KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

KUMASI

MODELLING DIRECT TAX REVENUE COLLECTION IN GHANA. A CASE STUDY OF INTERNAL REVENUE SERVICE, SEFWI WIAWSO TAX DISTRICT.

A thesis submitted to the Department of Mathematics of the College of Science, Kwame Nkrumah University of Science and Technology, Kumasi in partial fulfillment of the requirement for the Degree of

MASTER OF SCIENCE

(Industrial Mathematics)

Institute of Distance Learning (IDL)

BY

GEORGE MANTEY ASAMOAH

(PG3009409)

JULY 2011

DECLARATION

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere, except where due acknowledgement has been made in the text.

GEORGE MANTEY ASAMOAH	……	……
Candidate's Name	Signature	Date

Certified by:

MR. EMMANUEL HARRIS	……	……
Supervisor's Name	Signature	Date

Certified by:

MR. K. F. DARKWAH	……	……
Head of Department's Name	Signature	Date

Certified by:

PROF. I. K. DONTWI	……	……
Dean of Institute of Dist. Learning's Name	Signature	Date

ABSTRACT

Tax revenue generation in any country constitutes an integral component of any government fiscal policy and many countries therefore, depend mainly on taxation as a means of generating the required resources for funding government expenditure. The provision of social programs, infrastructural developments, maintenance of law and order, defense against external aggression and payment of civil and public servants are usually funded by government through tax revenue. Tax revenue analysis is very important as it enables tax authorities and government to forecast or make future tax revenue projections.

In this study, time series analysis is used with greater emphasis on Box-Jenkins model theory approach to analysis the direct tax revenue collection data from Internal Revenue Service, Sefwi Wiawso tax district. Autoregressive Integrated Moving Average (*ARIMA*(2,1,0)) is used to model the data.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Mr. E. Harris, lecturer at the Mathematics Department, Kwame Nkrumah University of Science and Technology, Kumasi for supervising this work.

Special thanks go to Mr. E. L. Amankwata, District Manager, Ghana Revenue Authority (Formerly Internal Revenue Service), Sefwi Wiawso for his encouragement.

Finally, I am very grateful to my wife and daughter for their encouragement and support.

DEDICATION

This work is dedicated to my wife, Belinda and daughter, Maame Serwaah.

TABLE OF CONTENTS

LIST OF FIGURES

PAGE

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Tax revenue generation in any country constitutes an integral component of any government fiscal policy and many countries therefore, depend mainly on taxation as a means of generating the required resources for funding government expenditure. The provision of social programs and infrastructural developments such as good roads, schools, portable water, electricity, health and sporting facilities are usually funded by government through tax revenue. It is also used for maintenance of law and order as well as defense against external aggression by means of resourcing the security and enforcement agencies with the requisite modern security apparatus. In Ghana for instance, payment of civil and public servants as well as provision of social welfare services such as National Youth Employment Programme and National Health Insurance Scheme are also funded through tax revenue.

Taxation as a means of generating the required revenue for developmental projects and provision of social services can be considered as an important economic policy tool which has several benefits but can however be a disincentive if not properly administered.

Taxation is often defined as the levying of compulsory contributions by public authorities having tax jurisdiction, to defray the cost of their activities. It is also a means by which government implements decisions to transfer resources from the private to the public sector which serves as a major instrument of social and economic policy (Ali-Nakyea, 2008, p. 4). Taxation also plays an essential role in the redistribution of wealth as those in higher income brackets are encouraged to pay more compared to those with lower incomes,

and allows government to regulate certain aspects of the country's economy by discouraging certain business activities such as the sale of alcoholic beverages, tobacco and other foreign goods that compete unfairly with local products in order to safeguard the indigenous industries.

There are basically two main categories of tax namely Direct and Indirect tax. The distinction originates from an administrative point of view and could sometimes be difficult to distinguish them. Indirect tax is a tax which is levied on one person or organisation in the expectation that the tax will be shifted or passed on to another (Ali-Nakyea, 2008, p. 11). Examples are excise duty, custom duty and value added tax. The administering authorities in the country for the period under study were Customs, Excise and Preventive Service and the Value Added Tax Service.

On the other hand, direct tax is intended to be paid by a person or organisation on whom/which it is actually levied, the impact (initial burden of payment to tax authorities) and incidence (final or ultimate economic burden of the tax) being on the same person or organisation (Ali-Nakyea, 2008, p. 10). In this case, the person or organisation that pays the tax bears the final economic burden of the tax and cannot transfer or pass it on to another person or organisation. The Internal Revenue Service was the administering authority of direct tax in the country for the period under study. This work however focuses only on tax revenue collection from direct taxes in Ghana, a case study of Sefwi Wiawso Tax District and therefore restricts the study as such.

Direct tax has the advantage of being able to know the number of persons or organisations that are suppose to pay the tax usually referred to as Taxpayer(s). At the beginning of each calendar year (usually within the first quarter), each taxpayer is given

what we called provisional Assessment. The provisional assessment is like a letter that states categorically the amount of tax to be paid (usually in four equal installments) in a calendar year and the taxpayer has the right under the tax laws to object to the assessment by stating clearly in a letter to the tax authorities why the tax burden is too much or stating precisely the grounds for the objection within nine (9) months of the commencement of the basis period to which the provisional assessment relates in each calendar year. On the basis of the reasons stated in the objection letter by the taxpayer, the tax authorities will then decide among other things to reduce, maintain or increase the tax. This system assists the tax authorities to have a rough estimate of amount of tax to be paid by each taxpayer however, most of the taxpayers do not pay their taxes within the stipulated period and are always in arrears.

Direct tax revenue has been growing steadily since the introduction of this tax in the country in 1943. In 2007, direct tax proceeds (excluding mineral royalties) accounted for 29.9% of the total tax revenue in Ghana (Tax Justice Country Report Series, Ghana Report, 2009). It is a generally held principle that direct tax is directly proportional to population growth, meaning as the working population of a country grows, wealth automatically increases thereby expanding the taxable capacity of a country and enables a larger share of the private sector's resources to be transferred to the government in the form of taxes. Another characteristic feature is that it is progressive, indicating those with higher incomes pay high taxes as a share of income than those with lower incomes.

A typical demerit of direct tax is its high administration cost in terms of stationery, convenience (getting closer to taxpayers – building more offices), transportation (field collection exercise) and so on. Direct tax administration also has the potential of disturbing

the survival of new business enterprises and companies, as well as reducing savings if not properly managed.

Revenue that accrues from direct tax consists of Income Tax, Stamp Duty, Gift Tax, Capital Gains Tax and other special taxes such as National Fiscal Stabilisation Levy and Airport tax.

Income tax consists of Corporate tax, Personal income tax (Self Employed tax) and Pay-As-You-Earn (PAYE). Corporate tax is a tax paid by companies on their profits in the year. The rate has been reduced to 25% since 2006. Personal income tax which is also called Self Employed tax requires persons or individuals to pay income tax at graduated rates in four equal installments. The Pay-As-You-Earn (PAYE) contributions are tax withholdings from salaries of employees in order to satisfy their income tax responsibilities (www.irs.gov.gh/pages/downloads).

Stamp Duty is administered under the Stamp Duty Act, 2005 (Act 689) as amended by Act 764 of 2008. The stamp duty is not a tax on transactions but on documents brought into being for the purposes of recording transactions. It is therefore a tax on documents or specific instruments which have legal effect. Duty is imposed on the wide range of instruments listed in the first schedule to the law. For example, there are stamp duty implications whenever any interest in landed property (or buildings) is transferred, a trustee is appointed, a power of attorney is given, property is mortgage and so on.

Documents not listed in the first schedule are not dutiable, for example a Will (www.irs.gov.gh/pages/downloads).

Gift tax is also a tax payable by a recipient on the total value of taxable gifts received in a year of assessment. The total value of taxable gift(s) must exceed GH¢50.00

in a year of assessment. Assets on which tax is imposed include Land, Building, Money including Foreign Currency, Shares, Bonds and Securities, Business and Business Assets. The rate of tax is 5% (www.irs.gov.gh/pages/downloads).

Capital Gains tax is a tax paid on the gains made from the realisation/ sale of a chargeable asset where gain exceeds GH¢50.00. Assets on which tax is imposed include Land, Building, Business Assets including Goodwill and Shares of a resident company. The rate of tax is 15% (www.irs.gov.gh/pages/downloads).

There are several ways of collecting income taxes and one special way of collecting these taxes is through Withholding Taxes method. The law requires a person or organisation effecting payment to another person or organisation to deduct the exact tax at source and subsequently pay the withheld tax to the tax authorities within fifteen (15) days after the end of the month in which the tax was withheld. A withholding agent is a person or organisation that withholds the tax at source and failure to withhold the tax means the withholding agent is personally liable to pay the tax that should have been withheld. There are different types of withholding taxes with different rates (www.irs.gov.gh/pages/downloads).

Vehicle Income Tax (VIT) is another way of collecting taxes from commercial vehicle owners on quarterly basis. Tax stamp is also a tax collected from small scale self employed persons in the informal sector on quarterly basis. The small scale business activities in the informal sector include dressmakers, susu collectors, chop bar operators, butchers and so on (www.irs.gov.gh/pages/downloads).

This work however contributes to the statistical understanding of the dynamics of monthly direct tax revenue collection in Ghana, a case study of Sefwi Wiawso tax district over a 10-year period (120 months) from 2000 to 2009.

Sefwi Wiawso tax district operates within five (5) political district assemblies namely Sefwi Wiawso District Assembly, Bibiani-Anhwiaso-Bekwai District Assembly, Juaboso District Assembly, Akontombra District Assembly and Bia District Assembly. The tax district has its main office at Sefwi Wiawso, a sub-office at Bibiani and a collection point at Oseikojokrom near the Ivory Coast – Ghana boarder in the Bia District. Tax revenue from the main office largely comes through PAYE and withholding taxes from district assemblies, health insurance schemes, rural banks, clinics or hospitals, timber companies and few construction firms. The sub office also depends mainly on PAYE and withholding taxes from mining companies particularly from Chirano Gold Mines Limited and its sister company Red Back Mining (Ghana) Limited. Oseikojokrom collection point performs quite well in the collection of withholding taxes particularly from Bia District Assembly and Bia District Mutual Health Insurance Scheme.

1.1.1 History Of Taxation In Ghana

Taxation was first introduced in Ghana, then the Gold Coast, in 1943 by the British Colonial Government at the time when World War II was raging. It should be noted, however, that, before the introduction of income tax in 1943, several attempts had already been made, for example as far back as April 1852, the Poll Tax Ordinance was passed to raise money to finance the increased cost of British Administration (Ali-Nakyea, 2008, p. 3).

Under the Ordinance, every man, woman or child residing in districts under British protection was to pay one shilling per head per year. These early experiments of the introduction of direct taxation failed because of weaknesses inherent in the system of collection and the fact that the first proceeds were mainly used to pay the increased salaries of British Officials and not for the construction of roads and schools (Ali-Nakyea, 2008, p. 3).

The first Income tax Law was thus the Income Tax Ordinance (No. 27), 1943. This Ordinance was modelled to a large extent on the general principles underlying the Income Tax Act then in force in the United Kingdom. It imposed the tax generally on incomes having their sources in Ghana so that foreign source income was not liable unless it was remitted in Ghana. One characteristic feature of this Ordinance was the numerous personal reliefs and deductions that it contained (Ali-Nakyea, 2008, p. 3).

Over the years the Income Tax Law has seen several changes through amendments, and modifications, such as the Income Tax (Amendment) Ordinance 1952. The first consolidated edition of the Income Tax Ordinance was published in March, 1953. The following Acts then introduced amendments to the consolidated edition Act 68 in 1961, followed by Acts 178 and 197 in 1963 and Act 312 in 1965. The second consolidated edition was published in September, 1966, that is, the Income Tax Decree, 1966 (No. 78). The Income Tax Decree 1975, SMCD 5, which was published in December, 1975, was the third consolidated edition. The current Income Tax Law is the Internal Revenue Act, 2000 (Act 592) and is the fourth consolidated edition (Ali-Nakyea, 2008, pp. 3 - 4).

On 31st December, 2009 an assent was given to the Ghana Revenue Authority Act, 2009 (Act 791). The Act was passed to establish the Ghana Revenue Authority to replace

the Internal Revenue Service, the Customs, Excise and Preventive Service and the Value Added Tax Service for the administration of taxes and to provide for related purposes in the country. Under this Act, the Internal Revenue Service and the Value Added Tax Service have been merged to become Domestic Tax Revenue Division whereas Customs, Excise and Preventive Service has also been renamed as Customs Division of the Ghana Revenue Authority (Ghana Revenue Authority Act, 2009, Act 791).

1.2 PROBLEM STATEMENT

The Statistics Units of all the fifty four tax districts of the Internal Revenue Service are required to prepare monthly baseline revenue projection(s) according to tax types before the end of each calendar year to enable the statistics unit at the head office have a consolidated baseline projection(s) for onward submission to the Ministry of Finance and Economic Planning to be factored in the government budget. At the beginning of each year, the statistics units of all the tax districts are again required to come out with their final monthly estimate(s) according to tax types.

Over the years the statistics units have relied on economic indicators such as inflation, growth and others in the preparation of the estimate(s) or use the crude method by taking the average tax revenue growth for the past two or three years and just increasing the previous year's actual collection by that percentage. It should be noted that these methods usually result in wild or very high deviation(s) from the actual collection.

However, having identified the problem, the question now is what appropriate statistical tool can produce estimate(s) that can give small variance. This study however considers time series analysis with greater emphasis on Box-Jenkins model(s) as an

alternative statistical tool to generate the baseline projection(s) that can be adjusted if there is any government intervention such as change in tax rates and so on, to arrive at the estimate(s) for each calendar year.

1.3 OBJECTIVES OF THE STUDY

The objectives of the study are:

- To extensively describe and understand the dynamics of the Internal Revenue Service (IRS), Sefwi Wiawso tax district monthly direct tax revenue collection data over a 10-year period (120 months) from January, 2000 to December, 2009.
- To fit a suitable model to the monthly direct tax revenue data and compute forecasts for the next twelve (12) months.

1.4 METHODOLOGY

Time series analysis is the main statistical tool that will be employed in the analyses of the monthly direct tax revenue data from Sefwi Wiawso tax district over the 10-year period (120 months) from 2000 to 2009. Graphical descriptions of the data will also be used for easy understanding and interpretation at the various stages of the analysis.

The Box-Jenkins time series model theory will be considered in the modelling of the data. Models to be considered in the formulation include Autoregressive process of order p (abbreviated to an AR(p) process), Moving Average process of order q (MA(q)), mixed Autoregressive/Moving-Average process (ARMA(p,q)) and Autoregressive Integrated Moving-Average of order p,d,q (ARIMA(p,d,q)) depending on the nature of the

sample autocorrelation function, sample partial autocorrelation function and other theoretical indicators to select the appropriate model. Some useful diagnostic checks will also be performed on the selected models or equations to identify the most appropriate one which best fit the direct tax revenue data.

R software package will be used to sketch the graphs, estimate the model parameters, perform the diagnostic checks and also compute the forecasts.

1.4.1 DATA COLLECTION

The data used throughout the study are secondary data and consist of monthly total direct tax revenue collection. These data values were collected from the Statistics Unit of Internal Revenue Service, Sefwi Wiawso Tax District Office. The data covered the period from January 2000 to December 2009 (120 months). Direct tax system in Ghana comprises Corporate tax, Self Employed tax, Pay-As-You-Earn (PAYE), Stamp Duty, Gift tax, Capital Gains tax and other special taxes such as National Fiscal Stabilisation Levy and Airport tax. However, the monthly direct tax revenue from Sefwi Wiawso tax district over the period under study consisted mainly of Corporate tax, Self Employed tax and PAYE with only tiny shares accruing from Stamp Duty, Gift tax and Capital Gains tax.

Some of the direct tax revenue data were recorded in the old currency (¢), that is from January, 2000 to June, 2007. The data values recorded in the old currency (¢) were converted to the new currency (GH¢). The conversion from the old currency (¢) to the new currency (GH¢) has made the unit (currency) of measurement uniform and it would also help reduce errors in differencing.

1.5 JUSTIFICATION OF THE STUDY

Tax revenue models comprise mainly of micro-simulation and econometric types because of the restrictions on tax revenue data. However, this work contributes to the statistical understanding of analysing direct tax revenue collection in Ghana, a case study of Sefwi Wiawso tax district over the period spanning 120 months (January, 2000 to December, 2009) by fitting a suitable model to the revenue data and computing forecast(s). The data are analysed using time series analysis with greater emphasis on Box-Jenkins model theory approach. This study will serve as a guide that would enable the Statisticians to have a fair idea on using time series analysis to build monthly estimate(s) that can be relatively close to monthly actual tax collection(s) with the best minimum deviation(s).

1.6 SCOPE AND LIMITATIONS OF THE STUDY

The study aims at analysing monthly direct tax revenue collection in Ghana, a case study of Sefwi Wiawso tax district for the period from 2000 to 2009 (120 months). The main idea is to analyse the various tax types such as Pay-As-You-Earn (PAYE), Corporate tax, Self Employed tax and the other taxes that constitute direct tax separately in order to increase precision (reduce the error margin in using one model for all the tax types) but restrictions on acquisition of the tax data have limited the work to the analyses of the monthly total collections without knowing the actual contribution or share of the various tax types.

1.7 ORGANISATION OF THE STUDY

The study is organised in five main chapters. The first chapter covers introduction that highlights background of the study, problem statement, objectives of the study, methodology, justification of the study, scope and limitations of the study and organisation of the study. Chapter two deals with the review of relevant literature of the study. The review focuses on the approaches that have been adopted by previous researchers and limitations of their methods, as well as a discussion of the results from previous studies.

The third chapter discusses extensively the mathematical or statistical methods and procedures used in the analysis of the monthly direct tax revenue data. The fourth chapter also deals with the analysis of the monthly direct tax revenue data over the 10-year period (120 months) from 2000 to 2009. The interpretations and discussions are also presented in this chapter. The last chapter covers conclusion and recommendations.

CHAPTER 2

LITERATURE REVIEW

This chapter covers the review of relevant literature which is aimed at finding out what has already been done, written down or printed on applications of time series analysis of tax revenue data. There is limited literature on applications of time series in modelling tax revenue data and even in studies where tax revenue models have been developed, they usually comprise mainly of econometric models because of the restrictions on tax revenue data. However, this review focuses on the various approaches that have been adopted by previous researchers and limitations of their methods, as well as a discussion of the results from previous studies.

In a research by Nikolov (2002) on "Tax Revenues Forecasting with Intervention Time Series Modeling" using monthly tax revenue collection from Republic of Macedonia for the period January, 1998 to July, 2002; the modelling started with the Box-Jenkins model selection approach and continued with the effects of the intervention analyses. In his conclusion, he recommended that the model could be used for forecasting but only for a few time units ahead because the variance of the forecasts in time series models becomes large in time.

In comparing time series Box-Jenkins approach and other time series methods, Clower and Weinstein (2006) studied quarterly sales tax revenue from 1994 to 2005 in the city of Arlington, Texas. In predicting future sales tax growth, three main statistical approaches were employed; Autoregressive Integrated Moving Average (ARIMA) model, Seasonal Exponential Smoothing and Ordinary Least Squares (OLS) Regression Analysis based on employment data. "Though not wildly different, each forecasting technique

produced slightly different outcomes". The ARIMA model was considered as suggesting slower growth in total revenues and its forecasts were judged to be at a lower bound for potential revenue growth compared to the other methods.

Fullerton (1989) also conducted a study on Forecasting State Government Revenues: A case study of Idaho Sales tax for the period 1967 to 1985. In the study, econometric model, ARIMA model as well as Composite model (combination of econometric and ARIMA models) were employed and found that a composite model built with econometric and univariate ARIMA projections of Idaho retail sales tax receipts provided better forecasts than either single model because the combined forecast variance was generally smaller than that of any of the single forecast methods considered.

In the study of Slobodnitsky and Drucker (2005) on "VAT Revenue Forecasting in Israel" using monthly VAT revenues since 1987; this study compared the results of ARIMA and Cointegration estimation methods and decided over the one best suited for VAT revenue forecasting. It was found out that the quarterly ARIMA specification performed the best with the least absolute deviation. "Monthly ARIMA came a close second and both budget forecast and cointegration were much less precise".

Braun (1988) conducted a research on "Measuring Tax Revenue Stability with Implications for Stabilization Policy: A Note". The main purpose of the study was to investigate whether aggregate state tax revenue data are characterised by a deterministic trend or a stochastic drift model using tax revenue data from the state of Georgia for the years 1950 to 1984, except for the general sales tax whose observations range from 1952 to 1984. The research revealed that forecasting tax revenue data using stochastic model was more robust and provided less errors compared to the deterministic trend model which

provided a poor measure or forecast because it always assumed the tax data (observations) to be in the state of stationarity.

CHAPTER 3

METHODOLOGY

The review of statistical or mathematical methods employed in the analysis of the direct tax revenue data is presented in this chapter. However, before we discuss the methodology we shall first look at some basic definitions.

3.1 DEFINITIONS

The following are some important definitions relating to time series:

3.1.1 Time Series

Granger and Newbold (1986) describe a time series as ‘‘. . . a sequence of observations ordered by a time parameter.’’

3.1.2 Component of Time Series

All time series contain at least one of the following components:

- Trend – the long-term tendency of a series to rise or fall (upward or downward movements) over a period of time.
- Seasonal – the periodic fluctuations in a time series within a certain time period. These fluctuations form a pattern that tends to repeat from one seasonal period to another.
- Cyclical – long departures from the trend due to factors other than seasonality. Cycles generally occur over a long time interval and the lengths of time between successive peaks or troughs of a cycle are not necessarily the same.

- Irregular – the movement or component left after accounting for trend, seasonal and cyclical movements; random noise or error in a time series. (Kirchgässner and Wolters, 2007, p. 3).

3.1.3 Stochastic Processes

A model that describes the probability structure of a sequence of observations is called a Stochastic Process (Box et al., 1994, p. 19).

3.1.4 Stationary Processes

A stochastic process $\{X_t\}$ is said to be strictly stationary if the joint distribution of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ is the same as the joint distribution of $X_{t_1-k}, X_{t_2-k}, \ldots, X_{t_n-k}$ for all $t_1, t_2, \ldots, t_n$ and lag k which is the time difference. In this case, $E(X_t) = E(X_{t-k})$, $Var(X_t) = Var(X_{t-k})$ for all t and k, and also $Cov(X_t X_s) = Cov(X_{t-k}, X_{s-k})$ for all t, s and k (Cryer and Chan, 2008, p. 16).

In practice, it is rare to come across strict stationary process and therefore the need for a less restricted definition that is practicable. A process $\{X_t\}$ is called Second-order stationary (weakly stationary) if its mean function is constant over time and its autocovariance function depends only on the lag, so that $E(X_t) = \mu$ and $\gamma_{t,t-k} = \gamma_{0,k}$ (Chatfield, 2004, p. 36).

3.1.5 White Noise Processes

A process $\{a_t\}$ is called a White Noise if it is a sequence of uncorrelated random variables from a fixed distribution with constant mean $E(a_t) = \mu_a$, usually assumed to be zero, constant variance $Var(a_t) = \sigma_a^2$ and $\gamma_k = Cov(a_t, a_{t+k}) = 0$ for all $k \neq 0$ (Wei, 2006, p. 15).

3.1.6 Differencing

Differencing simply means relating the present or current value to its previous values. Consider for example a series $\{y_2, \ldots, y_n\}$ which is formed from the original observed series, $\{x_2, \ldots, x_n\}$. First-order differencing is given as

$$y_t = x_t - x_{t-1} = \nabla x_t \quad \text{for } t = 2,3,\ldots,n \tag{3.1.1}$$

Second-order differencing is also expressed as

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1} = x_t - 2x_{t-1} + x_{t-2} \tag{3.1.2}$$

At most, second-order differencing is sufficient to attain stationarity (Chatfield, 2004, p. 19).

3.1.7 Sample Autocorrelation Function

Consider an observed series $x_1, x_2, \ldots, x_N$. We can form $N-1$ pairs of observations, namely $(x_1, x_2), (x_2, x_3) \ldots, (x_{N-1}, x_N)$, where each pair of observation is separated by one time interval. Taking the observations in each pair as separate variables, we can compute the correlation coefficient between x_t and x_{t+1} as

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \tag{3.1.3}$$

which is autocorrelation coefficient at lag k. A plot of r_k against the lag k for $k = 0,1,2,\ldots, M < N$ is called Correlogram or sample Autocorrelation Function (ACF.) (Chatfield, 2004, pp. 22 - 23).

3.1.8 Sample Partial Autocorrelation Function

The partial autocorrelation is the correlation between x_t and x_{t-k} after removing the effect of the intervening variables $x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}$. It is usually called the Partial Autocorrelation (PACF) at lag k and denoted by ϕ_{kk} and is defined as

$$\phi_{kk} = Corr(x_t, x_{t-k} | x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}) \tag{3.1.4}$$

That is, ϕ_{kk} measures the correlation between x_t and x_{t-k} given $x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}$ (or the correlation between x_t and x_{t-k} after adjusting for the effects of $x_{t-1}, x_{t-2}, \ldots, x_{t-k+1}$ (Box et al., 1994, pp. 64 -67; Kirchgässner and Wolters, 2007, pp. 52-54; Cryer and Chan, 2008, p. 115).

3.1.9 Moving Average Processes

Consider a white noise process $\{a_t\}$ with zero-mean and variance σ_a^2, then a process $\{X_t\}$ is called a Moving Average process of order q (abbreviated to $MA(q)$ process) if

$$X_t = a_t - \beta_1 a_{t-1} - \beta_2 a_{t-2} - \cdots - \beta_q a_{t-q} \tag{3.1.5}$$

where the β_i 's are constants. The mean, $E(X_t) = 0$ and $Var(X_t) = \sigma_a^2(1 + \sum_{i=1}^{q} \beta_i^2) = \gamma_0$.

$$\gamma_k = Cov(X_t X_{t+k}) = \begin{cases} \sigma_a^2(-\beta_k + \sum_{i=1}^{q-k} \beta_i \beta_{i+k}), & k = 1,2,\ldots,q \\ 0, & k > q \end{cases} \quad (3.1.6)$$

and

$$\rho_k = \begin{cases} \dfrac{-\beta_k + \beta_1\beta_{k+1} + \beta_2\beta_{k+2} + \cdots + \beta_{q-k}\beta_q}{1 + \beta_1^2 + \beta_2^2 + \cdots + \beta_q^2}, & k = 1,2,\ldots,q \\ 0, & k > q \end{cases} \quad (3.1.7)$$

(Box et al., 1994; Chatfield, 2004; Cryer and Chan, 2008).

3.1.10 Autoregressive Processes

Consider a white noise process $\{a_t\}$ with zero-mean and variance σ_a^2, then a process $\{X_t\}$ is said to be an Autoregressive process of order p (abbreviated to an $AR(p)$ process) if

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + a_t \quad (3.1.8)$$

This model resembles a multiple regression model but the dependent variable X_t is regressed on past values of X_t rather than separate independent or predictor variables. We can however derive that

$$\gamma_k = \alpha_1\gamma_{k-1} + \alpha_2\gamma_{k-2} + \cdots + \alpha_p\gamma_{k-p} \quad (3.1.9)$$

and

$$\rho_k = \alpha_1\rho_{k-1} + \alpha_2\rho_{k-2} + \alpha_3\rho_{k-3} + \cdots + \alpha_p\rho_{k-p} \quad \text{for } k \geq 1 \quad (3.1.10)$$

Thus,

$$\gamma_0 = \alpha_1\gamma_1 + \alpha_2\gamma_2 + \cdots + \alpha_p\gamma_p + \sigma_a^2 \quad (3.1.11)$$

but we know that $\rho_k = \gamma_k / \gamma_0$, hence the variance may be written as

$$\gamma_0 = \frac{\sigma_a^2}{1 - \alpha_1\rho_1 - \alpha_2\rho_2 - \cdots - \alpha_p\rho_p} \tag{3.1.12}$$

(Chatfield, 2004; Cryer and Chan, 2008).

3.1.11 Mixed Autoregressive Moving Average Processes (ARMA)

This model is a combination of *AR* and *MA* processes. That is, partly autoregressive and partly moving average and it is represented as

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + a_t - \beta_1 a_{t-1} - \beta_2 a_{t-2} - \cdots - \beta_q a_{t-q} \tag{3.1.13}$$

where $\{X_t\}$ is a Mixed Autoregressive Moving Average process of orders p and q (abbreviated to $ARMA(p,q)$) (Chatfield, 2004, pp. 46-48; Cryer and Chan, 2008, pp. 77-79).

3.1.12 Integrated ARMA (or ARIMA) Models

Suppose a series $\{X_t\}$ is nonstationary. If the *dth* difference of the series $W_t = \nabla^d X_t$ is a stationary $ARMA(p,q)$ process, then $\{X_t\}$ is said to be an $ARIMA(p,d,q)$ process. For illustration, consider an $ARIMA(p,1,q)$ process. If the first-order difference is $W_t = X_t - X_{t-1}$, then we can express the new (differenced) series as

$$W_t = \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + \cdots + \alpha_p W_{t-p} + a_t - \beta_1 a_{t-1} - \beta_2 a_{t-2} - \cdots - \beta_q a_{t-q} \tag{3.1.14}$$

or, in the original form,

$$X_t - X_{t-1} = \alpha_1(X_{t-1} - X_{t-2}) + \alpha_2(X_{t-2} - X_{t-3}) + \cdots + \alpha_p(X_{t-p} - X_{t-p-1}) + a_t - \beta_1 a_{t-1} - \beta_2 a_{t-2} - \cdots - \beta_q a_{t-q}$$

Simplifying, we have

$$X_t = (1+\alpha_1)X_{t-1} + (\alpha_2 - \alpha_1)X_{t-2} + (\alpha_3 - \alpha_2)X_{t-3} + \cdots + (\alpha_p - \alpha_{p-1})X_{t-n} - \alpha_p X_{t-p-1} + a_t - \beta_1 a_{t-1} - \beta_2 a_{t-2} - \cdots - \beta_q a_{t-q} \quad (3.1.15)$$

Equation (3.1.15) is called the Difference Equation Form which also resembles nonstationary $ARMA(p+1, q)$ process (Cryer and Chan, 2008, p. 92).

3.1.13 The Method of Moments Estimation

The method consists of equating sample moments to corresponding theoretical moments of random variables and solving the resulting equations to obtain estimates of any unknown parameters (Cryer and Chan, 2008, p. 149).

3.1.14 The Least Squares Estimation

The principle of the method is to find estimates $\hat{\alpha}_p$, $p = 1,2,\ldots$ called least squares estimators for the parameters α_p, $p = 1,2,\ldots$ such that the error sum of squares, example, $\sum_{t=2}^{n}[(X_t - \mu) - \alpha(X_{t-1} - \mu)]^2$ is minimum ((Cryer and Chan, 2008, p. 154).

3.1.15 The Maximum Likelihood Estimation

For any set of observations, $X_1, X_2, X_3, \ldots, X_n$, time series or not, the likelihood function L is defined to be the joint probability density of obtaining the data actually observed. However, it is considered as a function of the unknown parameters in the model

with the observed data held fixed. The likelihood function is for example given as $L(x;\alpha) = \prod_{i=1}^{p} f(x;\alpha)$ (Cryer and Chan, 2008, p. 158).

3.1.16 Akaike's Information Criterion (AIC)

$$AIC = \ln \hat{\sigma}_k^2 + \frac{n+2k}{n} \tag{3.1.16}$$

where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$ and k is the number of parameters in the model and n is the sample size.

The value of k yielding the minimum AIC specifies the best model (Shumway and Stoffer, 2006, p. 53).

3.1.17 AIC, Bias Corrected (AICc)

$$AIC_c = \ln \hat{\sigma}_k^2 + \frac{n+k}{n-k-2} \tag{3.1.17}$$

where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$ and k is the number of parameters in the model and n is the sample size (Shumway and Stoffer, 2006, p. 54).

3.1.18 Schwarz's Information Criterion (SIC) or Bayesian Information Criterion (BIC)

$$SIC \quad or\ BIC = \ln \hat{\sigma}_k^2 + \frac{k \ln n}{n} \tag{3.1.18}$$

where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$ and k is the number of parameters in the model and n is the sample size (Shumway and Stoffer, 2006, p. 54).

3.2 METHODOLOGY

Time series analysis is the main statistical tool employed in the analyses of the direct tax revenue collection data. The data used for the analysis are secondary data and consist of monthly direct tax revenue collection. These data were collected from the Statistics Unit of Internal Revenue Service (IRS), Sefwi Wiawso tax district office. The data covered the period from January, 2000 to December, 2009 (120 months). The software used for the analysis is R.

In the first place, the direct tax revenue data values were arranged vertically from January, 2000 to December, 2009 in an Excel worksheet and then imported into R software using the R-Commander. As a requirement in R, the data were converted into time series data values using the R command 'dataset=ts(data,start=2000,frequency=12)'.

A descriptive analysis of the direct tax revenue data was carried out by displaying a graphical representation of the monthly tax revenue series and also computing the summary statistics such as the mean, standard deviation, variance and so on.

The direct tax revenue was plotted against time (months) using the R command 'plot(dataset,xlab="Time(Month)",ylab="Tax Revenue")' in order to identify salient features such as trend, seasonality, outliers, discontinuities and stationarity in the dataset. The time plot depicted casual upward and downward behaviour with a general increasing linear trend which clearly indicated that the direct tax series was non-stationary. In order to confirm this, the sample autocorrelation function of the direct tax revenue data was computed and realised, as expected, that the autocorrelation coefficients at low lags were all 'large' and positive, and did not 'die out' or come down quickly to zero but declined gradually. The R command used for computing the autocorrelation function was

'acf(dataset,36)'. Because the tax revenue series was not stationary, differencing was therefore applied to transform the tax revenue data in order to attain the stationarity assumption, which is a prerequisite in Box-Jenkins modelling technique.

Consequently, a first-order differencing was performed with the R command 'difftax=diff(dataset)' to remove the trend component in the series. However, the number of observations was reduced from 120 to 119 because of the differencing. Using the rule of thumb that values exceeding $\pm 2/\sqrt{N}$ are significantly different from zero, where N is the number of terms in the differenced series (in this case, $N = 119$), the sample autocorrelation function of the differenced tax revenue series (first-order differencing) revealed that the series had actually attained an appreciable stationarity. This means that apart from few autocorrelation coefficients at the low lags (1 or 2) which were 'large' or exceeded the $\pm 2/\sqrt{N}$ limits, the rest of the autocorrelation coefficients were close to zero or not significantly different from zero.

After achieving the stationarity assumption, the next task was to specify the order of the model. The behaviour of the sample autocorrelation function (ACF) and the sample partial autocorrelation function could be used to identify the model and the order that describes the stationary time series data. The order of an *AR* model could be assessed from the behaviour of the sample partial autocorrelation function (PACF). Theoretically, we expect 95% of the values of the partial autocorrelation coefficients, ϕ_{kk} to fall within the limits $\pm 2/\sqrt{N}$ and values outside the range are significantly different from zero. The implication is that the sample partial autocorrelation function (PACF) of an $AR(p)$ model 'cuts off' at lag p so that values beyond p are not significantly different from zero. The

sample partial autocorrelation function (PACF) of the differenced tax revenue series was computed using the R command 'pacf(difftax,36)'.

However, the order of an $MA(q)$ model is usually clear from the sample autocorrelation function (ACF). The theoretical autocorrelation function of an $MA(q)$ process 'cuts off' at lag q and values beyond q are not significantly different from zero. Similarly, the R command used for computing the autocorrelation function of the differenced tax revenue series was 'acf(difftax,36)'. After specifying the orders for the *AR* and *MA* models using the sample partial autocorrelation function (PACF) and the sample autocorrelation function (ACF) respectively, three models were suggested.

Once we have specified the models, the next task was to estimate the parameters of each of the three models suggested using the stationary series (differenced tax revenue series) as well as to perform their corresponding diagnostic checks. The diagnostics actually helped us to evaluate the goodness-of-fit of the three models initially suggested.

On the basis of the diagnostic checks, the best-fit model was finally selected and subsequently used to make a 12-month forecast into the future of the Internal Revenue Service direct tax revenue collection.

CHAPTER 4

DATA ANALYSIS

In the previous chapter, we discussed the statistical methods employed in the analysis of the data. In this chapter, we are actually concerned with the analysis and the discussion or interpretation of the results of the analysis of the monthly direct tax revenue collection data recorded at Internal Revenue Service (IRS), Sefwi Wiawso tax district over the period from January, 2000 to December, 2009 (120 months), which would also lead us to have conclusive decisions on the data.

4.1 DESCRIPTIVE ANALYSIS OF THE DIRECT TAX REVENUE DATA

In this section, a brief discussion of the summary statistics as well as a time plot of the direct tax revenue collection data from January, 2000 to December, 2009 are presented to enable us identify trend and other significant features of the data.

The dataset has a mean value of 179,075 (GH¢) and a standard deviation value of 137,315 (GH¢) (details of the summary statistics are presented in appendix B). The time plot is given as

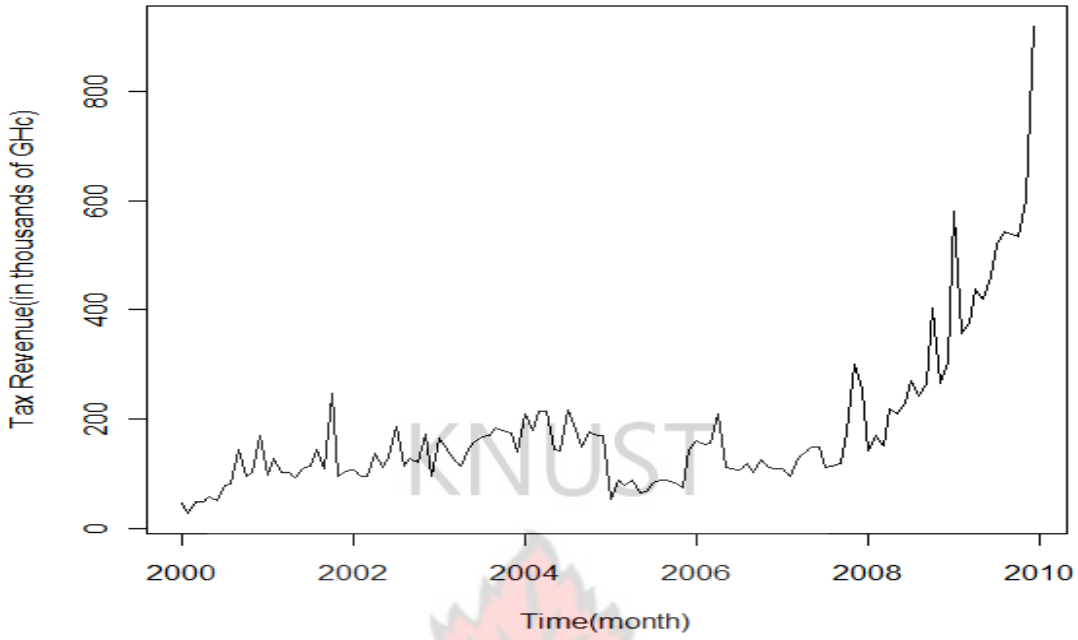


Figure 4.1: Time Plot of IRS, Sefwi Wiawso Tax District Direct Tax Revenue Collection data, with monthly totals in thousands of GH¢.

Figure 4.1 reveals that there is no apparent periodic or seasonal variation but rather irregular changes (upward and downward movements) in the time plot. The irregular changes are somewhat confined within a certain range, that is from the beginning of 2000 to the end of 2008. There was however a sharp increase at the beginning of 2009 but declined afterwards and started increasing again at a gradual rate until the end of 2009 when another sharp increase was recorded. From the time plot we could say that the series is non stationary due to the existence of linear trend component.

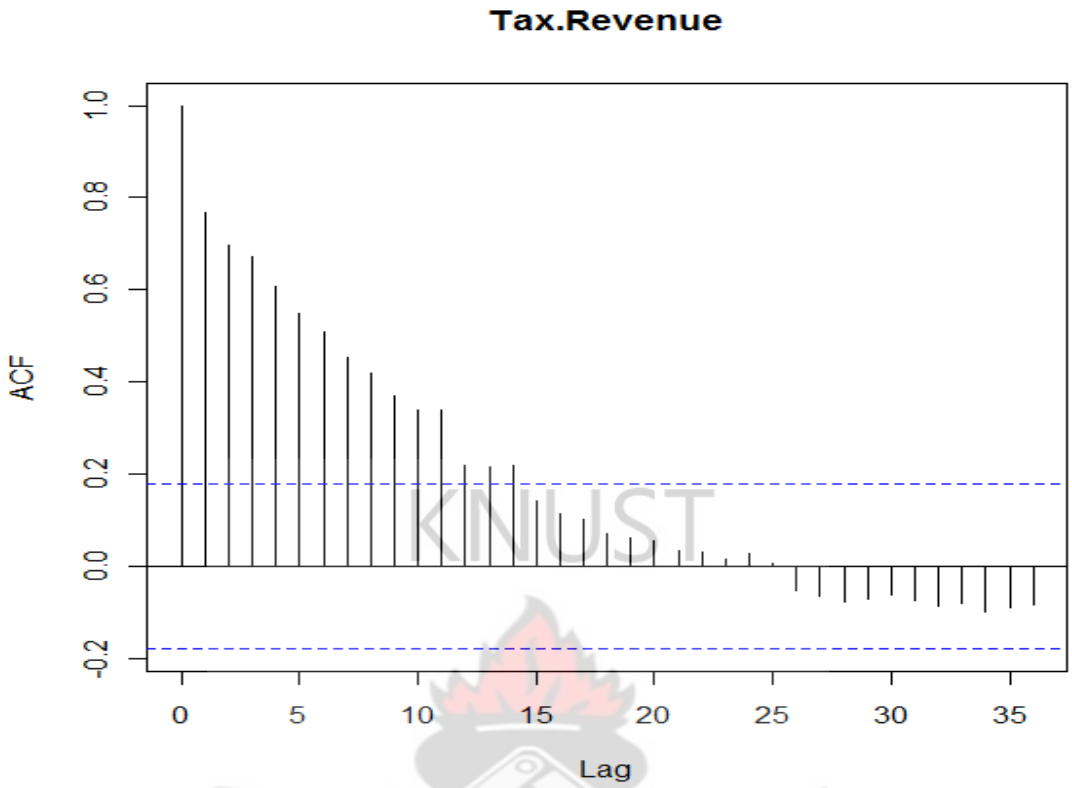


Figure 4.2: Autocorrelation Function of IRS, Sefwi Wiawso Tax District Direct Tax Revenue Collection

The autocorrelation function in figure 4.2 confirms the non stationarity of the direct tax revenue series and also explains the correlation between the values of the direct tax revenue collection data at different points apart in time or as a function of the two times or of the time difference. The sample autocorrelation function is decreasing gradually and does not 'die out' or come down quickly to zero which portrays the survival of trend component in the IRS direct tax revenue collection data.

4.2 TREND DIFFERENCING

In order to remove the trend component in the IRS direct tax revenue collection data, first-order differencing method is performed.

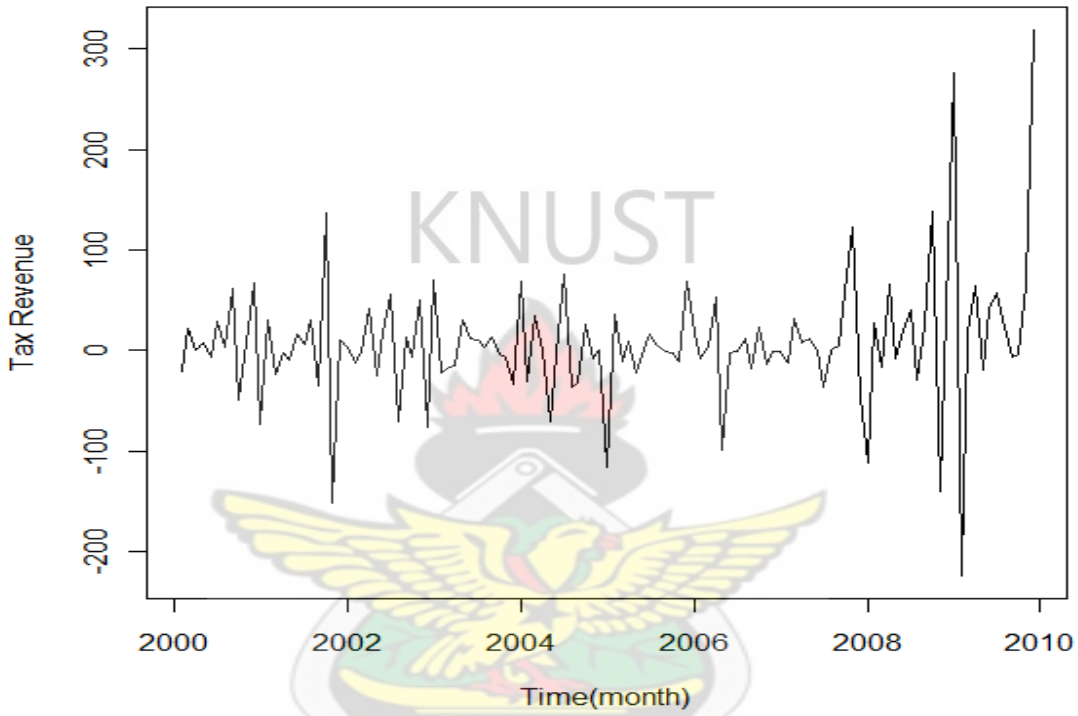


Figure 4.3: First-Order Differencing of IRS Direct Tax Revenue Collection Data

From figure 4.3, we could observe that the observations move irregularly without any obvious trend but revert to its mean value and the variability is also approximately constant. The differenced IRS direct tax revenue collection series now appears to be approximately stable or is said to have attained an appreciable stationarity.

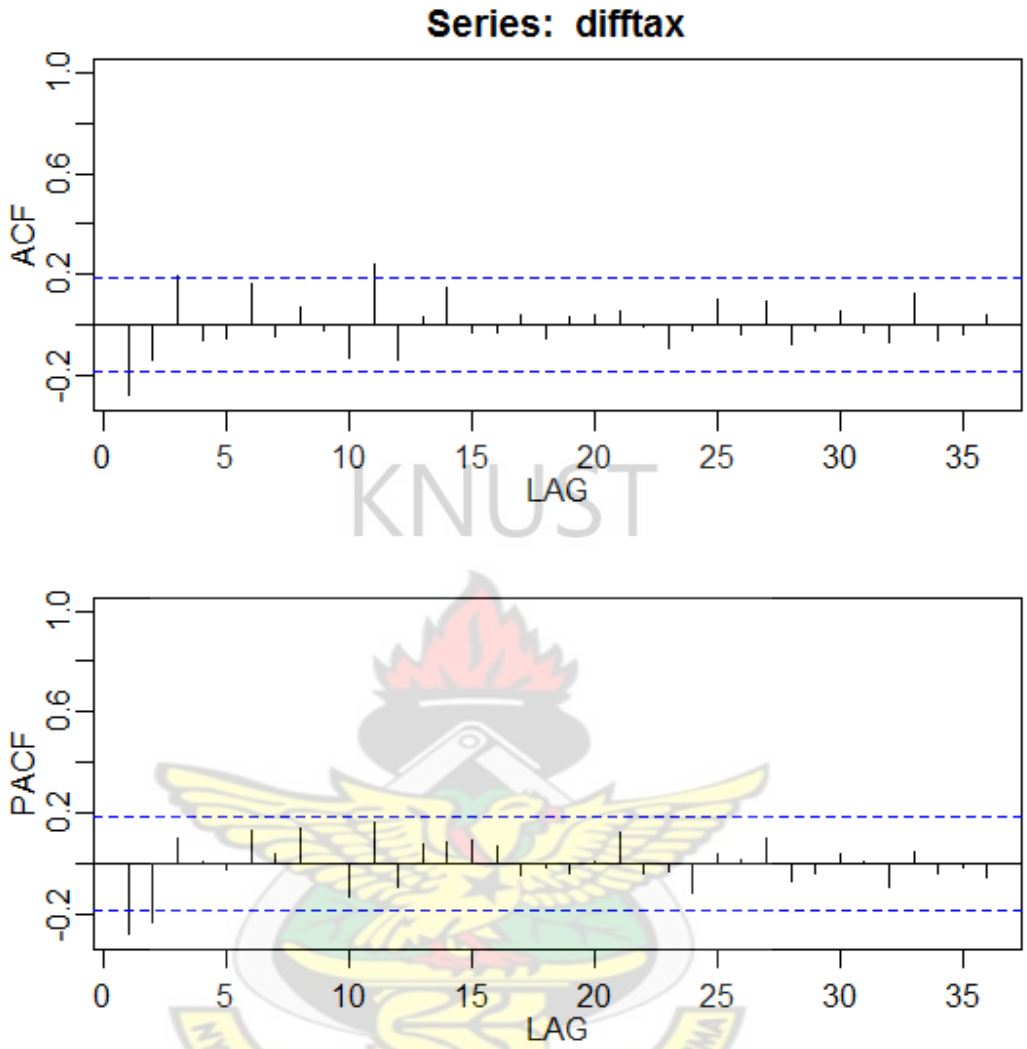


Figure 4.4: ACF and PACF of the First-Order Differencing of IRS Direct Tax Revenue Collection Data

The first diagram of figure 4.4 exhibits the sample autocorrelation function (ACF) of the first-order differencing of the IRS direct tax revenue collection data at different lags and the second diagram is the sample partial autocorrelation function (PACF) of the first-order differencing of the IRS direct tax revenue collection data also at various lags. Values outside the dotted lines are considered significantly different from zero. The sample ACF

produced significant coefficients at lags 1 and 11. The coefficient at lag 11 is somewhat significant but it is ignored because it does not have any physical importance to the study and could only be attributed to few outliers in the data. The sample PACF has large values at lags 1 and 2, then, is essentially zero for higher order lags. This is because the PACF measures the correlation between two variables after eliminating or controlling the effects of any intervening variables; hence, for lags greater than 2, the sample PACF should cut to zero or statistically not different from zero.

The following models are suggested after vigilantly examining the behaviour of both the sample ACF and the sample PACF of the first-order differencing of the IRS direct tax revenue collection data:

- *ARIMA*(2,1,0)
- *ARIMA*(0,1,1)
- *ARIMA*(2,1,1)

The estimated parameters of each of the three models as well as their corresponding diagnostic checks of the residuals and the AIC, AIC_c and BIC are presented in the next section to enable us select the best model for forecasting into the future.

4.3 MODEL SELECTION

In the previous section, three models were specified and we now estimate their respective parameters as well as to perform their diagnostic checks.

4.3.1 Parameter Estimates and Diagnostics of *ARIMA*(2,1,0) Model

Call:

arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),

xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc, REPORT = 1,

reltol = tol))

Coefficients:

	ar1	ar2	xmean
	-0.4941	-0.3669	6101.403
s.e.	0.1011	0.1005	2808.074

sigma^2 estimated as 3.206e+09: log likelihood = -1471.42, aic = 2950.85

$AIC

[1] 22.93873

$AICc

[1] 22.95848

$BIC

[1] 22.00879

The parameters based on the t-value estimate are statistically significant since all the t-values are greater than or equal to 2 in absolute value ($t \geq |2|$) (detailed calculations are presented in appendix D).

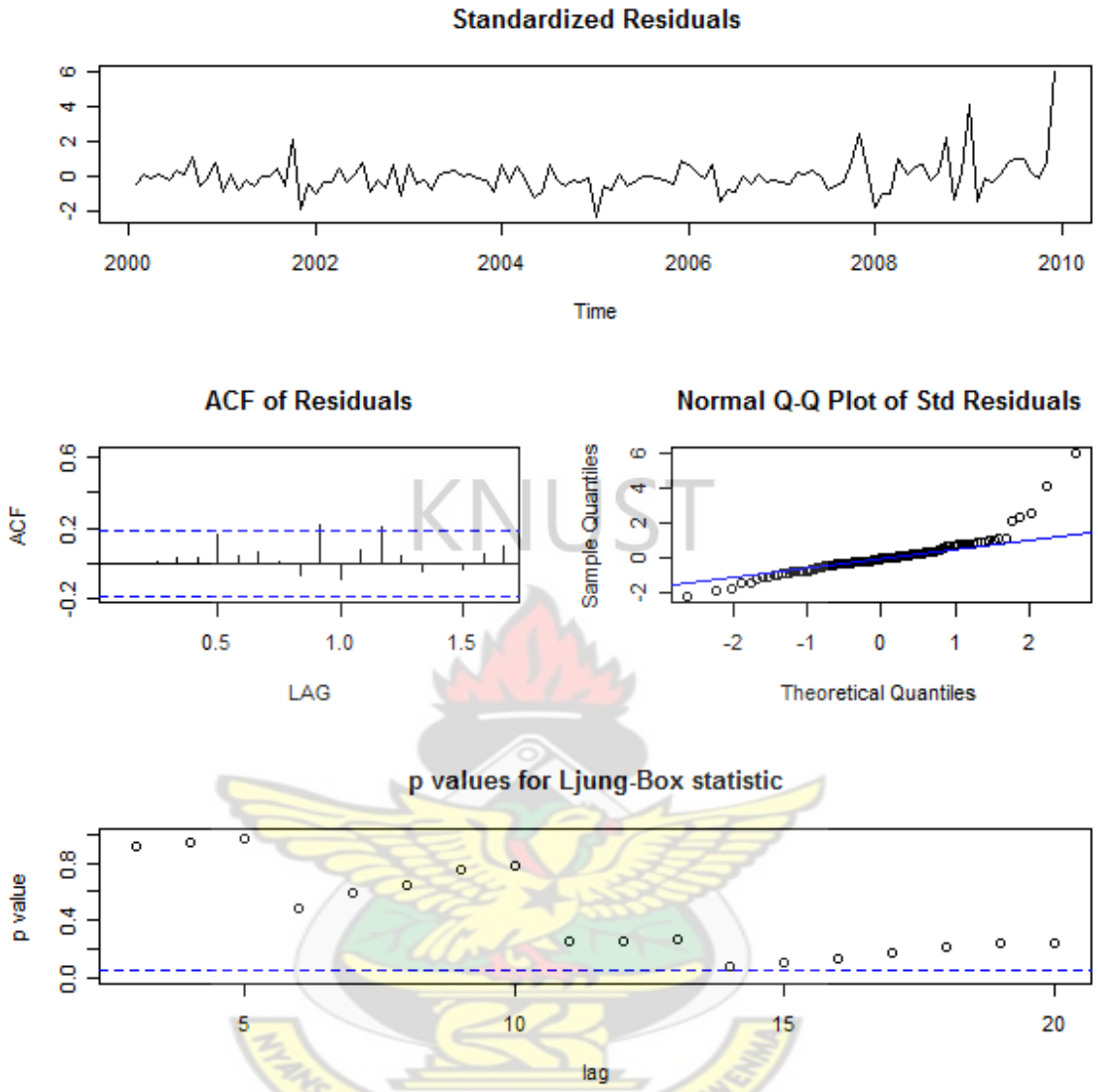


Figure 4.5: Diagnostics of *ARIMA*(2,1,0)

The diagnostics of the residuals from *ARIMA*(2,1,0) is exhibited in figure 4.5 above. The top part is the time plot of the standardized residuals of *ARIMA*(2,1,0). There are at least two (2) or three (3) residuals at the tail of the series with magnitudes larger than three (3) which is very unusual in standard normal distribution. This could be attributed to very large increases in tax revenue collection in some of the months as a result of bonuses paid to workers and also payment of accumulated withholding taxes which could be best

described as outliers in the data. However, inspection of the standardized residuals plot shows no obvious pattern and looks like an independently identically distributed sequence of zero mean with some few outliers.

The middle part of the diagnostics is the plot of the ACF of residuals. We notice two significant (outside the dotted lines) correlations in the plot but they are not meaningful because the lags which they occur do not relay any information and could be attributed to only the outliers. The ACF of the standardized residuals however, shows no apparent departure from the model assumptions.

In the middle part of the diagnostics, we could also find the normal Q-Q plot of standardized residuals at the right side. Most of the residuals are located on the straight line except for some few extreme residual values at the tails deviating from normality. The deviation from normality at the tails indicates that the outliers are quite prominent in the data, however, the normality assumption looks to be satisfied and so the residuals appear to be normally distributed.

The bottom part of the diagnostics is the time plot of the Ljung-Box statistics. It is observed that the Ljung-Box statistics plot is never significant at any positive lag.

4.3.2 Parameter Estimates and Diagnostics of *ARIMA*(0,1,1) Model

Call:

arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),

xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc, REPORT = 1,

reltol = tol))

Coefficients:

	ma1	xmean
	-0.4920	6005.913
s.e.	0.0856	2731.900

sigma^2 estimated as 3.353e+09: log likelihood = -1474.02, aic = 2954.04

$AIC

[1] 22.96683

$AICc

[1] 22.98539

$BIC

[1] 22.01353

The parameter based on the t-value test is statistically significant because the t-value is greater than or equal to 2 in absolute value $(t \geq |2|)$.

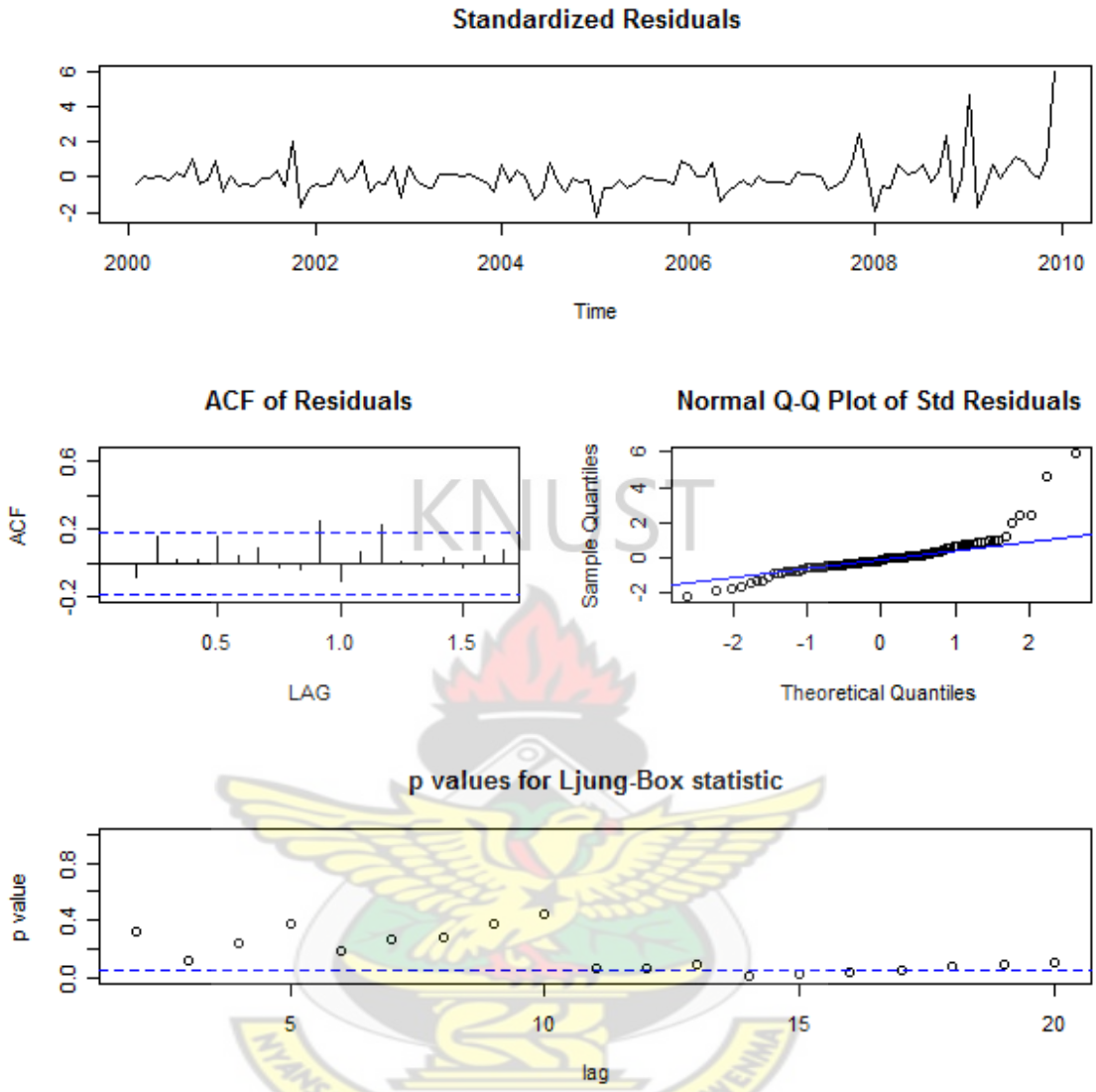


Figure 4.6: Diagnostics of *ARIMA*(0,1,1)

Figure 4.6 displays the diagnostics of the residuals from *ARIMA*(0,1,1). The top part is the time plot of the standardized residuals of *ARIMA*(0,1,1). There are outliers, however, with a few values exceeding 3 standard deviations in magnitude at the tail of the series. Inspection of the standardized residuals plot shows no apparent pattern and looks like an independently identically distributed sequence of zero mean with some few outliers.

The first diagram in the middle part of the diagnostics is the plot of the ACF of residuals. There are two significant (outside the dotted lines) correlations in the plot but they are not meaningful because the lags which they occur do not convey any information and could be attributed to only the outliers. The ACF of the standardized residuals however, shows no apparent departure from the model assumptions.

At the right side of the middle part of the diagnostics is the normal Q-Q plot of standardized residuals. Most of the residuals are located on the straight line except for some few extreme residual values at the tails deviating from normality which could be attributed to only outliers. However, the normality assumption looks to be satisfied and so the residuals appear to be normally distributed.

The bottom part of the diagnostics is the time plot of the Ljung-Box statistics. We notice that the Ljung-Box statistics plot is significant at some of the positive lags.

4.3.3 Parameter Estimates and Diagnostics of *ARIMA*(2,1,1) Model

Call:

arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),

xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc, REPORT = 1,

reltol = tol))

Coefficients:

	ar1	ar2	ma1	xmean
	-0.5444	-0.3837	0.0588	6137.408
s.e.	0.2635	0.1260	0.2879	2873.467

sigma^2 estimated as 3.205e+09: log likelihood = -1471.4, aic = 2952.81

$AIC

[1] 22.95521

$AICc

[1] 22.97648

$BIC

[1] 22.04862

The parameters based on the t-value estimate are not statistically significant because not all the t-values are greater than or equal to 2 in absolute value ($t \geq |2|$).

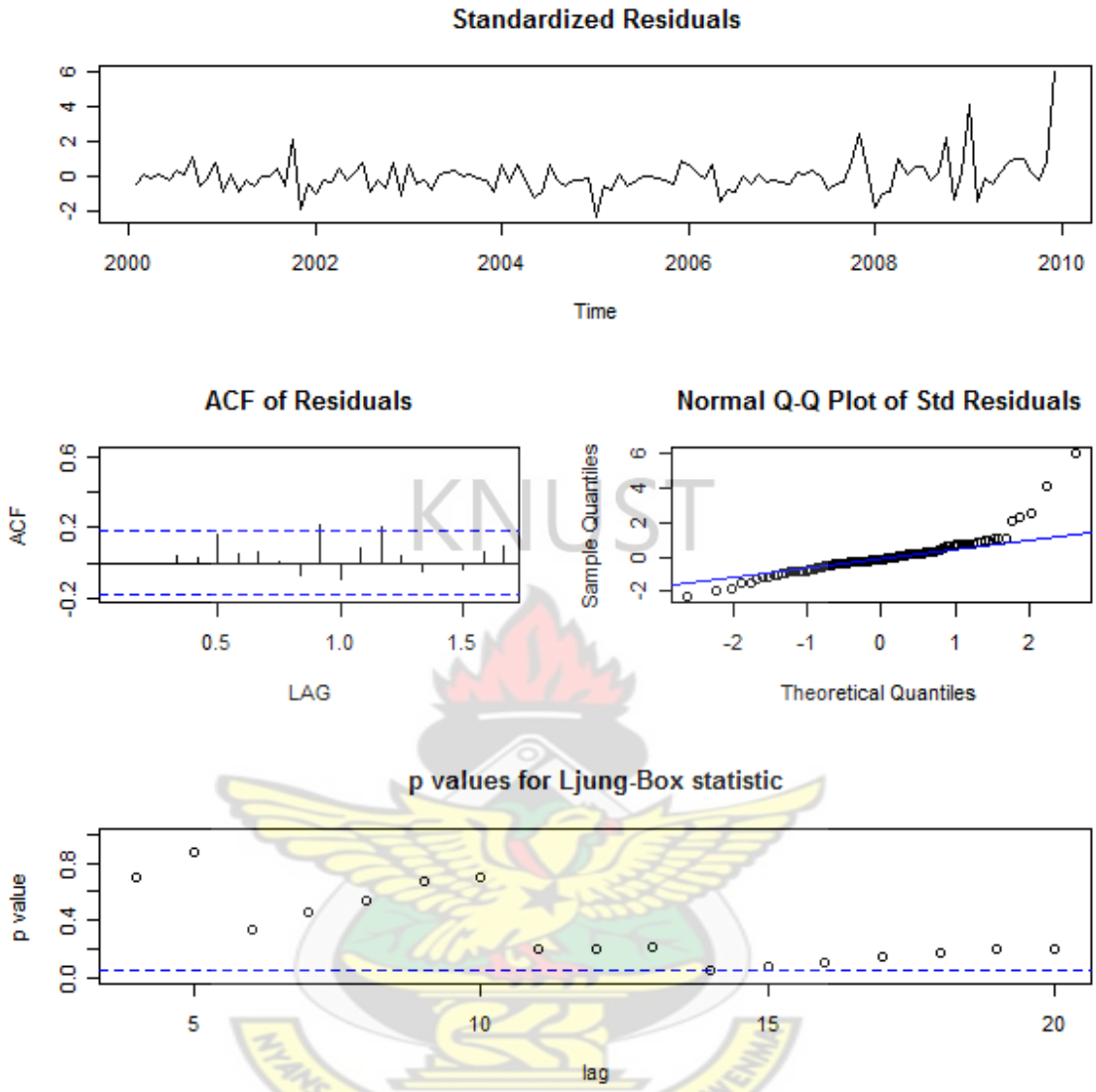


Figure 4.7: Diagnostics of *ARIMA*(2,1,1)

The diagnostics of the residuals from *ARIMA*(2,1,1) are displayed in figure 4.7 above. The top part is the time plot of the standardized residuals of *ARIMA*(2,1,1). There are outliers, however, with a few values exceeding 3 standard deviations in magnitude at the tail of the series. Inspection of the standardized residuals plot shows no obvious pattern and looks like an independently identically distributed sequence of zero mean with some few outliers.

The first diagram in the middle part of the diagnostics is the plot of the ACF of residuals. We notice two significant (outside the dotted lines) correlations in the plot but they are not meaningful because the lags which they occur do not express any information and could be attributed to only the outliers. The ACF of the standardized residuals however, shows no apparent departure from the model assumptions.

Again, at the middle part of the diagnostics, we could find the normal Q-Q plot of standardized residuals at the right side. Most of the residuals are located on the straight line except for some few extreme residual values at the tails deviating from normality which could be attributed to only outliers. However, the normality assumption looks to be satisfied and so the residuals appear to be normally distributed.

The bottom part of the diagnostics is the time plot of the Ljung-Box statistics. It is observed that at least one of the Ljung-Box statistics plot is significant at the positive lags.

The standardized residual plots of all the three models discussed above depicted that there are few outliers in the dataset; however, with a few values exceeding 3 standard deviations in magnitude, the models appeared to be independently and identically distributed with zero mean. The autocorrelation functions of the residuals showed evidence of significant correlations at two different lags in all the models but these lags are not meaningful and could be attributed to only the outliers. The residuals exhibited no obvious departure from the normality assumptions and appeared to be normally distributed in all the models. In the Ljung-Box statistics plot, we noticed significant values at some positive lags except in one model which the plot failed to indicate significant value at any positive lag.

Apart from the parameters in the models *ARIMA*(2,1,0) and *ARIMA*(0,1,1) which are significant at 5% level of significance and could be used for prediction purpose that of *ARIMA*(2,1,1) are not significant and could have a negative effect on any forecast if used.

To select the final model, we compare the *AIC*, AIC_c and *BIC* values for *ARIMA*(2,1,0) and *ARIMA*(0,1,1) models but all the *AIC*, AIC_c and *BIC* values favour *ARIMA*(2,1,0) model. We therefore conclude that *ARIMA*(2,1,0) model is the best-fit model for forecasting into the future of the IRS direct tax revenue collection.

To explicitly state the best-fit model, *ARIMA*(2,1,0), we should recall for instance that *ARIMA*(*p*,1,*q*) model could be written as a nonstationary *ARMA*(*p*+1,*q*) model (Cryer and Chan, 2008, p. 92). Thus, *ARIMA*(2,1,0) could be expressed as a nonstationary *ARMA*(3,0). If the first-order difference of the direct tax revenue data is denoted as $W_t = X_t - X_{t-1}$, then *ARIMA*(2,1,0) could be expressed in the notational form as

$$W_t = \theta + \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + a_t \tag{4.3.1}$$

or in the original form as

$$\begin{aligned} X_t - X_{t-1} &= \theta + \alpha_1 (X_{t-1} - X_{t-2}) + \alpha_2 (X_{t-2} - X_{t-3}) + a_t \\ X_t &= \theta + (1+\alpha_1) X_{t-1} + (\alpha_2 - \alpha_1) X_{t-2} - \alpha_2 X_{t-3} + a_t \end{aligned} \tag{4.3.2}$$

where the constant, $\theta = \mu(1-\alpha_1-\alpha_2)$.

From the R output for *ARIMA*(2,1,0), $\hat{\alpha}_1 = -0.4941$, $\hat{\alpha}_2 = -0.3669$ and $\hat{\mu} = 6{,}101.403$; hence $\hat{\theta} = 6{,}101.403[1-(-0.4941)-(-0.3669)] = 11{,}354.71$. Thus, the best-fit model is explicitly stated as

$$\hat{X}_t = 11{,}354.71 + 0.5059 X_{t-1} + 0.1272 X_{t-2} + 0.3669 X_{t-3} \tag{4.3.3}$$

In order to mitigate the effect of rounding errors, any forecast value generated by equation (4.3.3) should be rounded to the nearest figure.

4.4 FORECASTING

Forecasting into the future is one of the main objectives of any time series analysis. Thus, we now forecast or predict 12 months into the future of the IRS direct tax revenue collection using the best-fit model *ARIMA*(2,1,0) selected in the previous section.

$pred

Time Series:

Start = 121 (January, 2010)

End = 132 (December, 2010)

Frequency = 1

[1] 749798.8 728357.9 813034.9 790542.8 782079.0 805989.2 808759.9 810101.3

[9] 819901.9 826047.4 830896.3 837726.3

$se

Time Series:

Start = 121 (January, 2010)

End = 132 (December, 2010)

Frequency = 1

[1] 56622.24 63456.81 67065.28 75831.71 81990.38 86717.00 92171.40

[8] 97191.76 101696.67 106179.21 110490.33 114576.80

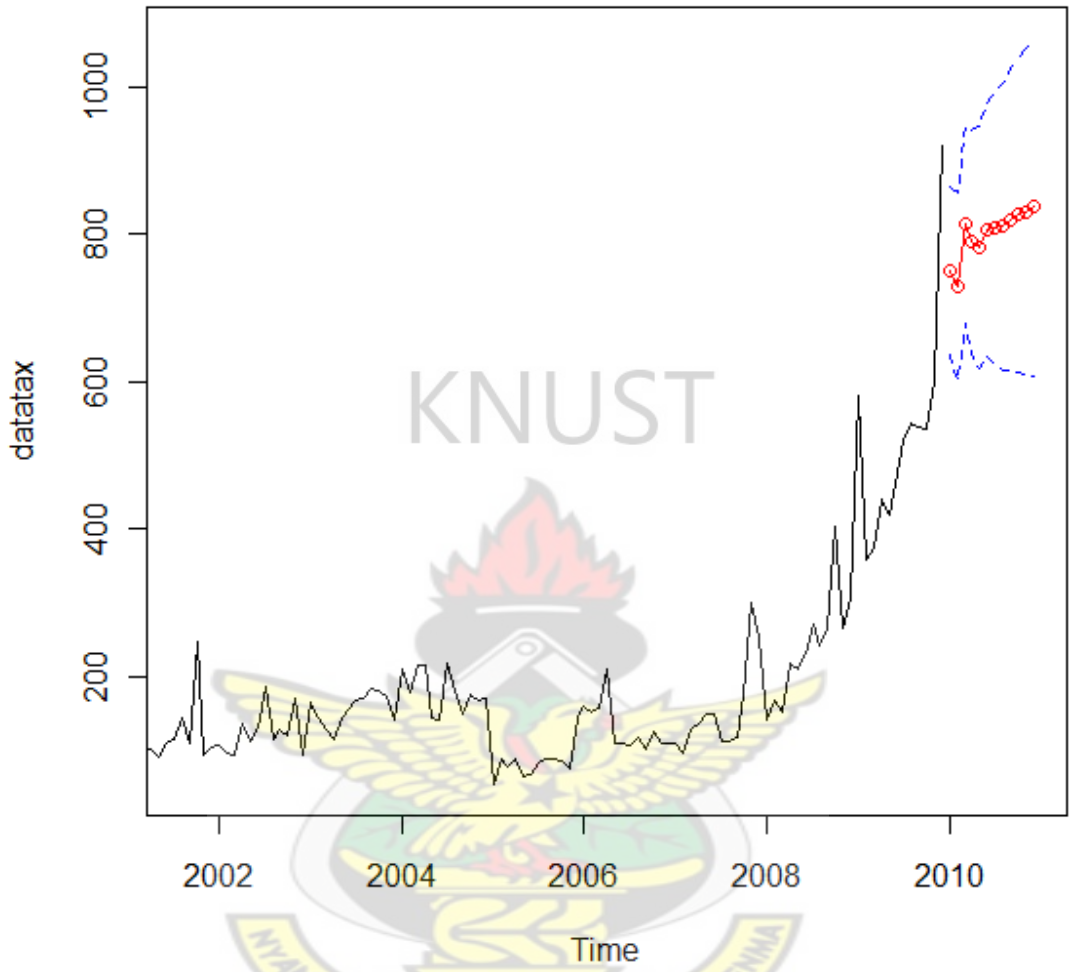


Figure 4.8: Graph of IRS, Sefwi Wiawso Tax District Office Direct Tax Revenue Collection, its forecasts and confidence intervals, with monthly totals in thousands of GH¢.

Figure 4.8 displays the diagrammatic representation of the original IRS, Sefwi Wiawso tax district direct tax revenue collection, with monthly totals in thousands of GH¢ (black line), its forecasts (red line) and confidence intervals (blue short dashes lines).

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

This chapter presents the conclusion and recommendations thereof after critically examining the findings of the study.

5.1 CONCLUSION

The study illustrates the application of time series analysis with greater emphasis on Box-Jenkins approach in modelling tax revenue collection. In this research, monthly direct tax revenue collection data from Internal Revenue Service (IRS), Sefwi Wiawso tax district over a 10-year period (120 months) from 2000 to 2009 are used to do the illustration. The behaviour of the sample autocorrelation function (ACF) of the monthly direct tax revenue collection revealed the existence of trend component in the data which led us to perform first-order differencing on the data in order to remove the trend for the series to attain an appreciable stability.

Based on the behaviour of the sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) of the differenced tax revenue series, three models are initially suggested but after the diagnostic checks, *ARIMA*(2,1,0) model is finally picked as the best-fit model that could be used to forecast into the future of the IRS direct tax revenue collection.

5.2 RECOMMENDATIONS

The *ARIMA*(2,1,0) model is recommended for forecasting into the future of the Internal Revenue Service (IRS), Sefwi Wiawso tax district direct tax revenue collection but the following precautionary measures should be taken into consideration:

- The model should not be used to forecast long time ahead (preferably a maximum of 24 months). This is because forecasting long time periods could lead to arbitrary large forecast values.
- Tax exemption to key taxpayers or contributors in the tax district should be done with care. The reason being that if the tax authorities (government) give tax exemption to key contributors to the tax revenue in the tax district, tax revenue collection would automatically decline and this could lead to wild deviation from the forecast value(s).
- The tax authorities should also broaden the tax base by means of bringing more individuals or enterprises and companies into the tax net to generate enough revenue as it is better to exceed set projection(s) rather than inability to achieve them.

Finally, it is also recommended that further research should be conducted to look for a more appropriate model(s) that could take care of drastic government interventions in future.

REFERENCES

Ali-Nakyea, A. (2008). *Taxation in Ghana, Principles, Practice and Planning*. 2nd Ed. Black Mask Limited, Cantonments – Accra.

Bowerman, B. L., O'Connell, R. T. and Hand, M. L. (2001). *Business Statistics in Practice*. 2nd Ed. McGraw-Hill/ Irwin Companies, Inc. 1221 Avenue of the Americas, New York, NY, 10020, pp. 626-665.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd Ed. Prentice-Hall International, Inc., New Jersey. ISBN 0-13-060774-6.

Braun, B. M. (1988). Measuring Tax Revenue Stability with Implications for Stabilization Policy: A Note. *National Tax Journal*, Vol. 41, no. 4, pp. 595-98.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. 2nd Ed. Springer-Verlag New York, Inc., New York, USA. ISBN 0-387-95351-5.

Chatfield, C. (2004). *The Analysis of Time Series. An Introduction*. 6th Ed. Chapman & Hall/ CRC, A CRC Press Company, New York. ISBN 1-58488-317-0.

Clower , T. L. and Weinstein, B. L. (2006). *Sales Tax Revenue in the City of Arlington, Texas: Historical Review and Projections*. Center for Economic Development and Research, University of North Texas, Denton, Texas.

Cryer, J. D. and Chan, K. S. (2008). *Times Series Analysis with Applications in R*. 2nd Ed. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA. ISBN 978-0-387-75958-6.

Davies, N. and Newbold, P. (1979). Some Power Studies of a Portmanteau Test of Time Series Model Specification. *Biometrika*, Vol. 66, pp. 153–155.

Fullerton, T. M., Jr. (1989). A Composite Approach to Forecasting State Government Revenues: Case study of the Idaho Sales Tax. *International Journal of Forecasting*, Vol. 5, pp.373-380, North- Holland.

Ghana Revenue Authority Act, 2009, Act 791.

Granger, C. W. J. and Newbold, P. (1986). *Forecast Economic Time Series*. 2nd Ed. Academic Press, New York, p 1.

Hamilton, J. D. (1994). *Times Series Analysis*. Princeton University Press, Princeton, New Jersey. ISBN 0-691-04289-6.

Internal Revenue Service, Taxpayer Publications, http://www.irs.gov.gh/pages/downloads (accessed: February, 10, 2011).

Kirchgässner, G. and Wolters, J. (2007). *Introduction to Modern Time Series Analysis*. Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-73290-7.

Nikolov, M. (2002), Tax Revenues Forecasting with Intervention Time Series Modeling. *Bulletin*, Ministry of Finance, Republic of Macedonia.

Paradis, E. (2005). *R for Beginners*. Institut des Sciences de l'Évolution, Université Montpellier II, France.

Prichard, W. (2009). Taxation and Development in Ghana: Finance, Equity and Accountability. *The Tax Justice Network*, United Kingdom.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications with R Examples*. 2nd Ed. Springer Science+Business Media, LLC, 233 Spring Street,

New York, NY 10013, USA. ISBN-10: 0-387-29317-5, ISBN-13: 978-0387-29317-2.

Slobodnitsky, T. and Drucker, L. (2005). *VAT Revenue Forecasting in Israel*. Ministry of Finance, State Revenue Administration

Wei, W. W. S. (2006). *Time Series Analysis. Univariate and Multivariate Methods*. 2nd Ed. Pearson Education, Inc., New York. ISBN 0-321-32216-9.

APPENDICES

APPENDIX A: DETAILS OF DIRECT TAX REVENUE COLLECTION DATA FROM INTERNAL REVENUE SERVICE, SEFWI WIAWSO TAX DISTRICT (2000 – 2010)

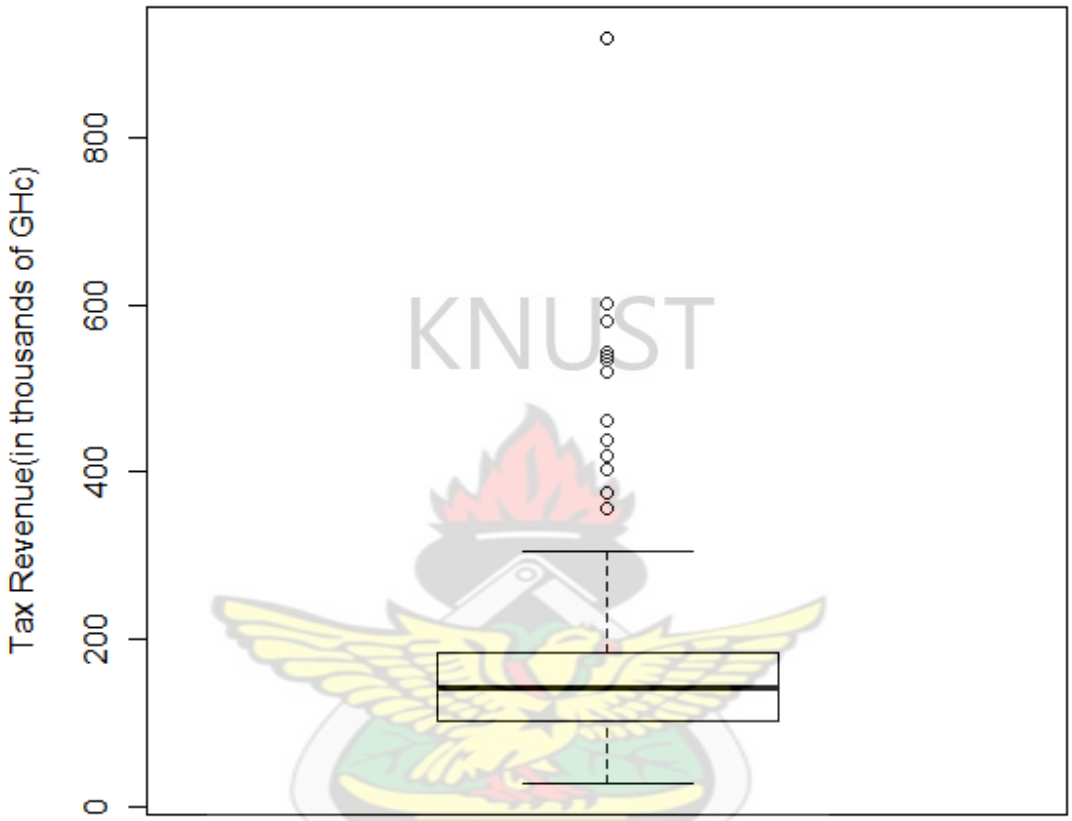
MONTH	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
JANUARY	46,751.00	96,534.00	107,870.00	165,696.00	209,137.00	53,076.00	160,994.00	108,703.00	141,713.00	581,714.00	614,085.00
FEBRUARY	26,720.00	127,333.00	95,334.00	144,204.00	178,836.00	88,702.00	153,018.00	96,226.00	168,747.00	357,321.00	878,675.00
MARCH	48,097.00	103,244.00	94,603.00	128,450.00	214,190.00	77,957.00	156,810.00	128,379.00	151,842.00	374,842.00	566,396.00
APRIL	48,617.00	101,582.00	136,517.00	114,111.00	214,479.00	86,924.00	209,794.00	136,583.00	217,851.00	439,296.00	471,283.00
MAY	56,851.00	91,997.00	111,361.00	144,621.00	144,555.00	65,286.00	110,363.00	148,530.00	209,860.00	420,353.00	1,017,198.00
JUNE	50,410.00	108,362.00	129,323.00	157,019.00	141,043.00	67,298.00	108,210.00	148,756.00	230,139.00	462,517.00	829,070.00
JULY	78,924.00	114,047.00	185,235.00	167,077.00	217,126.00	83,485.00	106,924.00	112,198.00	270,086.00	519,659.00	713,073.00
AUGUST	81,982.00	144,090.00	114,307.00	169,945.00	181,218.00	88,177.00	118,556.00	112,773.00	241,300.00	543,677.00	559,520.00
SEPTEMBER	144,351.00	108,643.00	127,247.00	183,706.00	149,597.00	87,355.00	101,072.00	118,344.00	265,800.00	537,768.00	390,455.00
OCTOBER	94,970.00	246,071.00	121,242.00	180,205.00	176,370.00	84,464.00	124,974.00	176,822.00	404,115.00	534,740.00	1,148,691.00
NOVEMBER	102,318.00	93,953.00	171,107.00	173,951.00	168,933.00	74,256.00	110,742.00	299,493.00	264,616.00	600,788.00	743,919.00
DECEMBER	169,860.00	104,479.00	94,761.00	139,888.00	169,661.00	143,550.00	109,379.00	253,043.00	305,409.00	920,497.00	884,271.00
TOTAL	**949,851.00**	**1,440,335.00**	**1,488,907.00**	**1,868,873.00**	**2,165,145.00**	**1,000,530.00**	**1,570,836.00**	**1,839,850.00**	**2,871,478.00**	**6,293,172.00**	**8,816,636.00**

APPENDIX B: SUMMARY STATISTICS OF DIRECT TAX REVENUE COLLECTION DATA (2000 – 2009)

Variable	N	Mean	SE Mean	StDev	Variance	CoefVar	Sum	Sum of Squares	Minimum
Tax Revenue	120	179075	12535	137315	18855388895	76.68	21488977	6.09193E+12	26720

Variable	Q1	Median	Q3	Maximum	Range	IQR	Skewness	Kurtosis	MSSD
Tax Revenue	102550	141378	184853	920497	893777	82303	2.54	8.07	2003559079

APPENDIX C: BOXPLOT OF THE DIRECT TAX REVENUE DATA (2000 – 2009)



APPENDIX D: T-VALUE CALCULATIONS

- ARIMA(2,1,0)

  ar1: $t-value = \dfrac{-0.4941}{0.1011} = |4.89|$

  ar2: $t-value = \dfrac{-0.3669}{0.1005} = |3.65|$

  xmean: $t-value = \dfrac{6{,}101.403}{2{,}808.074} = |2.17|$

- ARIMA(0,1,1)

  ma1: $t-value = \dfrac{-0.4920}{0.0856} = |5.75|$

  xmean: $t-value = \dfrac{6{,}005.913}{2{,}731.900} = |2.20|$

- ARIMA(2,1,1)

  ar1: $t-value = \dfrac{-0.5444}{0.2635} = |2.07|$

  ar2: $t-value = \dfrac{-0.3837}{0.1260} = |3.05|$

  ma1: $t-value = \dfrac{0.0588}{0.2879} = |0.20|$

  xmean: $t-value = \dfrac{6{,}137.408}{2{,}873.467} = |2.14|$

APPENDIX E: R CODES USED IN THE ANALYSIS

```
library(Rcmdr)

data

dataset=ts(data,start=2000,frequency=12)

plot(dataset,xlab="Time(Month)",ylab="Tax Revenue)")

acf(dataset,36)

difftax=diff(dataset)

plot(difftax,xlab="Time(Month)",ylab="Tax Revenue")

source(url("http://www.stat.pitt.edu/stoffer/tsa2/Rcode/itall.R"))

acf2(difftax,36)

sarima(difftax,2,0,0)

sarima(difftax,0,0,1)

sarima(difftax,2,0,1)

sarima.for(dataset,12,2,1,0)

history()
```