the residuals should be independent and contain no elements are predictable, the ACF and PACF of the residuals should all lie between the approximate 95% confidence interval limits. The residuals of ACF and PACF indicate that the residuals are not forecastable and this suggests that the $ARIMA(2, 3, 2)$ is adequate considering this criterion.

4.2.2 The Box-Pierce statistic (Portmanteau test)

The purpose of this test is used to verify if the residuals are independent and the null hypothesis is that the residuals are independent and the alternative is they are not independent. From the Box-Ljung test, we fail to reject the null hypothesis ($\chi^2 = 16.1004, df = 16, p-value = 0.446$) which indicates the residuals are independent.

4.2.3 Normality of residuals

The residuals have been assumed to be normally distributed throughout the model. Quantile - Quantile plots (QQ) plots are an effective tool for assessing the normality of residuals. From the plot in Figure(4.7), it can be easily observed that the Q-Q plot is approximately normally distributed.

4.2.4 Significance of parameters

To test the significance of the parameters, a standard error of the parameter estimates are computed and roughly speaking, the parameters of a model are accepted as significant if the estimated values of the parameter is twice the standard error of this estimate or more. If a parameter shows up as not statistically significant, it will be removed. But in this study, all the parameters are significant.

4.2.5 The cumulative periodogram

The cumulative periodgram is a very useful tool for describing a time series data set in identification and diagnostic test of model development. It is useful especially when the data set is small. The observed data set in the cumulative periodgram of Figure(4.8) is within the confidence interval of 95% confidence limit and the model is adequately fit.

4.2.6 Trend of the observed time series data

The trend in Figure (4.9) revealed that the HIV/AIDS prevalence was increasing in alarming rate from approximately mid 1990s and reached its climax in the years 2002 to 2004 and decreased onwards.

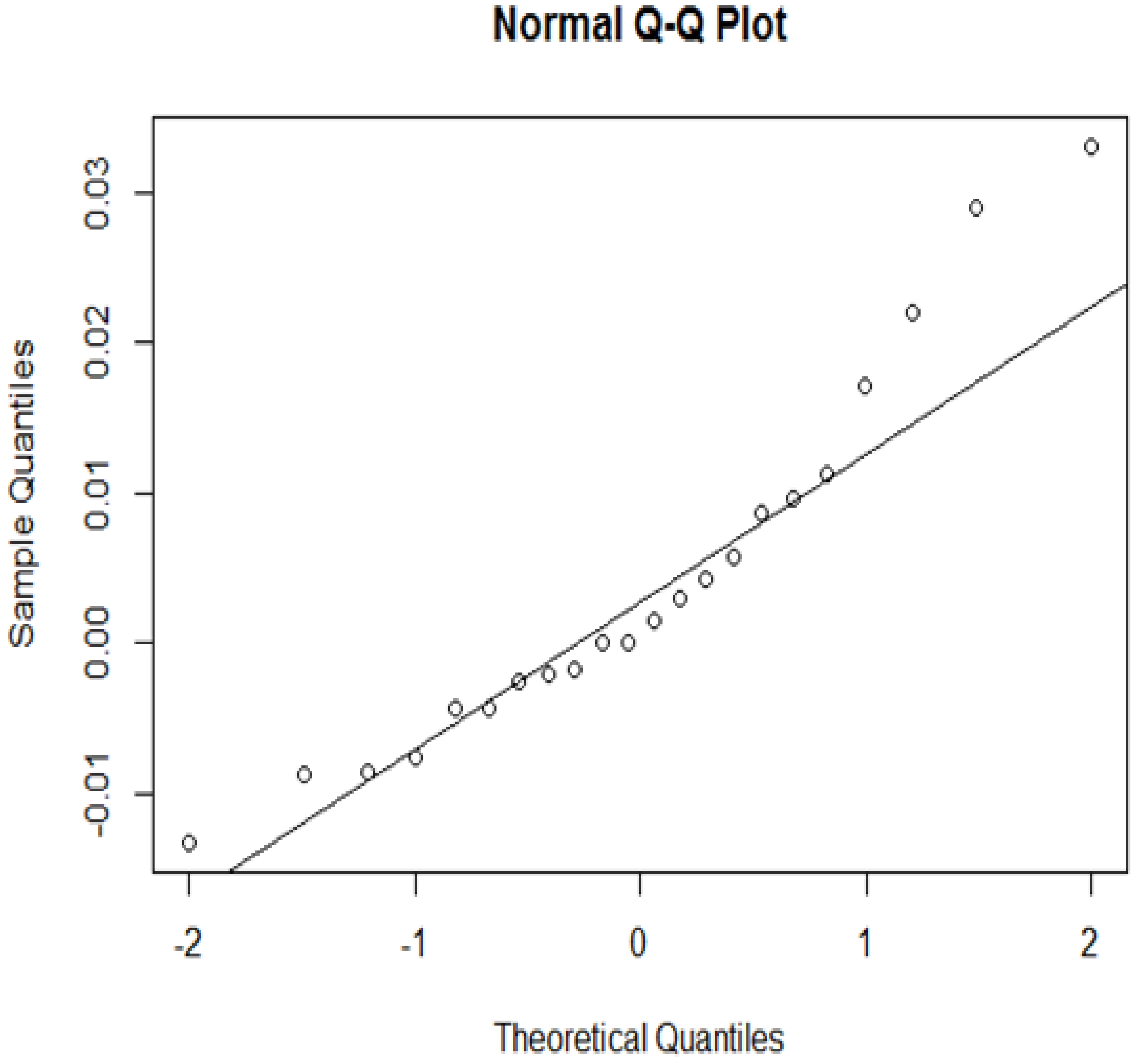

Figure 4.7: The Normal Q-Q Plot of the HIV/AIDS data

4.3 Forecasting (Prediction)

Time series forecasting tells future values of time series variables by extrapolating trends and patterns of past values of the series or by extrapolating the effect of other variables of the series [13]. It is often overarching to fit a trend curve to successive value and extrapolate yearly and non-seasonal data for long-term forecasting [18]. The

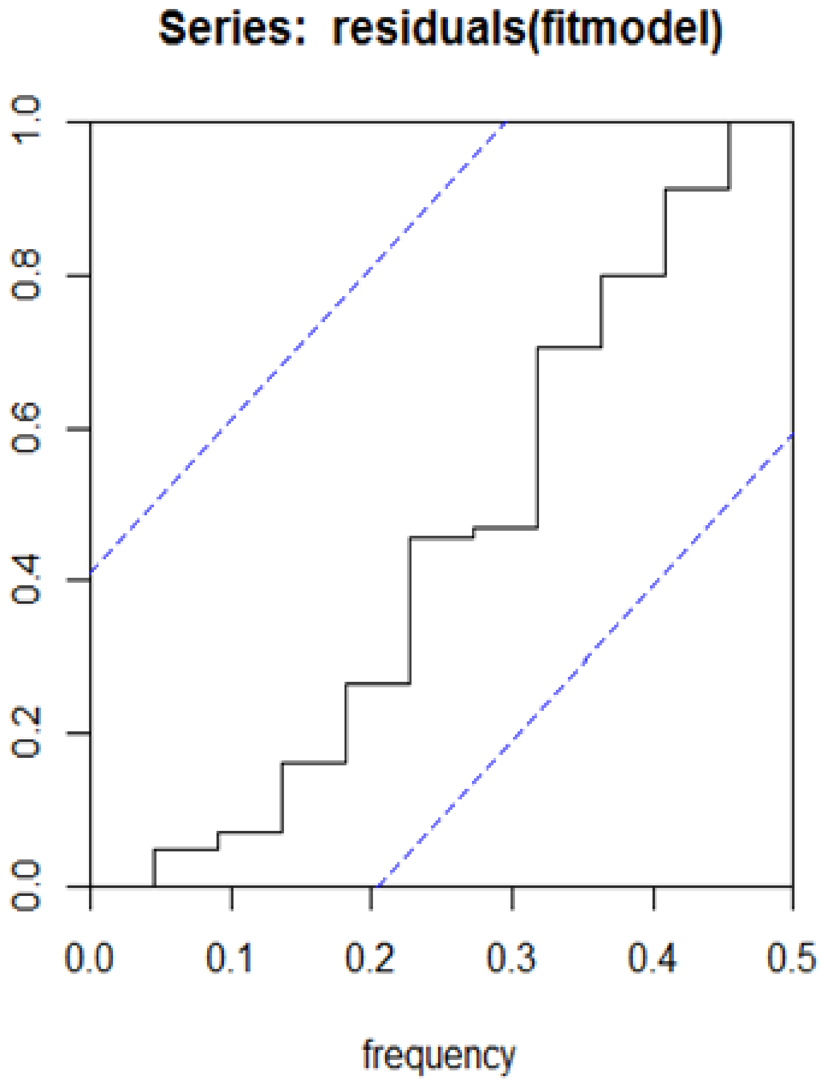

Figure 4.8: The Cumulative periodogram of the HIV/AIDS data

prediction of the observed time series data from the differenced data of the fitted model of this study for the next 5 years from year 2013 is given in the Table (4.4).

4.3.1 Forecast accuracy measures

Calculating the forecast accuracy measures using test data is mandatory and it is recommended to use mean absolute error measures for the same scale forecasting and when comparing forecast methods on a single data set, the mean absolute error accuracy measure is popular and mostly used as it is easy to understand and compute. Forecast accuracy can only be determined from the original data portion which is not

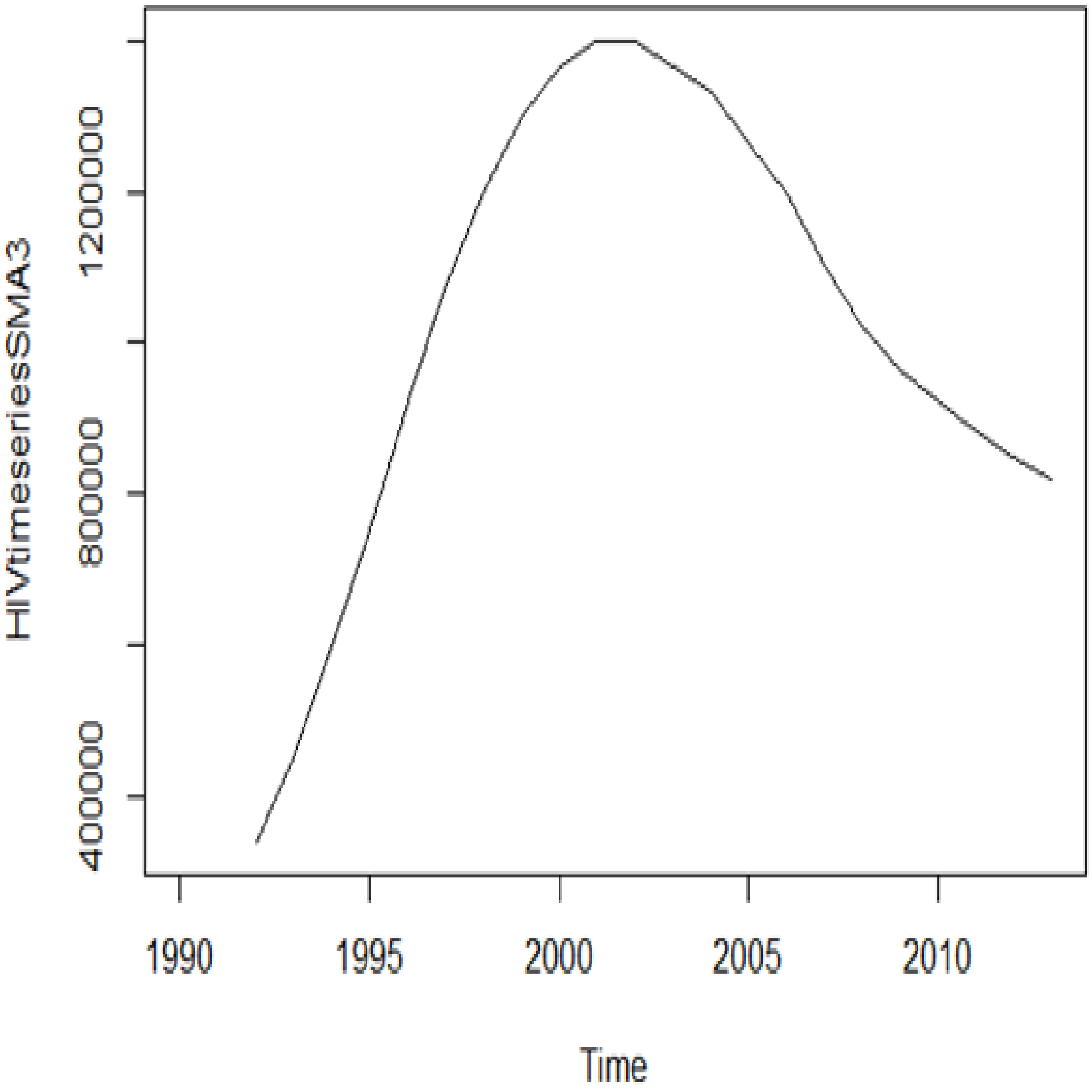

Figure 4.9: The trend of the observed HIV/AIDS data points

Table 4.4: The predicted value of HIV/AIDS infected people for the next 5 years since 2013 from the differenced data of the fitted model

Year	Predicted value	Standard error
2014	-18415.53	83025.78
2015	-17235.55	82934.56
2016	-25266.90	109693.37
2017	-36943.16	124587.15
2018	-41271.42	137773.34

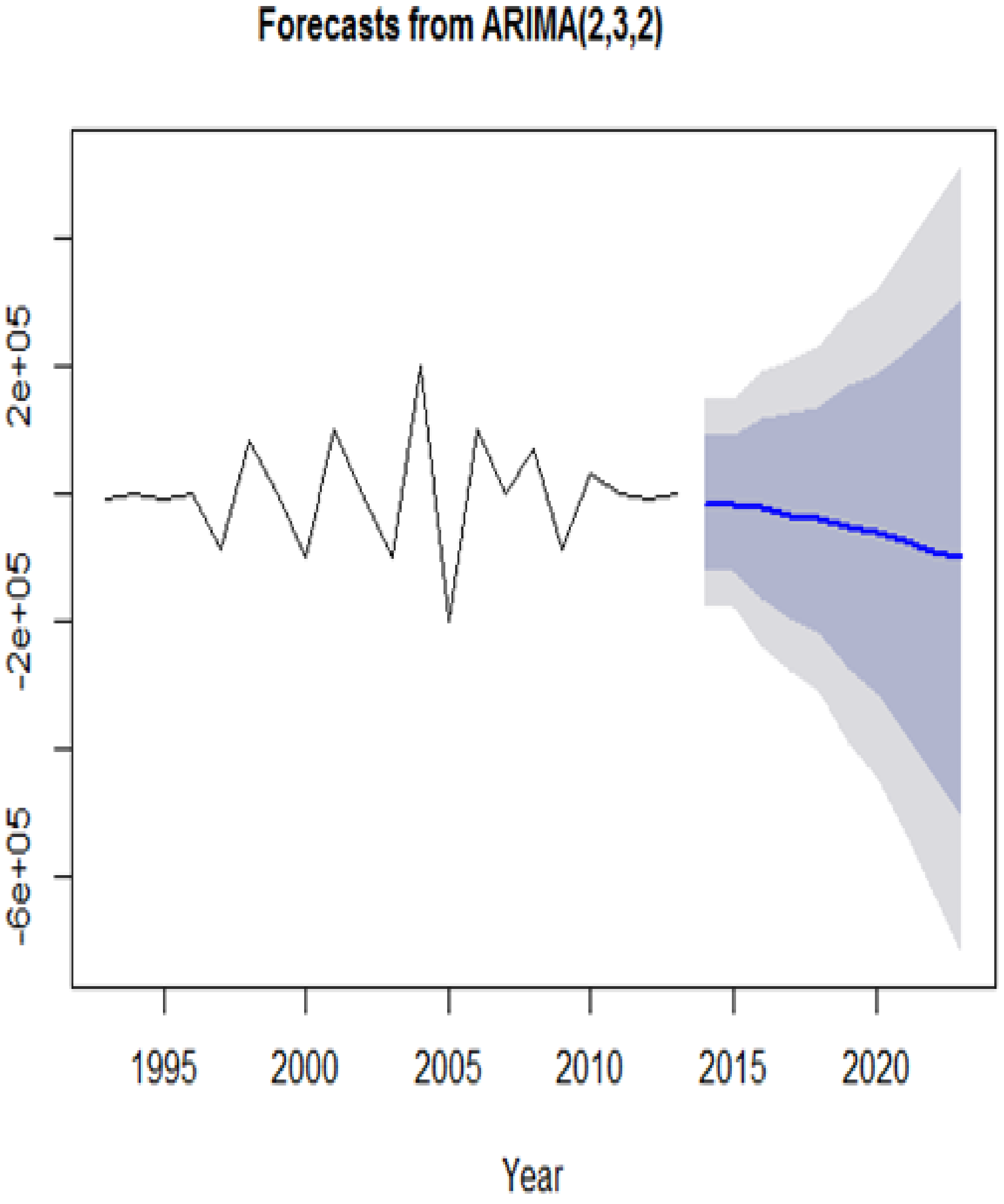

Figure 4.10: The prediction of the observed time series HIV/AIDS data in Ethiopia

used for fitting the model [17]. The mean absolute error based on equation (3.25) of the actual predicted value of the next 5 years is given in the Table (4.5). Since the calculated mean absolute error accuracy measures for this study is based on a single time series data observation, it is a bit intractable to compare its size whether it is big or not. The mean absolute error accuracy indicates that the mean magnitude of the errors in set of forecasts without taking in to consideration their direction.

Table 4.5: The mean absolute error accuracy measure of the HIV/AIDS infected people for the next 5 years since 2013

Year	Actual Predicted value	Forecasted value	Mean absolute error (MAE) (Actual Value -Forecasted Value)
2014	771584.47	18415.53	733168.94
2015	754348.92	17235.55	737113.37
2016	729082.02	25266.90	703815.12
2017	692138.86	36943.16	655195.7
2018	650867.44	41271.42	609596.02
		Total MAE	687777.83

4.4 Discussion

The gross data on the HIV/AIDS infected people in Ethiopia was analyzed using a time series, ARIMA model through achieving its stationarity after making the data three times differencing. The forecasting graph of the differenced data is given in Figure (4.10) indicating that the disease is in the prospect of declining for the next 5-10 years. The prediction of the actual values is based on the differenced data and trend values. But, in non-seasonal and non-stationary ARIMA model, the trend component of the time series will be removed with changing the non-stationary to stationary in the transformation mechanism like differencing and is left with irregular component.

Figure (4.10) showed that the differenced time series and the forecasted value for the next 10 years from 2013. As the predicted value considered the standard error (Table (4.4)), the actual predicted value for such a data set is given as the sum of the original data and the predicted value. The negative sign in the predicted value indicates that it decreases as it can be seen from the forecasted graph (4.10). For example, the actual predicted value for year 2014 becomes the original data in 2013 plus the predicted value for 2014. The actual predicted value for 2015 is given by the actual predicted value for 2014 plus the predicted value for year 2015 and it goes on like this till year 2018.

The actual predicted value from year 2014–2018 is given in Table (4.6). The actual predicted value in Table (4.6) is calculated taking in to account Figure (4.10) and the

Table 4.6: The actual predicted value of HIV/AIDS infected people for the next 5 years since 2013

Year	Actual predicted value
2014	771584.47
2015	754348.92
2016	729082.02
2017	692138.86
2018	650867.44

original HIV/AIDS infected people data of which time series plot is given in Figure (4.1).

Chapter 5

Conclusions and Recommendations

5.1 Conclusion

From this study, $ARIMA(2,3,2)$ appeared to be providing the best fit for HIV/AIDS epidemic in Ethiopia. The trend shows that the HIV/AIDS prevalence was increasing in alarming rate from approximately mid 1990s and reach its climax in the year 2002 to 2004 and decreases onward. The prediction shows that the prevalence of HIV/AIDS will decrease in Ethiopia for next 5 years.

5.2 Recommendation

ARIMA does not deal non linear relationships efficiently, it would be more practical and accurate if a combined model is used to capture different patterns equally. So, using a hybrid of unified model is highly recommendable. To come up with a comprehensive and perfect conclusion on prevalence of HIV/AIDS in Ethiopia, further investigation including research on significant contributing factors as a predictors of the disease will be necessary. Moreover, it would be recommendable if this model can be compared with other model that are developed for HIV/AIDs epidemics. It is also recommended if the time series analysis will be done on HIV/AIDS infected people categorized with age groups. It seems that it has been found out from this study a seminal result which is very much different from the previous similar studies leading to further work to be done or improving the model.

References

[1] Gould, W.S. (2009). HIV/AIDS in Developing Countries. *International Encyclopedia of Human Geography.* pp. 173-179

[2] Rauner, M.S., Brailsford, S.C., Flessa, S.(2005). Use of discrete-event simulation to evaluate strategies for the prevention of mother-to-child transmission of HIV in developing countries.*Journal of the operational research society* **56**:222-233.

[3] UNAIDS. (2010). Report on the global AIDS epidemic.

[4] MOH. (2006). *AIDS in Ethiopia.* 6^{th} edition. Addis Ababa. Ministry of Health.

[5] Gouws, E., Stanecki, K.A., Lyerla, R., Ghys, P.D. (2008). The epidemiology of HIV infection among young people aged $15-24$ years in Southern Africa. *AIDS*&#8203;**22**(4):s5-16. doi: 10.1097/01.aids.0000341773.86500.9d.

[6] CSA. (2005).Demographic and Halth Survey. Central Statistical Agency. Addis Ababa. Ethiopia. ORC Macro, Calverton, MD, USA.

[7] Biadgilign, S., Deribew, A., Amberbir, A., Escudero, H.R., Deribe, K. (2011). Factors associated with HIV/AIDS diagnositc, disclosure to HIV infected children receiving HAART: A multicenter study in Addis Ababa, Ethiopia. PLoS One **6**:e17572.

[8] Green, K.C., Scott Armstrong, J. (2012). Demand Forecasting: Evidence-based methods. https://marketing.wharton.upenn.edu/profile/226/printFriendly. Accessed on July 03, 2015.

[9] Walker, N., Grassly, N. C., Garnett, G. P., Stanecki, K. A., Ghys, P. D. (2004). Estimating the global burden of HIV/AIDS: what do we really know about the HIV pandemic? The *Lancet* **363**:2180-85.

[10] Abogaye-Sarfo, P., Cross, J., Mueller, U. (2010). Trend analysis and short-term forecast of incident HIV infection in Ghana. *African Journals of AIDS Research***9**(2): 165-173.

[11] Zhang, G.P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**:159-175.

[12] Stover, J. (2009). AIM: A Computer program for making HIV/AIDS projections and examining the demographic and social impacts of AIDS. Washington, DC: Futures Group International, Health Policy Initiative, Task Order 1.

[13] Sibanda, W., Pretorius, P.D. (2014). Trend analysis of HIV prevalence rates amongst Gen X and Y pregnant women attending Clinics in South Africa between 2001 and 2010. *Mediterranean Journal of Social Sciences* **5**(21).

[14] Box, G.E.P., Jenkins, M. (1976).*Time series analysis forecasting and control.* Holden-Day Inc.

[15] Wei, William, W.S. (2006). *Time series analysis: Univariate and Multivariate Methods* 2^nd *ed.*Pearson Addison Wesley.

[16] Cryer, J.D., Chan, Kung-Ski. (2008). *Time Series Analysis with Application in R,* 2^nd *ed.*

[17] Hyndman, R., Athanasopoulos, C. J. (2014). *Forecasting: Principle and Practice* (online version). https://www.otexts.org/book/fpp. Accessed from July 1-15, 2015.

[18] Chatfield, C. (2004) *The Analysis of Time Series: An introduction.* New York, Chapman and Hall/CRC Press.

[19] Durbin, J. (1960). The fitting of time series models. *Review of the International Institute of Statistics* **28**: 233-244.

[20] Levinson, N. (1947). The Weiner RMS error criterion in filter design and prediciton. *Journal of Mathematics Physics* **25**:261-278.

[21] World Bank. (2005). *World Development Report.* Washington, D.X. International bank for reconstruction and development and world bank.

[22] Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30A**: 9-14.

[23] Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **66**:237-242.

Appendix

R-code

```
>getwd()
>setwd("C:/Users/Demissew/Documents/Biometry/HIV Project/Data")
>HIV<-read.table("HIVdata.txt", header=TRUE)
>HIVtimeseries<-ts(HIV, start=c(1990))
>HIVtimeseries
>plot.ts(HIVtimeseries, ylab="Number of people infected", main="HIV/AIDS in Ethiopia")
>par(mfrow=c(1,2))
>acf(HIVtimeseries)
>pacf(HIVtimeseries)
# The resulting time series of first differences (above) does not appear to be stationary in mean. Therefore, we can difference the time series twice, to see if that gives us a stationary time series:
>HIVtimeseriesdiff2<- diff(HIVtimeseries, differences=2)
>plot.ts(HIVtimeseriesdiff2, main="HIV/AIDS in Ethiopia")
>HIVtimeseriesdiff3<-diff(HIVtimeseries, differences=3)
>plot.ts(HIVtimeseriesdiff3,main="HIV/AIDS in Ethiopia")
# To extract the trend component of a non-seasonal time series that can be described using an additive model, it is common to use a smoothing method, such as calculating the simple moving average of the time series. The SMA() function in the "TTR" R package can be used to smooth time series data using a simple moving average. To use this function, we first need to install the "TTR" R package:
>install.packages("TTR")
>library("TTR")
```

```
>HIVtimeseriesSMA3<-SMA(HIVtimeseries, n=3)
>plot.ts(HIVtimeseriesSMA3)
>HIVtimeseriesSMA8<-SMA(HIVtimeseries,n=8)
>plot.ts(HIVtimeseriesSMA8, main="HIV/AIDS Trend")
# Install "forecast" package
>install.packages("forecast")
>library("forecast")
# ARIMA models
>HIVtimeseriesdiff1<- diff(HIVtimeseries, differences=1)
>plot.ts(HIVtimeseriesdiff1, ylab="Number of people infected", main="HIV/AIDS in Ethiopia")
# To test stationarity, Dickey-Fuller test for variable will be used:
>library(tseries)
>adf.test(HIVtimeseriesdiff2, alternative="stationary")
>adf.test(HIVtimeseriesdiff2, alternative="explosive", k=0)
>adf.test(HIVtimeseriesdiff3, alternative="stationary")
# DF and ADF tests for differenced variable
>adf.test(HIVtimeseriesdiff2, k=0)
>adf.test(HIVtimeseriesdiff2)
# Selecting appropraite ARIMA models involves examining the correlogram and partial correlogram of the stationary time series and plotting autocorrelation correlogram
>acf(HIVtimeseriesdiff2, lag.max=20) # plot a correlogram
>acf(HIVtimeseriesdiff2, lag.max=20, plot=FALSE) # get the autocorrelaton values
>acf(HIVtimeseriesdiff2, lag.max=20,main="Autocorrelation Plot")
>acf(HIVtimeseriesdiff2, lag.max=20, plot=FALSE)
>acf(HIVtimeseriesdiff3, lag.max=20,main="Autocorrelation Plot")
>acf(HIVtimeseriesdiff3, lag.max=20, plot=FALSE)
# Plotting partialautocorrelation correlogram
>pacf(HIVtimeseriesdiff2, lag.max=20) # plot a partial correlogram
>pacf(HIVtimeseriesdiff2, lag.max=20, plot=FALSE) # get the partial autocorrelation values
>pacf(HIVtimeseriesdiff2, lag.max=20, main="Partial autocorrelation Correlogram")
```

```
>pacf(HIVtimeseriesdiff2, lag.max=20, plot=FALSE)
>pacf(HIVtimeseriesdiff3, lag.max=20, main="Partial autocorrelation Correlogram")
>pacf(HIVtimeseriesdiff3, lag.max=20, plot=FALSE)
>par(mfrow=c(1,2))
>plot(acf,data=HIVtimeseriesdiff3)
>plot(pacf, data=HIVtimeseriesdiff3)
>library("forecast")
>fit1<-Arima(HIV, order=c(0,0,2))
>fit1
>fit2<-Arima(HIVtimeseriesdiff2, order=c(0,0,2))
>fit2
>fit3<-Arima(HIV, order=c(0,0,1))
>fit3
>fit4<-Arima(HIVtimeseriesdiff3, order =c(2,3,1))
>fit4
>fit6<-Arima(HIVtimeseriesdiff3, order =c(2,3,0))
>fit6
>fit7<-Arima(HIVtimeseriesdiff3, order =c(1,3,0))
>fit7
>fit8<-Arima(HIVtimeseriesdiff3, order=c(0,3,1))
>fit8
>fit9<-Arima(HIVtimeseriesdiff3, order=c(2,3,2))
>fit9
>fit10<-Arima(HIVtimeseriesdiff3, order=c(0,3,0))
>fit10
>fit11<-Arima(HIVtimeseriesdiff3, order=c(3,2,2))
>fit11
>fit12<-Arima(HIVtimeseries,order=c(3,2,2))
>fit12
# The ACF plot of the residuals from the ARIMA(2,3,2) model shows all correlations
within the threshold limits indicating that the residuals are behaving like white noise.
# A portmanteau test returns a large p-value, also suggesting the residuals are white
```

noise.

```
>par(mfrow c=(1,2))
>acf(residuals(fit9))
>pacf(residuals(fit9))
# Portmaneau test if the residuals are independent (Diagnostic checking)
>Box.test(residuals(fit9), lag=20, fitdf=4, type="Ljung")
# Forecasting: the code for forecasting
>library("forecast")
>plot(forecast(fit4), xlab="Year")
>plot(forecast(fit9), xlab="Year")
# Plot of residuals
>fitmodel<-arima(HIVtimeseriesdiff3, order=c(2,3,2))
>fitmodel
>tsdiag(fitmodel)
# Normality of residuals
>qqnorm(residuals(fitmodel))
>qqline(residuals(fitmodel))
>qqnorm(residuals(fitmodel1))
>qqline(residuals(fitmodel1))
# Prediction of the fitted model
>predict (fit9,n.ahead=10)
```