Forecasting Annual Tax Revenue of the South African Taxes Using Time Series Holt-Winters and ARIMA/SARIMA Models.

by

Mangalani P. Makananisa

Submitted in accordance with the requirements for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor: Mr R. Ssekuma

Co-Supervisor: Prof P. Ndlovu

October 2015

Abstract

This study uses aspects of time series methodology to model and forecast major taxes such as Personal Income Tax (PIT), Corporate Income Tax (CIT), Value Added Tax (VAT) and Total Tax Revenue(TTAXR) in the South African Revenue Service (SARS).

The monthly data used for modeling tax revenues of the major taxes was drawn from January 1995 to March 2010 (in sample data) for PIT, VAT and TTAXR. Due to higher volatility and emerging negative values, the CIT monthly data was converted to quarterly data from the first quarter of 1995 to the first quarter of 2010.

The competing ARIMA/SARIMA and Holt-Winters models were derived, and the resulting model of this study was used to forecast PIT, CIT, VAT and TTAXR for SARSfiscal years 2010/11, 2011/12 and 2012/13. The results show that both the SARIMA and Holt-Winters models perform well in modeling and forecasting PIT and VAT, however the Holt-Winters model outperformed the SARIMA model in modeling and forecasting the more volatile CIT and TTAXR.

It is recommended that these methods are used in forecasting future payments, as they are precise about forecasting tax revenues, with minimal errors and fewer model revisions being necessary.

Key terms

SARS, Personal Income Tax (PIT), Corporate Income Tax (CIT), Value Added Tax (VAT), Total Tax Revenue (TTAXR), Holt-Winters, Autoregressive integrated moving averages.

Contents

Acknowledgments

I would like to extend my sincere appreciation and gratitude to God my creator.

Special thanks go to the following people:

- Mr Rajab Ssekuma and Prof Ndlovu for their support and dedication in critically evaluating my work chapter by chapter.

- Professor J. Fresen of the University of Missouri, former head of the Department of Statistics at the University of Limpopo (MEDUNSA), who made me believe that statistics is one of the greatest tools which is applicable in all fields.

- Professor P.R. Gopalraj from the School of Mathematics and Applied Mathematics at the University of Limpopo for always encouraging me to like Mathematics and Mathematical statistics.

- English editor, Jennifer Lindsey-Renton, I appreciate all the work he has done to this dissertation.

- Finally, I would like to thank all whose direct and indirect advice helped me to complete this dissertation.

Declaration by Student

I declare that the submitted work has been completed by me, the undersigned, and that I have not used any other than permitted referenced sources and materials, or engaged in any plagiarism.

All references and other sources used by me have been appropriately acknowledged in the work.

I further declare that the work has not been submitted for the purpose of academic examination, either in its original or similar form, anywhere else.

Declared on the (date):

Signed:

Name: **Mangalani P. Makananisa**

Chapter 1

Introduction

1.1 Background

The revenue collected by the South African Revenue Services (SARS) plays an important role in government expenditure, in that it contributes a large portion of the country Gross Domestic Product (GDP). SARS was given a mandate by the South African Revenue Service Act 34 of 1997 to collect all revenues that are due, to ensure maximum compliance with the legislation, and to provide a customs service that will maximise revenue collection, border protection and facilitate trade processes (SARS Annual Report (2014)).

Tax revenue has increased at an average of 11.8% every year since the 1995 financial year. SARS collected R742,7 billion in tax revenue in year 2011/12, R4 billion more than its targeted for the 2012 budget. This figure represents nominal growth in revenue of 10% over the previous fiscal year. The tax revenue was expected to be around R810.2 billion for the 2012/13 financial year, an estimated growth of 9.1%. The forecast of R810.2 billion by SARS excludes non tax revenue, Southern African Customs Union (SACU) payments and Amnesty Proceeds SARS and National Treasury (2012).

The total South African tax revenue as a percentage of GDP at market price increased from 22.6% in 1995/96 to 24.6% in 2011/12, with the major contributors being Personal Income Tax (PIT), Corporate Income Tax (CIT) and Value Added Tax (VAT). The country's economy is a major driver of revenue collection SARS and National Treasury (2012).

1.1.1 SARS regular forecasting techniques

SARS depends on the National Treasury for their fiscal year forecasts for all their tax types. The aggregate yearly forecasts are then distributed monthly by the SARS Revenue Analysis and Reporting Unit. Following this, the monthly forecasts are distributed into daily forecasts by the Cash Flow Unit. The distribution of monthly and daily targets involves the use of recent historical weights and professional judgement adjustments. SARS initial financial year forecasts are released in March, which are followed by two revisions that are produced at the Medium Term Budget Policy Speech (MTBPS) forecasts in October and the final forecast revision in February the following year (SARS Annual Report (2014)). With the availability of historical data, SARS is currently using the following approaches to forecast revenue:

- The Constant Trend Growth Model;
- The Macro Simulation Model;
- The Explanatory Model; and
- The Professional Judgement Model.

1.1.1.1 The Constant trend growth model

In this approach, the forecast for the current year is based on the assumption that the rate of growth over the financial year remains uniform. The fiscal year forecast is derived from the following equation

$$\hat{Y_t} = \frac{YT_t}{YT_{t-1}} \times Y_{t-1} \tag{1.1.1}$$

where $\hat{Y_t}$ represents the forecast of the fiscal year t, YT_t represents the year to date actuals of the fiscal year t, YT_{t-1} represents the year to date actuals for the fiscal year $t-1$ and Y_{t-1} is the actuals of the fiscal year $t-1$.

The use of the Constant Trend Growth Model has a limitation in that the method is based on the assumption that the rate of growth over the financial year remains uniform, which may not always be the case given the factors that drive growth in the South African economy at a specific time interval. This calls for the need for a better forecasting model.

1.1.1.2 Macro simulation model

This method uses detailed data at an individual or transaction level to calculate tax liability from individuals who file income tax returns. This is done through combining internal variables and information from tax surveys conducted. This method also allows for individual taxpayer analyses by age group, level of education, gender and so on. The tax laws are applied to each individual record in the database in order to arrive at the total tax liability. Most importantly, this approach is based on a study by Van Heerden and Schoeman (2010), and has the ability to simulate alternative policy proposals such as changing of tax brackets and tax rates. Some references used by Van Heerden and Schoeman (2010) includes the works of Engle and Granger (1986), Davies (2009), Kakwani (1977) and others. The equation used calculates the average tax allowance ratio, τ_{allow}, as follows:

$$\tau_{\text{allow}} = \frac{y_i - \nu_i}{y_i} \tag{1.1.2}$$

where y_i represents the gross income of the i^{th} individual and ν_i is the taxable income of the i^{th} individual/entity. The ratio per bracket or income group is then applied to each individual/entity to derive individual allowances using the following equation:

$$Allow_i = y_i \times \tau_{allow} \tag{1.1.3}$$

The taxable income is defined as gross income with allowances removed using the following equation:

$$\nu_i = y_i - Allow_i \tag{1.1.4}$$

Therefore to calculate the Personal Income Tax liability (PIT_i) for the i^{th} individual, one needs to know the tax bracket or income group, and allowance/exemptions using the official tax codes to taxable income.

The limitation of the micro-simulation models is that it does not estimate the actual cash flow, but rather the liability due within a specific financial year. The difference between liability due and actual collections can be distorted by time differences, administrative efficiency, compliance and tax evasion. This model is thus dependent on detailed data at a transactional level; when there are more missing observations, the model will give skewed results. This could lead to a bias in average tax allowance ratio, taxable income calculation per tax bracket and tax liability.

1.1.1.3 Explanatory models

SARS uses this method as a major revenue forecasting tool. This approach forecasts revenue using the relationship that exists between individual tax type and Gross Domestic Product (GDP) components (so-called tax type base). For example, the relationship between PIT and compensation of employees and inflation, CIT and Gross Operating Surplus (GOS), VAT and Consumption and Fixed Investments, and Total Tax Revenue and GDP (Boonzaaier (2012)). Some references used by Boonzaaier (2012) includes the works of Sobela and Holcombe (1996), Wolswijk (2007), Hendry and Nielsen (2007) and others. Explanatory models are represented by the following sets of equations.

Set I:

$$\begin{aligned} \nabla ln(PIT_t) &= \beta_0 + \beta_1 \nabla ln(C_t^*) + e_t \\ C_t^* &= C_t \times \Upsilon_t \end{aligned} \tag{1.1.5}$$

where ∇ is the first difference operator, β_i is the i^{th} coefficient, C_t is compensation of employees at time t, ln is the natural logarithm, e_t is the error term at time t and Υ_t represents the maximum PIT tax rate at time t to account for policy changes.

Set II:

$$\begin{aligned} \nabla ln(CIT_t) &= \beta_0 + \sum_{i=1}^{4} \beta_i \nabla ln(G_{t-i}^*) + e_t \\ G_t^* &= G_t \times \ell_t \end{aligned} \tag{1.1.6}$$

where G_t is the Gross operating surplus at time t, ℓ_t represent CIT tax rate at time t to account for policy changes, and G_t^* is the derived variable which multiplies G and ℓ at time t.

Set III:

$$\begin{aligned} \nabla ln(VAT_t) &= \beta_0 + \beta_1 \nabla ln(h_t^*) + \beta_2 \nabla ln(I_t^*) + e_t \\ h_t^* &= h_t \times \Im_t \\ I_t^* &= I_t \times \Im_t \end{aligned} \tag{1.1.7}$$

where h_t is the total household consumption at time t, I_t is Fixed Investments, $\Im_t$ is VAT tax rate at time t to account for policy changes and I_t^* is the derived variable which multiplies I and ℓ at time t.

Set IV:

$$\nabla ln(TTAX_t) = \beta_0 + \beta_1 \nabla ln(GDP_{t-1}) + e_t \tag{1.1.8}$$

where $TTAX_t$ represent Total Tax Revenue at time t, GDP_{t-1} the Gross Domestic Product at time $t-1$.

From equations (1.1.5), (1.1.6), (1.1.7) and equation (1.1.8) we observe that explanatory models depend on the construction of quarterly models which are then distributed to monthly forecasts by using the monthly weighted average derived from recent historical actuals.

This is a limitation in that it may create a bias in the results of the explanatory models, which comes as a result of omitting important explanatory variable(s). Moreover, the out of sample forecasts for the explanatory variables are collected externally from different sources. This implies that under/over estimation of explanatory variables by the external sources could lead to the under/over estimation of revenue. This calls for the need of a better forecasting model.

1.1.1.4 Professional judgement model

This plays a significant role in revenue forecasting and is based on expert assessments. Professional judgement forecasts are conducted by a revenue committee which meets once in a month. Forecasts are based on the estimated outcome from the models described above and on information that relates to cash flow, administrative changes and other special factors which directly or indirectly affect revenue collection (Boonzaaier (2012)).

Since professional judgement comprises of forecasts of different models and is based on different scenarios, there could be problems when the anticipated scenario does not hold economically. For example, if the forecast for a specific tax type is reduced or increased based on the cash flow information, or if some individuals or companies expected to pay tax within a given time interval do not pay, the outcome could be over estimation of tax revenue. Similarly, if individuals or companies decide to pay tax in advance based on their future forecasts, the tax forecast outcome could be an under estimation of tax revenue. This is a big limitation of the professional judgment techniques used by SARS in forecasting tax revenues, which calls for the use of a more scientific approach such as time series to forecast tax revenues.

1.1.2 Aim and objectives of the thesis

Time series methods tend to be ignored by SARS analysts who are involved in revenue forecasting because of the methods do not use the relationship between revenue and the explanatory variables. The time series models used in this study are good for short-term forecasts; every model has some advantages and disadvantages, it depends on the use of the models.

The main objective of this study is to introduce the use of Holt-Winters and autoregressive integrated moving average (ARIMA)/Seasonal autoregressive integrated moving average (SARIMA) time series methods in modeling and forecasting the annual tax revenue collections of three tax types: Personal Income Tax (PIT), Corporate Income Tax (CIT), Value Added Tax (VAT) and Total Tax Revenue (TTAXR) using research software R.

In this study, tax revenues of the above mentioned taxes are modeled with the rationale of finding a times series model that could lead to a better forecasting technique for South African tax revenues. The particular objectives follows.

- To find a suitable ARIMA/SARIMA and Holt-Winters model that fits the data for accurately forecasting the annual tax payments. The model found may be used by South African authorities for planning and decision making purposes.
- To test if the selected model ARIMA/SARIMA or Holt-Winters is the most suitable. If so, then separate model for tax type is used for forecasting SARS fiscal year 2010/11, 2011/12 and 2012/13, respectively. The selection involves checking the accuracy of the model using different measures such as the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and so forth.
- To compare the model forecasts with the actual realisation for the two competing models; the one that performs better than the other is then recommended for forecasting the continuation of tax payments.

1.1.3 Data

The monthly time series data was obtained from SARS (internal data) for the three taxes (PIT, CIT and VAT) and overall Total Tax Revenue. SARS internal data is stored in SARS systems and the revenue figures for each month are considered preliminary until the 15th of the following month, when they are adjusted to be final figures. The sample used for PIT, VAT and TTAXR were drawn from January 1995 to March 2012. The volatility and emergence of some negative

values in the monthly CIT data results were converted to CIT quarterly data, commencing from the first quarter of 1995 to the first quarter of 2010. The use of this monthly/quarterly time series data has the advantages which follow.

- The sample size was bigger to allow for assumptions of classical normal distribution to be satisfied.
- The data patterns were visible for trend, seasonal and cyclical components to be observed.
- It allowed for revision of the model within a year if needed.

1.2 Organisation of the thesis

This dissertation consists of five chapters. Chapter 1 gives a background to the South African tax revenue, the methods that are used at SARS, their limitations, the objective of the study, and information about the data sources. Chapter 2 covers the literature review regarding the traditional methods used by SARS in forecasting tax revenue, and the time series methods of the Holt-Winters and ARIMA/SARIMA models in forecasting tax revenue. In Chapter 3, the theory of the Holt-Winters methods and the ARIMA/SARIMA models is discussed in detail. Chapter 4 covers the application of the methods defined in chapter 3 on the South African major tax types (PIT, CIT, VAT and TTAXR), while Chapter 5 provides a conclusion and recommendations.

Chapter 2

Literature Review

2.1 Introduction

Forecasting future outcomes will always be limited to the predictive power of the method or model used, while the perfection of future forecasts is also determined by a deeper knowledge of individuals regarding the data or variable of interest. Quantitative methods are currently commonly used, as data availability is expanding due to increasing efficiency in data capturing and technology used. Beyond the understanding of some factors that could minimise predictive power, there are the unknown occurrences that could also introduce uncontrollable errors to future forecasts. Although we expect some historical occurrences to continue into the future (Hyndman et al. (1998)), the future could bring unexpected changes as it is not constant over time. This supports the idea of short-term forecast rather than a long-term one.

Time series models tend to apply the rule that "no one knows you better than yourself", although there could be some individuals who know you. This mean that the time series methods assume that explanatory variables are captured in the historical occurrences of the variable of interest Shumway and Stoffer (2006). It is thus important to revise the model as the actual realisation for the variable being studied becomes available; this could better the future expected outcome. However, if by any chance the historical data used is incorrect, the future expectation will also follow the wrong path.

In the case of SARS, the use of incorrect data could be due to human error when capturing SARS cash flow or transactions, as well as a lack of tax knowledge (the combination of sub taxes to form main taxes, i.e. PIT, VAT or CIT). Before constructing time series models, data cleaning

and verification will add value to the future forecasts. The correction of monthly data used in this study was done internally when the finalised figures or the corrected/adjusted figures for the prior month were released around the 15th of the next month. This includes the distribution of unallocated revenues to their respective tax types. The tax types of interest are seasonal as this leads to no omission of the peak months as they form part of seasonal changes.

2.2 Review studies on ARIMA and Holt-Winters models

Modeling tax revenue with higher precision is important to a country as it leads to a budget distribution which is closer to the reality of the unknown future. This enables the state to not under or over spend revenue relative to actual collections. The Holt-Winters and ARIMA/SARIMA models are known to perform well for short-term forecasts, but are disadvantaged when it comes to explaining the components that influence or drive the variable of interest (Pindyck and Rubinfeld (1998)). This chapter looks at some of the literature reviews on the Holt-Winters and ARIMA/SARIMA models and their applications to several environments or fields.

Besides reporting on growth in the country′s GDP, the collection of all tax revenues for the state (Total Tax) is generally one component that drives the performance of the country against its expenditure; inappropriate budgeting of the state accounts could result in a deficit or a surplus. The revenue collection of all taxes is expected to grow in order for the country to meet its basic needs and the smooth running of the state affairs.

Several studies on modeling data series using time series models have been done in different fields, including finance,tourism and transportation (Pelinescu et al. (2010), Koirala (2012), Brojba (2010), Jayesekara and Passty (2009), and Slobodnitsky and Drucker (2008)). These types of models have become useful tools in modeling univariate data, thanks to their precision when it comes to predicting in-sample and out of sample values. The application of the Holt-Winters and ARIMA models on state revenues has increased over the past few years, with several authors supporting the use of these time series models:

Pelinescu et al. (2010) analysed the Romanian local budget with the aim of assisting officials to create efficient plans and manage local income and expenditure using a good strategic management tool. This arose as a result of the local authorities finding it difficult to predict future revenues

to construct their annual budgets. Using the historical data from the first quarter of 2000 to the third quarter of 2010, the authors' study applied the Holt-Winters multiplicative and additive models to forecast total local revenue and own revenue of local authorities. The E-views software was used to build and run the Holt-Winters equations and to select a model that minimised the Root mean squared error (RM SE). The study recommended the use of Holt-Winters models as a tool for multi-annual budget forecasting, because it is user-friendly and provides stable forecasts. A similar study for Romania was conducted by Brojba (2010) who used ARIMA models on monthly earning data for the period 2007 to 2008 (the economic crisis period) to model the total budget revenue. The ARIMA models used captured the data movement during the economic crisis because the data contained or showed the trend and seasonality. The fitted values were close to the actuals and the study concluded that ARIMA models can be used to set targets and sound future developments. However, the model has its limitations as the parameters are sensitive to sample selection, with the most accurate forecasts being for the short-term (Brojba (2010)).

The use of the Holt-Winters and ARIMA models on Total Tax Revenue monthly or quarterly data appears to be adoptable for revenue forecasting, as shown in the literature reviewed above. In addition, Brojba (2010) used a simple ARIMA model for Total Tax Revenue. However, Total Tax, being a combination of all taxes, is generally bound to be a seasonal series, which could lead to a seasonal ARIMA model as shown by Koirala (2012). The use of a simple ARIMA model was justified by the minimal period used (economic crisis period). Similarly, this model could also be applied to model the South African total tax revenue because the monthly to yearly data is available, and the only difference that can be expected could be the difference in the fitted parameters.

Jayesekara and Passty (2009) used ARIMA models, which included the dummy variables for seasonal adjustment as a univariate benchmark model,to forecast the net income tax revenue for Cincinnati, Ohio. The monthly data was obtained from the Cincinnati income tax division (CITD) for the period January 1970 to April 2009. The data used to carry out the estimation was then reduced to start from 1989 due to changes in tax rates. In order to reach the two final ARIMA models, the best fit models were selected based on the model with minimal Akaike information criterion (AIC) , the minimum root mean squared error ($RMSE$), and the model with the highest R-Squared (R^2). The ARIMA models fitted predicted the Cincinnati net income tax well, capturing the seasonality in the data throughout the sample used, and the within sample estimates were comparable with the actuals for the period 2006 to 2009. Furthermore, the ARIMA model was considered to forecast the net income tax starting from January 2008 (a portion of the in-sample)

to verify the effectiveness of the model. Moreover, data were converted to bi-monthly in order to construct a bi-monthly model. Jayesekara and Passty (2009) thus recommended the use of ARIMA to the CITD for short-term forecasts of net income tax.

Similarly, Chatagny and Soguel (2009) used the ARIMA model to estimate tax revenues (an addition of PIT and CIT) for all 28 cantons or districts. The main aim of the study was to prove that forecast bias can be reduced by using univariate time series models. Tax revenue data for the period 1944 to 2006, together with the observed official forecasts, were obtained from the districts. The time series data were divided into two samples (1944 to 2006 and 1976 to 2006) due to some districts not having recorded historical data in some years. To assess the ARIMA model's performance against the observed forecasts for the two sample periods, the mean percentage error was used to classify the over, under and zero error per canton or district. The results from the mean percentage error showed that the observed forecasts under estimated tax revenue in almost all cantons and ARIMA models had the Mean Percentage Error (MPE) close to zero error in the two sample periods. The study concluded that bias from the observed forecast can still be improved by using simple univariate models (ARIMA). The only limitation in using the ARIMA model is the lack of explanation of what causes the bias, as it can be due to other factors (see Pindyck and Rubinfeld (1998)). However, the models were considered useful for forecasting purposes as they do not need explanatory variables, thus they are costless in terms of information gathering.

Slobodnitsky and Drucker (2008) constructed VAT (revenue and refunds) monthly and quarterly ARIMA models for Israel's state revenue. Their aim was to find a model that best fit the VAT for revenue forecasting purposes. Data from between January 1987 and December 2006 were used for modeling and forecasting, however for the quarterly ARIMA models, there were some explanatory variables included such as the tax rate, consumption, sector GDP etc. The ARIMA ex-post forecasts from the year 2000 were compared to the forecasts from the co-integration model and official tax projections obtained from Israel's annual budget book. The quadratic loss function was used as a measure of estimate accuracy against the actuals and the quarterly net VAT ARIMA model was found to best resemble the actuals.

Brew and Wiah, (2012) analysed and built a VAT revenue model for the Tarkwa-Nsuaem Municipality in the western region of Ghana. Questionnaires and interviews were used to study 520 businesses in the municipality, with the interviews covering businesses that always issued VAT receipts to customers and businesses which did not often issue VAT receipts. VAT revenue data

for four years (January 2007 to December 2010) was obtained from the head office in the municipality. The rate of increase in VAT revenue in each year was compared with the results from the interviews to assess the efficiency in the mode of collection. Firstly, random sampling was used to select businesses, from which 520 were interviewed on how often they issued VAT invoices to their customers. VAT revenue was considered a dependent variable and businesses issuing VAT as independent variables for regression purposes. Autocorrelation and partial autocorrelation correlograms were generated using SPSS statistical software to help in the identification of the appropriate parameters for ARIMA model building. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) suggested an ARIMA(1,1,0) structure represented by the following equation to model VAT revenue:

$$Y_t = \mu + Y_{t-1} + \phi(Y_{t-1} - Y_{t-2})$$

This study established that the efficiency in the mode of collection of VAT revenue in the Tarkwa-Nsuaem Municipality was beyond average. It was also observed that the simple ARIMA model provided an evolution equation with a simple interpretation where VAT revenue could be estimated before the actual collection was done. This assisted the revenue authorities in Tarkwa-Nsuaem to determine the relationship between the new and the prior VAT collected, to predict VAT payments in advance, and to check the efficiency in the mode of future VAT collections.

Some researchers believe that econometric regression models are more precise than univariate time series models. For example, Corvalao et al. (2010) applied a regression model on VAT collection in Santa Catharina, Brazil, for the period January 1995 to December 2001. The mean absolute percentage error (MAPE) was used to compare the precision of the fitted values from the econometric regression and ARIMA model done by the budget department in Santa Catharina, which was found to be 2.51% and 4.63% as compared to actuals for regression and the ARIMA model, respectively. Short-term forecasts were generated for the period January to April 2002. This study did not convincingly show the leading econometric regression models against ARIMA as the MAPE was calculated for only a one year period (January to December 2001) and not for the entire sample. Moreover, to obtain the short-term forecasts some of the explanatory variables of sample forecast were generated using the ARIMA model.

Fomby (2008) applied Holt-Winters models to construct forecasts on the Plano sales tax in Texas. The monthly Plano sales tax data was obtained for the period February 1990 to November 2006

and divided into two groups, namely the in-sample data consisting of 167 observations (February 1990 to December 2003) and the model validating data (January 2004 to November 2006) for 23 observations. Holt-Winters methods for the Plano sales tax were used and the forecasts were derived from four competing models which were examined for three horizons, namely one-step, three steps, and six steps ahead ($\mathbf{h} = 1, 3, 6$). The $MAPE$ criterion was used as the measure of accuracy, and the additive Holt-Winters with trend and seasonality performed well on the three forecast horizons ($\mathbf{h} = 1, 3, 6$), with minimal $PMAE$ of 4.3, 2.84 and 3.5 for the three horizons respectively.

The study concluded that understanding the characteristics of the time series data helps in finding the accurate exponential smoothing forecasts, as shown in the case of Plano's sales tax. Furthermore, if one has to forecast more lines within minimal time, exponential smoothing for some of the data series can be used. However, if time is not one of the limiting constraints, then ARIMA models can be considered as they require thorough model identification.

Cote et al. (2010) used alternative methods or models to forecast industrial property and valuations of non-residential real commercial properties to improve property tax forecasts in El Paso City. In 2008, around 64.2% of tax revenue in El Paso was obtained from property taxes as a primary revenue source. The annual Personal Income data series were obtained from 1996 to 2006. An ARIMA model and three other models were fitted on commercial and industrial property data, which were then compared with the random walk and the random walk with a drift (as bench mark models) for accuracy and reliability of alternative models. The inequality U-coefficient was used to assess the predictive accuracy of the models. The ARIMA model performed better than the random walk but was found to be less effective when compared to the random walk with a drift model. However, the error differential results for the random walk with a drift model were found to be inconclusive for commercial purposes. The study suggested that further research should be conducted on different data samples to provide the worthiness of the ARIMA model and three other models.

Berwick and Malchose (2012) referred to fuel tax and license and registration fees as a major source of revenue for North Dakota's Transport Department. A survey was conducted to test the state of North Dakota's fuel tax revenue, where it was found that some state model sub-components of fuel tax revenue such as gasoline and diesel consumption, motor vehicle fuel and so forth can be modelled using time series ARIMA. The yearly time series data from 1951 to 2010 was used to model and forecast revenue for the period 2008 to 2013, with 2008 to 2010 forecasts constructed

to test the efficacy of the model. Simple ARIMA (1,1,1) and ARIMA (1,2,1) were fitted for fuel tax revenue and license and registration fees respectively. The ARIMA models fitted captured the movement of the two revenue sources with minimal errors against the actual and the forecast generated from the fitted models. Therefore, Berwick and Malchose (2012) concluded that the model estimate needs to be evaluated against actual outcome as it will tell if there is a need for models modifications in case of changes in North Dakota's economic system.

In the case of Croatia as one of the transitional countries, forecasts for both fiscal and the main macroeconomic variables are done for medium-run forecasts (i.e. three year forecasts) by the Croatian Ministry of Finance. However, the fiscal forecasts from the Ministry of Finance are based on simple trend extrapolation methods, expert judgement and forecasts of micro-economic variables. On average, the original budget forecasts underestimate the fiscal year actuals and the forecasts revision overestimates the actuals. This is due to the fact that official forecasts contain political bias Botric and Vizek (2012). ARIMA models and other formal econometric methods were introduced by Botric and Vizek (2012), for fiscal forecasting of both direct and indirect taxes in Croatia to improve the existing literature. Using the data from 1995 to 2006 (except for VAT which started in 1998), the methods were applied on the seven fiscal revenues, namely income tax, Corporation Tax, VAT, Property Tax, Import Duties, Excises and Social Contributions. After the modeling stage, the Thiels inequality U-coefficient and $MAPE$ were used to compare the estimate's accuracy against the actuals up to 2006, and the formal econometric methods were also used to forecast 2007 and 2008 revenue. The forecasts from the formal methods were also compared to the forecast from the Croatian Ministry of Finance. The ARIMA models were applied on annual growth rate series and the forecasts were more accurate than the official forecasts. This study concluded that Croatia's Ministry of Finance could benefit from using the formal (time series) methods, because they show accurate forecasts in a transitional Croatia through replacing expert judgement, simple trend extrapolation and political bias. This could also give government authorities more information on the future direction of Croatia's economy.

Silvestrini et al. (2008) estimated annual budget deficits in France using the monthly data from government revenue and expenditure, dating from January 1996 to December 2004, using direct and indirect tax revenues variables such as VAT, Income Tax, Corporate Tax, Tax on oil products and other tax revenues combined. On the expenditure side, the variables of interest were debt interest payments, wages and pensions, functioning expenditures, interventions, civil capital expenditures and military expenditures. The aim was to build a statistical univariate model to

identify or detect the closing and widening of the government deficit in advance, which could be used to advise the government on decision making and deficit regulation.

The reduced sample ending December 2001 was used to model all revenue and expenditure variables using seasonal ARIMA models. Data for 2004−2005 were reserved for model forecast comparison with the actuals (model validation). The monthly estimates were then updated as the new data became available. The monthly forecasts were also summed to form the (i.) monthly cumulative forecasts for the two validation years, and (ii) the temporary aggregated annual ARIMA models, constructed for one step ahead yearly forecasts. The monthly cumulative forecasts, aggregated annual forecasts and the French official forecast (traditional forecasts) were compared with the 2004 and 2005 yearly realisations (actuals). It was observed that the temporary aggregated annual ARIMA forecasts were close to the actuals when compared to the traditional forecasts.

Kudrle (2008) analysed the European Union's (EU) tax havens' liabilities data to gauge the impact of the Organisation for Economic Co-operation and Development's (OECD) harmful tax project, which previously lacked sharing of information on tax havens from foreign entities to reduce tax avoidance, as it appeared that there was PIT evasion and CIT avoidance. This study mainly explored the Cayman Islands data, however all tax havens agreed on information exchange in 2004. Data starting from the first quarter of 1975 to the last quarter of 2005 for both offshore and tax havens were analysed. ARIMA models were applied to the time series data, interventions were conducted, and the results revealed that there was no impact from the OECD project and that it required more than the OECD's demands to reduce the tax haven problem. As a result, Kudrle (2008) suggested automatic sharing of information, which covers sufficient financial instruments from different governments to minimising tax evasion or avoidance.

Suwanvijit (2013) used an additive Holt-Winters model to predict Indonesian, Malaysian and Thai tourism growth trends (IMT-GT). The aim was to predict a more reliable level of demand for decision making for long-term arrivals. Monthly data stretching from January 2002 to December 2011 obtained from Thailand's IT department's marketing database for model fitting. The criterion statistic used for model selection was $MAPE$ and the monthly forecasts for the period 2012 to 2017 were generated. The study concluded that an average annual increase of the overall arrivals to IMT-GT will be around 17% on the out-of-sample forecasts, and the results will contribute to the development and understanding of tourism forecasts.

2.3 Review studies on SARIMA models

Etuk and Igbuda (2013) fitted a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to the Nigerian Naira-British Pound exchange rate (NPER). Their paper mainly focused on analysing the characteristics of the NPER series and fitting the appropriate seasonal ARIMA model to the data. The monthly exchange rate data for the period 2004 to 2011 was obtained from the Central Bank of Nigeria's website (www.cenbank.org). The data were graphically visualised and stationarised, and ACF and PACF were studied to derive the final model for the series which is $ARIMA((0,1,0)(2,1,1)12$. The model used explained 61% of the variation of the NPER data and was found to be adequate, as the visualisation of the fitted values followed the actuals path throughout the sample used.

Otu et al. (2014) used the Seasonal ARIMA model to forecast inflation rates in Nigeria. The study used the sampled data from November 2003 to October 2013 (120 observations). The main objective was to get a structure that represented the data well enough and to forecast the Nigerian inflation rate for the period November 2013 to October 2014 (12 point forecast). After analysing the properties of the inflation series (including model residual analysis), $ARIMA(1,1,1)(0,0,1)_12$ was found to be the best fitting model for the inflation rate. The 12 month forecasts showed a decreasing inflation trend for the period from November 2013 to October 2014. The recommendation from the study forecast results was to assist policy makers in Nigeria on decision making.

Revenue forecasts in Nepal are conducted by two major institutes, the Ministry of Finance (MOF) and the Nepal Rastra Bank (NRB). Yet the methods used to forecast Nepalese revenues were not efficient in capturing revenue flows and there was a lack of solid documented methodology for revenue forecasting Koirala (2012). Koirala used Nepal Rastra Bank's monthly data from August 1997 to August 2012 to estimate and forecast Nepal's total revenue for the financial years 2012/13 and 2013/14 (Nepal's financial years start in August and end in July). In addition, five methods were constructed to assist with this exercise, namely Holts method, Winters method, the Decomposition method, the Seasonal ARIMA method and the Growth method. From all the methods, the Winters (with seasonal component) and the Seasonal ARIMA were found to best fit Nepal's total revenue, which was also shown by their smaller Mean Percentage Error (MPE) and MAPE as compared to the remaining three methods. The two methods were recommended to forecast the total revenue in Nepal as they reduced forecasting errors.

The Box-Jenkins methodology is widely used in different fields, including for natural climatic phe-

nomena such as rainfall. Nirmal and Sundaram. (2010) used SARIMA($SARIMA(0,1,1)(0,1,1)_{12}$) to model the average Tamilnadu rainfall in India. The study used sample data from 1871 to 2006, which was obtained from the Indian Institute of Tropical Methodology (IITM) in Pune, India. The performance evaluation criteria used was the mean absolute percentage error (MAPE). The study concluded that SARIMA models are useful time series models for forecasting Tamilnadu's monthly rainfall.

A similar study was done in Dhaka, Bangladesh, where seasonal ARIMA was used to model and forecast rainfall. The study aim was to assist water authorities to prioritise and manage water demands. Monthly rainfall data from 1981 to June 2010 was obtained, and the RMSE and AIC criterion were used to select the best representative model of the actual rainfall. The $ARIMA(0,0,1)(0,1,1)_1 2$ model was judged to be adequate in explaining or representing the rainfall time series data for the selected sample. Model adequacy techniques were used to confirm the closeness of the fitted values compared to the actual values. A Seasonal ARIMA model fitted was then used to forecast two years (July 2010 to June 2012), Mahsin et al. (2012).

Tourism plays an important part in the economic growth of every country, thus the more tourism inflow to a country, the more the GDP expands. The time series plays an important role in forecasting the future tourism inflow. Singh (2013) built ARIMA models with seasonal effect (SARI M A) to predict the number of international tourist arrivals to Bhutan, India. The study was the first attempt to use the SARIMA in modeling Bhutan tourist arrivals, since there was no literature on modeling tourism arrival for Bhutan. Monthly international tourist arrivals data for the period January 1983 to December 2012 was used to build several SARIMA models. Beside the R^2, $RMSE$, $MAPE$, BIC statistic, the model with residuals that are white noise $ARIMA(0,1,1)(1,1,1)$ was selected, and used to generate monthly forecasts for 2013 and 2014. The study concluded that the forecast from the SARIMA model could provide useful information for tourism arrivals in Bhutan.

2.4 Conclusion

Forecasting future revenue with maximum precision is important for every country's economy, as it leads to a better overall distribution of future budgets. From the above literature review it can be seen that time series models have proven to be useful methods to forecast tax revenues, and are applicable to bigger and smaller tax types. However, there is a need to have a recorded quantitative data sample of the same historical path. These types of methods are more precise for short-term

forecasts (two to three years) because their precision declines over a longer period. To obtain or generate long-term forecasts which are more precise, more knowledge on the variables of interest must be obtained, the model must be well defined, and some statistic such as root mean squared error, mean absolute percentage error, Akaike information criterion, quadratic lost function and many others must be considered. Over and above this, monitoring of the derived, forecasts need to be considered for revision purposes if necessary. The following chapter illustrates the theory on the Holt-Winters and ARIMA/SARIMA-models, looking at model specification and test statistics for forecasting purposes.

Chapter 3

Overview of SARIMA, model identification, estimation, Exponential smoothing, and forecasting methods

3.1 Introduction

This chapter discusses the theory of the SARIMA models, followed by model identification in section 3.3. Section 3.4 covers model estimation methods, while choosing the best model among competing ARIMA/SARIMA models occurs in section 3.5. Testing the significance of parameters in the ARIMA/SARIMA model takes place in section 3.6, a diagnostic check for adequacy in section 3.7, exponential smoothing methods in section 3.8, and finally forecasting in section 3.9.

3.2 Theory of ARIMA/SARIMA modeling

In this section the researcher briefly discusses the basic types of ARIMA/SARIMA models, namely:

- The autoregressive model of order p, denoted by $AR(p)$
- The moving average model of order q, denoted by $MA(q)$
- The autoregressive-moving model of order (p,q), denoted by $ARMA(p,q)$

- The autoregressive integrated moving average model of order (p,d,q), $ARIMA(p,d,q)$, and
- The seasonal $ARIMA$ model of order (p,d,q)(P,D,Q), $ARIMA(p,d,q)(P,D,Q)$

3.2.1 Autoregressive models of order p, $AR(p)$

The $AR(p)$ model for a stationary y_t series is given by:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{3.2.1}$$

where μ is the constant term, p is a non negative integer, $\phi_1, \phi_2, ..., \phi_p$ are model parameters to be estimated. In this case p lags of y_t are deemed to be important in determining the time series behaviour of y_t. The ε_t is the disturbance or error term at time t with zero mean and constant variance σ^2. Equation (3.2.1) can also be written using a backshift operator as:

$$\phi(\mathbf{B}) y_t = \mu + \varepsilon_t \tag{3.2.2}$$

where $\phi(\mathbf{B}) = 1 - \phi_1 \mathbf{B} - \phi_2 \mathbf{B}^2 - ... - \phi_p \mathbf{B}^p$ is a polynomial in $\mathbf{B}$ of order p. The back shift operator ($\mathbf{B}$) is defined in the way that $\mathbf{B}^k \ y_t = y_{t-k}$ where $k = 0, \pm 1, \pm 2, ...$ is the lag period.

An AR(p) process is said to be stationary provided that the absolute roots of the polynomial in $\mathbf{B}$, $\phi(\mathbf{B}) = 0$, are all greater than 1 (Wei (2006)).

An AR(p) process is detected by observing the autocorrelation function (ACF) and/or the partial autocorrelation function (PACF) . With time series data the ACF for y_t can be calculated as:

$$\rho_k = \frac{cov(y_t; y_{t-k})}{var(y_t)} = \frac{\gamma_k}{\gamma_0} \tag{3.2.3}$$

where $cov(y_t; y_{t-1}) = E\left[(y_t - E(y_t))(y_{t+k} - E(y_{t+k}))\right]$ is the covariance between the two variables and $var(y_t)$ is the variance of y_t (University of Pretoria (2013)).

The partial autocorrelation (ρ_{kk}) measures correlation between (time series) observations that are k times apart after controlling for correlations at intermediate lag. If the series is an AR(p) process, it must satisfy the following conditions:

1. The ACF of the process decays exponentially with lag k. That is, the values of ρ_1, ρ_2.... decrease in a steady fashion.
2. The PACF has a non zero lag 1, 2,...p, and has a zero partial autocorrelation at all lags after lag p. This means that it cuts off at lag p.

3.2.2 Moving average models of order q, MA(q)

The MA(q) model for a time series y_t is given by

$$y_t = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - ... - \theta_q\epsilon_{t-q} \tag{3.2.4}$$

where q is a non-negative integer, $\theta_1, \theta_2, ..., \theta_q$ are model parameters to be estimated, and ϵ_t is a series of random errors each with zero mean and constant variance σ^2. Alternatively, in terms of the back shift operator, the MA(q) model may be written as:

$$y_t = \theta(\mathbf{B})\epsilon_t \tag{3.2.5}$$

where $\theta(\mathbf{B}) = 1 - \theta_1\mathbf{B} - \theta_2\mathbf{B}^2 - ... - \theta_q\mathbf{B}^q$ is a polynomial in $\mathbf{B}$ of order q. An MA process of order q is invertible if the roots of the polynomial in $\mathbf{B}$, $\theta(\mathbf{B}) = 0$, all lie outside the unit circle (Box and Jenkins (1970)).

Among the conditions for identification of MA(q) process include;

1. The ACF has a none zero autocorrelation at lag 1, 2, ...q and has a zero autocorrelation at all lags after q. This means that the process cuts off after lag q. That is:

$$\rho_k \neq 0 \quad for \quad k = 1, 2, ..q \tag{3.2.6}$$

$$\rho_k = 0 \quad for \quad k > q \tag{3.2.7}$$

2. The PACF decays exponentially. This implies that the values of $\rho_{11}, \rho_{12}, ...$ decreases in a steady fashion.

3.2.3 Autoregressive moving average models, ARMA(p, q)

The model for the series y_t can be an AR(p) model or an MA(q) model or a combination of both the AR(p) and the MA(q) models. The latter model is called an autoregressive moving average of order (p, q), denoted by ARMA(p, q), and is given by:

$$y_t = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t - \theta_1\epsilon_{t-1} - ... - \theta_q\epsilon_{t-q} \tag{3.2.8}$$

Where $\phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q$ are model parameters to be estimated, and ϵ_t is a series of random errors each with zero mean and constant variance σ^2 (Box and Jenkins (1976)).
Alternatively, the ARMA(p, q) model may be written as:

$$\phi(\mathbf{B})y_t = \theta(\mathbf{B})\epsilon_t \tag{3.2.9}$$

Where $\phi(\mathbf{B}) = 1 - \phi_1\mathbf{B} - \phi_2\mathbf{B}^2 - ... - \phi_p\mathbf{B}^p$ is a polynomial in $\mathbf{B}$ of order p, and $\theta(\mathbf{B}) = 1 - \theta_1\mathbf{B} - \theta_2\mathbf{B}^2 - ... - \theta_q\mathbf{B}^q$ is a polynomial in $\mathbf{B}$ of order q. It can be noted that the AR(p) and MA(q) models are special ARMA(p, q) models. For example, the AR(p) is the $ARMA(p, 0)$ model, and the MA(q) model is $ARMA(0, q)$ model.

The ACF of an ARMA(p, q) process decays exponentially after lag 1. That is:

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\theta_1\phi_1} \tag{3.2.10}$$

Where ϕ_1, θ_1 are parameters of AR(1) and MA(1) respectively. From equation (3.2.10), it can be deduced that if the series y_t is generated by an ARMA(p, q) process, then the sample ACF of the series y_t generally attenuates as the lag increases rather than 'cutting off' at some lag. For example, in case ρ_1 is significantly different from zero, the subsequent ρ_k gets closer and closer to zero (Chatfield (2004)).

3.2.4 Autoregressive integrated moving averages, ARIMA(p, d, q)

In practice, some time series are non-stationary (Gujarati and Porter (2003)). This means that ARMA(p, q) models are not practically applicable to some time series or econometrics data. This implies that we have to first stationalise the series and then fit ARMA(p, q) models to the resultant stationary series. One way is to use variance stabilisation by applying transformations on the data, for example; logarithmic, square root and so forth. The other way of stationalising a time series is differencing (subtracting a specific lag from a time series) when this is done, the resultant series is ARIMA(p,d,q).

The general model for ARIMA(p, d, q) is expressed as:

$$(1 - \phi_1\mathbf{B} - \phi_2\mathbf{B}^2 - ... - \phi_p\mathbf{B}^p)(1 - \mathbf{B})^d Y_t = \vartheta_0 + (1 - \theta_1\mathbf{B} - \theta_2\mathbf{B}^2 - ... - \theta_q\mathbf{B})\epsilon_t \tag{3.2.11}$$

Alternatively, as a backshift operator, the ARIMA(p ,d ,q) is expressed as:

$$\phi(\mathbf{B})y_t = \phi(\mathbf{B})(1 - \mathbf{B})^d Y_t = \theta(\mathbf{B})\epsilon_t \tag{3.2.12}$$

where $\phi(\mathbf{B}) = 1 - \phi_1\mathbf{B} - \phi_2\mathbf{B}^2 - ... - \phi_p\mathbf{B}^p$, $\theta(\mathbf{B}) = 1 - \theta_1\mathbf{B} - \theta_2\mathbf{B}^2 - ... - \theta_q\mathbf{B}^q$, d is the degree

of differencing and ϵ_t is a series of random errors each with zero mean and constant variance σ^2. Model (3.2.12) for the series Y_t is referred to as the autoregressive integrated moving average model of order (p, d, q), and is denoted by ARIMA(p, d, q), (see Diggle (1990)).

Note that the AR(p), MA(q) and ARMA(p, q) models are special ARIMA(p, d, q) models. For example, the ARMA(p, q) model is the $ARIMA(p, 0, q)$ model, the AR(p) model is the $ARIMA(p, 0, 0)$ model and the MA(q) model is the $ARIMA(0, 0, q)$ model.

3.2.5 $ARIMA(p, d, q)(P, D, Q)_s$

Time series data can be non-stationary, and/or, exhibit the rise and fall on the fixed points for an observed period. Such series are called seasonal time series data and can be seasonally differenced to obtain stationarity. Seasonal differenced series is represented by the following equation.

$$\nabla^d \nabla_s^D y_t = (1 - \mathbf{B^d})(\mathbf{1} - \mathbf{B^s})^{\mathbf{D}} \mathbf{y_t} \tag{3.2.13}$$

which, on re-arranged to becomes:

$$\nabla^d \nabla_s^D y_t = (1 - \mathbf{B^d} - \mathbf{B^{sD}} + \mathbf{B^{sD+d}}) \mathbf{y_t} \tag{3.2.14}$$

where ∇^d is normal differencing operator and ∇_s^D is the seasonal difference operator. Assume that y_t is a monthly series ($s = 12$) exhibiting seasonal pattern with first seasonal difference ($D = 1$) and is a homogeneous non-stationary series of order 1 ($d = 1$), then equation (3.2.14) above can be expressed as follows:

$$\begin{aligned} \nabla^1 \nabla_{12}^1 y_t &= (1 - \mathbf{B} - \mathbf{B^{12}} + \mathbf{B^{13}}) \mathbf{y_t} \\ &= y_t - y_{t-1} - y_{t-12} + y_{t-13} \end{aligned} \tag{3.2.15}$$

Similarly for quarterly series ($s = 4$), equation (3.2.14) is expressed as follows;

$$\nabla^1 \nabla_4^1 y_t = y_t - y_{t-1} - y_{t-4} + y_{t-5} \tag{3.2.16}$$

The general seasonal ARIMA models is represented by $ARIMA(p, d, q)(P, D, Q)_s$ Shumway and Stoffer (2006), using a back shift operator, it is then expressed as;

$$\phi(\mathbf{B}) \mathbf{\Phi}(\mathbf{B^s}) \mathbf{x_t} = \delta + \theta(\mathbf{B}) \mathbf{\Theta}(\mathbf{B^s}) \eta_\mathbf{t} \tag{3.2.17}$$

where:
$\mathbf{x_t} = (1 - \mathbf{B^d})(\mathbf{1} - \mathbf{B^s})^{\mathbf{D}} \mathbf{y_t} = (\mathbf{1} - \mathbf{B^d} - \mathbf{B^{sD}} + \mathbf{B^{sD+d}}) \mathbf{y_t}$ is the product of seasonal differencing D

and non-seasonal differencing d, $\mathbf{s}$ is the series seasonality which takes the value 4 for quarterly time series data and 12 for monthly time series data, δ is the constant term and η_t is the disturbance or error term at time t.

Furthermore, Yurekli et al. (2005) express $\phi(\mathbf{B}), \mathbf{\Phi}(\mathbf{B^s}), \theta(\mathbf{B}) \mathbf{and} \mathbf{\Theta}(\mathbf{B^s})$ as follows:

$\phi(\mathbf{B}) = \mathbf{1} - \phi_\mathbf{1}\mathbf{B} - \phi_\mathbf{2}\mathbf{B^2} - ... - \phi_\mathbf{p}\mathbf{B^p}$ is non-seasonal AR components of order p.
$\Phi(\mathbf{B^s}) = \mathbf{1} - \mathbf{\Phi_1}\mathbf{B^s} - \mathbf{\Phi_2}\mathbf{B^{2s}} - ... - \mathbf{\Phi_P}\mathbf{B^{Ps}}$ is seasonal AR components of order P.
$\theta(\mathbf{B}) = \mathbf{1} - \theta_\mathbf{1}\mathbf{B} - \theta_\mathbf{2}\mathbf{B^2} - ... - \theta_\mathbf{q}\mathbf{B^q}$ is the non-seasonal MA components of order q, and
$\Theta(\mathbf{B^s}) = \mathbf{1} - \mathbf{\Theta_1}\mathbf{B^s} - \mathbf{\Theta_2}\mathbf{B^{2s}} - ... - \mathbf{\Theta_Q}\mathbf{B^{Qs}}$ is the seasonal MA components of order Q.

3.3 Model Identification

Once a tentative ARIMA model for a given time series has been fitted using the ARIMA as discussed in section 3.2.4, the next step is to identify the fitted model. Model identification refers to the methodology in identifying the required transformation, such as variance stabilizing transformation and/or difference transformation, the decision to include the deterministic parameter ϑ when $d \geq 1$, and the proper orders of p and q for the model Gooijer et al. (1985).

To illustrate model identification, we consider a time series modeled by the general ARIMA(p,d,q) in section 3.2.4 and follow the following steps:

- Plot the time series data and choose proper transformation. Through careful examination of the fitted plot, we get an idea on whether the series contains trend, seasonality, outliers, non-constant variance or is generally non-stationary. If a series is non-stationary, we difference it or perform variance stabilization techniques such as taking natural logarithm (ln) to reduce it to stationarity.

- The research compute and examine the sample ACF and the sample PACF of the original series to further confirm a necessary degree of differencing so that the differenced series is stationary. In this, if the sample ACF decays very slowly and the sample PACF cuts off at lag 1, then taking the first difference $(1 - \mathbf{B})Y_t$ is adequate. Alternatively, we could perform a unit root test as proposed by Dickey and Fuller (1979).

- Compute and examine the sample ACF and PACF of the properly transformed and/or differenced series to identify the orders of p and q, where p is the highest order in autoregressive

polynomial $(1 - \phi_1\mathbf{B} - \phi_2\mathbf{B}^2 - ... - \phi_p\mathbf{B}^p)$ and q is the highest order in the moving average polynomial $(1-\theta_1\mathbf{B}-\theta_2\mathbf{B}^2-...-\theta_q\mathbf{B}^q)$. The following table gives the summary of indicators of orders p and q (Bowerman et al., 2004).

Table 3.1: Characteristics of theoretical ACF and PACF for stationary process

Process	ACF	PACF
AR(p)	Tail off as exponential decay of damped sine wave	cuts off after lag p
MA(q)	cuts off after lag q	Tail off as exponential decay of damped sine wave
ARMA(p,q)	Tails off after lag (q-p)	Tails off after lag (p-q)

We then identify the orders p and q by matching the patterns in the sample ACF and PACF with the theoretical patterns of known models.

- Test for deterministic trend ϑ_0 when $d > 0$. This is done by comparing the sample mean $\overline{W}$ of the differenced series $W_t = (1-\mathbf{B})^d Y_t$ with its approximate standard error $S_{\overline{W}}$. In general the standard error $S_{\overline{W}}$ is given by:

$$S_{\overline{W}} = \left[\frac{\hat{\gamma}_0}{n}(1 + 2\hat{\rho}_1 + 2\hat{\rho}_2 + ... + 2\hat{\rho}_k)\right]^{\frac{1}{2}} \tag{3.3.1}$$

where $\hat{\gamma}_0$ is the sample variance and $\hat{\rho}_1, \hat{\rho}_2, ..., \hat{\rho}_k$ are the first k significant sample ACFs of differenced series.

3.4 Model estimation methods

Once a tentative $ARIMA/SARIMA$ model has been identified, the next step is to estimate and test for significance of the model parameters. Typical methods of estimating the model parameters are either the least squares method and/or the maximum likelihood methods (Bowerman et al. (2004)).

These methods are briefly reviewed under the assumption that the identified tentative ARIMA model for a given time series y_t model in equation (3.2.9) can be re-rewritten as:

$$\epsilon_t = \frac{\phi(\mathbf{B})}{\theta(\mathbf{B})} y_t \tag{3.4.1}$$

Where $\phi(\mathbf{B}) = 1 - \phi_1\mathbf{B} - \phi_2\mathbf{B}^2 - ... - \phi_p\mathbf{B}^p$, $\theta(\mathbf{B}) = 1 - \theta_1\mathbf{B} - \theta_2\mathbf{B}^2 - ... - \theta_q\mathbf{B}^q$ and ϵ_t is a series of random errors each with zero mean and constant variance σ^2.

3.4.1 Least squares method

The least squares estimates of the parameters of the ARIMA model (3.2.11) and SARIMA model (3.2.13) are values of $(\underline{\phi}^T, \underline{\theta}^T)^T$ and $(\underline{\phi}^T, \underline{\theta}^T, \underline{\Phi}^T, \underline{\Theta}^T)^T$ which minimises the respective error sum of squares:

$$SSE(\phi, \theta) = \sum_{i=1}^{n} \epsilon_t^2 \tag{3.4.2}$$

and

$$SSE(\phi, \Phi, \theta, \Theta) = \sum_{i=1}^{n} \varepsilon_t^2 \tag{3.4.3}$$

This method is used especially where the unknown parameters are linear functions known constants. The parameters' estimates are obtained by taking the first derivative with respect to the parameters of equations (3.4.2) and (3.4.3), equating the derivatives to zero, and numerically solving the resultant equations to obtain the least squares parameter estimates.

3.4.2 Maximum likelihood method

Under the assumption that the error terms (ϵ or ε) in section 3.4.1 are independent normally distributed random variables with mean zero, and variance δ^2, ($iid\ N(0, \delta^2)$, the maximum likelihood estimate of the parameters of the SARIMA model (3.2.13) are values of $(\underline{\phi}^T, \underline{\Phi}^T, \underline{\theta}^T, \underline{\Theta}^T)^T$ which maximise likelihood function given by:

$$L(\underline{\phi}^T, \underline{\theta}^T, \underline{\Phi}^T, \underline{\Theta}^T, \delta^2) = (2\pi\delta^2)^{-\frac{n}{2}} e^{(-\frac{1}{2\delta^2}\sum_{i=1}^{n}\epsilon_t^2)} \tag{3.4.4}$$

where:

$\underline{\phi} = (\phi_1, \phi_2, ..., \phi_p)^T$,

$\underline{\theta} = (\theta_1, \theta_2, ..., \theta_q)^T$,

$\underline{\Phi} = (\Phi_1, \Phi_2, ..., \Phi_P)^T$, and

$\underline{\Theta} = (\Theta_1, \Theta_2, ..., \Theta_Q)^T$

To obtain the estimates using the maximum likelihood function, we take the natural logarithm on both sides of the equation (3.4.4) and then equate the first derivative of the resultant function to zero.

3.5 Choosing the best model among competing SARIMA models

There may be several competing ARIMA/SARIMA models that adequately describe the given time series. The problem becomes that of choosing the best among the competing models. Two criteria of choosing the best model which are commonly used are the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

The AIC is defined by the following equation.

$$AIC = -2\ln(\text{Likelihood}) + 2r \tag{3.5.1}$$

Where the Likelihood is the maximum likelihood given in equation (3.4.4) evaluated at the maximum likelihood estimates of the model parameters, and r denotes the number of model parameters. The AIC increases with the number of model parameters (r), and the best model is one with the smallest AIC.

The BIC is an extension of the AIC, and is given by equation:

$$BIC = -2\ln(\text{ Likelihood}) + r\ln(N) \tag{3.5.2}$$

Where N is the number of observations in the differenced series (to stationarise the original series). As with the AIC, the best model among competing ARIMA/SARIMA models is one with the smallest BIC. This study shall employ both criteria in assessing the best model that fits tax revenue data.

3.6 Testing the significance of the parameters in the SARIMA model

Provided that the series is long enough and that the residuals are white noise, the significance or insignificance of the parameters in the model are tested using the t-test given by:

$$t = \frac{\text{estimate}}{se(\text{estimate})} \tag{3.6.1}$$

which has an asymptotic standard normal distribution Chatfield (2004). The decision rule of the test is that a parameter is insignificantly different from zero at the $\alpha \in (0, 1)$ level of significance if $|t| > z_{\alpha/2}$ - the $(1-\alpha)100$ percentile of the $N(0, 1)$ distribution. Alternatively, we could use the p-value judgement. This procedure considers parameters to be significant if its p-value is less than the set value of significance. Parameters which are insignificant in the model are then removed and the reduced model refitted to the series. This process is repeated until the best model is obtained. Once we have identified the model assumed to be the best fit, we then do diagnostic checks to see if the identified or selected model is appropriate.

3.7 Diagnostic check for adequacy

A good way to check model adequacy is to analyse the residuals of the series obtained from the model. If the model is correctly specified, and the parameters are reasonably close to the true value, the residuals should have nearly the properties of white noise. This means that they should behave roughly like independent, identically distributed normal variables with zero mean and common variance. Hence, residual will be stationary in both the mean and variance. There are several tests for stationarity of residuals. Among the prominent ones are:

- ⋆ Graphical analysis,
- ⋆ Autocorrelation function (ACF) and correlogram analysis; and
- ⋆ Portmanteau test

3.7.1 Graphical Analysis

In this analysis we plot the series against time to examine the nature of a time series. The aim is to find out whether in the time series there is a possibility of:

- An upward or downward trend,
- The mean varying with time, and/or
- The variance being constant over time

From the analysis of the nature of plot, we then determine whether the residuals are stationary or not. If the plot suggests a rectangular scatter plot around a zero horizontal level with no trends whatsoever, then the residuals are stationary.

3.7.2 Autocorrelation function (ACF) and correlogram

Another test is based on the autocorrelation function (ACF). The ACF at lag k ($k = 1, 2, ...$), denoted by ρ_k is defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{covariance\ at\ lag\ k}{variance} \tag{3.7.1}$$

The value of the ACF lies between -1 and 1, that is, $-1 \leq \rho_k \leq 1$ and is unitless because γ_k and γ_0 are measured in the same units . The plot of ρ_k against k is known as the population correlogram.

Since in practice information on the population is not always available, we focus on the use of the sample autocorrelation function (SACF), denoted as $\hat{\rho}_k$. This is defined as:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \tag{3.7.2}$$

$$\hat{\gamma}_k = \frac{\sum (Y_0 - \overline{Y})(Y_{t+k} - \overline{Y})}{n} \qquad \hat{\gamma}_0 = \frac{\sum (Y_t - \overline{Y})^2}{n} \tag{3.7.3}$$

where n is the sample size and $\overline{Y}$ is the sample mean.

The plot of $\hat{\rho}_k$ against k is called the sample correlogram. Based on the sample correlogram, the series is considered to be stationary, if it has no significant spike at lag k. Otherwise, it is non-stationary. Sample correlogram of non-stationary series usually have high values of the auto-correlation coefficient at various lags. In drawing the sample correlogram, it is usually important to select the best lag length. The choice of the lag length can be selected in any of the following ways:

1. Using the rule of the thumb as indicated by Gujarati and Porter (2003). This requires computing the ACF up to one third to one-quarter the length of the time series.
2. Using statistical tests like the Bartlett test, Box-Pierce Q-statistic, and Ljung-Box (LB) test.

- **Bartlett test**

According to the Bartlett test, in large samples, if a time series is purely random, that is, white noise, the sample autocorrelation (SACF) coefficients $\hat{\rho}_k$ are approximately normal with mean zero and variance equal to one over the sample size, that is, $\hat{\rho}_k \sim N(0; \frac{1}{n})$.

This implies that we test hypothesis of $H_0 : \rho_k = 0$ against $H_1 : \rho_k \neq 0$ at each individual k lag, and estimate ρ_k using the confidence interval estimate:

$$\hat{\rho}_k \pm z_{\frac{\alpha}{2}} \times Se_{\hat{\rho}_k} \tag{3.7.4}$$

where $z_{\frac{\alpha}{2}}$ is the multiplier from the z-standard normal tables and $Se_{\hat{\rho}_k}$ is the standard error of $\hat{\rho}_k$. From equation (3.7.4), if the confidence interval estimate includes zero, we don't reject H_0 at the set level of significance (α). Otherwise we reject H_0. The rejection of H_0 implies that the specified SACF is significant at the specified lag k. Hence, we select the k^{th} lag length.

The setback for this test is that it can be used only on large ($n > 30$) samples size, however this was not a limitation to our study because our data set has greater than 30 observations.

- **Box-Pierce Q-statistic**

Instead of testing for the individual autocorrelation, as indicated by the Bartlett test, we can test the joint hypothesis that all the ρ_k up to a certain lag are equal to zero using the Box and Pierce Q-statistic denoted as:

$$Q = \sum_{k=1}^{m} \hat{\rho}_k^2 \tag{3.7.5}$$

where n is the sample size and m is the lag length. This test is also used to test whether a time series is white noise. If sample sizes are large, this test follows a Chi-square distribution with m degrees of freedom. The null hypothesis $H_0 : \hat{\rho}_k = 0$ is rejected against $H_1 : \hat{\rho}_k \neq 0$ if the computed test statistic value Q exceeds the critical value (χ^2_m) from the Chi-square table with m degrees of freedom.

- **Ljung-Box test**

The alternative way, we could test for the joint hypothesis that all the ρ_k up to a certain lag are equal to zero, to use that Ljung-Box (LB) statistic. This has a test statistic given by:

$$LB = n(n+2)\sum_{k=1}^{m}(\frac{\hat{\rho}_k^2}{n-k}) \tag{3.7.6}$$

This statistic also follows a chi-square (χ^2_m) distribution with m degrees of freedom. In large samples both the Q and LB test follow the Chi-square distribution, however, the LB test is considered to be the most powerful. From this we draw that if sample sizes are small, more research needs to be done on which tests are best to test for white noise in small samples. Nevertheless this was also not a limitation as the data set used was considered to be large enough.

3.7.3 Portmanteau test

If the ARIMA/SARIMA model adequately fits the given time series, then the residuals from fitting the model should be white noise. For this reason, a plot of the ACF of the residuals should show no significant spikes (correlations) at all lags k.

One formal test of the null hypothesis that the residuals from fitting the ARIMA/SARIMA model are white noise is the Portmanteau test Chatfield (2004). The test for white noisiness of the residuals uses the sample ACF of the residuals to test the null hypothesis:

$$h_0 : \rho_1 = \rho_2 = ... = \rho_k = 0, \text{ for some } k > 1 \tag{3.7.7}$$

The test statistic for the hypothesis is:

$$Q = N(N-2)\sum_{j=1}^{k} \frac{\hat{\rho}_j^2}{(N-k)} \sim \chi^2_{k-p-q} \tag{3.7.8}$$

Where N is the number of observations in the differenced series (to stationarise the original series). The null hypothesis that the residuals series is white noise is rejected at the $\alpha \in (0, 1)$ level of significance if $Q > \chi^2_{k-p-q}$ - the $(1-\alpha)100$ percentile of the χ^2_{k-p-q} distribution.

Once diagnostic checks have been done on the identified model, we can then use the selected model to forecast future values.

The following section will discuss alternative models under Exponential Smoothing Method.

3.8 Exponential Smoothing

Exponential smoothing provides a forecasting method that is most effective when the components (trend and seasonal factors) of a time series may change over time. It is a method that weights the observed time series value unequally. This is accomplished by using one or more smoothing constants, which determines how much weight is given to each observation. Historically the exponential smoothing methods were intuitive methods not based on any formal statistical models, however work on state space models with a single source by Hyndman et al. (2002) provided a statistical framework for the exponential smoothing methods. Among these methods include:

- Simple exponential smoothing,
- Holt's trend corrected exponential smoothing; and
- Holt-Winters methods

3.8.1 Simple exponential smoothing

Simple exponential smoothing is used for forecasting a time series when there is no trend or seasonal pattern but the mean (or level) of the time series (y_t) is slowly changing over time. This method gives the most recent observation more weight compared to the successive older observations. This enables the forecaster to update the estimate of the level of the time series, so that changes in the level can be detected and incorporated into the forecasting system. In summary, simple exponential smoothing works as follows:

Given a time series $y_1, y_2, y_3, ...y_n$, we begin by calculating the initial level (l_0) as:

$$l_0 = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.8.1}$$

Next, assume that at the end of time period $T-1$, we have an estimate l_T for the level of the time series, we compute the update estimate by using the smoothing equation:

$$l_T = \alpha y_T + (1-\alpha) l_{T-1} \tag{3.8.2}$$

Where α is the smoothing constant between 0 and 1.

Given a simple exponential smoothing is used for forecasting a time series when there is no trend or seasonal pattern, it may not be applicable in this study because the trends of PIT, CIT, VAT and TTAXR change with time. The researcher therefore explored another possible exponential smoothing technique - Holts trend corrected exponential smoothing.

3.8.2 Holt's trend corrected exponential smoothing

Holt's trend corrected smoothing is applied when a time series displays changing level (mean) and the growth rate (slope). In this technique, the estimated level in time period T uses the smoothing constant α and is given by the:

$$l_T = \alpha y_t + (1-\alpha)(l_{T-1} + b_{T-1}) \tag{3.8.3}$$

where α is the smoothing constant between 0 and 1. The growth rate (b_T) of the time series T uses the smoothing constant γ and is given by:

$$b_T = \gamma \left[l_T - l_{T-1} \right] + (1-\gamma) b_{T-1} \tag{3.8.4}$$

Where γ is the smoothing constant between 0 and 1.

Despite the fact that Holts corrected smoothing is used to forecast a time series that displays changing level and changing growth rate, it may not be applicable to SARS data sets because of the data seasonality. Correction from the mean of data in literature leads to a loss of vital information Bowerman et al. (2004), therefore the research explore another possible exponential smoothing technique - Holt-Winters exponential smoothing.

3.8.3 Holt-Winters methods

The Holt-Winters methods are designed for a time series that exhibits linear trend and seasonality. These methods include;

- Additive Holt-Winters method; and
- Multiplicative Holt-Winters method

3.8.3.1 Additive Holt-Winters method

The additive Holt-Winters method is appropriate when a time series has a linear trend with an additive seasonal pattern, for which the level (mean), the growth rate and the seasonal pattern may be changing. This model can be described as:

$$y_t = (\beta_0 + \beta_1 t) + Sn_t + \epsilon_t \tag{3.8.5}$$

Where β_0 is the growth rate and Sn_t the fixed seasonal pattern. The additive Holt-Winters method can be summarised as follows:

Suppose that the time series $y_1, y_2, ...y_n$ exhibits linear trend locally and has a seasonal pattern with constant (additive) seasonal variation and the level, growth rate and seasonal pattern may be changing. In his case the estimate l_T for the level, the estimate b_T for the growth rate, and the estimate S_{nT} for the seasonal factor of the time series in time period T are given by the smoothing equation:

$$l_T = \alpha(y_T - Sn_{T-1}) + (1-\alpha)(l_{T-1} + b_{T-1}) \tag{3.8.6}$$

$$b_T = \beta(l_T - l_{T-1}) + (1-\beta)b_{T-1} \tag{3.8.7}$$

$$S_{nT} = \gamma(y_T - l_T) + (1-\gamma)Sn_{T-1} \tag{3.8.8}$$

Where α, β and γ are smoothing constants between 0 and 1, l_{T-1} and b_{T-1} are estimates in time period $T-1$ for the level and growth rate respectively, and Sn_{T-1} is the estimate in time $T-1$

for the seasonal factor.

The application of these methods requires that a time series have a linear trend with an additive seasonal pattern. Unfortunately this is not always the case for SARS data because of changes in economic performance. The additive Holt-Winters model may be suitable for modeling some taxes in the South African economy, however some tax evaluations tend to change in a multiplicative manner. The researcher thus turned to the multiplicative Holt-Winters methods discussed below.

3.8.3.2 Multiplicative Holt-Winters methods

If a time series has a linear trend with a fixed growth rate b and a fixed seasonal pattern Sn_t with increasing (multiplicative) variation, it may be described by a multiplicative model:

$$y_t = (\beta_0 + \beta_1 t) \times Sn_t \times IR_t \tag{3.8.9}$$

The multiplicative Holt-Winters method is appropriate when a time series has a linear trend with multiplicative seasonal pattern for which the level, growth rate and the seasonal pattern may be changing rather than fixed.

In this method, the estimate l_T for the level, the estimate for the growth rate (b_T), and the estimate for the seasonal factor of the time series in time period T are given by following smoothing equations:

$$l_T = \alpha(\frac{y_T}{Sn_{T_L}}) + (1-\alpha)(l_{T-1} + b_{T-1}) \tag{3.8.10}$$

$$b_T = \beta(l_T + l_{T-1}) + (1-\beta)b_{T-1} \tag{3.8.11}$$

$$Sn_T = \gamma(\frac{y_T}{l_T}) + (1-\gamma)Sn_{T-L} \tag{3.8.12}$$

Where α, β and γ are smoothing constant between 0 and 1, l_{T-1} and b_{T-1} are estimates in time period T-1 for the level and the growth rate respectively, Sn_{T-L} is the estimate in time period $T-L$ for the seasonal factor.

3.9 Forecasting

The predictions of future events and conditions are called forecasts, and the act of making such predictions is called forecasting. In this study, forecasts discussed below will depend on whether the technique chosen is ARIMA/SARIMA or smoothing technique. In case of smoothing techniques, the point forecast and their confidence interval are derived.

3.9.1 Forecasts for SARIMA models

For a given time series, they may be several competing SARIMA models for forecasting. According to Wei (2006), the models can be compared for goodness-of-forecasting using four criteria described below. Fit the SARIMA models to the $t-l$ $(0 < l \leq t)$ observations of the time series and use the fitted models to forecast the last l observed values of the series. Calculate:

$$\epsilon_t(t-l+j) = Y_{t-l+j} - \hat{Y}_{t-l+j}, j = 1, 2, ..., l \tag{3.9.1}$$

where for $j = 1, 2, ..., l$ $\hat{Y}_{t-l+j}$ is the forecast of Y_{t-l+j} using any of the competing models; and compute

Mean percentage error/bias given by

$$MPE = \frac{1}{l}\sum_{j=1}^{l}\frac{\epsilon_t(t-l+j)}{Y_{t-l+j}} \times 100\% \tag{3.9.2}$$

Mean square error given by

$$MSE = \frac{1}{l}\sum_{j=1}^{l}\epsilon_t^2(t-l+j) \tag{3.9.3}$$

Mean absolute error given by

$$MAE = \frac{1}{l}\sum_{j=1}^{l}|\epsilon_t(t-l+j)| \tag{3.9.4}$$

and

Mean absolute percentage given by

$$MAPE = \frac{1}{l}\sum_{j=1}^{l}\left|\frac{\epsilon_t(t-l+j)}{Y_{t-l+j}}\right| \times 100\% \tag{3.9.5}$$

for each of the competing models. The best model for forecasting is the one with the smallest MPE, MSE, MAE or $MAPE$, depending on the criterion/criteria which one chooses to use. The MPE, MSE, MAE or $MAPE$ criteria are also applicable to the exponential smoothing methods for model selection.

3.9.2 Forecasts for exponential smoothing

In this section, we discuss point forecasts and derive their respective confidence interval for the smoothing techniques discussed in section 3.2

3.9.2.1 Forecast for simple exponential smoothing

In simple exponential smoothing, a point forecast at time T of any future value $y_{T+\tau}$ of a time series is the last estimate l_T for the mean of the time series, because there is no trend or seasonal pattern to exploit. Such a forecast made in time period T for $y_{T+\tau}$ is given by:

$$\hat{y}_{T+\tau}(T) = l_T \quad for \;\; (T = 1, 2, 3..) \tag{3.9.6}$$

If T=1, the a 95% prediction interval computed in time period T for $y_{T+\tau}$ is

$$[l_T \pm Z_{0.025}s] \tag{3.9.7}$$

If T=2, the a 95% prediction interval computed in time period T for $y_{T+\tau}$ is

$$\left[l_T \pm Z_{0.025}s\sqrt{1+\alpha^2}\right] \tag{3.9.8}$$

In general, for any time T, a 95% prediction interval computed in time period T for $y_{T+\tau}$ is

$$\left[l_T \pm Z_{0.025}s\sqrt{1+(T-1)\alpha^2}\right] \tag{3.9.9}$$

Where the standard error s at time T is given by:

$$s = \sqrt{\frac{SSE}{T-1}} = \sqrt{\frac{\sum \left[y_T - l_{T-1}\right]^2}{T-1}} \tag{3.9.10}$$

3.9.2.2 Forecast for Holt's trend corrected exponential smoothing

In simple exponential smoothing, a point forecast at time T of any future value $y_{T+\tau}$ of a time series is given by:

$$\hat{y}_{T+\tau} = l_T + \tau b_T \quad \text{for} \;\; T = 1, 2, ... \tag{3.9.11}$$

If $\tau = 1$, the a 95% prediction interval computed in time period T for $y_{T+\tau}$ is computed from:

$$[(l_T + b_T) \pm Z_{0.025}s] \tag{3.9.12}$$

If $\tau = 2$, the a 95% prediction interval computed in time period T for $y_{T+\tau}$ is computed from:

$$\left[(l_T + 2b_T) \pm Z_{0.025}s\sqrt{1+\alpha^2(1+\gamma^2)}\right] \tag{3.9.13}$$

If $\tau = 3$, the a 95% prediction interval computed in time period T for $y_{T+\tau}$ is computed from:

$$\left[(l_T + 3b_T) \pm Z_{0.025}s\sqrt{1+\alpha^2(1+\gamma^2)+\alpha^2(1+2\gamma)^2}\right] \tag{3.9.14}$$

In general, for any time $\tau \geq 2$, a 95% prediction interval computed in time period T for $y_{T+\tau}$ is

$$\left[(l_T + \tau b_T) \pm Z_{0.025} s \sqrt{1 + \sum \alpha^2 (1 + j\gamma)^2}\right] \tag{3.9.15}$$

Where the standard error s at time T is given by:

$$s = \sqrt{\frac{SSE}{T-2}} = \sqrt{\frac{\sum \left[y_t - (l_T + b_{T-1})\right]^2}{T-2}} \tag{3.9.16}$$

3.9.2.3 Forecast for additive Holt-Winters method

A point forecast made in time period T for $y_{T+\tau}$ is:

$$\hat{y}_{T+\tau}(T) = l_T + \tau b_T + Sn_{T+\tau-L} \tag{3.9.17}$$

where $Sn_{T+\tau-L}$ is the "most recent" estimate of the seasonal factor for the season corresponding to time period $T + \tau$.

A 95% prediction interval computed in time period T for $y_{T+\tau}$ is:

$$[\hat{y}_{T+\tau}(T) \pm Z_{0.025} s \sqrt{c_\tau}] \tag{3.9.18}$$

for $\tau = 1$ and $c_\tau = 1$.

If $2 \leq T \leq L$ then

$$c_\tau = \left[1 + \sum_{j=1}^{T-1} \alpha^2 (1 + j\gamma)^2\right] \tag{3.9.19}$$

If $L \leq T$ then

$$c_T = 1 + \sum_{j=1}^{T-1} \left[\alpha(1 + j\gamma) + d_{j,L}(1 - \alpha)\delta\right]^2 \tag{3.9.20}$$

where $d_{j,L} = 1$ if j is an integer multiple of L and 0 otherwise.

The standard error s is computed in time period T is:

$$s = \sqrt{\frac{SSE}{T-3}} = \sqrt{\frac{\sum_{t=1}^{T} \left[y_T - \hat{y}_T(T-1)\right]^2}{T-3}} = \sqrt{\frac{\sum_{i=1}^{T} \left[y_T - (l_{T-1} + b_{T-1} + Sn_{T-L})\right]^2}{T-3}} \tag{3.9.21}$$

3.9.2.4 Forecast for multiplicative Holt-Winters method

The point forecast made in time period T for $y_{T+\tau}$ is:

$$\hat{y}_{T+\tau}(T) = (l_T + b_T) Sn_{T+\tau-L} \quad for \;\; T = 1, 2, ... \tag{3.9.22}$$

where $Sn_{T+\tau-L}$ is the "the most recent" estimate of the seasonal factors for this season corresponding to time period T.

An approximate 95% confidence interval estimate computed in time period T for $y_{T+\tau}$ is:

$$[\hat{y}_{T+\tau}(T) \pm Z_0.025 s(\sqrt{c_\tau})(Sn_{T+\tau-L})] \tag{3.9.23}$$

If $\tau = 1$ then $c_1 = (l_T + b_T)^2$

If $\tau = 2$ then $c_2 = \alpha^2(1+\gamma)^2(l_T + b_T)^2 + (L_T + 2b_T)^2$

If $\tau = 3$ then $c_3 = \alpha^2(1+2\gamma)^2(l_T + b_T)^2 + \alpha^2(1+\gamma)^2 + (L_T + 2b_T)^2 + (l_T + 3b_T)^2$

Chapter 4

Application of SARIMA and Holt-Winters models on South African taxes

4.1 Introduction

This chapter focuses on the application of the SARIMA models and the Holt-Winters methods on the major taxes, namely PIT, CIT, VAT and TTAXR at SARS, using internal monthly data from January 1995 to March 2010. As specified in Chapter 1, the statistical R-software is used for model fitting and forecasting. Figure 4.1(a.) to (d.) below are graphic representations of the major taxes and TTAXR . The data was loaded into R using the following codes:

```
mydata = read.csv("data path with forward slash/data.csv",
header = T, sep =';', dec =';')
```

The data of interest (mydata) contains the numeric variables PIT, CIT, VAT and TTAXR recorded in millions rand and are converted to time series object of monthly occurrences.

```
pit=ts(with(mydata,PIT),start=c(1995,1),end = c(2010,3),freq=12)
cit=ts(with(mydata,CIT),start=c(1995,1),end = c(2010,3),freq=12)
vat=ts(with(mydata,VAT),start=c(1995,1),end = c(2010,3),freq=12)
ttaxr=ts(with(mydata,TTAXREV),start=c(1995,1),end = c(2010,3),freq=12)
```

The following R-codes were used to obtain graphic representation for major taxes and TTAXR in Figure 4.1.

```
par(mfrow = c(2,2)) # display a 2 by 2 R-Graphics window
plot(pit,ylab = 'Rand Million', main ='(a.) Personal Income Tax (PIT)',
xlab = 'Period')
plot(cit,ylab = 'Rand Million', main ='(b.) Corporate Income Tax (CIT)',
xlab = 'Period')
plot(vat,ylab = 'Rand Million', main ='(c.) Value Added Tax (VAT)',
xlab = 'Period')
plot(ttaxr,ylab = 'Rand Million', main ='(d.) Total Tax Revenue(TTAXR)',
xlab = 'Period')
```

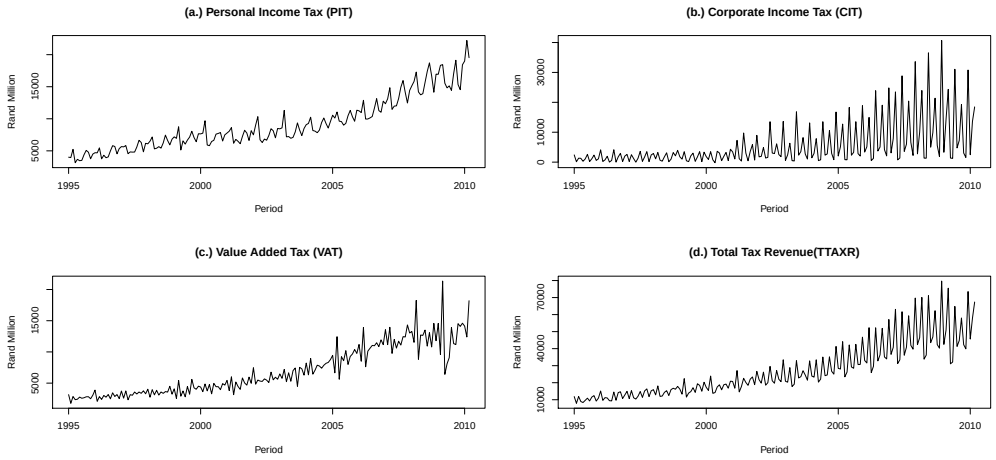

Figure 4.1: Major Taxes and Total Tax Revenue

It is important to note that the SARS fiscal year runs from April to March. The monthly frequency data above was used for the model and the fitted values were aggregated to obtain the yearly fitted values. The initial sample was then used to forecast SARS fiscal year 2010/11. Since it is known that time series models are good for modeling short-term forecasts, the initial sample was increased to the end of March 2011 in order to forecast the 2011/12 SARS fiscal year. Lastly, the forecasts for 2012/13 were obtained using the sample ending March 2012. The output can be assessed by first analysing the results of PIT, then VAT, followed by CIT 5and TTAXR.

4.2 Personal Income Tax

Personal Income Tax (PIT) is the largest source of revenue in the South African economy, contributing approximately 34% of the Total Tax Revenue. Individuals generally receive most of their incomes as salary/wages, pension/retirement payments and investment income (interests and dividends). Some individuals may also have a business income which is taxable as a personal income, for example, sole proprietors and partners SARS and National Treasury (2012).

```
tsdisplay(pit,main ='Level values of PIT, ACF and PACF')
```

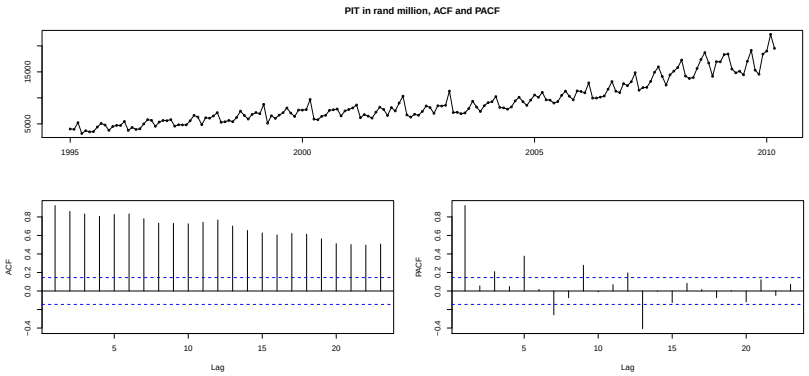

Figure 4.2: Personal Income Tax (PIT), ACF and PACF

Figure 4.2 clearly shows the gradual increasing trend in PIT, indicating an additive seasonality predictable over time provided that there are no shocks or structural changes. This movement in PIT is influenced by wages and salaries which normally increase once a year, and as a result an additive increase in the PIT time series is expected. The following section focuses on modeling and forecasting PIT using the SARIMA and Holt-Winters methods respectively.

4.2.1 SARIMA model for PIT

As discussed in Chapter 3, in order to fit the SARIMA model the data need to be stationary in mean with a constant variance. From section 4.2 above, it can be conclude that the PIT series at level or original values is not stationary, because its mean and variance change over time. The natural logarithmic transformation is used to minimize variation in the PIT time series data. The research then re-examine the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the transformed $ln(PIT)$, series to verify the non-stationarity of the data.

4.2.1.1 ln transformed Personal Income Tax ACF and PACF

```
tsdisplay(log(pit),main ='Level values of ln(PIT), ACF and PACF')
```

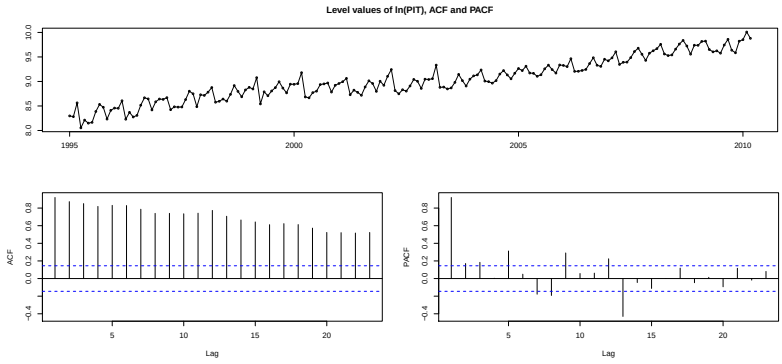

Figure 4.3: Personal Income Tax ln(PIT) Time Series Display at level values

All the ACF from Figure above are outside the interval limits of $\pm\frac{1}{\sqrt{T}}$ and has a damped sine wave; the PACF shows the spike on some few observations. This means that ln(PIT) data is non-stationary at level (mean) values, therefore the research consider differencing the data to obtain stationarity. Differencing the transformed series is done using the following R-codes:

```
tsdisplay(d(log(pit)),main ='Level values of ln(PIT), ACF and PACF')
```

The output of which is as follows:

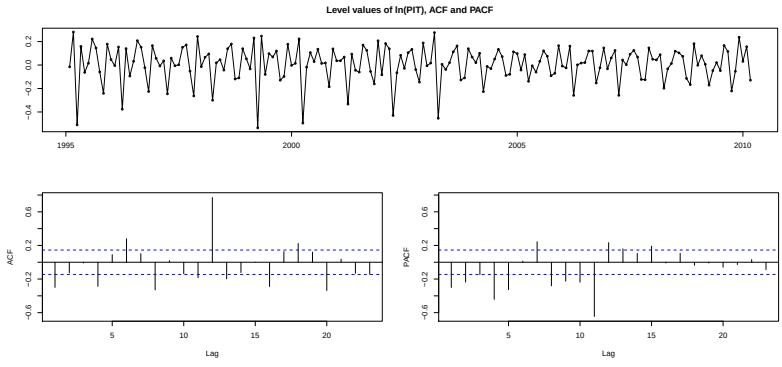

Figure 4.4: dln(PIT) time series display at first differenced values

The differenced ln(PIT) values in Figure 4.4 show that the data is stationary around the mean. The ACF and PACF also show significant spikes on lag 12, which signalled seasonality in the series. The autocorrelation function from the differenced ln(PIT) data reduced in most lagged points as compared to the level values, with some of the PACF outside the zero value confidence interval bound, that is the SARIMA model could be considered to model dln(PIT) time series data.

4.2.1.2 PIT SARIMA model output

Several competing seasonal ARIMA models were computed on the initial sample (Jan 1995 to March 2010) together with the statistic such as Akaike information criterion (AIC), the BIC which work the same as AIC but adding more penalties, the Root Mean Squared Error (RMSE). The table below shows the various competing seasonal ARIMA models built for the PIT time series.

Table 4.1: SARIMA Models for Personal Income Tax (PIT)

	Model	R^2	L	AIC	BIC	$LB(p-Value)$
1	$ARIMA(0,1,1)(0,0,1)_{12}$	0.9593	150.35	-294.69	-285.08	$2.20 * 10_{12}^{-16}$
2	$ARIMA(1,1,0)(1,0,0)_{12}$	0.9936	224.73	-443.46	-433.85	$2.18 * 10^{-06}$
3	$ARIMA(0,1,1)(1,0,1)_{12}$	0.9903	251.14	-494.29	-481.47	0.6471
4	$ARIMA(2,1,3)(1,0,0)_{12}$	0.9952	247.06	-480.11	-457.68	0.4815
5	$ARIMA(2,1,3)(1,0,1)_{12}$	0.9922	254.60	-493.20	-467.56	0.9800
6	$ARIMA(2,1,3)(2,0,1)_{12}$	0.9932	257.17	-496.35	-467.51	0.9550
7	$ARIMA(2,1,3)(1,0,2)_{12}$	0.9918	257.81	-497.62	-468.79	0.9199

The model that minimises the RMSE is considered, the R-squared (R^2) which is a well-known statistic for the explanatory model and select the model based on the model with higher value R^2. The log likelihood statistic (L) and finally the Ljung-Box statistic which tests the independence of the residuals. The statistics were observed to select the best model that will then be used to forecast. It is important to note that the R^2 on the time series model for R-softwares is not pre-programmed as the R^2 is seen as a statistic commonly used in explanatory models. However, the R-software give us a platform to program or to compute the statistic which were not included (See Appendix C on how R^2 was computed).

The seven competing SARIMA models in Table 4.1 above have higher R^2 ranging from 95% to very close to 99.6% , i.e. this model explains huge variations in PIT. Holding other statistics and selecting the best model on the basis of R^2 model four ($ARIMA(2,1,3)(1,0,0)_12$) will be

selected as it has the highest R^2 of 99.52%. The Log likelihood statistics chose the seventh model with the maximum likelihood of 257.81%, while model four also contained the minimum AIC and BIC statistics. This could encourage one to select model four, since it had many statistics that supported its significance for predicting PIT. Hence the best SARIMA to model PIT is model five $(ARIMA(2,1,3)(1,0,1)_12)$, because the model has the residuals which are highly independent from one another with the Ljung-Box (LB) $p-value = 0.9800$. This satisfied one condition of the SARIMA models, which emphasise that the residuals from the fitted model should be independent. For this reason, model five was used to forecast the continuation of the PIT historical patterns. Model five could be mathematically represented as follows:

$$(\mathbf{1} - \phi_1\mathbf{B} - \phi_\mathbf{2}\mathbf{B^2})(\mathbf{1} - \mathbf{\Phi_1}\mathbf{B^{12}})\mathbf{w_t} = (\mathbf{1} - \theta_\mathbf{1}\mathbf{B} - \theta_\mathbf{2}\mathbf{B^2} - \theta_\mathbf{3}\mathbf{B^3})(\mathbf{1} - \mathbf{\Theta_1}\mathbf{B^{12}})\epsilon_\mathbf{t} \tag{4.2.1}$$

where $w_t = ln(PIT_t) - ln(PIT_{t-1})$

ϕ_i is the i^{th} autoregressive $(AR(i))$ coefficient

Φ_1 is the first seasonal autoregressive $(SAR(1))$ coefficient

θ_j is the j^{th} moving average $(MA(j))$ coefficient

Θ_1 is the first seasonal moving average $(SMA(1))$ coefficient

$\mathbf{B}$ is a back shift operator with $\mathbf{B^i}\ \mathbf{w_t} = \mathbf{w_{t-i}}$ and $\mathbf{B^j}\ \epsilon_\mathbf{t} = \epsilon_\mathbf{t-j}$, and

$i = 1, 2$ and $j = 1, 2, 3$, ϵ_t being an error term at time t.

Table 4.2 present the maximum likelihood parameters estimation for the SARIMA model in equation (4.2.1) fitted to PIT time series, generated by R command $cwp(sarima.pit)$.

Table 4.2: PIT SARIMA model parameters estimation

Parameter	Coefficient	Standard error	t-Value	p-Value
$AR(1)$	-1.8374	0.0357	-51.3997	0.0000
$AR(2)$	-0.9279	0.0338	-27.4124	$1.95 * 10^{-165}$
$MA(1)$	1.1531	0.0566	20.3828	$2.38 * 10^{-92}$
$MA(2)$	-0.3921	0.0912	-4.3005	$1.70 * 10^{-05}$
$MA(3)$	-0.7367	0.0528	-13.9530	$3.02 * 10^{-44}$
$SAR(1)$	0.9778	0.0118	83.0119	0.0000
$SMA(1)$	-0.4939	0.1041	-4.7458	$2.08 * 10^{-06}$

using the coefficients in Table 4.2, equation (4.2.1) can be re-written as:

$$(\mathbf{1}+1.84\mathbf{B}+\mathbf{0.93B^2})(\mathbf{1}-\mathbf{0.98B^{12}})\mathbf{w_t} = (\mathbf{1}+\mathbf{1.15B}-\mathbf{0.39B^2}-\mathbf{0.74B^3})(\mathbf{1}-\mathbf{0.49B^{12}})\epsilon_\mathbf{t} \quad (4.2.2)$$

4.2.1.3 PIT SARIMA model residual analysis

The residuals from PIT SARIMA model are within the boundary $\pm\frac{1}{\sqrt{T}}$, where T is the number of series observations. Based on the residual's 95% confidence intervals, the residuals are assumed to be not far from the zero line, that is, they come from a well defined model (See the residual plot from the PIT SARIMA model below).

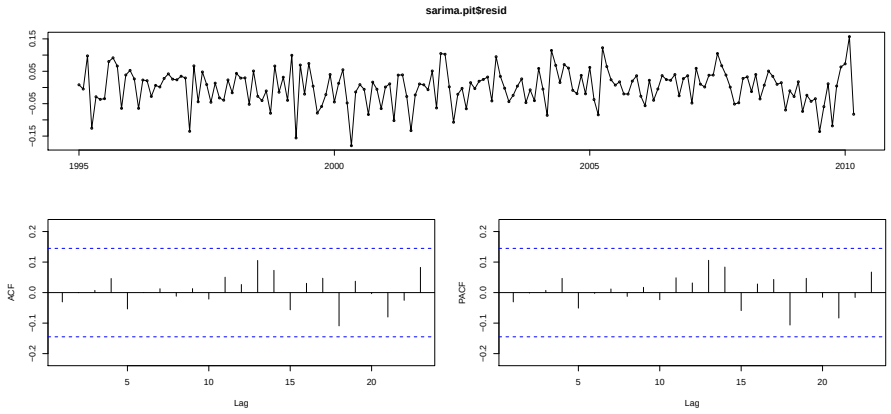

Figure 4.5: arima.pit model residuals, ACF and PACF

From the selected $ARIMA(2,1,3)(1,0,1)_{12}$ model the histogram and the q-q plot further confirm the independence of the residuals, as the two plots showed the distribution of the residual overtime, with the most of the values centred around zero (mean zero). There was some skewness.

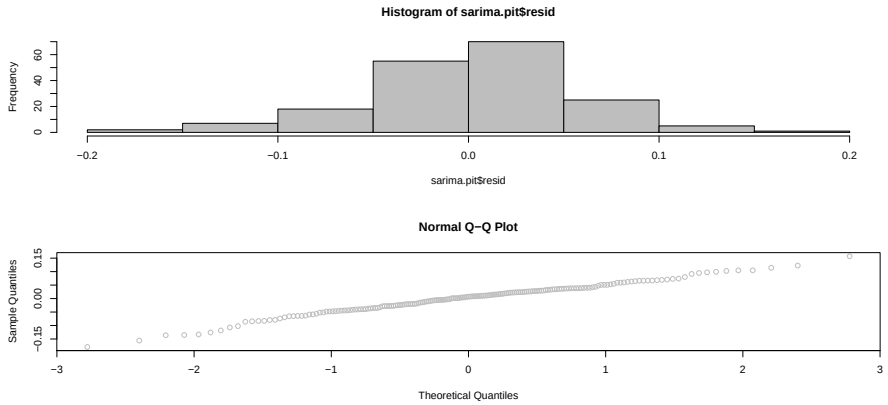

Figure 4.6: Residual Histogram and Q-Q plot from sarima.pit model

The histogram and the q-q plot above are generated by the *R* commands:

```
par(mfrow=c(2,1)) # 2 by 1 graphic display
hist(sarima.pit$resid,col = 16); qqnorm(sarima.pit$resid,col = 16)
```

The more general statistics to verify the residual white noise come from the Ljung-Box (LB) test, which was calculated on the sampled residuals of the model fitted. This was produced by the following code in *R*:

```
Box.test(sarima.pit$resid, lag =20, type= 'Ljung') #$
        Box-Ljung test
data:  sarima.pit$resid
X-squared = 9.2293, df = 20, p-value = 0.98 #$
```

The Ljung-Box test with a chi-squared of 9.2293 from 20 degrees of freedom gave a p-value of 0.98. This shows that the residuals are independent or uncorrelated and assumed to be coming from a well specified model. The researcher then fitted the values of PIT using SARIMA, as per the following section.

4.2.1.4 PIT SARIMA model fitted values

Figure 4.7 shows the monthly actual and the fitted values from PIT seasonal ARIMA model for the initial sample. Monthly fitted values were generated using the following *R*-codes:

```
pit.e = exp(fitted(sarima.pit)) # re-transform the ln transformed PIT fitted
ts.plot(pit,pit.e,col = c(1,2), main = 'Actual PIT Vs. SARIMA Fitted',
xlab ='Period', ylab ='Rand Million')
legend(1995,20000,ncol =2, c('PIT','Fitted'),fill = c(1,2))
```

The output from the *R*-codes above becomes the figure below:

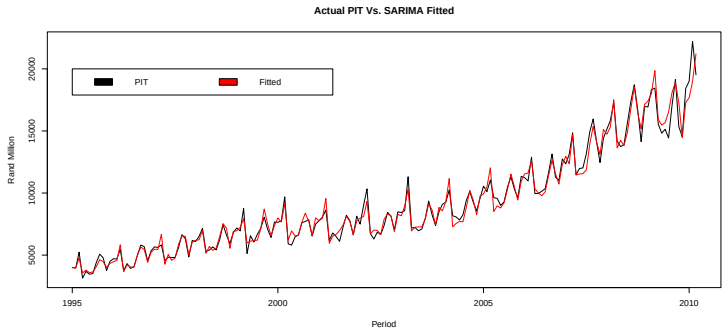

Figure 4.7: PIT Actuals and SARIMA Fitted Values

It can be seen that the fitted values closely follow the pattern of the actual values. The estimates are close to the actuals, from which one can deduce that the model fits the PIT data set. As the aim of the study was to fit and forecast the yearly payments, the monthly actuals fitted values were aggregated to form SARS fiscal year values, as shown in Table B.1 of Appendix B. The following section focuses on the forecasting of PIT payments using a SARIMA model.

4.2.1.5 PIT SARIMA model forecast values

PIT SARIMA monthly forecasts for the period April 2010 to March 2011 were generated using the following *R*-codes:

```
lpit.f11 = predict(sarima.pit, n.ahead = 12)# forecast 12 out of sample observations
pit.f11 = exp(lpit.f$pred) # re-transform the ln transformed PIT forecast
pit.f11 #$ Print the forecast values
```

These forecasts were aggregated to form the forecast of R228,8bn for fiscal year 2010/11 against the actual of R226,9bn, that is, a percentage error of -0.82%.

When forecasting the monthly Figures of the fiscal year 2011/12, the initial sample was then increased by another 12 months until the end of March 2011. That is, 195 observations were included. The monthly forecasts for April 2011 to March 2012 amounted to R254,8bn for 2011/12, an increase of 1.76% above the actual (R250,4bn) for the same period. The monthly forecasts for 2011/12 were generated using the following codes;

When forecasting the monthly figures for fiscal year 2011/12, the initial sample was increased by another 12 months until the end of March 2011, i.e. 195 observations were included. The monthly forecasts for April 2011 to March 2012 amounted to R254,8bn for 2011/12 an increase of 1.76% above the actual (R250,4bn) for the same period. The monthly forecasts for 2011/12 were generated using the following codes:

```
pit=ts(with(mydata,PIT),start=c(1995,1),end = c(2011,3),freq=12)
# expand the initial sample
sarima.pit = arima(log(pit), order = c(2,1,3),
 seasonal = list(order = c(1,0,1),period = 12)) # re-fit the SARIMA model
summary(sarima.pit) # output the model summary
lpit.f12 = predict(sarima.pit, n.ahead = 12) # forecast 12 out
of sample observations
pit.f12 = exp(lpit.f12$pred) # re-tranform the logarithm treansformed PIT forecast
pit.f12 #$ print the forecast
```

The researcher then used a similar approach to generate the aggregate forecast for 2012/13.

Table 4.3 summaries the results of Personal income tax (PIT) actual payments and their SARIMA model aggregated forecasts for 2010/11, 2011/12 and 2012/13.

Table 4.3: PIT Actuals Vs. SARIMA Forecast in Rand Million

Fiscal Year	Sample Used	Number of Observations	PIT Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	226,927	228,777	-0.82%
2011/12	Jan 1995 - Mar 2011	195	250,400	254,807	-1.76%
2012/13	Jan 1995 - Mar 2012	207	275,823	276,345	-0.19%

The following section summarises the results of additive Holt-Winters methods for Personal income tax (PIT).

4.2.2 Holt-Winters method for PIT time series

Unlike SARIMA models, Holt-Winters models look at a time series data of interest, separated into three components which are; the level value, trend and seasonal, and gives each components weights on an interval of zero to one, to be able to fit the model and forecast the future values. The PIT series was described as having a gradual increasing trend and an additive seasonality predictable over time (see section 4.2). This implies that an additive Holt-Winters model suited the PIT time series, the results of which were generated by the following R command:

```
pit10 = window(pit,start = c(1995,1), end = c(2010,3))
pit.ahw = HoltWinters(log(pit10), seasonal ='additive');pit.ahw
```

Table 4.4 represents the smoothing constants, level and trend estimation for the additive Holt-Winters model in equation (4.2.3) fitted to PIT time series.

Table 4.4: PIT additive Holt-Winters model constants estimation

Smoothing constants	Coefficient	Level and trend	Coefficient
alpha(α)	0.1681941	β_0	9.80113063
beta (β)	0.0832148	β_1	0.00823491
gamma (γ)	0.4792593		

The additive Holt-Winters model initial seasonality factor Sn_t values for 12 months are shown in Table 4.5.

Table 4.5: Initial values for seasonal factors from PIT additive Holt-Winters model

Seasonal smooth	s_1	s_2	s_3	s_4	s_5	s_6
Coefficient	-0.0853	-0.1155	-0.1119	-0.0934	0.03696	0.1235
Seasonal smooth	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
Coefficient	-0.0459	-0.1424	0.0442	0.0520	0.1322	0.1023

$$y_t = (9.801 + 0.008t) + Sn_t + \epsilon_t \tag{4.2.3}$$

Equation (4.2.3) shows the results of an additive Holt-Winters model for PIT in Table 4.4. The additive Holt-Winters model represented by equation (4.2.3) has an R^2 value of 0.9897, which implies that the fitted model explains about 98.97% of the movement in PIT actuals series.

The estimate average value l_T for the level, the estimate b_T for growth rate and the estimate sn_T for the seasonal factor of the data in time period T are given by the smoothing equation:

$$l_T = 0.168(y_T - Sn_{T-1}) + (1 - 0.168)(l_{T-1} + b_{T-1}) \tag{4.2.4}$$

$$b_T = 0.083(l_T - l_{T-1}) + (1 - 0.083)b_{T-1} \tag{4.2.5}$$

$$Sn_T = 0.479(y_T - l_T) + (1 - 0.479)Sn_{T-1} \tag{4.2.6}$$

where $\alpha = 0.168$, $\beta = 0.083$ and $\gamma = 0.479$ are smoothing constants ranging between 0 and 1. The l_{T-1} and b_{T-1} are estimates in time period $T - 1$ for the level and growth rate, respectively and Sn_{T-1} is the seasonal factor. The application of these methods requires data to have a linear trend with an additive seasonal pattern.

4.2.2.1 PIT Holt-Winters fitted values

An initial sample with data ending March 2010 was used to fit the model given in equation (4.2.3), and its fitted values were aggregated to form fiscal years fitted for the in-sample observations.

Figure 4.8 shows the actual and the fitted values for PIT additive Holt-Winters model.

From the above figure it can be shown that the fitted values are not far away from actual values. This confirms that an additive Holt-Winters model capture the movement of PIT data, therefore the model fitted can be used to forecast the future values.

4.2.2.2 PIT additive Holt-Winters forecast values

Table 4.6 show the PIT actual payments and an additive Holt-Winters model forecast for fiscal year 2010/11, 2011/12 and 2012/13, respectively.

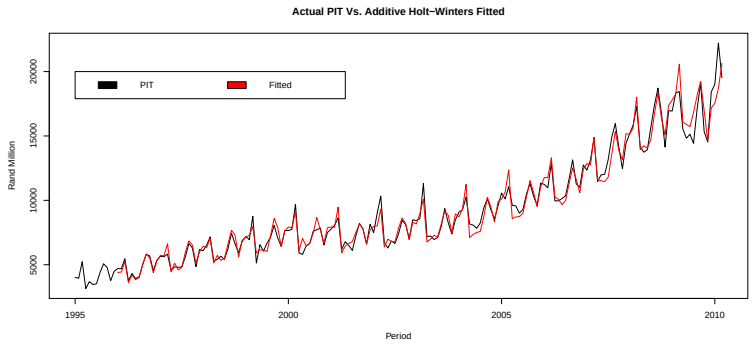

Figure 4.8: PIT Actuals and Additive Holt-Winters Fitted Values

Table 4.6: PIT sample used, actuals and forecasts for Holt-Winters model (Rand million)

Fiscal Year	Sample used	Number of observations	PIT Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	226,927	228,189	-0.56%
2011/12	Jan 1995 - Mar 2011	195	250,400	252,531	-0.85%
2012/13	Jan 1995 - Mar 2012	207	275,823	276,067	-0.09%

The monthly forecast from the initial PIT additive Holt-Winters model in 2010/11 amounted to R228,2 bn, which was 0.56% above the actual for the same period of R226,9bn. Still using the additive Holt-Winters model, the sample was increased by one year to end in March 2011 to forecast the fiscal PIT total payments for 2011/12, which amounted to R252,5bn, an increase of about 0.85% as compared to the actual of R250,4bn for the same period. Finally, the sample was increased by one year (207 data points were considered) to forecast the 2012/13 fiscal year. The forecast amounted to R276,1bn and the observation for the same period was R275,1bn, a difference of less than R1billion. The R-codes used are as shown below:

```
# Forecast April 2010 to March 2011 (FY2010/11)
lpit.f11 = predict(pit.ahw,n.ahead= 12,
prediction.interval = T, level = 0.95)
pit.f11 = exp(lpit.f11[,1])
pit.f11

# Forecast April 2011 to March 2012 (FY2011/12)
```

```
pit11 = window(pit,start = c(1995,1), end = c(2011,3))
pit.ahw12 = HoltWinters(log(pit11), seasonal ='additive')
lpit.f12 = predict(pit.ahw12,n.ahead= 12,
prediction.interval = T, level = 0.95)
pit.f12 = exp(lpit.f12[,1])
pit.f12

# Forecast April 2012 to March 2013 (FY2012/13)
pit12 = window(pit,start = c(1995,1), end = c(2012,3))
pit.ahw13 = HoltWinters(log(pit12), seasonal ='additive')
lpit.f13 = predict(pit.ahw13,n.ahead= 12,
prediction.interval = T, level = 0.95)
pit.f13 = exp(lpit.f13[,1])
pit.f13
```

4.2.3 PIT models comparison and conclusion

The seasonal SARIMA and additive Holt-Winters methods were used to model PIT time series data. In both methodologies, models were fitted on log transformed data. Table 4.7 gives the results of the measure of accuracy of SARIMA and additive Holt-Winters methods on PIT time series.

Table 4.7: PIT SARIMA and additive Holt-Winters measure of accuracy

Methodology	R^2	ME	RMSE	MAE	MPE	MAPE	MASE
SARIMA	0.9952	0.00088	0.05491	0.04278	0.00699	0.47489	0.37603
Holt-Winters	0.9897	4.537815	4.53781	4.57353	49.9909	50.3905	40.1935

Based on the results of the indicators of accuracy in Table 4.7, one can be tempted to conclude that SARIMA model was more accurate in estimating the PIT series than additive Holt-Winters model. This accuracy results is attributed to the fact that SARIMA models are generated using difference data, which is generally stationary. However, based on the point forecast percentage errors (see Table 4.3 and 4.6), it cannot be concluded that additive Holt-Winters model performed slightly better than SARIMA model in forecasting PIT series, even though forecasts from both methods were not far from the actual PIT series. The results (fitted and forecast) from the two time series methods used show that both SARIMA and Holt-Winters model perform well against

the PIT actuals and can be used as interval to which the actual realisation can fall.

4.3 Value Added Tax

VAT is an in-direct tax which is levied on consumption of goods or services. It is the second largest tax, with a fixed tax rate of 14% SARS and National Treasury (2012). VAT contributes around 26% to Total Tax Revenue, 6.7% to nominal GDP and has an increasing trend, which shows a spike in March every year. When modeling VAT data one should consider the increasing trend of the series and the seasonal components (March spike), which is increasing every year (See Figure 4.9).

```
tsdisplay(vat,main ='Level values of VAT, ACF and PACF')
```

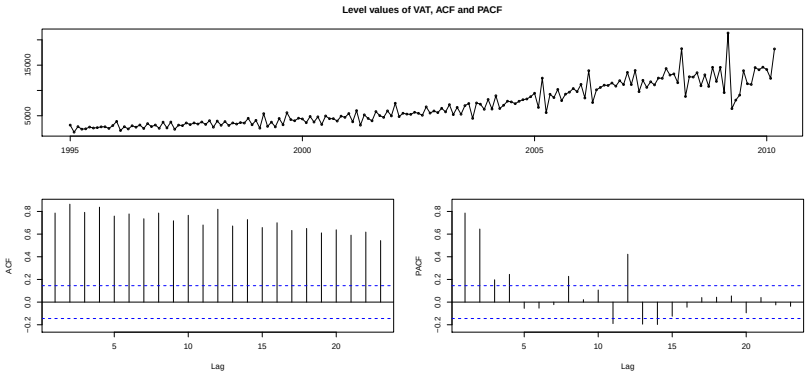

Figure 4.9: Value Added Tax (VAT), ACF and PACF

4.3.1 SARIMA model for VAT

This section covers modeling of monthly VAT using the time series SARIMA model, which was derived of fitted values from the model, an analysis of model residuals and forecasting of 2010/11, 2011/12 and 2012/13 VAT payments.

4.3.1.1 Natural logarithm of VAT, ACF and Partial PACF

The natural logarithm transformation (ln) is normally used to stabilize the variances of financial variables. Similarly, the same transformation can also be applied to the SARS payments data to minimise payment fluctuations to obtain better model. The ln transformation was used in the

VAT data, which is represented as ln(vat). VAT data display, its Autocorrelation function (ACF) and Partial autocorrelation function(PACF) were obtained using the following R commands:

```
tsdisplay(log(vat),main ='ln(vat), ACF and PACF')
```

The result of which is of the following output.

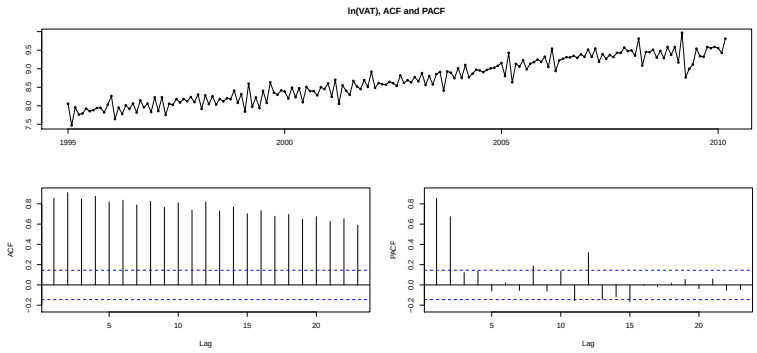

Figure 4.10: ln(VAT) Time Series Display at level values

From the time series display of ln(VAT), it is clear that the series is non-stationary with changing mean and variance over time. This is supported by the ACF and PACF plot for ln(VAT). We now consider data differencing to obtain stationarity at the mean as follows:

```
tsdisplay(diff(log(vat)),main ='d(ln(VAT)), ACF and PACF')
```

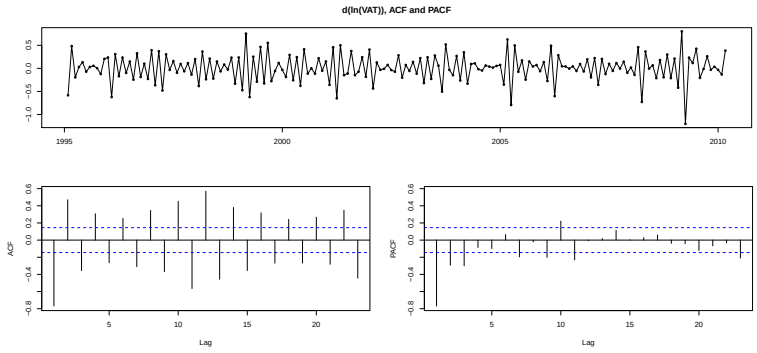

Figure 4.11: dln(VAT), time series display at first differenced values

Transformed VAT appears to be stationary in the mean with the ACF and PACF that are not white noise. That is, we can now use differenced historical information to fit and to forecast

VAT payments data. The intention is to fit the model such that more variation is covered and a white-noise model residuals is achieved.

4.3.1.2 SARIMA output for VAT

This subsection looks into modeling of VAT time series data using SARIMA model. The previous section concluded that VAT is seasonal, with a spike or a consistence peak in March of every year. This allows us not to even look at the ordinary ARIMA model but to search the best fitting SARIMA model. Four SARIMA models were fitted to the VAT data and the results are shown in Table 4.8.

Table 4.8: SARIMA models for VAT

	Model	R^2	L	AIC	BIC	$LB(p-Value)$
1	$ARIMA(1,1,1)(0,0,1)_{12}$	0.9370910	79.39	-150.78	-137.96	0.1055
2	$ARIMA(1,1,1)(1,0,0)_{12}$	0.9613275	91.37	-174.74	-161.92	0.2065
3	$ARIMA(1,1,1)(1,0,1)_{12}$	0.9877714	108.09	-206.18	-190.16	0.7469
4	$ARIMA(0,1,2)(1,0,1)_{12}$	0.9812438	108.68	-207.35	-191.33	0.8442

SARIMA models were fitted using the bottom-up approach, that is, by starting with a few lag combinations of autoregressive, seasonal autoregressive, seasonal and non-seasonal moving averages, and gradually increasing the number of those components. The model that best fits the VAT data selected is $ARIMA(0,1,2)(1,0,1)_{12}$ or model number 4 from Table 4.8, because the residuals from this model appear to be highly uncorrelated. The highly uncorrelated residuals are shown by the Ljung-Box test with a p-value of 0.8442 against the null hypothesis, which emphasises that the residuals are highly correlated. This is also supported by the highest AIC, BIC, log-likelihood and R^2 of 98%. The model selected shows that the current value of VAT is highly related to the closest lagged values of its own historical path rather than values far apart. Model four could be mathematically represented as follows:

$$(1-\Phi_1\mathbf{B^{12}})\mathbf{w_t} = (\mathbf{1}-\theta_\mathbf{1}\mathbf{B}-\theta_\mathbf{2}\mathbf{B^2})(\mathbf{1}-\mathbf{\Theta_1}\mathbf{B^{12}})\epsilon_\mathbf{t} \tag{4.3.1}$$

where $\mathbf{w_t} = \mathbf{ln(vat_t)} - \mathbf{ln(vat_{t-1})}$

Φ_1 is the first seasonal autoregressive ($SAR(1)$) coefficient

θ_j is the j^{th} moving average ($MA(j)$) coefficient, $j = 1, 2$

Θ_1 is the first seasonal moving average ($SMA(1)$) coefficient

$\mathbf{B}$ is a back shift operator with $\mathbf{B^i}\ \mathbf{w_t} = \mathbf{w_{t-i}}$ and $\mathbf{B^j}\ \epsilon_\mathbf{t} = \epsilon_\mathbf{t-j}$,

ϵ_t is an error term at time t

Table 4.9 presents the maximum likelihood parameters estimation for the SARIMA model of equation (4.3.1), fitted to VAT time series and were generated by R-command $cwp(vat.sarima)$.

Table 4.9: VAT SARIMA model parameters estimation

Parameter	Coefficient	Standard error	t-Value	p-Value
$MA(1)$	-1.1055	0,0697	-15,8653	$1.10 * 10^{-56}$
$MA(2)$	0.2947	0,0715	4,1201	$3.79 * 10^{-05}$
$SAR(1)$	0.9747	0,0164	59,3121	0.0000
$SMA(1)$	-0.7444	0,0767	-9,6997	$3.03 * 10^{-22}$

Using the coefficients in Table 4.9, equation (4.3.1) can be re-written as;

$$(\mathbf{1 - 0.97B^{12}})\mathbf{w_t} = (\mathbf{1 - 1.105B + 0.29B^2})(\mathbf{1 - 0.74B^{12}})\epsilon_\mathbf{t} \tag{4.3.2}$$

The following section analyses the residuals for the model represented by equation (4.3.2).

4.3.1.3 VAT SARIMA model residual

Residuals from the VAT SARIMA model or model 4 in Table 4.8 appear to be white noise, as they are all within the 95% confidence interval $\pm\frac{1}{\sqrt{T}}$, where T is the number of series observation (See Figure 4.12). These confirms that the residuals are uncorrelated and that they are from a well-defined model. Residuals from VAT SARIMA model are assumed to be not far from zero line.

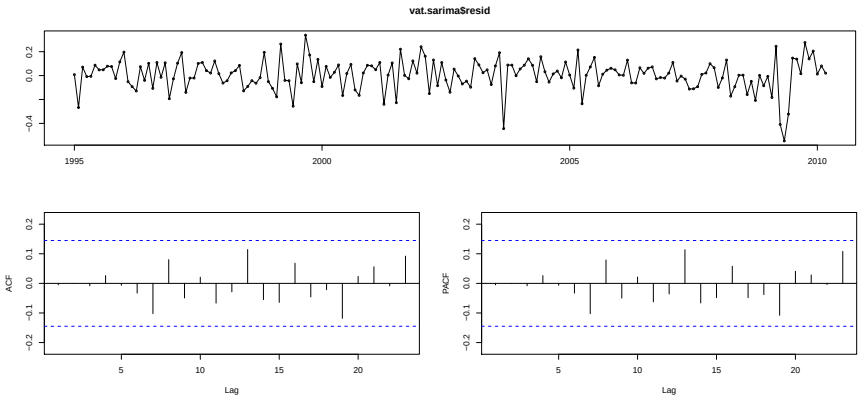

Figure 4.12: VAT SARIMA model residuals, ACF and PACF

When confirming the normality of the model residuals, we look at the residual plot of VAT SARIMA model below. Although the residual are a bit skewed to the left, they look slightly normally distributed with the mean zero and variance (δ_2), (see the histogram and the q-q plot below).

```
par(mfrow=c(2,1)); hist(vat.sarima$resid, col =16); qqnorm(vat.sarima$resid,
col =16)
```

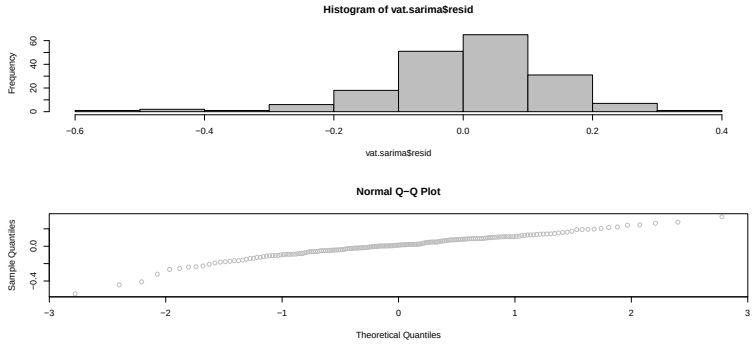

Figure 4.13: Residual Histogram and Q-Q plot from vat.sarima model

The more general statistics to verify the residual white noisiness is the Ljung-Box (LB) statistic, which is calculated on the sampled residuals of the fitted model. This is produced using the following command in *R*:

```
Box.test(vat.sarima$resid, lag =20, type= 'Ljung') #$
```

```
        Box-Ljung test
data:  vat.sarima$resid
X-squared = 13.7239, df = 20, p-value = 0.8442 #$
```

The Ljung-Box test with a chi-squared of 13.7239 from 20 degrees of freedom gave a p-value of 0.8442. This shows that the residuals from VAT SARIMA model are independent or uncorrelated and assumed to be coming from a well-defined model. We then fit the values of VAT using SARIMA model in the following section.

4.3.1.4 VAT SARIMA model fitted values

The monthly fitted values were obtained from the VAT SARIMA model using the initial sample starting January 1995 to March 2010 and were aggregated to form the yearly fitted values for VAT shown in Table B.2 of Appendix B. The fitted values were obtained using the following R-command:

```
vat.fit = exp(fitted(vat.sarima))# re-transform the ln transformed VAT fitted
vat.fit # Print the fitted values
```

VAT SARIMA model performed well on the initial sample as it shows the percentage error which is less that 5%. For the fiscal year 1995/96 to 2009/10, the model shows the highest percentage error of around 4% in 1995/96, 2004/05 and 2008/09 only. The in-sample actuals and fitted values are shown below.

```
ts.plot(vat,vat.fit,col=c(1,2), main='Actual VAT Vs. SARIMA Fitted',xlab='Period',
ylab ='Rand Million');legend(1995,20000,ncol =2, c('VAT','Fitted'),fill = c(1,2))
```

The R- codes above output the Figure below

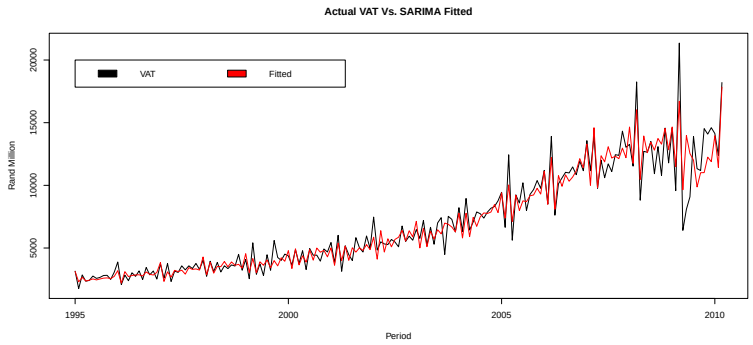

Figure 4.14: Actuals VAT and SARIMA Fitted Values

It can be observed that the fitted values closely follow the patterns of the actual values. Since the in-sample estimates are closed to the actuals, which deduce that the model fits the VAT data set.

4.3.1.5 VAT SARIMA model forecast values

Table 4.10 shows the value Added Tax (VAT) actual payments and the SARIMA model forecast for the fiscal year 2010/11, 2011/12 and 2012/13, respectively.

Table 4.10: VAT Actuals Vs. SARIMA Forecast in Rand Million

Fiscal Year	Sample Used	Observations	VAT Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	183,571	169,281	7.78%
2011/12	Jan 1995 - Mar 2011	195	191,020	204,239	-6.92%
2012/13	Jan 1995 - Mar 2012	207	215,023	210,522	2.09%

The forecast from the VAT SARIMA model for fiscal years 2010/11 and 2011/12 was around R169,3bn and R204,3bn respectively. That is, a percentage error of 7.8% and -6.9% respectively against the actuals of R183,6bn and R191,0bn for the same period. The higher percentage error of around 7% for the two fiscal years was due to the 2008/09 economic recession, which impacted negatively on VAT payments starting from 2009/10. The model performed exceptionally well in fiscal year 2012/13, forecasting R210,5bn to be collected compared to the actual realisation amount of R215,0bn, which was an error of 2.1%. The smaller error shows that the actual VAT payment recovered to a normal trend after the recession effect. The model will be expected to perform better when forecasting future values, assuming that there will be no shocks such as another economic

recession.

R-commands used to generate forecasts for the fiscal years of interest are as follows:

```
# Forecast 2010/11
lvat11 = predict(vat.sarima, n.ahead = 12) Forecast in ln scale
vat11 = exp(lvat11$pred);vat11 #$ re-transform and print the forecast

# Forecast 2011/12
vat=ts(with(mydata,VAT),start=c(1995,1),end = c(2011,3),freq=12)
vat.sarima = arima(log(vat), order = c(0,1,2),
seasonal = list(order = c(1,0,1),period = 12))
lvat12 = predict(vat.sarima, n.ahead = 12)
vat12 = exp(lvat12$pred);vat12 #$

# Forecast 2012/13
vat=ts(with(mydata,VAT),start=c(1995,1),end = c(2012,3),freq=12)
vat.sarima = arima(log(vat), order = c(0,1,2),
seasonal = list(order = c(1,0,1),period = 12))
lvat13 = predict(vat.sarima, n.ahead = 12)
vat13 = exp(lvat13$pred);vat13 #$
```

The next section summarises the results of Holt-Winters for Value added tax (VAT)

4.3.2 Additive Holt-Winters method for VAT)

VAT is assumed to have additive seasonality with a spike in March every year. However, more weight of a seasonal smoothing parameter form Holt-Winters models is able to handle the increase in the March spike, $\gamma = 0.313$ (See Table 4.11). VAT sample data were transformed using natural logarithm (ln) to minimise variation for the purpose of fitting a better model. The Holt-Winters additive method assumes that the ln(VAT) time series data was represented by the model in equation (4.3.3), which is obtained by using the following R-codes:

```
vat.ahw = HoltWinters(log(vat), seasonal ='additive');vat.ahw
```

Table 4.11 presents the smoothing constants, level and trend constants for the additive Holt-Winters model.

Table 4.11: VAT additive Holt-Winters model parameters estimation

Smoothing parameter	Coefficient	Level and trend	Coefficient
alpha(α)	0.1172597	β_0	9.446352944
beta (β)	0.004434004	β_1	0.008246231
gamma (γ)	0.3129413		

The additive Holt-Winters model initial seasonality factor Sn_t values for 12 months are given in Table 4.12.

Table 4.12: Initial values for seasonal factors from VAT additive Holt-Winters model

Seasonal smooth	s_1	s_2	s_3	s_4	s_5	s_6
Coefficient	-0.3479	-0.0650	-0.0617	0.1185	-0.0219	0.0357
Seasonal smooth	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
Coefficient	0.0699	0.1573	0.0775	0.1684	-0.0809	0.3819

From Tables 4.11 and 4.12, we can write the additive Holt-Winters equation for PIT as follows.

$$y_t = (9.446 + 0.008t) + Sn_t + \epsilon_t \tag{4.3.3}$$

where y_t represent ln(vat)

Equation (4.3.3) shows the results of an additive Holt-Winters model for VAT in Table 4.11.

The estimate average value l_T for the level, the estimate b_T for growth rate and the estimate sn_T for the seasonal factor of the data in time period T are given by the smoothing equation:

$$l_T = 0.118(y_T - Sn_{T-1}) + (1 - 0.118)(l_{T-1} + b_{T-1}) \quad (4.3.4)$$

$$b_T = 0.004(l_T - l_{T-1}) + (1 - 0.004)b_{T-1} \quad (4.3.5)$$

$$Sn_T = 0.323(y_T - l_T) + (1 - 0.313)Sn_{T-1} \quad (4.3.6)$$

The additive Holt-Winters model explained around 95% (model R^2) of the variation in VAT actuals series, thus only 5% of the data variation was unexplained by the model.

Subsection 4.3.2.1 summarises the model fitted values from the initial sample which ending March 2010.

4.3.2.1 VAT Holt-Winters fitted values

The Figure 4.15 below shows the VAT additive Holt-Winters model fitted values from the initial sample (January 1995 to March 2010), which were aggregated to form fiscal year fitted values (1996/07 to 2009/10).

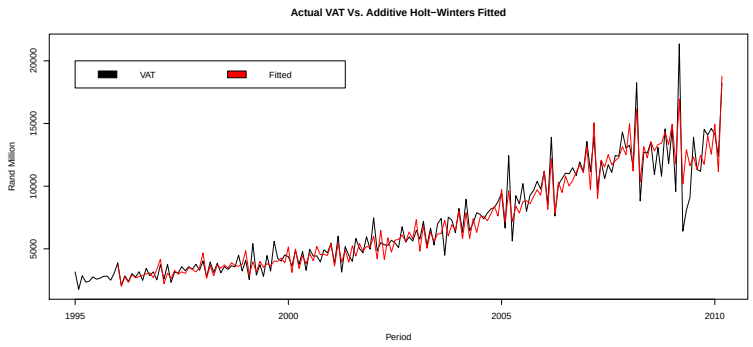

Figure 4.15: Actuals VAT and Holt-Winters Fitted Values

From the above figure, it can be observed that the fitted values are not far away from actual values. This confirms that additive Holt-Winters model VAT series performs well. The observed and fitted were plotted using the commands below.

```
ts.plot(vat,exp(fitted(vat.ahw)[,1]),col = c(1,2),
main = 'Actual VAT Vs. Additive Holt-Winters Fitted',
xlab ='Period', ylab ='Rand Million')
legend(1995,20000,ncol =2, c('VAT','Fitted'),fill = c(1,2))
```

4.3.2.2 VAT additive Holt-Winters model forecast values

Table 4.13 shows the sample used, VAT actual payments, the additive Holt-Winters model forecast and the percentage error for fiscal year 2010/11, 2011/12 and 2012/13.

Table 4.13: VAT Actuals Vs. Holt-Winters Forecast in Rand Million

Fiscal Year	Sample Used	Observations	VAT Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	183,571	169,202	7.83%
2011/12	Jan 1995 - Mar 2011	195	191,020	202,069	-5.78%
2012/13	Jan 1995 - Mar 2012	207	215,023	212,536	1.16%

The forecasts for fiscal year 2010/11 and 2011/12 were 7.83% below the actual value and 5.78% above the actual value respectively. However, for the fiscal year 2012/13 with the sample ending March2012 (207 observations), the model forecasted the collection of around R212,5bn, which was 1.16% below the actual of R215,0bn. This poor forecast with an error above 5% in the 2010/11 and 2011/12 SARS fiscal years was related to the economic recession which impacted VAT collection from 2009/10, with a recovery to a normal trend in 2010/11. Assuming that there will be no economic recession effect, the additive Holt-Winters model for VAT is expected to perform well in predicting future VAT payments. To illustrate that the additive Holt-Winters will hold with the assumption that there will be no recession effect, the forecast payments for fiscal year 2013/14 (beyond the scope of this study) can be done using the same models. The actual for the fiscal year 2013/14 amounted to R237,8bn and the model forecast was R237,4bn. This was an error of 0.15%.

The *R*-commands/codes used to generate the forecasts are as follows:

```
# Forecast 2010/11
vatf_ahw = predict(vat.ahw,n.ahead = 12,prediction.interval = T,
level = 0.95)
exp(vatf_ahw[,1])

# Forecast 2011/12
vat=ts(with(mydata,VAT),start=c(1995,1),end = c(2011,3),freq=12)
vat.ahw = HoltWinters(log(vat), seasonal ='additive')
```

```
vatf_ahw12 = predict(vat.ahw,n.ahead = 12,prediction.interval = T,
level = 0.95)
exp(vatf_ahw12[,1])

# Forecast 2012/13
vat=ts(with(mydata,VAT),start=c(1995,1),end = c(2012,3),freq=12)
vat.ahw = HoltWinters(log(vat), seasonal ='additive')
vatf_ahw13 = predict(vat.ahw,n.ahead = 12,prediction.interval = T,
level = 0.95);exp(vatf_ahw13[,1])
```

4.3.3 VAT Models Comparisons and conclusion

A comparison of VAT Holt-Winters and SARIMA models using accuracy measurements ME, RMPE, MAE, MPE, MAPE and MASE will be unrealistic, as the Holt-Winters model is fitted on the non-stationary series. The SARIMA model takes into consideration the stationarity of the data, i.e. most of the time the series will have the expected mean of zero. Conversely, the Holt-Winters calculates the mean level of the data at all points as the data increases. If one considers the R^2 comparison, ARIMA with $R^2 = 98.1\%$ will be selected as the best model against the additive Holt-Winters of $R^2 = 95.5\%$. However, the percentage error from the out of sample forecast shows that additive Holt-Winters is the model of choice as it performs better than the SARIMA model. For financial year actual and fitted values comparison, refer to Table B.2 in Appendix B.

The poor performance from the two models in the fiscal year 2010/11 and 2011/12 was due to the economic recession which affected the VAT payments, hence the models are taken to be the true structure which capture most of the fluctuation in VAT payments. This study recommends the use of an additive Holt-Winters model when forecasting.

4.4 Corporate Income Tax

The third largest souce of tax revenue is CIT, which contributes around 20% to Total Tax Revenue (TTAXR) on average. This is an income tax levied on companies at a rate of 28% SARS and National Treasury (2012). Figure 4.16 shows the CIT series from January 1995 to March 2010.

```
tsdisplay(cit,main ='Level values of CIT, ACF and PACF')
```

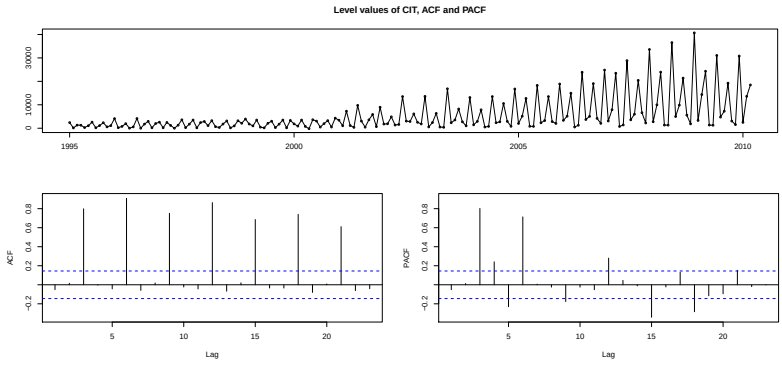

Figure 4.16: Corporate Income tax (CIT), ACF and PACF

The CIT data in Figure 4.16 shows an unclear trend, that is, there was no obvious increasing or decreasing trend. This assumes the stationarity in the mean of the CIT data. The data also revealed multiplicative increasing variance over the years, which assumes the non-stationary variance on CIT data. The volatility on monthly CIT data introduces the negative values in some of the months as a result of more refunds being given for those specific months. Due to surfacing of the negative observations and the volatility in the monthly CIT, the data were then converted to quarterly CIT, as per the figure below.

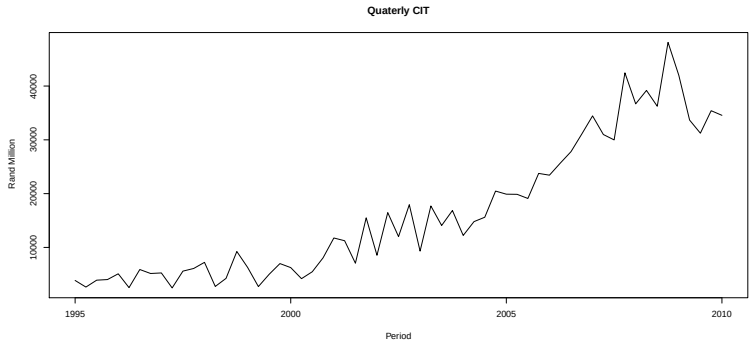

Figure 4.17: Quarterly Corporate Income Tax

The *R*-commands for quarterly transformed CIT data are as follows:

```
citq_data=read.csv("data path",header = T, sep=';',dec =';')
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2010,1),freq = 4)
plot(citq, main = "Quaterly CIT", xlab = "Period", ylab =" Rand Million" )
```

The function *seasonalplot*() from the library forecast in *R* allowed the researcher to plot the seasonality of the series and enables visualisation of the peaks occurring regularly on the fixed months/quarterly CIT data over the years. This was done using the following *R*-commands:

```
citq_data = read.csv("data path",header = T, sep=';',dec =';')
par(mfrow=c(2,1))
seasonplot(cit, main ='Monthly CIT Seasonal Plot',ylab='Rand Million',col=16)
seasonplot(citq, main ='Quarterly CIT seasonal Plot',ylab='Rand Million',col=16)
```

The results from the above *R*-commands produce the monthly and quarterly CIT seasonal plot shown in the figure below:

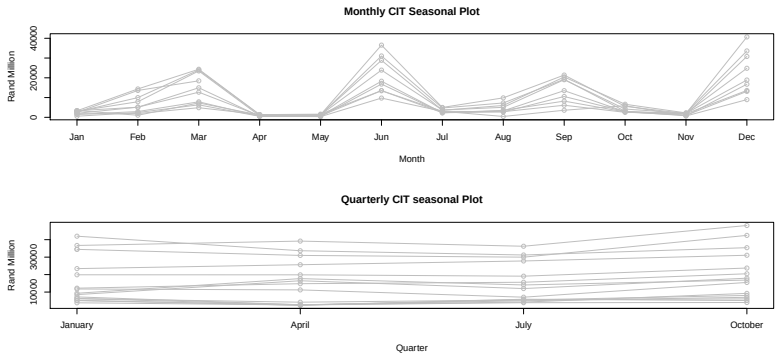

Figure 4.18: Monthly and Quarterly CIT Seasonal Plot

The monthly CIT seasonal plot shows peaks in the months of March, June and September, however the volatility on the monthly CIT data led the researcher to convert to quarterly data, allowing a simpler model to be fitted. As the purpose of the study was to forecast annual revenue collections for South African taxes, the data aggregation will not deviate the aim of the study. The following section summarises the SARIMA model fit for CIT.

4.4.1 SARIMA Model for CIT

This section focuses on the CIT time series data, modeling and residual analysis. The section also looks at the CIT forecast analysis and the best model is selected to be used in forecasting CIT - the SARIMA or the Holt-Winters model.

4.4.1.1 Transformed CIT, ACF and PACF

The quarterly CIT data reduced the higher variation and eliminated the negative observation that surfaced when monthly data were used. The CIT quarterly data still clearly showed an increasing trend, unstable movements and multiplicative movements. The researcher focused on fitting the SARIMA structure that captured the CIT movements and forecast the future CIT payments by first looking at the CIT quarterly data stationarity. This was done using the following *R*-commands:

```
library(forecast)
tsdisplay(log(citq),main = 'ln(cit)')
```

The results from the above *R*-commands will then be the following Figure.

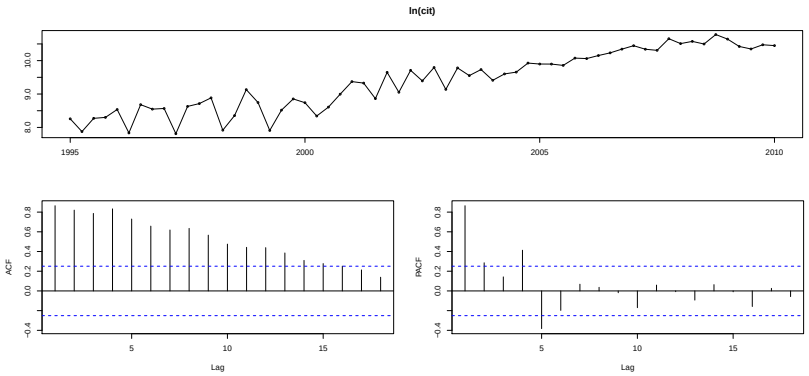

Figure 4.19: Quarterly ln(cit), ACF and PACF

The time series display on the above figure shows the quarterly log transformed CIT data with higher fluctuation from 1995 up until just before 2005, and the series became slightly smoother from 2005 onwards with an increasing trend. The ACF and the ACF showed several spikes which were significantly outside the 95% confidence interval. This led into stationarising the series as follows:

```
tsdisplay(diff(log(citq)),,main = 'd(ln(cit))')
```

The *R*-command above will results in to the following Figure.

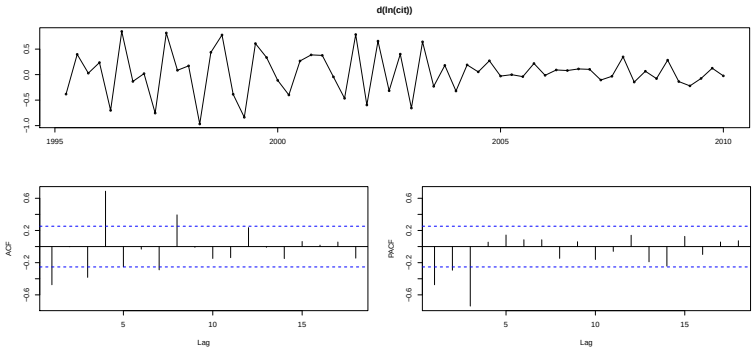

Figure 4.20: Quarterly d(ln(cit)), ACF and PACF

4.4.1.2 Quarterly CIT SARIMA model

Due to the volatility of the CIT data, obtaining several models became complex and required more research be done. The complexity of the CIT data was due to the uncontrollable companys behaviour. The Seasonal ARIMA model ($ARIMA(3,1,0)(0,0,1)_4$) with dummy variable $cdq3$ is used below to cover some unexplained parts of the series for the third quarter and to minimise the model error. The model was fitted to the log transformed CIT data as follows:

```
arima.citq = arima(log(citq), order = c(3,1,0),seasonal = list(order = c(0,0,1),
period = 4), xreg = cdq3)
```

Thus the mathematical representation of the model becomes equation (4.4.1)

$$(\mathbf{1} - \phi_{\mathbf{1}}\mathbf{B} - \phi_{\mathbf{2}}\mathbf{B^2} - \phi_{\mathbf{3}}\mathbf{B^3})\mathbf{w_t} = \mathbf{cdq3} + (\mathbf{1} - \mathbf{\Theta_1 B^4})\epsilon_{\mathbf{t}} \tag{4.4.1}$$

Where $\mathbf{w_t} = \mathbf{ln(cit_t)} - \mathbf{ln(cit_{t-1})}$
ϕ_i is the i^{th} autoregressive ($AR(i)$) coefficient, $i = 1, 2, 3$
Θ_1 the seasonal moving average of order 1 ($SAR(1)$) coefficient,
$\mathbf{B}$ is a back shift operator with $\mathbf{B^i}\ \mathbf{w_t} = \mathbf{w_{t-i}}$ and $\mathbf{B^j}\ \epsilon_{\mathbf{t}} = \epsilon_{\mathbf{t-j}}$,
$cdq3$ is the third quarter dummy variable for CIT, and
ϵ_t is an error term at time t

From the $ARIMA(3,1,0)(0,0,1)_4$ model we can deduce that the first difference of log transformed quarterly CIT data structure was captured by only three lags of autoregressive components, first

order seasonal moving average, and a dummy variable which was equal to one for every third quarter and zero elsewhere. The model fitted capturing around 99.4% movement of the log transformed CIT.

Table 4.14 presents the maximum likelihood parameters estimation for the SARIMA model of equation (4.4.1), fitted to CIT time series. These were generated by $cwp(cit.sarima)$ R-command.

Table 4.14: CIT SARIMA model parameters estimation

Parameter	Coefficient	Standard error	t-Value	p-Value
$cdq3$	-0.0888	0.03347	-2.6523	0.00799
$AR(1)$	-0.6819	0.10995	-6.2015	$5.59 * 10^{-10}$
$AR(2)$	-0.5168	0.12745	-4.0550	$3.5 * 10^{-05}$
$AR(3)$	-0.5164	0.12052	-4.848	$1.83 * 10^{-05}$
$SMA(1)$	0.4507	0.12586	3.5807	0.000034

Using the coefficients in Table 4.14, equation (4.4.1) can be re-written as;

$$(\mathbf{1 + 0,68B + 0,52B^2 + 0,52B^3})\mathbf{w_t} = \mathbf{-0,089} + (\mathbf{1 + 0,4507B^4})\epsilon_\mathbf{t} \tag{4.4.2}$$

The following section summarises the residual analysis from the quarterly CIT SARIMA model above

4.4.1.3 Quarterly CIT SARIMA model residual analysis

In order to use the quarterly CIT SARIMA-model to forecast future CIT payments, the researcher considered analysing the model residuals or investigating if the residuals were white noise or not correlated. If the model residuals are white noise, then the model can be used to forecast the continuation of the historic path.

Figure 4.21 is the plot of the residuals from $ARIMA(3,1,0)(0,0,1)_4$ model with dummy variable $cdq3$ for the third quarter. This Figure also shows the ACF and the PACF for the residuals series. These is generated by a $tsdisplay(sarima.citq\text{resid})$ R-command.

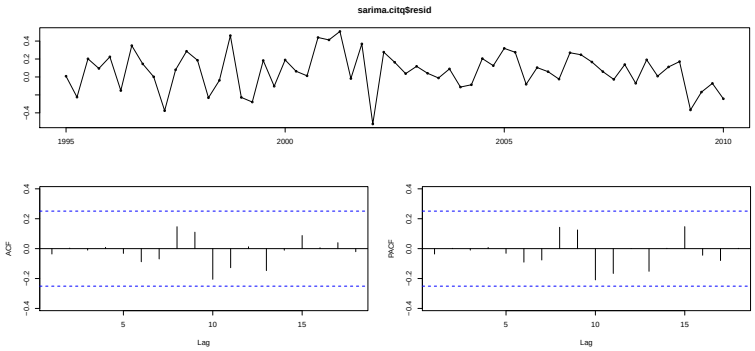

Figure 4.21: sarima.citq model residuals, ACF and PACF

From the graph above it can be seen that the ACF and the PACF from the model residuals were captured within the 95% confidence lines. This shows that the residuals of the $ARIMA(3,1,0)(0,0,1)_4$ model with dummy $cdq3$ are white noise or independent from each other. The researcher further examined the residuals by using the formal Q-statistics or the Ljung-Box (LB) statistic, which took a sample of a residual for a white noise testing. This was done using the following *R*-command:

```
Box.test(sarima.citq$resid, lag =20, type= 'Ljung') #$
```

The command results will then be the output below.

```
        Box-Ljung test
data:  sarima.citq$resid
X-squared = 10.5469, df = 20, p-value = 0.9571 #$
```

The Ljung-Box test with a Chi-squared value of 10.5469 from 20 degrees of freedom gave a p-value of 0.9571 or 95.7 percent. Since the p-value was greater than 0.05, it can be said that the residuals series of a quarterly SARIMA model are independent/uncorrelated and assumed to be coming from a well-specified model. The model residual independence is supported by the histogram and the q-q plot below that resembles a normal distributed series with mean zero.

The *R*-commands below generate the two-by-one figure containing the model residuals histogram and a q-q plot.

```
par(mfrow=c(2,1))# 2 by 1 graphic display
hist(sarima.citq$resid, col =16)# plot resid histogram using light grey colour
```

```
qqnorm(arima.citq$resid, col =16)# Q-Q plot, light grey colour
```

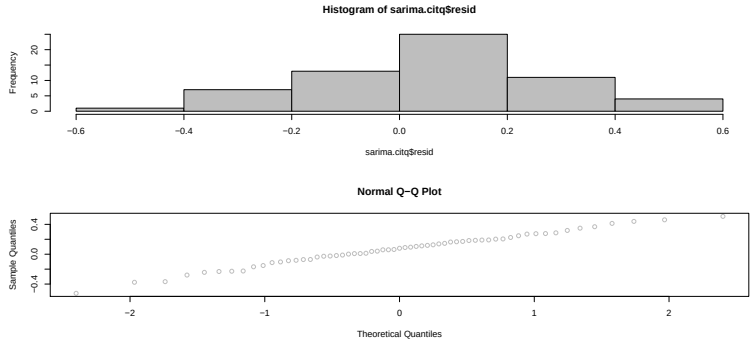

Figure 4.22: Residual Histogram and Q-Q plot from sarima.citq model

The histogram and the q-q plot in Figure 4.22 clearly indicate that the residuals from $ARIMA(3,1,0)(0,0,1)_4$ model are normally distributed with mean zero and variance δ^2. The model is then considered to forecast CIT future payments. The following section focuses on the model in-sample fitted values.

4.4.1.4 Quarterly CIT SARIMA model fitted values

Fitted values from quarterly CIT SARIMA model were obtained and plotted against the actuals for the in-sample using the following R commands:

```
cit_fit = fitted(arima.citq);ts.plot(citq,exp(cit_fit), main =
'Actual CIT Vs. Fitted', xlab = 'Period', ylab = 'Rand Million', col = c(1,2))
```

From the commands above, Figure 4.23 is then obtained.

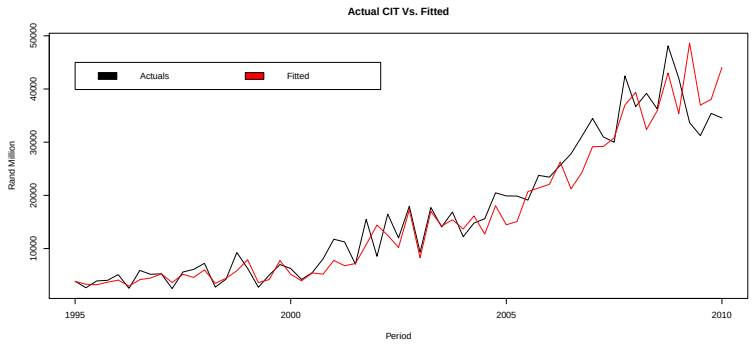

Figure 4.23: CIT Actuals and Fitted Values

Fitted values from the quarterly SARIMA model for CIT seems to follow the CIT actual movements though there data looks volatile over the years. Towards the end of 2010 the fitted values show an increasing trend and actuals drop. This is due to the lag effect of recession which started 2007/08 leading to drastic decrease in CIT tax payments for 2009/10 fiscal year.

The following section summarises quarterly CIT SARIMA model forecasts.

4.4.1.5 Quarterly CIT SARIMA model forecast values

CIT payments suffered the effect of recession more than PIT and VAT in 2008/09 and 2009/10 fiscal years. The CIT trend showed no improvements at the end of 2009/10 as a results of recession impact (See Appendix C for the three tax type trend plot).

The forecast for the fiscal years 2010/11, 2011/12 and 2012/13 are shown in Table 4.15.

Table 4.15: CIT Actuals Vs. SARIMA Forecast in Rand Million

Fiscal Year	Sample Used	Observations	CIT Actual	Forecast	% Error
2010/11	1995Q1 - 2010Q1	61	132,902	117,714	11.4%
2011/12	1995Q1 - 2011Q1	65	151,627	139,953	7.7%
2012/13	1995Q1 - 2012Q1	69	159,259	155,776	2.2%

The effect of recession was also seen in the fiscal years 2010/11 and 2011/12, resulting in CIT SARIMA model forecasts that were above 5% error compared to actual values. The series volatility and the recession made it difficult to capture most of the series fluctuation, although the fitted

model explained around 97% of the sample used. However, when comparing the forecasts to the actuals, we see the improvements in percentage error of 11.4% from the fiscal year 2010/11 to 2.2% in 2012/13 (Table 4.15).

The following R-codes were used to obtain quarterly forecasts for the three fiscal years in Table 4.15:

```
# Forecast 2010/11
citq_f = predict(arima.citq, n.ahead = 4, newxreg = c(0,0,1,0))
exp(citq_f$pred)#$;citq_f11 = forecast(arima.citq,4) ;citq_f11

# Forecast 2011/12
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2011,1),freq = 4)
newxreg = c(cdq3,0,0,1,0)
arima.citq = arima(log(citq), order = c(3,1,0),
seasonal = list(order = c(0,0,1),period = 4), newxreg )
citq_f12 = predict(arima.citq, n.ahead = 4,
newxreg = c(0,0,1,0));exp(citq_f12$pred)#$

# Forecast 2012/13
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2012,1),freq = 4)
newxreg = c(cdq3,0,0,1,0,0,0,1,0)
arima.citq = arima(log(citq), order = c(3,1,0),
seasonal = list(order = c(0,0,1),period = 4), newxreg )
citq_f13 = predict(arima.citq, n.ahead = 4, newxreg = c(0,0,1,0))
exp(citq_f13$pred) #$
```

The following summarise multiplicative Holt-winters method for quarterly CIT data.

4.4.2 Holt-Winters method for CIT time series

The quarterly CIT time series assumes one of the highly volatile multiplicative seasonality series that need extra effort when modeling and forecasting future outcomes. Holt-Winters with multiplicative seasonality was fitted to the square root transformed CIT for further minimisation of series variation as follows:

```
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2010,1),freq = 4)
```

```
citq.mhw = HoltWinters(citq^0.5, seasonal ='multiplicative');citq.mhw
```

Tables 4.16 and 4.17 were obtained from the command above. The two tables summarise the results of the multiplicative Holt-Winters smoothing constants, level, trend and initial values of seasonal factors respectively.

Table 4.16: CIT multiplicative Holt-Winters model Smoothing constants, level and trend estimation

Smoothing constants	Coefficient	Level and trend	Coefficient
alpha(α)	0.2234804	β_0	178.809322
beta (β)	0.000000	β_1	1.816682
gamma (γ)	0.8557064		

Table 4.17: CIT multiplicative Holt-Winters model initial values of seasonal factor

Seasonal smooth	s_1	s_2	s_3	s_4
Coefficient	1.005643	0.975048	1.067638	1.046920

From Table 4.16 we can mathematically represent the CIT multiplicative Holt-Winters as equation (4.4.3).

$$\sqrt{y_t} = (178.81 + 1.82t) \times Sn_t \times IR_t \tag{4.4.3}$$

where y_t represents quarterly CIT ($citq$) at time t and IR_t represents irregularities or error term at time t

The estimate average value l_T for the level, the estimate b_T for growth rate and the estimate sn_T for the seasonal factor of the data in time period T are given by the following equations:

$$l_T = 0.223(y_T - Sn_{T-1}) + (1 - 0.223)(l_{T-1} + b_{T-1}) \tag{4.4.4}$$

$$b_T = 0.00(l_T - l_{T-1}) + (1 - 0.00)b_{T-1}$$

$$b_T = b_{T-1} \tag{4.4.5}$$

$$Sn_T = 0.856(y_T - l_T) + (1 - 0.856)Sn_{T-1} \tag{4.4.6}$$

where l_{T-1} and b_{T-1} are estimates in time period $T-1$ for the level and growth rate component respectively. The Sn_{T-1} is the estimate in time $T-1$ for the seasonal factor. The trend constant or parameter β was assign zero weight, thus $b_T = b_{T-1}$ for $T = 1, 2, 3, ...$

Multiplicative Holt-Winters model represented by equation (4.4.3) explains around 91.4% of the variation in transformed CIT series. Subsection 4.4.2.1 summarises the model fitted values from the initial sample ending first quarter in 2010.

4.4.2.1 CIT multiplicative Holt-Winters fitted values

The square root transformed CIT multiplicative Holt-Winters model fitted values from the initial sample (Quarter 1, 1995 to Quarter 1, 2010) were obtained and were re-transformed to the original CIT using the following R commands:

```
fitted(vat.ahw) # fitted values of ln(vat)
exp(fitted(vat.ahw)[,1]) # convert ln(vat) to normal scale of vat
```

From the above R commands, the actual and fitted values were plotted against each other as follows:

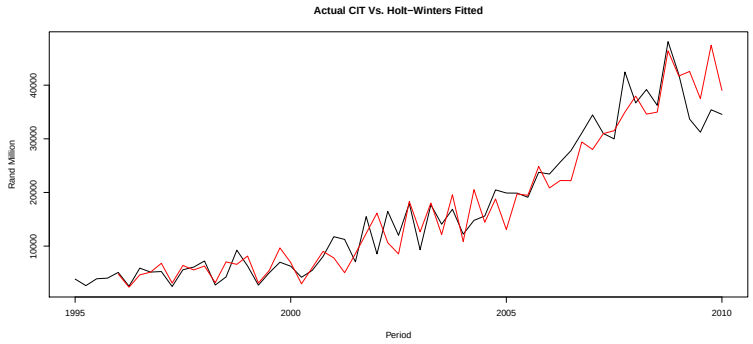

Figure 4.24: Quarterly CIT Actuals and Fitted Values

Fitted values on the figure above are not far from the volatile CIT actuals, even though there was recession effect at the end of the sample used. The quarterly observed and fitted values were aggregated to obtain the in-sample observed and fitted values for the fiscal year 1996/07 to 2009/10 (see Table B.3 in Appendix B).

The following section summarise the three years forecasts from the model in equation (4.4.3).

4.4.2.2 CIT multiplicative Holt-Winters forecast values

The forecast for the fiscal year 2010/11, 2011/12 and 2012/13 are shown in Table 4.18.

Table 4.18: CIT Actuals Vs. Holt-Winters Forecast in Rand Million

Fiscal Year	Sample Used	Observations	CIT Actual	Forecast	% Error
2010/11	1995Q1 - 2010Q1	61	132,902	141,289	-6.3%
2011/12	1995Q1 - 2011Q1	65	151,627	141,455	6.7%
2012/13	1995Q1 - 2012Q1	69	159,259	163,645	-2.8%

The multiplicative Holt-Winters model forecasts for three fiscal years (2010/11, 2011/12 and 2012/13) were R141,3bn, R141,3bn and R163,5bn respectively. The percentage forecast error of -6.3%, 6.7% and -2.8% were observed against the actuals for the same period. The errors of slightly above 5% in 2010/11 and 2011/12 were attributed to the longer recession effect on CIT payments, with the recovery in the forecast observed for fiscal year 2012/13.

The *R*-command used to obtain quarterly forecasts aggregated to the fiscal years forecasts in Table 4.18 above are given as follows:

```
# Forecast 2010/11
citqf_11 = predict(citq.mhw ,n.ahead= 4,prediction.interval = T, level = 0.95);
citqf_11 = citqf_11[,1]; citqf_11^2

# Forecast 2011/12
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2011,1),freq = 4);citq
citq.mhw = HoltWinters(citq^0.5, seasonal ='multiplicative');citq.mhw
citqf_12 = predict(citq.mhw ,n.ahead= 4,prediction.interval = T, level = 0.95);
citqf_12 = citqf_12[,1]; citqf_12^2

# Forecast 2012/13
citq = ts(with(citq_data,CITQ),start=c(1995,1),end = c(2012,1),freq = 4);citq
citq.mhw = HoltWinters(citq^0.5, seasonal ='multiplicative');citq.mhw
citqf_13 = predict(citq.mhw ,n.ahead= 4,prediction.interval = T, level = 0.95);
citqf_13 = citqf_13[,1]; citqf_13^2
```

4.4.3 CIT Models Comparisons

Due to the series volatility and the recession, CIT was the most challenging time series data to work with in this study because of the companies inconsistent behaviour and payments patterns. However, the SARIMA and the multiplicative Holt-Winters models were able to cover most of the variations in CIT. The impact of the recession on CIT was higher compared to that of the PIT and VAT series. Based on the models performance against the actuals on the in-sample and the forecast horizon, multiplicative Holt-Winters is recommended to predict future CIT payments. For annual fitted values against the actuals, see Table B.3 in Appendix B.

4.5 Total Tax Revenue

Total Tax Revenue (TTAXR) is the collection of all direct and indirect taxes, with 80% of the revenue contributed by two direct taxes (PIT and CIT) and one indirect tax (VAT). The remaining 20% include taxes such as Fuel Levies, Customs Duties, Air Departure Taxes, Electricity Levies, Diamond Export Levies and others. The TTAXR in Figure 4.5 shows a predictable increasing trend and an increasing variance and seasonality over time.

Figure 4.25 was generated using the following *R* command.

```
tsdisplay(ttaxr,main ='Level values of Total tax revene, ACF and PACF')
```

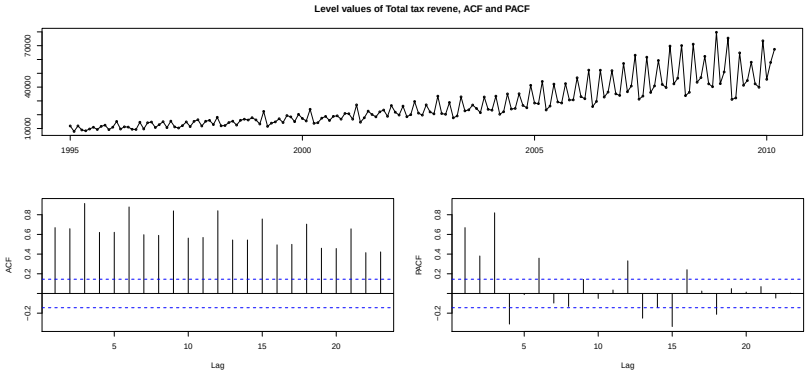

Figure 4.25: Total tax revenue (TTAXR), ACF and PACF

The following sub-sections summarise modeling and forecasting of TTAXR using SARIMA method.

4.5.1 TTAXR SARIMA Model

The natural logarithm transformed TTAXR was used to minimise the series variation and for better model fit.

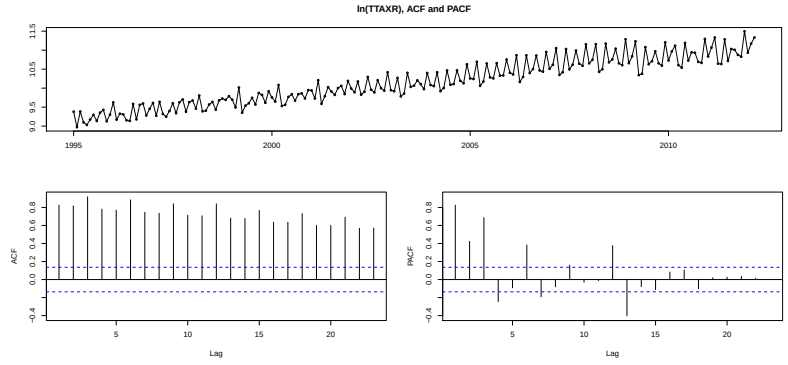

Figure 4.26: ln(TTAXR) Time Series Display at level values

From Figure 4.26 it can be seen that the natural *log* transformed TTAXR is non-stationary at level values. The non-stationarity of the transformed TTAXR requires differencing to obtain the data stationarity.

The figure below shows the differenced log-transformed series for TTAXR.

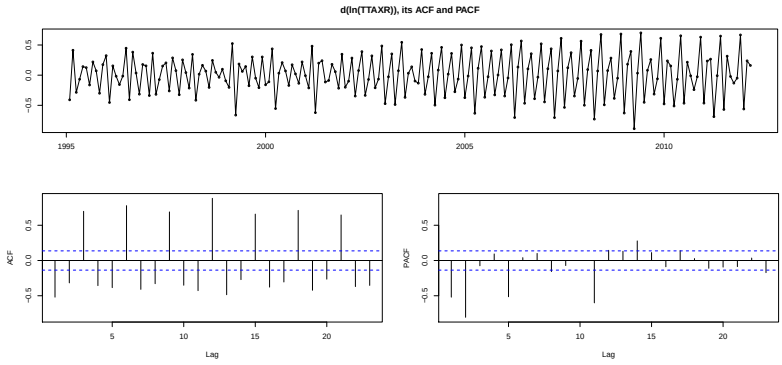

Figure 4.27: dln(ttaxr),ACF and PACF'

The first differenced TTAXR series shows stationarity around the mean and variance (δ^2). Most of the ACF and PACF are beyond the 95% boundary.

The following section summarises the SARIMA for the TTAXR series.

4.5.1.1 TTAXR SARIMA model output

Table 4.19 presents three competing TTAXR SARIMA models. TTAXR, being a collection of all taxes, is bound to be seasonal because the three taxes which are seasonal (PIT, CIT and VAT) contribute around 80% to TTAXR.

Table 4.19: SARIMA Models for Total tax revenue(TTAXR)

	Model	R^2	L	AIC	BIC	$LB(p-Value)$
1	$ARIMA(1,1,1)(1,0,1)_{12}$	0.9860883	189.68	-369.37	-353.35	0.02309
2	$ARIMA(2,1,1)(1,0,0)_{12}$	0.985655	192.52	-373.05	-353.82	0.2529
3	$ARIMA(3,1,3)(1,0,1)_{12}$	0.9888463	198.68	-379.35	-350.52	0.7541

Model 3, which is the $ARIMA(3,1,3)(1,0,1)_{12}$ is showing to be the best fitting model, with the p-value from Ljung-Box test of 0.7541 signifying the white noise residuals. The model power to predict/forecast the TTAXR is supported by the higher log-likelihood of 198.7 and the $R^2 = 98.9\%$ compared to the other two competing SARIMA models. The selected model can be mathematically represented by:

$$(\mathbf{1} - \phi_1\mathbf{B} - \phi_\mathbf{2}\mathbf{B^2} - \phi_\mathbf{3}\mathbf{B^3})(\mathbf{1} - \mathbf{\Phi_1}\mathbf{B^{12}})\mathbf{w_t} = (\mathbf{1} - \theta_\mathbf{1}\mathbf{B} - \theta_\mathbf{2}\mathbf{B^2} - \theta_\mathbf{3}\mathbf{B^3})(\mathbf{1} - \mathbf{\Theta_1}\mathbf{B^{12}})\epsilon_\mathbf{t} \quad (4.5.1)$$

Where $\mathbf{w_t} = ln(ttaxr_t) - ln(ttaxr_{t-1})$

ϕ_i is the i^{th} autoregressive ($AR(i)$) coefficient, $i = 1, 2 and 3$

Φ_1 is the first seasonal autoregressive ($SAR(1)$) coefficient

θ_j is the j^{th} moving average ($MA(j)$) coefficient, $j = 1, 2 and 3$

Θ_1 is the first seasonal moving average ($SMA(1)$) coefficient

$\mathbf{B}$ is a back shift operator with $\mathbf{B^i}\ \mathbf{w_t} = \mathbf{w_{t-i}}$ and $\mathbf{B^j}\ \epsilon_\mathbf{t} = \epsilon_\mathbf{t-j}$ and,

ϵ_t is an error term at time t

Table 4.20 presents the maximum likelihood parameters estimation for the SARIMA model in equation (4.5.1) fitted to TTAXR time series, generated by $cwp(sarima.ttaxr)$, R command.

Table 4.20: Total tax revenue SARIMA model parameters estimation

Parameter	Coefficient	Standard error	t-Value	p-Value
$AR(1)$	-1.8117	0.1586	-11.4226	$3.22 * 10^{-30}$
$AR(2)$	-1.4011	0.1467	-9.5487	$1.31 * 10^{-21}$
$AR(3)$	-0.4158	0.0848	-4.9023	$9.47 * 10^{-07}$
$MA(1)$	0.8176	0.1722	4.7469	$2.07 * 10^{-06}$
$MA(2)$	-0.3029	0.1036	-2.9243	0.0035
$MA(3)$	-0.5714	0.1419	-4.0266	$5.66 * 10^{-05}$
$SAR(1)$	0.9556	0.0187	51.0238	0.0000
$SMA(1)$	-0.2470	0.0955	-2.5871	0.0097

Using the coefficients in Table 4.20, equation (4.5.1) can be re-written as:

$$(\mathbf{1}+1.81\mathbf{B}+\mathbf{1.40B^2}+\mathbf{0.42B^3})(\mathbf{1}-\mathbf{0.96B^{12}})\mathbf{w_t} = (\mathbf{1}+\mathbf{0.82B}-\mathbf{0.30B^2}-\mathbf{0.57B^3})(\mathbf{1}-\mathbf{0.25B^{12}})\epsilon_\mathbf{t} \tag{4.5.2}$$

The following section summarises the residual analysis of the model generated in equation (4.5.2).

4.5.1.2 TTAXR SARIMA model residual analysis

Another way of reviewing the model goodness is to analyse the relationship between the data series of interest (log transformed TTAXR) and the model residuals. If this relationship between transformed TTAXR and the residuals is highly correlated, then there could be some significant

information unexplained or not covered by the model fitted. Otherwise, if relationship is insignificant, the model is assumed to be good for forecasting.

Figure 4.28 is a scatter plot of the log transformed TTAXR against the model residuals generated using the following *R* commands:

```
plot(log(ttaxr),sarima.ttaxr$resid,col = c(1,2),main='ln(ttaxr) Vs. Residuals')
legend(9,0.18,ncol =2, c('ln(ttaxr)','Residuals'),fill = c(1,2)) #$
```

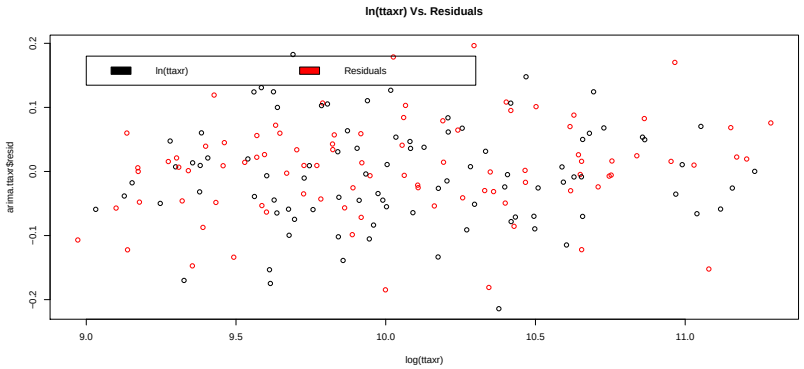

Figure 4.28: ln(ttaxr) Vs. Residuals

The scatter plot shows no clear relationship between the log transformed TTAXR and the model residuals. The more general statistic to confirm the weaker correlation between the observed series and model residuals will be the computation of the correlation coefficient, which was found to be around 0.109. There is a weak relationship between the log transformed TTAXR and the model residual series. The correlation coefficient is obtained by the following *R* command:

```
cor(log(ttaxr),sarima.ttaxr$resid) #$
0.1092146
```

4.5.1.3 TTAXR SARIMA model fitted values

Residuals from TTAXR model were found to be uncorrelated. This permits us to use the model to forecast Total tax revenue future values/payments.

Monthly fitted values were obtained as follows:

```
arima.ttaxrf=exp(fitted(sarima.ttaxr))# re-transform the log transformed fitted
sarima.ttaxrf # Print the fitted values
```

The researcher then plotted the fitted values against the observed shown in the Figure 4.5.1.3 as follows:

```
ts.plot(ttaxr,arima.ttaxrf,col = c(1,2),main = 'Actual TTAXR Vs. SARIMA Fitted',
xlab ='Period', ylab ='Rand Million')
legend(1995,70000,ncol =2, c('Total Tax','Fitted'),fill = c(1,2))
```

The *R*-commands above generate the Figure below:

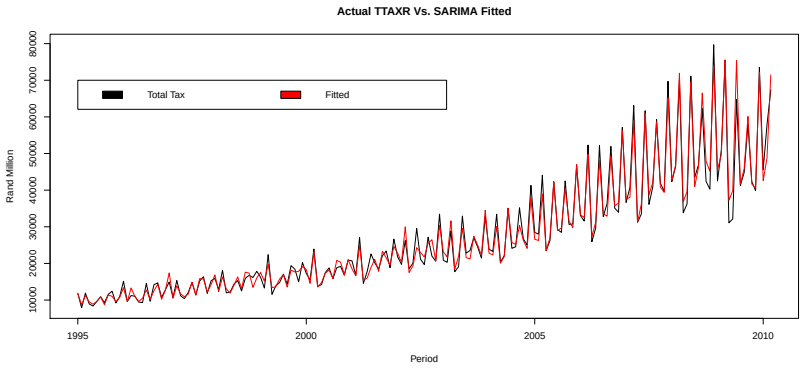

Figure 4.29: Actual TTAXR and SARIMA Fitted Values

From Figure 4.29 it can be observed that the fitted values follow the pattern of the actuals. This implies that the model fits the TTAXR data set and can be used to forecast future values. Monthly fitted values were then aggregated to form fiscal year fitted values for the period 1995/96 to 2009/10 (See Table B.4 in Appendix B for annual fitted values).

The following section presents the forecasts for three fiscal years from the fitted SARIMA model.

4.5.1.4 TTAXR SARIMA model forecast values

Table 4.21 summarises the results of TTAXR actual payments and the SARIMA model aggregated forecasts for the fiscal year 2010/11, 2011/12 and 2012/13.

Table 4.21: TTAXR Actuals Vs. SARIMA Forecast in Rand Million

Fiscal Year	Sample Used	Observations	TTAXR Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	674,183	617,497	8.41%
2011/12	Jan 1995 - Mar 2011	195	742,650	738,800	0.52%
2012/13	Jan 1995 - Mar 2012	207	813,826	811,812	0.25%

The three fiscal year forecasts followed the observed actuals (have small percentage error) except for the year 2010/11. This was attributed to recession disturbance of the year 2009/10. Below are the *R*-commands used on the TTAXR data to generate the forecasts:

```
# Forecast 2010/11
lttaxr.f = predict(arima.ttaxr, n.ahead = 12) # Forecast in log scale
ttaxr.f = exp(lttaxr.f$pred);ttaxr.f #$ re-transform and print the forecast

# Forecast 2011/12
ttaxr=ts(with(mydata,TTAXREV),start=c(1995,1),end = c(2011,3),freq=12)
arima.ttaxr = arima(log(ttaxr), order = c(3,1,3),
seasonal = list(order = c(1,0,1),period = 12))
lttaxr.f12 = predict(arima.ttaxr, n.ahead = 12)
ttaxr.f12 = exp(lttaxr.f12$pred); ttaxr.f12 #$

# Forecast 2012/13
ttaxr=ts(with(mydata,TTAXREV),start=c(1995,1),end = c(2012,3),freq=12)
arima.ttaxr = arima(log(ttaxr), order = c(3,1,3),
seasonal = list(order = c(1,0,1),period = 12))
lttaxr.f13 = predict(arima.ttaxr, n.ahead = 12)
ttaxr.f13 = exp(lttaxr.f13$pred);ttaxr.f13 #$
```

4.5.2 Holt-Winters method for TTAXR time series

TTAXR is a combination of all taxes which are additive and multiplicative. The multiplicative Holt-winters was assumed to model TTAXR data. These results were generated by the following *R* command:

```
ttaxr.mhw = HoltWinters(ttaxr, seasonal ='multiplicative'); ttaxr.mhw
```

Table 4.22 represents the smoothing constants and level and trend coefficients estimation for the multiplicative Holt-Winters model in equation (4.5.3) fitted to TTAXR time series.

Table 4.22: Total tax revenue multiplicative Holt-Winters model coefficients estimation

Smoothing constants	Coefficient	Level and trend	Coefficient
alpha(α)	0.1042284	β_0	49391
beta (β)	0.2100435	β_1	45.341
gamma (γ)	0.8944451		

The multiplicative Holt-Winters model initial seasonality factor Sn_t values for 12 months are given in Table 4.23.

Table 4.23: Initial values for seasonal factors from Total tax revenue Holt-Winters model

Seasonal smooth	s_1	s_2	s_3	s_4	s_5	s_6
Coefficient	0.6156	0.6473	1.3125	0.8367	0.9110	1.1944
Seasonal smooth	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
Coefficient	0.8675	0.8190	1.5261	0.9268	1.1464	1.3740

From Table 4.22, multiplicative Holt-Winters can be represented by the following equation.

$$y_t = (49391 + 45.3t) \times Sn_t \times IR_t \tag{4.5.3}$$

where y_t represent TTAXR, and IR_t irregularities or error term

Equation (4.5.3) shows the results of multiplicative Holt-Winters model for TTAXR in Table 4.22. The estimate average value l_T for the level, the estimate b_T for growth rate and the estimate sn_T for the seasonal factor of the data in time period T are given by the following smoothing equation:

$$l_T = 0.104(\frac{y_T}{Sn_{T_L}}) + (1 - 0.104)(l_{T-1} + b_{T-1}) \tag{4.5.4}$$

$$b_T = 0.210(l_T + l_{T-1}) + (1 - 0.210)b_{T-1} \tag{4.5.5}$$

$$Sn_T = 0.894(\frac{y_T}{l_T}) + (1 - 0.894)Sn_{T-L} \tag{4.5.6}$$

where l_{T-1} and b_{T-1} are estimates in time period $T-1$ for the level and growth rate respectively, and Sn_{T-1} is the estimate in time $T-1$ for the seasonal factor. The multiplicative Holt-Winters model represented by equation (4.5.3) explains around 99% of the variation in Total tax revenue series. The following section summarises the model fitted values.

4.5.2.1 TTAXR Holt-Winters fitted values

Multiplicative Holt-Winters monthly fitted value were obtained as follow:

```
fitted(ttaxr.mhw)[,1] # ttaxr fitted values
```

The Figure below shows the fitted values against the actuals on the in-sample:

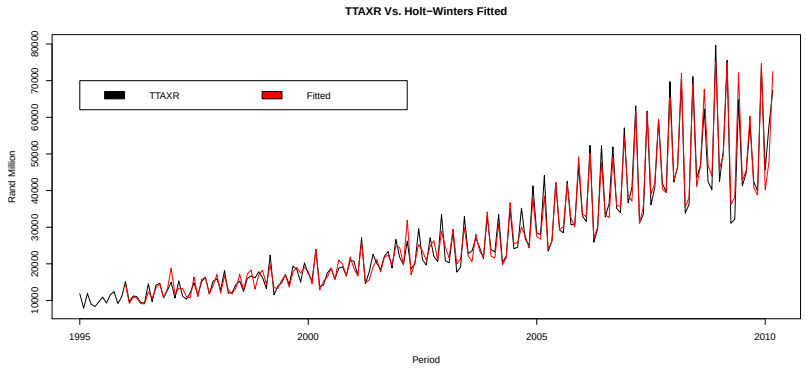

Figure 4.30: Actual TTAXR and Holt-Winters Fitted Values

Figure 4.30 shows that the fitted values are not far away from the actuals, which clearly indicates that the model performs well on the in-sample. This allows us to look into the model forecast in the following section.

4.5.2.2 TTAXR multiplicative Holt-Winters forecast values

The TTAXR payments and the multiplicative Holt-Winters model forecast for the fiscal year 2010/11, 2011/12 and 2012/13 are presented in Table 4.24.

The decrease in TTAXR to R598,7bn in 2009/10 from R625,1bn in 2008/09 was due to the economic recession, resulting in a model forecast value of R605,3bn in 2010/11. This was an error of 10.21% below the actual of R674,2bn for the same period. The forecast for the fiscal years 2011/12

Table 4.24: Total Tax Actuals Vs. Holt-Winters Forecast in Rand Million

Fiscal Year	Sample Used	Observations	TTAXR Actual	Forecast	% Error
2010/11	Jan 1995 - Mar 2010	183	674,183	605,340	10.21%
2011/12	Jan 1995 - Mar 2011	195	742,650	734,690	1.07%
2012/13	Jan 1995 - Mar 2012	207	813,826	816,522	-0.33%

and 2012/13 were 0.07% and -0.33% respectively, as shown in Table 4.24. The model performs well on the out of sample forecasts.

The *R*-codes used to obtain the monthly forecast for the three fiscal years are as shown below:

```
# Forecast 2010/11
ttaxr_mhwf = predict(ttaxr.mhw,n.ahead= 12,prediction.interval = T,
level = 0.95); ttaxr_mhwf[,1]

# Forecast 2011/12
ttaxr =ts(with(mydata,TTAXREV),start=c(1995,1),end = c(2011,3),freq=12)
ttaxr.mhw = HoltWinters(ttaxr, seasonal ='multiplicative')
t.mhw12 = predict(ttaxr.mhw , n.ahead = 12,prediction.interval = T,
level = 0.95); t.mhw12[,1]

# Forecast 2012/13
ttaxr =ts(with(mydata,TTAXREV),start=c(1995,1),end  c(2012,3),freq=12)
ttaxr.mhw = HoltWinters(ttaxr, seasonal ='multiplicative')
t.mhw13 = predict(ttaxr.mhw , n.ahead = 12,prediction.interval = T,
level = 0.95); t.mhw13[,1]
```

4.5.3 TTAXR models comparisons and conclusion

Both the SARIMA and multiplicative models were fitted to TTAXR and performed well on the in-sample period. However, on the out of sample, the SARIMA model outperformed the Holt-Winters model, despite the trend disturbance on fiscal year 2010/11 from the impact of the economic recession which manifested on the TTAXR in 2009/10. When comparing the two models, the SARIMA

model is recommended when forecasting the continuity of TTAXR payments (See Table B.4 in Appendix B for year fitted values compared to actuals).

Chapter 5

Conclusion and Recommendations

This study uses aspects of time series methodology (Seasonal autoregressive integrated moving averages (SARIMA) and Holt-Winters) to model annual payments and to forecast 2010/11, 2011/12 and 2012/13 payments for the three major taxes, Personal Income Tax, Corporate Income Tax, Value Added Tax and Total Tax Revenue in the South African Revenue Service. The monthly data used for modeling tax revenues of the major taxes were drawn from January 1995 to March 2010 (in sample data) for the mentioned tax types. Due to higher volatility and emerging negative values, the Corporate Income Tax monthly data was converted to quarterly data from the first quarter of 1995 to the first quarter of 2010. The monthly and quarterly fitted/forecast values were aggregated to form annual fitted/forecast values.

The competing Seasonal Autoregressive Integrated Moving Averages and Holt-Winters models were fitted to the taxes mentioned above, and the measures of accuracy such as Mean Error, Root Mean Squared Error, Mean Absolute Error, Mean Percentage Error, Mean Absolute Percentage Error and Mean Absolute Squared Error were computed. Other measures of accuracy such as R^2, Akaike Information Criterion and Bayesian Information Criterion were used. The Ljung-Box statistic was also used for the SARIMA models to test if the residuals from the models were white noise.

The SARIMA and Holt-Winters models applied to the taxes mentioned above generally performed well, however the model that best fits the tax type (with minimal percentage error) on the in-sample and out of sample was recommended to be used for future forecasts, with the proviso of updating the series as the new information becomes available.

The model that best fit the tax types mentioned above between the Holt-Winters and SARIMA model is the following: the results show that both the models perform well against the Personal Income Tax and Value Added Tax data, however the Holt-Winters model outperformed the SARIMA model for the volatile Corporate Income Tax data, and for Total Tax Revenue, the SARIMA model out-performed the Holt-Winters model.

The comparison of the two methods forecasts with the actual realisation was done for Personal Income Tax, Corporate Income Tax, Value Added Tax and Total Tax Revenue for 2010/11, 2011/12 and 2012/13, and it was observed that the models performed well. However, some years forecast error against the actuals was higher than 5% due to the economic recession which impacted negatively on revenue collection (in-sample data used). The study concludes that the selected models are expected to perform better when forecasting future values, assuming that there will be no shocks such as an economic recession.

The SARIMA and Holt-Winters models use the historical patterns of the same series to forecast future values, i.e. they are seen as not considering factors that influence the movement of the variable of interest. However, the historical patterns from time series data of interest includes the effect of the explanatory variables. This is an indication that the time series methods eliminate the issue of biasness when fitting the model. The time series models used in this study are good for short term forecasts and they are limited for sensitivity analysis. Every model has some advantages and disadvantages these depend on the use of the models. In this study the time series aimed to forecast the continuation of the historical patterns of the South African taxes Personal Income Tax, Corporate Income Tax, Value Added Tax and Total Tax Revenue.

This study recommends the use of these methods when forecasting future payments. If tax recovery approaches do not change, these methods will be precise with limited bias in forecasting tax revenues with minimal error and fewer model revisions being necessary. This will assist the South African Revenue Service authorities in making decisions regarding future revenues.

Appendix A

Coefficient with p-values (cwp) for ARIMA Models

The coefficient with p-values (cwp) is the program used for all ARIMA/SARIMA models to calculate the p-values or significance of all the coefficients included in the model using the coefficients standard errors and the coefficients t-values and is given by the R-codes below;

```
 cwp <- function (object){
     coef <- coef(object)
        if (length(coef) > 0) {
         mask <- object$mask
         sdev <- sqrt(diag(vcov(object)))
         t.rat <- rep(NA, length(mask))
         t.rat[mask] <- coef[mask]/sdev
         pt <- 2 * pnorm(-abs(t.rat))
         setmp <- rep(NA, length(mask))
         setmp[mask] <- sdev
         sum <- rbind(coef, setmp, t.rat, pt)
         dimnames(sum) <- list(c("coef", "s.e.", "t ratio", "p-value"),
             names(coef))
         return(sum)
     } else return(NA)#$
}
```

Appendix B

In-ample Actuals Vs. Fitted

Table B.1: PIT Actuals Vs. SARIMA and Holt-Winters Fitted in Rand Million

Fiscal Year	PIT Actual	SARIMA Fitted	SARIMA % Error	AHW Fitted	AHW % Error
1995/96	51,179	50,952	0.44%		
1996/97	59,520	59,275	0.41%	59,740	-0.37%
1997/98	68,342	68,047	0.43%	69,183	-1.23%
1998/99	77,734	77,570	0.21%	78,130	-0.51%
1999/00	85,884	86,598	-0.83%	86,933	-1.22%
2000/01	86,478	89,895	-3.95%	90,010	-4.08%
2001/02	90,390	89,195	1.32%	88,896	1.65%
2002/03	94,337	94,335	0.00%	94,057	0.30%
2003/04	98,495	99,724	-1.25%	100,052	-1.58%
2004/05	110,982	109,202	1.60%	109,328	1.49%
2005/06	125,645	123,982	1.32%	124,869	0.62%
2006/07	140,579	138,886	1.20%	138,977	1.14%
2007/08	168,774	165,614	1.87%	166,243	1.50%
2008/09	195,146	195,966	-0.42%	197,422	-1.17%
2009/10	205,146	207,484	-1.14%	207,539	-1.17%

FY− Financial year, SARIMA− Seasonal Autoregressive Integrated Moving Averages

AHW− Additive Holt-Winters, MHW− Multiplicative Holt-Winters

Table B.2: VAT Actuals Vs. SARIMA and Holt-Winters Fitted in Rand Million

Fiscal Year	VAT Actual	SARIMA Fitted	SARIMA % Error	AHW Fitted	AHW % Error
1995/96	32,768	31,328	4.39%		
1996/97	35,903	35,186	2,00%	35,216	1.91%
1997/98	40,096	39,484	1.53%	39,616	1.20%
1998/99	43,985	43,829	0.35%	43,821	0.37%
1999/00	48,377	46,844	3.17%	47,566	1.68%
2000/01	54,455	53,609	1.55%	53,585	1.60%
2001/02	61,056	59,193	3.05%	59,623	2.35%
2002/03	70,150	69,984	0.24%	69,032	1.59%
2003/04	80,682	77,942	3.40%	77,755	3.63%
2004/05	98,158	94,107	4.13%	92,248	6.02%
2005/06	114,352	111,108	2.84%	109,362	4.36%
2006/07	134,463	133,193	0.94%	130,680	2.81%
2007/08	150,443	151,378	-0.62%	149,198	0.83%
2008/09	154,343	160,558	-4.03%	160,170	-3.78%
2009/10	147,941	147,428	0.35%	153,992	-4.09%

FY− Financial year

SARIMA− Seasonal Autoregressive Integrated Moving Averages

AHW− Additive Holt-Winters

MHW− Multiplicative Holt-Winters

Table B.3: CIT Actuals Vs. SARIMA and Holt-Winters Fitted in Rand Million

FY	CIT Actual	SARIMA Fitted	SARIMA % Error	MHW Fitted	MHW % Error
1995/06	15,667	14,226	9.2%		
1996/07	18,834	16,804	10.8%	18,929	-0.5%
1997/08	21,378	19,334	9.6%	21,340	0.2%
1998/09	22,523	21,597	4.1%	24,982	-10.9%
1999/00	20,972	20,693	1.3%	25,135	-19.9%
2000/01	29,492	22,319	24.3%	25,758	12.7%
2001/02	42,354	39,113	7.7%	42,148	0.5%
2002/03	55,745	48,254	13.4%	50,145	10.0%
2003/04	60,881	60,341	0.9%	60,615	0.4%
2004/05	70,782	61,386	13.3%	66,851	5.6%
2005/06	86,161	79,310	8.0%	84,893	1.5%
2006/07	118,999	100,944	15.2%	101,851	14.4%
2007/08	140,120	136,357	2.7%	135,418	34%
2008/09	165,539	146,696	11.4%	157,687	4.7%
2009/10	134,883	167,715	-24.3%	166,526	-23.5%

FY− Financial year

SARIMA− Seasonal Autoregressive Integrated Moving Averages

AHW− Additive Holt-Winters

MHW− Multiplicative Holt-Winters

Table B.4: Total tax Actuals Vs. SARIMA and Holt-Winters Fitted in Rand Million

Fiscal Year	TTAXR Actual	SARIMA Fitted	SARIMA % Error	MHW Fitted	MHW % Error
1995/96	127,278	126,528	0.59%		
1996/97	147,332	145,529	1.22%	147,382	-0.03%
1997/98	165,327	164,052	0.77%	166,065	-0.45%
1998/99	184,845	185,263	-0.23%	186,003	-0.63%
1999/00	201,386	202,030	-0.32%	202,679	-0.64%
2000/01	220,334	218,963	0.62%	220,030	0.14%
2001/02	252,298	251,328	0.38%	253,467	-0.46%
2002/03	282,210	285,538	-1.18%	283,847	-0.58%
2003/04	302,508	297,188	1.76%	296,299	2.05%
2004/05	354,980	338,250	4.71%	341,742	3.73%
2005/06	417,195	416,016	0.28%	421,820	-1.11%
2006/07	495,549	485,102	2.11%	483,320	2.47%
2007/08	572,815	574,744	-0.34%	574,972	-0.38%
2008/09	625,100	637,313	-1.95%	634,712	-1.54%
2009/10	598,705	617,332	-3.11%	609,685	-1.83%

FY− Financial year

SARIMA− Seasonal Autoregressive Integrated Moving Averages

AHW− Additive Holt-Winters

MHW− Multiplicative Holt-Winters

Appendix C

Seasonal-Trend Decomposition procedure based on Loess plots

Seasonal-Trend Decomposition procedure based on Loess (STL) this is the procedure that decompose time series data into trend, seasonal, and remainder components (Cleveland et al. (1990)). The STL function is included in *R* library called forecast by Hyndman (2008).

```
library(forecast)
plot(stl(pit,s.window = "periodic"),main ='Personal income tax (PIT) STL-plot')
```

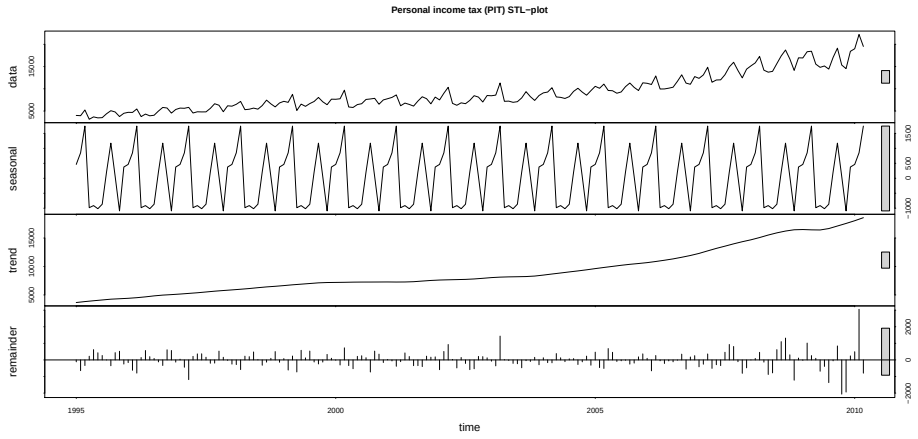

Figure C.1: Personal income tax (PIT) STL-plot

```
plot(stl(vat,s.window = "periodic"),main ='Value added tax (VAT) STL-plot')
```

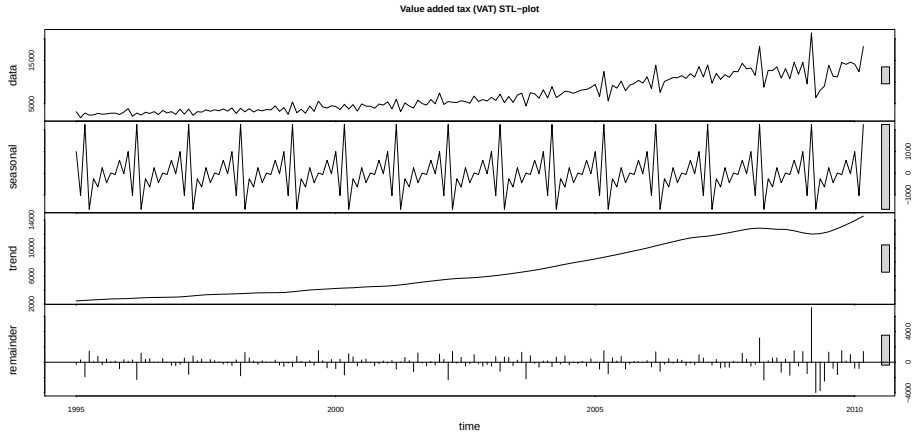

Figure C.2: Value added tax (VAT) STL-plot

```
plot(stl(cit,s.window = "periodic"),main ='Corporate income tax (CIT) STL-plot')
```

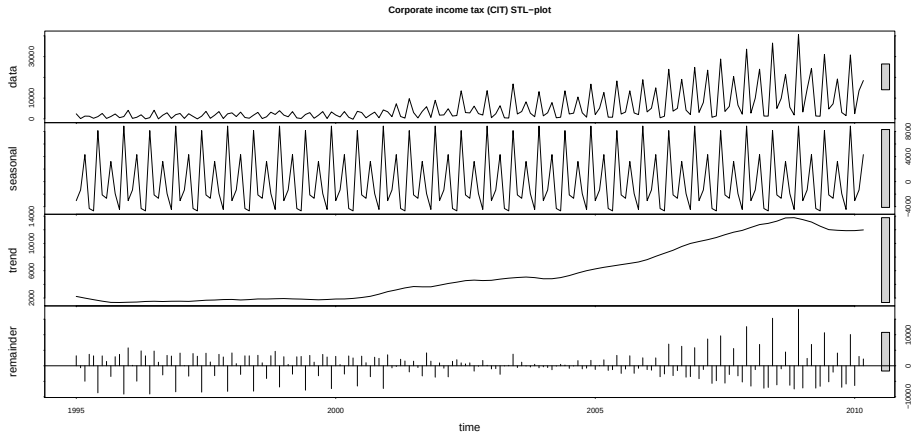

Figure C.3: Corporate income tax (CIT) STL-plot

R-squared (R^2) Computation

Let f be the fitted values of y and $\bar{y}$ the mean, then

$$R^2 = \frac{\sum(f - \bar{y})^2}{\sum(y - \bar{y})^2} \tag{C.0.1}$$

Which can be coded in R software as;

```
RSq = sum((f- mean(y))^2)/sum((y - mean(y))^2)
```

Bibliography

Berwick M., and Malchose D., 2012. Forecasting North Dakota fuel tax revenue and license and registration fee revenue. North Dakota. North Dakota State University Frago.

Bowerman B.L., O'Connell R.,and Koehler L., 2004. Forecasting Time series and regression. Fourth edition. Cengage Learning Brand.

Box G.E.P., and Jemkins G.M., 1970. Time Series Analysis: Forecasting and Control. San Francisco Calif: Holden Day.

Box G.E.P., and Jemkins G.M., 1976. Time Series Analysis: Forecasting and Control. San Francisco Calif: Holden Day.

Boonzaaier W., 2012.Revenue forecasting practices: current international developments and the case of south Africa. South African Revenue Service.

Botric V., and Vizek M., 2012. Forecasting Fiscal Revenues in a Transition Country: The Case of Croatia. Croatia. *Zagreb International Review of Economics and Business,* 15:23−36.

Brew L., and Wiah E.N., 2012. An Assessment of the Efficiency in the Collection of Value Added Tax Revenue in Tarkwa-Nsuaem Municipality (Ghana) Using Time Series Model. *British Journal of Arts and Social Sciences,* 6.

Brojba L.C., Dumitru C.G., and Belciug A.V., 2010. On the use of some optimal strategies of fiscal administration during economic crisis. Romania: *Romanian Journal of Economic Forecasting,* 1.

Chatagny F., and Soguel N.C., 2009. Tax Revenue Forecasting in the Swiss Cantons: A Time Series Analysis. Switzerland. IDHEAP.

Chatfield C., 2004. Analysis of time series: An introduction. Sixth edition Florida, CRC Press.

Cleveland R.B., Cleveland C.G., Belciug W.S., McRea J.E., and Terpenning I., 1990. STL: A seasonal-Trend decomposition Procedure Based on Loess Sweeden. *Journal of Official Statistics,* 6:3−73.

Corvalao D., E., Samohyl R., W., and Brasil G., H., 2010. Forecasting the Collection of the State Value Added Tax (Icms) in Santa Catarina: the General to Specific Approach in Regression Analysis. Santa Catarina, Brazil. *Brazilian Journal of Operations and Production Management.* 7: 105−121.

Cote K.N.A., Smith W.M.D., and Fullerton T.M., 2010. Municipal Non-Residential Real Property Valuation Forecast Accuracy. University of Taxas at El Paso. *International Journal of Business and Economics Perspectives,* 1.

Davies J.B., 2009. Combining Microsimulation with CGE and Macro Modelling for Distributional Analysis in Developing and Transition Countries*International Journal of Microsimulation,* 2(1):49−65.

Dickey D.A., and Fuller W.A., 1979. Distribution of the estimates for autoregressive time series with a root.*The American statistical association,*74:427−431.

Diggle, P.J., 1990. Time Series: a biostatistical introduction.,Oxford: Clarendon Press.

Etuk E.H., Igbuda R.C.,2005. A SARIMA fit to monthly Nigerian Naira-British pound exchange rates Nigeria. *Journal of computations and modeling,* 13:133−144.

Engle R.F., and Granger C.W.J., 1986. Co-integration and error correction: Representation, estimation, and testing.*Econometrica,*55(2):251−276.

Fomby T.B., 2008. Exponential Smoothing Models.Southern Methodist University Dallas Department of Economics.

Gooijer J.G., Abraham B., Jenning C.L. and Robinson L., 1985. Methods for determining the order of an autoregressive moving average; A survey. *Int. Statist, Rev, 53:301−329.*

Gujarati D.N., and Porter D.C., 2003. Basic Econometric:fifth addition *Higher Education, McGrawHill.*

Hendry D.F., and Nielse B., 2007. Econometric Modelling A Likelihood Approach: First Edition New Jersey: Princeton University Press.

Hyndman R.J., Makridakis S., and Wheelwright S.C., 1998. Forecasting methods and application: Third edition USA: John Wiley and Sons.

Hyndman R.J., Koehlor A.B., Synder R.D, and Grose S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. Australia. *International journal of forecasting,* 18:439−454.

Hyndman R.J., 2008. Time series and forecasting in R Australia: Monash University.

Jayesekara N., and Passty B.W., 2009. City of Cincinnati Income Tax Revenue: A Forecast Based on Economic Data and Causal Inference. *University of Cincinnati.*

John G., Nelson P., and Reet V., 2007. Unit Root Tests and Structural Breaks: A Survey with Applications. Universidad Pablo Olavide Sevilla. *Revista De Metodos Cuantitativos Para La Economia,* 3: 63−79.

Kalekar P.S., 2008. Time Series Forecasting Using Holt-Winters Exponential Smoothing. Kanwal Rekhi School of Information Technology.

Kakwani N.C., 1977. Measurement of tax progressivity: An international comparison.*Economic Journal,* 87:71−80.

Koirala T.P., 2012.Government Revenue Forecasting in Nepal. Nepal. Nepal Rastra Bank.

Kudrle R.T., 2008. The OECD Harmful Tax Competition initiative and Tax Havens: From Bombshell to Damp Squib. University of Minnesota. The Berkeley Electronic Press. *Global Economic Journal* 8.

Maindonald J., and Braun J., 2003. Data analysis and graphics using R: an example-based approach.Cambridge: Cambridge University Press.

Mahsin M.D., Akhter Y., and Begum M., 2012. Modeling rainfall in Dhaka division of Bangladesh using time series analysis . Dhaka, Bangladesh. *Journal of Mathematical Modeling and Application,*1:67−73.

Nirmal M., and Sundaram S.M., 2010. A seasonal ARIMA model for forecasting monthly rainfall in Tamilnadu. Tamilnadu, India. *National Journal of advances in building sciences and mathematics,* 1.

Ostertagova E., and Ostertag O., 2012. Forecasting Using Simple Exponential Smoothing Method. Kosice. *Acta Electrotechnica et Informatica,* 12:62−66.

Otu O.A., Ousji G.A., jude O., Ifeyinwa M.H., and Iheingwara A.I., 2014. Application of SARIMA models in modelling and forecasting Nigeria's inflation rates. America. *American Journal of Appleid Mzthematics and Statistics,* 2:16−28.

Paul S.K., 2011. Determination of Exponantial Smoothing Constant to Minimize Mean Absolute Deviation. Bangladesh University. *Global Journal of research in Engineering,* 11

Pelinescu E., Anton L.V., Ionescu R., and Tasca R., 2010. The analysis of local budgets and their importance in the fight against the economic crisis effects. Romania. *Romanian Journal of Economic Forecasting,*1.

Pindyck R.S., and Rubinfeld D.L, 1998. Econometric Models and Economic Forecast: Fourth edition. Singapore: McGraw-hill International Editions.

SARS., 2014. South African Revenue Service 201314 Annual report.South Africa: SA National publication.

SARS and National Treasury., 2012. South African Tax Statistics.South Africa: SA National publication.

Singh E.H., 2013. Forecasting Tourist Inflow in Bhutan using Seasonal ARIMA Bhutan,India. *International Journal of Science and Research,* 2:2319−7064.

Silvestrini A., Salto M., Moulin L., and Veredas D., 2008. Monitoring and forecasting annual public deficit every month: the case of France. France. *Empirical Economics,*34:493−534.

Sobela R.S., and Holcombe R.G., 1996. Measuring the growth and variability Of tax bases over the business cycle. *National Tax Journal,* 49(4):535−552.

Suwanvijit W., 2013. Indonesia−Malasia−Thailand Growth Triangle(IMT−GT) Tourism Forecasts: 2012−2017, Breckenridge, Colorado USA. *The cultural institute international academic conference,*

Botric V., and Vizek M., 2008. The Israeli Tax System: VAT Revenue Forecasting in Israel Israel. Ministry of Finance, State Revenue Administration.

Shumway R.H., and Stoffer D.S, 2006. Time Series Analysis and Its Applications With R Examples: Third edition.USA:Springer.

University of Pretoria., 2013.University of Pretoria:Notes on econometric analysis of cointegration.

Van Heerden Y., and Schoeman N.J., 2010. An empirical dissemination of the personal income tax regime in South Africa using a Microsimulation Tax Model. University of Pretoria Department of Economics,Working paper 25.

Wei W.W.S., 2006.Time series analysis: Univariate and Multivariate methods: Second edition.Pearson Addison Wesley.

Wolswijk G.,2007. Short and long run tax elasticities the case of the Netherlands. European Central Bank, 763.

Van Heerden Y., and Schoeman N.J., 2010. An empirical dissemination of the personal income tax regime in South Africa using a Microsimulation Tax Model. University of Pretoria Department of Economics,Working paper 25.

Yurekli K., Kurunc A., and Ozturk F., 2005. Testing the Residuals of an ARIMA Model on the Cekerk Stream Watershed in Turkey. Turkey. *Turkish Journal,* 29:61−74.