KWAME NKRUMAH UNIVERSITY OF SCINCE AND TECHNOLOGY, KUMASI

COLLEGE OF SCIENCE

DEPARTMENT OF MATHEMEMATICS

TIME SERIES ANALYSIS ON MEMBERSHIP ENROLMENT OF NATIONAL HEALTH INSURANCE SCHEME.

A CASE STUDY OF SAVELUGU/NANTON DISTRICT MUTUAL HEALTH INSURANCE SCHEME IN NORTHERN REGION

BY

ABUBAKARI FUSEINI

A THESI SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI, IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MATER OF SCIENCE DEGREE IN INDUSTRIAL MATHEMATICS

JUNE 2012

DECLARATION

I hereby declare that this submission is my own work towards Master of Science degree and that to the best of my knowledge, it contains no material previously published by another person nor materials which have been accepted for the award of any other degree of any university, except for references cited from extracts, scripts, text books, journals, papers and other sources which have been duly acknowledged.

ABUBAKARI FUSEINI …… ……

(PG3005609) Signature Date

I declare that I have supervised the candidate's research work and I confirmed that the student has my permission to present it for assessment.

NANA KENA FREMPONG …… ……

(Supervisor) Signature Date

Certified by:

MR. K.F.DARKWAH …….. ……

(Head of Department) Signature Date

Certified by:

Prof. I. K. Dontwi …….. ……

(Dean, IDL) Signature Date

DEDICATION

This project is wholly dedicated to Almighty Allah, my parents, family members and my wife for their immense support and inspirations received.

ACKNOWLEGEMENT

A research of this nature cannot be completed without certain amount of support, guidance and directions from others.

My outmost thanks first of all goes to the Almighty Allah without whose grace and mercy I would not have come from this far.

My heartfelt gratitude also goes to my supervisor, Nana Kena Frempong for his patient, good sense of direction, critical comments and suggestions during the period of this research work.

Special thanks and gratitude are also extended to the management of the Savelugu/Nanton MHIS for providing me data for this research work.

I also owe a great sense of gratitude to all the lecturers of the Mathematics department, KNUST for impacting in me further and advanced knowledge of Mathematics and its applications.

I also thank Mr. Aziz (Agbazo) and Mr. Anokye all of Kumasi Polytechnic for the support and directions received during my studies.

I cannot forget the contributions of Mr. Ahmed Imoro of NHIA, Mr. Rashid Tanko (Computer) of NHIA, Mr. Abdul-Rahman Gado of NHIA, Mr. Hudu Issah of NHIA, Mr. Adam Meisuna of NHIA and all the staff and management of the National Health Insurance Authority. God bless you all for you supporting me during the period of my studies.

Many thanks to my parents and all other family members especially Alhaji Abdul-Rahman of Nalco Saudi Company Ltd (Kingdom Of Saudi Arabia), Alhaji Baba (BA) Of Saudi Arabia and Abubaka Ali all of the PAPAYE'S family in Tamale for their immense support in my educational career. Allah blesses you all.

ABSTRACT

The National Health Insurance Scheme (NHIS) is a social policy; a kind of social re-engineering that caters for the healthcare needs of the most vulnerable and all residents in Ghana. It allows everybody residents in Ghana to register with a nearby district mutual health insurance scheme for their healthcare needs. The research was undertaken to study the trend/pattern of membership registration at the Savelugu/Nanton District Mutual Health Insurance Scheme and also to select the best ARIMA model that will be use in predicting future enrollment values for the scheme. Secondary data was obtained from the office of the Savelugu/Nanton Mutual Health Insurance Scheme (SNDMHIS) from the inception of the Scheme in the year 2005 up to 2010. A mathematical concept, Time Series

analysis was used to model the membership enrollment of the scheme and was used to make predictions of future enrollment into the scheme. The overall ARIMA model obtained was $y_t = -1.321\varepsilon_{t-1} + 0.321\varepsilon_{t-2}$. The model was used to make prediction for 2011 and 2012. The predicted values recorded were decreasing from month to month. Findings from the study also indicate that enrollment of individuals to the scheme had experienced an increase and a decrease linear trend from the year 2005 to 2010. The highest enrollment (4,213) from the inception of the scheme was recorded in December 2008. Thereafter, enrollments have been declining gradually from year to year. However, most of other few high enrollment values were recorded from August to December over the period under study and same was seen from the predicted values of 2011 and 2012.

LIST OF TABLES

Page

LIST OF FIGURES

Page

DIFINITION OF TERMS

Benefit Package: Health services and consumables that the Health Insurance Scheme provides to its members under the terms of the scheme.

Community: A collection of individuals living in the same geographical area united by common interest and sharing the same pre-occupation.

Coverage: The percentage of people who have registered in the National Health Insurance Scheme.

Exemptions: It is a term used to describe those who are registered under the Scheme free of charge and receive their healthcare needs free of charge. They include: Children under 18 years, the aged (70 years and above) and the indigents.

Core-poor: Adults who are unemployed and do not receive any identifiable and constant support from elsewhere for survival.

Quality healthcare: Receiving healthcare from a health facility as the priority source with qualified (trained) personnel and appropriate care provision.

Insured: A person who has registered and fully paid the annual contribution meets the scheme's membership criteria and is entitled to the approved benefits package.

Service providers: These are healthcare providers mostly public (GHS) and missionary (CHAG) health facilities.

Out-patient care: Part of the benefits package which involves the beneficiary being offered health services without having to be hospitalized.

TABLE OF CONTENT

Page

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

In recent times, healthcare financing has become a world-wide problem with increasing demand for healthcare services and relatively slow increase in its supply (McConnell, 1999).

In 1985, the Hospital Fees Regulations (LI 1313) mandated fees to be charged for health services rendered to patients. The introduction of the user fees at government hospitals and clinics, as part of cost sharing policy by government of Ghana in the health sector, put some financial burden on the citizens especially, the poor. This problem was compounded when the 'cash and carry" system was introduced in 1992 which was oriented towards full cost recovery of drugs and partial recovery for the cost of other services. The introduction of this policy resulted in a decline in utilization of health services in the country. To offset the negative effects of the "Cash and Carry" system, especially its consequences on the poor, the Government commissioned various studies into alternatives, principally insurance-based. Initially, a lot of efforts were vested into investigating the feasibility of a national health insurance scheme.

The National Health Insurance scheme (NHIS) was therefore introduced in 2003 by the National Health Insurance Act, 2003, Act 650 with the view to improving financial access of Ghanaians, especially the poor and the vulnerable, to quality basic healthcare services. Under the NHIS, the rich subsidizes the poor, the healthy subsidizes the sick and the economically active pays for children, the aged and the indigent.

The National Health Insurance scheme is a social policy, a kind of social re-engineering that caters for the most vulnerable in society through the principal equity, solidarity, risk sharing, cross-subsidization, re-insurance, subscriber/community ownership and value for money, good governance and transparency in the healthcare delivery.

1.1.1 Access to Healthcare

Generally, access to healthcare is low worldwide. The poor access to healthcare is due to geographical factors such as long distances to health facilities and poor nature of roads, financial due to poverty and the three main delays (delays in taking decision, delay in getting means to the facility and delay in getting treatment). As a result, countries have adopted measures, policies and legislations to improve access to healthcare by greater number of the population.

In America for example, such measures as " Play-or-pay" (proposals designed to increase employer-sponsored health insurance), tax credits and vouchers to help low income families afford healthcare and national insurance portability and Accountability Act (HIPAA) was enacted with the ultimate aim of increasing access to healthcare.

The World Health Organization (2001) estimates that, about 80% of the populations living in rural areas in developing countries depend on traditional medicines for their health needs.

The Ghana Poverty Reduction Strategy (GPRS, 2003) identifies that about 40% of Ghanaians are living in poverty and close to 27% are in extreme poverty. This confirms the 1999 Health Sector Review report which estimates that only one third (1/3) of the population use formal health sector services, leaving the two third (2/3) to the informal health system which consists mainly the traditional practitioners whose services are not only inadequate but also inappropriate.

1.1.2 Health Insurance in Ghana

The government of Ghana remains the main source of health funding through tax revenue. According to Kunfaa (1996), the central government is the main financier of health services in many developing countries including Ghana. He stated that on the average, the central government spends about 2.4% of the GDP on the healthcare needs of the people. Other sources such as donor agencies, user charges, income generating projects and local government contributions are becoming increasingly important to look at.

Prior to the law on the National Health Insurance Scheme, financing of healthcare in Ghana was on the basis of the 'cash and carry' system. Act 650, 2003, which established the National Health Insurance Scheme sought to change the earlier health financing system in the country.

The law sets out an elaborate governance and administrative set-up for the provisioning of health insurance in Ghana. The law currently has established 145 Mutual Health Insurance Schemes at the Districts, Municipals and Metropolitan areas all over the country. It established the National Health Insurance Authority to secure the implementation of a National Health Insurance policy that ensures access to basic healthcare services to all residents in Ghana.

1.1.3 Healthcare Financing

Financing healthcare is a major international concern as many governments are struggling to provide better services while available resources are diminishing. Some time ago, healthcare cost in public facilities were born entirely by the government. Out-of-pocket payment mechanism at the point and time of treatment was later introduced to reduce the ever increasing burden on government in various countries. This was done as means of reducing financial burden on the government and so, to sustain healthcare delivery. This rather deepened the problem of financial barrier of access to quality healthcare. In response, most countries throughout the world are adopting health insurance scheme(s) as means of financing healthcare and ensuring greater access to healthcare for all.

The Ministry of Health reports (2001), has indicated that, in Ghana the principal health sector financing mechanisms are government budgetary allocations and user fees. General tax revenue and donor pull fund account for 80% of public health expenditure and 20% from user fess.

According to the Ministry of Health (2004), financing healthcare has gone through a series of history in Ghana. At independent, healthcare in public facilities was 'Free' for all Ghanaians. This implied that there was no out-of-pocket payment at the point of consumption of healthcare in the public health facilities and healthcare financing in the public sector was therefore entirely through tax revenue.

According to the National Health Insurance Authority (2008), generally, for the mode of financing under the District Mutual Health Insurance Scheme (DMHIS), premiums payable are in line with one's ability to pay to enjoy a package of health services covering over 95% of diseases affecting Ghanaians. There are different contributions levels in both the formal and informal sectors of the economy. For the informal sector, community health insurance committees are to identify and categorize residents into social groups based on their socio-economic status. The policy framework of the NHIS (2004) proposed six main types of categories as Core-poor, Very poor, Poor, Middle income, Rich and very Rich. The core-poor pays nothing. Those in the paying category pays a minimum of GH¢ 7.2 per annum.

Formal sector workers pay 2.5% out of their SSNIT contribution and once both parents pay, all children less than 18 years of these contributors automatically benefit. Government has also introduced a 2.5% health insurance levy on some goods to cater for the indigents and the poor in society. A person can only benefit from insurance packages after some months of contribution to the scheme. This is what is termed as the waiting period.

1.1.4 The National Health Insurance Authority

The National Health Insurance Authority was also established by Act 650 to oversee and regulate the implementation of the National Health Insurance policy that ensures access to basic healthcare services to all residents in Ghana.

To operate as a health insurance scheme in Ghana, the scheme must first be registered under the companies code 1963 Act 179, as a company limited by guarantee or shares. Thereafter, the health insurance scheme has to apply to the NHIA for registration and licensing.

Some of the functions of the NHIA include:

1. Register, license and regulate health insurance schemes.

2. Supervise the operations of health insurance schemes.
3. Grant accreditation to healthcare providers and monitor their performances.
4. Ensure that healthcare services rendered to beneficiaries of schemes by accredited healthcare providers are of good quality.
5. Make proposals to the Minister for the formulation of policies on health insurance.
6. Maintain a register of licensed health insurance schemes and accredited healthcare providers;
7. Manage the National Health Insurance Fund.

1.1.5 Types of Health Insurance.

Generally, three main categories or types of health insurance have been identified by the health insurance Act (2003, ACT 650).

The first and the most popular one is the District Mutual Health Insurance Scheme (DMHIS). This is the one that is operational in Ghana. This is the public/non-commercial scheme that allows everyone resident in Ghana to register under this scheme. The District Mutual Health Insurance schemes that are operating in Ghana were registered as companies limited by guarantee and have their independent governing boards. Their operations are subsidized by the National Health Insurance Fund (NHIF) also established under Act 650.

The second category of health insurance is the Private Commercial Health Insurance (PCHIS) operated by approved companies. Under this scheme, one can just walk into any of such companies and buy the insurance for himself and his dependants just as you would for a car. The commercial insurance do not receive subsidy from the National Health Insurance Fund and they are required to pay a security deposit before they start operations.

The third category of health insurance is known as the Private Mutual Health Insurance Scheme (PMHIS). Under this, any group of people (say members of a church or social group) can come together and start making contributions to cater for their health needs, providing for services approved by the governing council of the scheme. This type of insurance does not also benefit from the National Health Insurance Fund.

1.1.6 Problems of the National Health Insurance Scheme.

The National Health Insurance Scheme (NHIS), just like any other organization, has its own unique problems and this section is devoted to the discussion of these problems.

After few years of functioning, the DMHISs have encountered serious governance, institutional, operational, administrative and financial problems.

One problem that faces the National Health Insurance Scheme (NHIS) is the fact that the District Mutual Health Insurance Schemes (DMHIS) depend on the National Health Insurance Authority for most of their funding, including bail outs.

Another problem has to do with the abuse of the system. Card bearers of the Scheme tend to over attend the health facilities thereby bringing very high claims bills to the Scheme for payment. There are some clients who have been found to have been attending health facilities three times within one week. Some of them do this by attending different facilities on different days within the same week.

Low standard of quality of healthcare has also been identified as one of the problems facing the National Health Insurance Scheme. It was observed that most accredited public facilities do not render the needed healthcare to the clients and this they saw is diverting the vision of bringing quality healthcare to the door step of the people and the reduction of poverty among the people of the country.

Also, Non adherence to the guidelines for prescription and dispensing as laid by the Ministry of Health has been identified as another challenge that has implications for the sustainability of the National Health Insurance Scheme (NHIS) and quality of healthcare for the subscribers.

Over-billing and non-adherence to tariffs have also been identified as another treat posing danger to the sustenance of the operations of the scheme. The present Tariff structure allows three visits within a 2 week period to promote quality of care. Unfortunately, some providers have been billing NHIS three times even though patients did not visit the facility three times within the two weeks period.

1.9 Problem Statement

It is globally accepted that the investment in the healthcare of the citizens and residents of any nation is very critical to the development of that nation. While many other economic, political and social advances must also be made, the importance of investment in the health of a nation is paramount.

Healthcare is a social problem and the solution to it lies outside individuals and their immediate environment. It is important for government to continue intervention through thoughtfully guided and collective actions. Health Insurance is one of such collective actions which advance the social welfare of any nation if properly managed and well resourced.

The implementation of Ghana's Health Insurance has increased facility attendance by clients as a result of the high number of people who have registered to get treated free of charge for ailments and conditions that would have cost them cedis under the cash and carry system. It is therefore important to undertake objective study on the recorded figures on the enrollment, analyze them and make predictions in order for management to know and prepare for the future.

1.3 Rational for the study

It is expected that the introduction of the National Health Insurance in Ghana will bring to the door step of Ghanaians and other non-Ghanaians the best quality and accessible healthcare no mater where they live.

This expectation could be achieved if the National Health Insurance Scheme could be embraced by everybody in Ghana more especially the informal sector workers, the indigents, the pregnant women and the vulnerable in society. Equally important is the understanding of modalities put in place by the scheme to reach the people to get them enrolled.

This research is therefore aimed at assessing the membership enrollment at the Savelugu/Nanton District Mutual Health Insurance Scheme since its inception in March 2005 to the year 2010 and making predictions for future registration or enrollment in order for management to know and prepare for the future.

1.4 Research Objective

This research is undertaken to:

1. Study the trend/pattern of membership registration at the Savelugu/Nanton District Mutual Health Insurance Scheme.
2. To select the best ARIMA model that will be use in predicting future registration values for the scheme.

1.5 Justification/Significant of the study

The research is intended to collate the registration figures of the Savelugu/Nanton from March 2005 to December 2010 and it is intended to serve the following purposes:

1. It will serve as a partial fulfillment of the requirement for the Master of Science Degree in **Industrial Mathematics** at the University of Science and Technology (KNUST), Kumasi.
2. It will help management of the Savelugu/Nanton Scheme to objectively plan ahead of schedule in making decisions that concern their enrollment, expected premiums to be paid and attendance of the registered clients at the various accredited healthcare centers for their healthcare needs. These will then serve as guideline for the evaluation of healthcare policies of the central Government.

1.6 Scope and Limitations

This study covers the operations of National health insurance at the Savelugu/Nanton District Mutual Health Insurance Scheme.

A research of this nature cannot be carried out without some limitations. Because of Time and financial constraint, the study will take a critical look at only the enrollment of members at the Savelugu/Nanton District Mutual Health Insurance Scheme (DMHIS) as one of the schemes in the Northern Region. As a result of this, the findings from this study may not apply to other Mutual Health Insurance Schemes as the Savelugu/Nanton Mutual Health Insurance Scheme may have its peculiar characteristics.

1.7 Study Area

The atlas of Ghana indicates Savelugu/Nanton District in the Northern Region. It is located along Tamale-Bloga road and it is about twenty-five kilometers away from Tamale metropolis.

Savelugu/Nanton District shares boundary with West Mamprusi District to the north of Kokobila. To the eastern part of the district, it shares boundary with Karaga District at Limo. The district shares boundary with the Tamale Metropolis to its south at Yilonayili and to the Western part of the district

is the Tolon/Kunbungu District at Moglaa. There are many ethnic groups in the District; these include Fulani, Frafra, Dagabas and Moshies who are residing at Pong-Tamale. Others are Ewes in the fishing communities at the river sites. Dagombas who are the original settlers are the major ethnic group in the District.

The projected population as at December, 2010 is 123,339. The total land mass of the District is 1, 790.7sqkm.

Statistics of the Ghana Education Service (GES) shows that the District has eight (8) educational circuits, twenty (20) Junior Secondary Schools and eighty-six (86) primary schools of which five (5) of them are private owned schools.

1.8 Organization of the Study

For the sake of convenience, this study has been arranged into five chapters. The First chapter constitutes the background of the study, statement of the problem, Aims and objectives of the study, significant of the study area, limitations of the study area and the organization of the study.

Chapter Two covers literature review on the theoretical and empirical works of previous researches. Chapter Three covers the research methodology which deals with the methods the research will use in achieving its target objective.

The Data presentations, interpretations and analysis of the data from the scheme are discussed in chapter Four.

Chapter Five, which is the conclusion part, contains summary of salient points, conclusion of the research work, suggestions and recommendations of the research work

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter is intended to review the works of several authors in textbooks, journals, articles, news papers, unpublished researches and reports that are related to Time Series analysis. Analysis of data by means of statistical methods with the aid of a software package is common in industry and administration and it is an integral part in the studies of Mathematics and Statistics.

Time series has become one of the most important and widely used tools in most analysis in the field of Mathematics and Statistics. Its application ranges from astrophysics and neurophysiology and it covers in areas like Statistics, Mathematics, Signal Processing, Econometrics, Financial Mathematics, study of biological data, control systems, communications, vibrations engineering and many other areas. It can be use to describe, explain, predict and control changes through time of selected variables.

2.1 Historical development of Time Series

Time series have played an important role in the early natural sciences. Babylonian astronomy used time series of the relative positions of stars and planets to predict astronomical events. The analysis of time series helps to detect regularities in the observations of a variable and derive 'laws' from them, and/or exploit all information included in this variable to better predict future developments. The basic methodological idea behind these procedures, which were also valid for the Babylonians, is that it is possible to decompose time series into a finite number of independent but not directly observable components that develop regularly and can thus be calculated in advance. For this procedure, it is necessary that we have different independent factors which will have an impact on the variable. In the middle of the 19th century, this methodological approach to astronomy was taken up by the economists CHARLES BABBAGE and WILLIAM STANLEY JEVONS.

Tinbergen (1939) constructed the first econometric model for the United States and thus started the scientific research programme of empirical econometrics. At that time, however, it was hardly taken

into account that chronologically ordered observations might depend on each other. The prevailing assumption was that, according to the classical linear regression model, the residuals of the estimated equations are stochastically independent from each other. For this reason, procedures were applied which are also suited for cross section or experimental data without any time dependence.

Cochrane and Orcutt (1949) were the first to notice that this practice might cause problems. They showed that if residuals of an estimated regression equation are positively auto-correlated, the variances of the regression parameters are underestimated and, therefore, the values of the F and t statistics are overestimated. This problem could be solved at least for the frequent case of first order autocorrelation by transforming the data adequately.

Almost at the same time, Durbin and Watson (1950), developed a test procedure which made it possible to identify first order autocorrelation. The problem seemed to be solved (more or less), and, until the 1970's, the issue was hardly ever raised in the field of empirical econometrics.

This did not change until Box and Jenkins (1970), published a textbook on time series analysis that received considerable attention. First of all, they introduced univariate models for time series which simply made systematic use of the information included in the observed values of time series. This offered an easy way to predict the future development of this variable.

Today, the procedure is known as *Box- Jenkins Analysis* and is widely applied or use for many field work analysis. One approach, advocated in the landmark work of Box and Jenkins (1970), develops a systematic class of models called autoregressive integrated moving average (ARIMA) models to handle time-correlated modeling and forecasting. The approach includes a provision for treating more than one input series through multivariate ARIMA or through transfer function modeling. The defining feature of these models is that they are multiplicative models, meaning that the observed data are assumed to result from products of factors involving differential or difference equation operators responding to a white noise input.

A more recent approach to the same problem uses additive models more familiar to statisticians. In this approach, the observed data are assumed to result from sums of series, each with a specified time series structure; for example, in economics, assume a series is generated as the sum of trend, a seasonal effect, and error. The state-space model that results is then treated by making judicious use of the celebrated Kalman filters and smoothers, developed originally for estimation and control in space applications.

It became even more popular when Granger and Newbold (1975), showed that simple forecasts which only considered information given by one single time series often outperformed the forecasts based on large econometric models which sometimes consisted of many hundreds of equations. Warren M. Persons (1919), came out with the decomposition into unobserved components that depend on different causal factors, as it is usually employed in the classical time series analysis.

2.2 General Applications of Time Series Analysis

Time Series analysis can be used to describe, explain, predict and control changes through time of selected variables. Geographers have utilized and developed techniques that add a distinctive spatial dimension to this analysis. In spectral analysis, time series are disaggregated into their individual components, breaking series down into waveforms and combinations of fundamental harmonics known as Fourier series. These are analyzed using power spectral (plots of variance in time series data vs. frequency), which may be applied to autoregressive time series, where correlations between observations at different time-distances apart in the series are calculated.

Cross-spectral analysis relates multiple time series through application of modeling and correlation techniques. More complex modeling exercises predict future values of a variable at a particular location based on its own values (autoregressive), lagged spatial diffusion effects and lagged exogenous or explanatory variables. A major spatial application time series analysis by geographers in the last two decades has been prediction of the return times for epidemics notably for outbreaks of

Measles. Influenza and Acquire Immune Deficiency Syndrome (AIDS). Based on detailed understanding of the relationship between size of community and temporal spacing of epidemics, predictive models have built on basic expansion diffusion models linked to regression analysis incorporating key explanatory variables.

According to R.E. Abdel-Aal, A.M. Mangoud (1998), in a research article, stated that two univariate time series analysis methods have been used to model and forecast the monthly patient volume at the family and community medicine primary healthcare clinic of King Faisal University, Al-Khob in the Kingdom of Saudi Arabia. Models were based on nine years of data and 2 years forecasts were made. They used the optimum ARIMA model of the fourth order operating on the data after differencing twice at the non-seasonal level and once at the seasonal level. It gives mean and maximum absolute percentage errors of 1.86 and 4.23% respectively over the forecasting interval. A much simpler method based on extrapolating the growth curve of the annual means of the patient volume using a polynomial fit gives the better figures of 0.55 and 1.17 respectively. This is due to the fairly regular nature of the data and the lack of strong random components that required ARIMA processes for modeling. A statistical time series analysis was used to study the interdependence between the primary and secondary pollutants in Taipei area by Kuang-Jung Hsu (1992) in his article of the "Time series Analysis of the Interdependence among Air pollutants". He made estimations using the vector auto-regression model (VAR) indicate that 2 and 4h time lags are sufficient to represent the observed values at two stations studied. The impulse response functions and variance decompositions of NO, NO2 and NO3 were derived using the vector moving average representations to examine the significance of one species on others. Influences of photochemistry and transport processes on these air pollutants at different locations were evaluated from the results. This technique may provide a simple tool for preliminary assessment of pollution problems.

Chi-JieLu et al. in (2009) wrote that financial time series are inherently noisy and non-stationary. It is regarded as one of most challenging applications of time series forecasting. Due to the advantages of generalization capability in obtaining a unique solution, support vector regression (SVR) has been

successfully applied in financial time series forecasting. In the modeling of financial time series using SVR, one of the key problems is the inherent high noise. Thus, detecting and removing the noise are important but difficult tasks when building an SVR forecasting model. To alleviate the influence of noise, a two-stage modeling approach using independent component analysis (ICA) and support vector regression is proposed in financial time series forecasting. ICA is a novel statistics signal processing technique that was originally proposed to find the latent source signals from observed mixture signals without having any prior knowledge of the mixing mechanism. The proposed approach first uses ICA to the forecasting variables for generating the independent components (ICs). After identifying and removing the ICs containing the noise, the rest of the ICs are then used to reconstruct the forecasting variables which contain less noise and served as the input variables of the SVR forecasting model. In order to evaluate the performance of the proposed approach, Nikkei 225 opening index and TAIEX closing index are used as illustrative examples. Experimental results show that the proposed model outperforms the SVR model with non-filtered forecasting variables and a random walk model.

Benjamin et al. in (2001) presented an extensive discussion on factors that may cause the reversal of the cross-country correlation coefficient through the application of Time series analysis. To account for cross-sectional heteroscedasticity and time-wise autocorrelation, Benjamin applied Prais-Winsten regressions to pooled time series, thereby neglecting possible non-stationary of the variables. She finds that the relationship between female labour participation and fertility becomes positive over time, although the timing of this shift depends on the country group (broadly reflecting role incompatibility).

Gauthier et al. in (1996) presented a historical review of the development of family policy in OECD countries using Time series analysis. She clustered countries into four different groups. First, in countries belonging to the pro-family/pro-natalist model the major concern is low fertility and because of this the main task of family policy is to encourage families to have children. This is done

by helping mothers reconcile work and family life. In this model, relatively high levels of support are provided for maternity leave and child-care facilities. Great emphasis is placed on cash benefits and more particularly, towards the third child. In the second, pro-traditional model the preservation of the family is the main concern. Government takes some responsibility for supporting families, but the most important sources of support are seen as the families themselves and voluntary organizations. Under this model, a medium level of state support for families is provided. The low provision of childcare does not give women the opportunity to combine employment and family responsibility easily. The third, pro-egalitarian model seeks to promote gender equality. Men and women are treated as equal breadwinners and equal careers and policy aims to support dual parent/worker roles. Liberal policies on marriage, divorce and abortion mean that there are few restrictions on how people can choose their family life. Fourth, in the countries belonging to pro-family but non-interventionist model the main concern is the families in need. The participation of women in the labour force is not discouraged, but limited benefits are provided by the state to support them. Families are viewed as basically self-sufficient and able to meet their own needs through the private market with only limited help from the state.

An article written by S Montani, C Larizza, R Bellazi and M Stefanelli (2002) titled "clinical time series", described the application of a method for intelligent analysis of clinical Time series in diabetes mellitus domain. Such a method is based on temporal abstraction and relies on the following steps:

(i) 'Pre-processing' of raw data through the application of suitable filtering techniques.

(ii) 'extraction' from the pre-processed data of a set of abstract episodes (temporal abstraction) and

(iii) 'post-processing' of temporal abstraction; the post-processing phase results in a new set of features that embeds high level information on patient dynamics. The derived features set are used to obtain new knowledge through the application of machine learning algorithms. The paper describes in detail the application of this methodology and presents some

results obtained on simulated data and on a data-set of four diabetics patients monitored for more than 1 year.

A paper by R. Almgren (2003) titled 'Time series analysis and statistical arbitrage or "How to use data to make trading and investment decisions. He stated that Time series analysis is the branch of statistics that studies processes evolving in time. All of statistics, like all of science, is about predicting the future from the past. He indicated that if the heights of 100 people in a town are measured, the model can be constructed for the distribution of the heights, and then before the 101st person will be measured, various probabilistic statements about what is expected of the person's height will be made. The difference in time series analysis is that the data points are explicitly ordered in time. According to him, this has a few immediate and general consequences:

• Joint distributions are important. When he looked at 100 people,

He doesn't expect successive measurements to have much relationship, since he probably selected the samples in a rather haphazard (let us not say "random") way. He doesn't expect the height of one person to depend on the height of the person he measured just before. He stated that, in time series, the data points are explicitly ordered in time and the distribution of one value may very likely be closely connected with the preceding value.

• Sample sizes are limited. In general, data points are divided into two classes: ones I have measured yet *vs.* ones I have not measured yet, and we try to make statements about the latter based on the former. In "standard" statistics, it is often true (though certainly not always true) that we can get larger samples if we really need them. If 100 people are not enough to calibrate our height model, then we can measure 1000 if we have enough resources and if the town is big enough. (If the town has very few people then we might as well measure all the people and then we don't need statistics.).

In a research article by H. Jaeger (2004) titled 'Echo-state networks (ESN)' examines the possibility of using an echo-state network for predicting the need for dialysis in the Intensive Care Unit (ICU). The study compared the performance in prediction obtained by the echo-state network with that obtained by two other time series analysis methods, namely a support vector machine (SVM) and a

Naive Bayes classifier (NB). To reach these objectives, diuresis and creatinine values were retrieved from the ICU database from a study population consisting of an observational cohort of 916 patients admitted consecutively to the ICU between May 31th 2003 and November 17th 2007, selected from a total of 9752 MICU/SICU patients admitted in this period after application of inclusion/exclusion criteria. Only diuresis and creatinine values of the first three days after ICU admission were retrieved.

The outcome parameter in this study was the prediction of dialysis between day 5 and day 10 after ICU admission. Eight thousand seven hundred and twenty five (8725) patients with a length of stay (LOS) on the ICU <10 days and 111 patients that received dialysis in the first five days of ICU admission were excluded from the analysis. Some further preprocessing of the diuresis and creatinine data was needed. Diuresis was only measured in 2 hour intervals while creatinine was measured once, twice or exceptionally three times a day. Hence, the interval between creatinine measurements was larger than the interval between two diuresis measurements. Since the input of time series need to contain measurements over regular time intervals which have to be the same for both parameters, interpolation of the data was the first preprocessing step. Furthermore, since the availability of both diuresis and creatinine measurements did not fully overlap, additional preprocessing was at hand. After preprocessing of the data of the 916 patients that had fulfilled the inclusion criteria, 830 patients in total were available with 60 interpolated measurements for both creatinine and diuresis. 62% of these patients were male, mean age of the study population was 58.6 years; total mortality was 17% and mean SAPS II score was 37.2. The echo-state network performance in predicting dialysis was measured by calculating the area under the receiver operating curve (AUC). For comparison, the AUC's for the same prediction problem, obtained by two other time series analysis methods consisting of a support vector machine (SVM) and a naive Bayes (NB) algorithm were calculated. Several parts of the algorithms had a stochastic nature, such as the random initialization of the reservoir weights. Therefore, the ESN, SVM and NB analyses were repeated three times each time using another initialization of the weights in the echo-state network, to see whether or not the variations seen in different analyses were caused by contingent network characteristics. Furthermore,

the computational complexity of the three methods was compared through their required execution times.

In time series analysis literature (Box et al. 1994), an intervention event is an input series that indicates the presence or absence of an event. An intervention event causes a time series process to deviate from its expected evolutionary pattern. It is assumed that the intervention event occurs at a specific time, has a known duration, and is of a particular type. The time of the intervention is when the event begins to cause deviation. The duration of the intervention is how long the event causes deviation. The type of the intervention is how the event's influence changes over time. The intervention response is how the intervention causes deviation. Generally speaking, intervention events are essentially dummy regressors or indicator variables (explicitly determined by the time, duration, and type) that are introduced through a transfer function filter (specified by the response) that results in an intervention effect. The term *i*ntervention sometimes applies to the event, response, and/or effect. Many promotions could be considered interventions because they can cause the aforementioned deviations.

In such cases, the practitioner usually knows the time, duration, and type of the promotion, which has occurred in the past or will occur in future. The response of the promotion is not generally known. However, the response may be identified by an analysis of the (similar) historical data and/or from the judgment of the practitioners and other experts based on their past experience, domain knowledge, and other statistical analyses outside the scope of this paper. If you know the intervention event and response, the intervention effect can be derived from the response parameter estimates. The end result of promotional analysis, using interventions, is that the practitioner has a better understanding of how past promotions affected the historical data. Armed with the knowledge gained by intervention analysis, the practitioner can better determine the value of past promotions and better forecast product and service demand by taking into account future promotions that are similar.

A paper by A. Palmer, J.J. Montano, F.J. Franconetti (1997) titled "Sensitivity Analysis Applied to Artificial Neural Networks for forecasting Time series'' presented a novel procedure known as sensitivity analysis applied to a multilayer perception (MLP) which allows the most relevant lagged terms in time series forecasting to be identified. Secondly, the paper conducted a comparison of forecasting accuracy between the neural network models resulting from applying the sensitivity analysis to the network model derived from the traditional procedure and the classic ARIMA model-using the time series corresponding to the number of passengers in transit through the Balearic Islands (i.e. their case study). Their findings established that a neural network derived from sensitivity analysis provides the greatest forecasting accuracy.

A paper by G. Hanski and M. Tourchin (1991) titled "Predation on a cycle prey" explained that predator–prey theory predicts that the actions of generalist predators will damp or suppress fluctuations in prey populations. It explained that ecological theory predicts that generalist predators should damp or suppress long-term periodic fluctuations (cycles) in their prey populations and depress their average densities. However, the magnitude of these impacts is likely to vary depending on the availability of alternative prey species and the nature of ecological mechanisms driving the prey cycles. They explained that these multispecies effects can be modeled explicitly if parameterized functions relating prey consumption to prey abundance, and realistic population dynamical models for the prey, are available. These requirements are met by the interaction between the Hen Harrier (Circus cyaneus) and three of its prey species, the Meadow Pipit (Anthus pratensis), the field vole (Microtus agrestis), and the Red Grouse (Lagopus lagopus scoticus). They used this system to investigate how the availability of alternative prey and the way in which prey dynamics are modeled might affect the behavior of simple trophic networks. They generated cycles in one of the prey species (Red Grouse) in three different ways: through (1) the interaction between grouse density and macroparasites, (2) the interaction between grouse density and male grouse aggressiveness, and (3) a generic, delayed density-dependent mechanism. Their results confirmed that generalist predation can damp or suppress grouse cycles, but only when the densities of alternative prey are low. They also

demonstrated that diametrically opposite indirect effects between pairs of prey species can occur together in simple systems. In this case, pipits and grouse are apparent competitors, whereas voles and grouse are apparent facilitators. Finally, they found that the quantitative impacts of the predator on prey density differed among the three models of prey dynamics, and these differences were robust to uncertainty in parameter estimation and environmental stochasticity.

An article written by Noel Cressie and Scott H. Holan (2011) titled "Time series in the environmental sciences" explained a special issue of time series and its applications in the environment. According to them, Environmental processes and ecosystem phenomena have deservedly received considerable attention over the last decade by both scientists and those involved in public-policy decisions. Many of the investigations in these areas are a result of potential climate change and its far-reaching impacts. Nonetheless, a common thread in environmental investigations is the need for methodological tools capable of describing and predicting these complex, and typically high-dimensional, processes.

The study of environmental time series is fundamental to the larger goal of sustainability and adaptation. Knowing how and ultimately why, environmental processes change over time gives governments and protectors of the commons a means for rational, informed decision making. The catchword is 'change', and while temporal variability is of primary importance to many environmental processes, there is typically a strong spatial component (in fact, the word 'environs' means 'around' in French) and thus, not surprisingly, these studies are often spatio-temporal. An important aspect in developing statistical models for environmental (and other) applications is to pay homage to the underlying science. For example, it is a basic law of nature that 'trees do not grow to the sky', and that trees make up forests and forests grow and recede according to many factors. Trees develop from seeds to saplings to mature trees, using nutrients and water in the soil and CO2, O2 and light from the atmosphere to grow. Superimposed on the nonlinear tree-growth process is the uncertainty of the trees' environment, leading to a statistical model in time series analysis for the growth and recession of the forest.

A paper by Arthur VanLear (2007) titled "Time Series Analysis in Communication Research" has indicated that, the impact of time series analysis on scientific applications within the field of communication can be partially documented by listing the kinds of communication research to which time series methods have been applied. In the area of mass communication research, time series analysis methods have been most commonly applied to the study of agenda-setting processes in a variety of contexts including AIDS policy, breast cancer screening, marijuana use among adolescents, drunk driving policy and behavior, global warming, consumer confidence and political judgments employed time series regression analysis to compare the contribution of news coverage of mammography screening and physician advice to the utilization of mammography by women 40 years and older in the United States between 1989 and 1991. Data on mammography-related national media attention between January 1989 and December 1991 were generated by analyzing the content of seven nationally and regionally prominent newspapers (the New York Times, Washington Post, Los Angeles Times, Chicago Tribune, Boston Globe, St. Petersburg Times, and USA Today). All relevant news stories appearing in these newspapers in a course of each month during the research period (N = 36 months) were aggregated to represent the volume of media attention to this issue in that particular month. Comparable national-level data on mammography utilization by women ages 40 and older and prevalence of physicians' advice to have a mammogram were compiled from the Behavioral Risk Factor Surveillance System (BRFSS) that is administered each month by the Centers for Disease Control and Prevention (CDC) to a representative cross-section of non institutionalized adults nationwide. The proportion of women 40 years and older in each month who had a mammogram in the year preceding the interview served as the dependent variable in the analysis. To estimate the prevalence of physician advice to have a mammogram in each month, the proportion of women 40 years and older indicating that having a mammogram in the past year was their physician's idea was used. Using time series analysis to examine the direction of influence between these three variables controlling for potential confounding variables, the researchers found that both channels of time Series Analysis communication (news coverage and physician advice) accounted together for

51% of the variance over time in mammography-seeking behavior by women 40 years and older. Moreover, he found that physician advice was particularly influential for women who had regular access to a physician, while news coverage of mammography was more influential among women who did not have regular access to a physician (mainly due to lack of health insurance).

In this and other similar studies, the typical approach taken by the researchers was to correlate national news coverage of issues over time with outcomes related to public opinion or public policy on these issues during the same time period. In virtually all cases, some form of aggregated data was used and most studies were limited to the investigation of the relationship between two time series. However, the time series methods employed in these studies vary considerably, ranging from trend analysis to time series regression and traditional ARIMA methods. He stated that, there is little doubt that time series analysis methodology has enriched agenda-setting research in a number of important ways, including the ability to describe and analyze the agenda-setting process and to correlate it with a host of hypothesized outcomes over time. For example, by applying these methods, researchers were able to estimate lagged effects of news coverage on individuals, groups, and institutions and calculated the rate at which these effects decay for different issues. They were also able to compare media effects across issues and populations as well as to examine indirect (or mediated) effects between the news, the policy agenda, and personal behavior over time. Perhaps more importantly, the use of these methods allows more rigorous tests of agenda-setting theory and facilitates multilevel theorizing and research.

A publication by Peter J. Hudson, Andy P. Dobson and Tim G. Benton (2002) titled "Trophic interactions and population growth rates: describing patterns and identifying mechanisms" indicated the importance of time series in modeling tropic interactions and the population growth rate. They indicated that the concept of population growth rate has been of central importance in the development of the theory of population dynamics, few empirical studies consider the intrinsic growth rate in detail, let alone how it may vary within and between populations of the same species. In an attempt to link theory with data they took two approaches. First, they addressed the question

'what growth rate patterns does theory predict they should see in time-series?' the models make a number of predictions, which in general are supported by a comparative study between time-series of harvesting data from 352 red grouse populations. Variations in growth rate between grouse populations were associated with factors that reflected the quality and availability of the main food plant of the grouse. However, while these results support predictions from theory, they provide no clear insight into the mechanisms influencing reductions in population growth rate and regulation. In the second part of the paper, they considered the results of experiments, first at the individual level and then at the population level, to identify the important mechanisms influencing changes in individual productivity and population growth rate. *Trichostrongylus tenuis* is found to have an important influence on productivity, and when incorporated into models with their patterns of distribution between individuals has a destabilizing effect and generates negative growth rates.

The hypothesis that negative growth rates at the population level were caused by parasites was demonstrated by a replicated population level experiment. With a sound and tested model framework they then explore the interaction with other natural enemies and showed that in general they tend to stabilize variations in growth rate. Interestingly, the models showed selective predators that remove heavily infected individuals can release the grouse from parasite-induced regulation and allow equilibrium populations to rise. By contrast, a tick-borne virus that killed chicks simply leads to a reduction in the equilibrium. When humans take grouse they do not appear to stabilize populations and this may be because many of the infective stages are available for infection before harvesting commences. In their opinion, an understanding of growth rates and population dynamics is best achieved through a mechanistic approach that includes a sound experimental approach with the development of models. They stated that models can be tested further to explore how the community of predators and others interact with their prey.

A journal by P.J. Hudson1, I.M. Cattadori1, B. Boag1 and A.P. Dobson (2006) titled "Climate disruption and parasite–host dynamics: patterns and processes associated with warming and the frequency of extreme climatic events" has indicated that, levels of parasitism and the dynamics of

helminth systems is subject to the impact of environmental conditions such that long term increases in temperature will increase the force of infection and the parasite's basic reproduction number, R0. They postulated that an increase in the force of infection will only lead to an increase in mean intensity of adults when adult parasite mortality is not determined by acquired immunity. Preliminary examination of long term trends of parasites of rabbits and grouse confirmed these predictions. They indicated that Parasite development rate increases with temperature and while laboratory studies indicate this is linear some recent studies indicate that this may be nonlinear and would have an important impact on R0. Warming would also reduce the selective pressure for the development of arrestment and this would increase R0 so that in systems like the grouse and Trichostrongylus tenuis this would increase the instability and lead to larger disease outbreaks. Extreme climatic events that act across populations appear important in synchronizing transmission and disease outbreaks, so they speculated that climate disruption will lead to increase frequency and intensity of disease outbreaks in parasite populations.

Pampel et al. in (2001) estimates the effect of female employment and country dummies by applying a fixed effects GLS model which also adjusts for autocorrelation and heteroscedasticity. Pampel uses all variables in levels, thereby disregarding the possibility of non-stationarity of the data. Averaged across all nations he finds negative effects of female employment on fertility depending on class and gender equality of the respective country group. Moreover, Pampel finds that initial increases in female labour participation strongly lower fertility, but continued increases have a progressively less negative influence on fertility.

Adsera et al. in (2004) explicitly tested for non-stationarity and estimated the effects of labour market arrangements on pooled fertility rates using levels, logs, and first differences of the variables depending on the test results. By applying a random effects model, Adsera ignored any country heterogeneity. By regressing different labour market indicators on fertility, Adsera did not explicitly

address the question of the factors behind the change in the correlation. She detected indirect evidence that labour market institutions shaped the changing correlation.

To avoid serial correlation of the data Kogel (2004) uses quinquennial data (i.e., only the data points 1960, 65, 70, 75, 80, 85, 90, 95 and 99). Since the data are not difference stationary, Kogel applies all variables in logarithms. Kogel showed that the time series association between TFR and FLP has not changed and offers two convincing elements that may explain the change in the cross-country correlation. These are the presence of unmeasured country specific factors and country heterogeneity in the magnitude of the negative time-series association between fertility and female employment.

A paper by Jan Chvosta, Donald J. Erdman and Mark little (2011) entitled "Measuring the performance and risk of holding financial assets" is an important aspect of any good financial management plan. The value of each asset in a portfolio depends on a set of economic values called risk factors. These specific risk factors can impact the individual asset value, and they can impact the whole asset group. For example, the volatility of an individual stock is a risk factor that affects only assets that depend on that stock. On the other hand, the Federal Reserve deciding to increase interest rates impacts the whole banking sector although it is unlikely to impact other sectors such as manufacturing and agriculture. The severity of the impact might also depend on the individual positions of firms and their business plans. Clearly, a portfolio that holds a single asset has a higher risk exposure than one that holds a group of assets. Creating a well-balanced portfolio is an effective tool for managing risk. A portfolio manager can use the fact that stock prices depend directly on individual volatilities and not directly on interest rates to generate what mathematically looks like a diversified portfolio. Unfortunately, the risk factors themselves are not independent. High stock volatility usually follows changes in interest rate policy. The correlation of the risk factors in a portfolio can be measured and modeled in different ways. Many of these methods rely on simple distributional (normality) assumptions; if these assumptions hold, the obtained measures perform reasonably well. For example, if the underlying distributions are multivariate normal, you can use the correlation matrix to obtain a measure of joint asset movement. You can also use multivariate

regression analysis to estimate the joint movement. However, in practice these simplifying assumptions often do not hold. Underlying marginal distributions might not be symmetric or normal, or they might have thicker tails. Any departure from normality of marginal distributions causes a problem for multivariate modeling that is based on the normality assumptions. It also causes the estimates of the dependent structure to become unreliable.

De jong et al in (2007) presented an extensive discussion on test of strict stationarity based on quantile indicators. He stated that test for serial correlation and stationarity are mostly applied to the level of a time series. For financial time series, movements in the variance are often of interest; so serial correlation and stationarity tests are applied to squares or absolute values (usually after first removing the mean). However, the distribution of a variable may display changes over time that are not captured by the level or variance. For example, there may be movements in skewness or the tails. Quantiles provide a more comprehensive description of the properties of a variable, and tracking changes in quantiles over time. He stated that, for most distributions with finite variance the indicator-based stationarity test is less powerful than the standard test based on residuals from the mean. This observation leads to generalizing the standard stationarity test to provide alternatives to the quantile indicator tests.

In a European Journal of Operational Research by JukkaRantala and HarriHietikko in 1988, wrote an article on "An Application of Time Series Methods to financial Guarantee Insurance" which comprises a summary of a study made in Finland concerning solvency issues in financial guarantee insurance. The time fluctuation of bankruptcy intensity is analyzed by fitting Box-Jenkins type models to empirical data, and this fluctuation is combined with the variation in the number of claims and the individual claim sizes, based on empirical claim size distribution. The estimated models are used to evaluate, for example, the variance of the claims ratio and of the solvency ratio of the financial guarantee insurer. The variation range of the solvency ratio and the appropriate premium level are discussed with numerical examples.

A paper by Michel Leonard (1994) titled "Promotional Analysis and Forecasting for Demand Planning: A Practical Time Series Approach" has indicated that many businesses use time series analysis for sales promotions to increase the demand for or visibility of a product or service.

These promotions often require increased expenditures (such as advertising) or loss of revenue (such as discounts), and/or additional costs (such as increased production). Business leaders need to determine the value of previous or proposed promotions. One way to evaluate promotions is to analyze the historical data using time series analysis techniques. In particular, intervention analysis can be used to model the historical data taking into account a past promotion. This type of promotional analysis may help determine how past promotions affected the historical sales and can help predict how proposed promotions may affect the future based on similar past promotions.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter deals with the definition of Time Series Analysis, applications of Time Series, the general aspect of Time Series and the review techniques that are useful for the analysis of Time series data.

The ability to model and perform decision modeling and analysis is an essential feature of many real-world applications ranging from emergency medical treatment in intensive care units to military command and control systems. Existing formalisms and methods of inference have not been effective in real-time applications where tradeoffs between decision quality and computational tractability are essential. In practice, an effective approach to time-critical dynamic decision modeling should provide explicit support for the modeling of temporal processes and for dealing with time-critical situations.

3.1 Theory and Definition of Time Series Analysis

Time series unlike the other statistical procedures such as the analysis of random samples of observations, is based on the assumption that successive values in a data file represent consecutive measurements taken at equally spaced time intervals. Time series analysis has the advantage of modeling both linear and nonlinear relationships between variables over time. Most standard data analysis methods used in communication research such as ANOVA and OLS regression assume linear association between variables of interest. Frequently, however, nonlinear functions provide better approximation of the true relationship between communication-related variables.

Time series analysis is an analytical technique that is broadly applicable to a wide variety of domains. In domains in which data are collected at specific (usually equally spaced) intervals, a Time series analysis can reveal useful patterns and trends related to time. Industrial as well as governmental

agencies rely on Time series analysis for both historical understanding as well forecasting and predictive modeling.

There are two main goals of time series analysis: (1) identification of the nature of the phenomenon represented by the sequence of observations, and (2) forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices). Regardless of the depth of the understanding and the validity of our interpretation (theory) of the phenomenon, we can extrapolate the identified pattern to predict future events.

Generally, Time series can be defined as a set of quantitative observations arranged in a chronological order of observations. In Time series, time is assumed be discrete variable. A time series is also defined mathematically as a set of observed values taken at specified times. The set of values is typically denoted "Y", and the set of times as "t1, t2, t3, ... etc". In other words, Y is a function of "t", and the goal of a time series analysis is to find a function that describes the movement of data. It should be noted that a time series is often graphed, and trends and patterns are visually apparent.

A Time Series, such as electric signals and voltage that can be recorded continuously in time is said to be continuous. On the other hand, a Time Series, such as interest rates, yields and volumes of sales which are taken only at specific time intervals, is said to be discrete. Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their education level, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time

series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

The annual crop yield of sugar-beets and their price per ton for example is recorded in a chronological order with time. The newspapers' business sections report daily stock prices, weekly interest rates, monthly rates of unemployment and annual turnovers. Meteorology records hourly wind speeds, daily maximum and minimum temperatures and annual rainfall. Geophysics is continuously observing the shaking or trembling of the earth in order to predict possibly impending earthquakes. An electroencephalogram traces brain waves made by an electroencephalograph in order to detect a cerebral disease, an electrocardiogram traces heart waves. The social sciences survey annual death and birth rates, the number of accidents in the home and various forms of criminal activities. Parameters in a manufacturing process are permanently monitored in order to carry out an on-line inspection in quality assurance. The above series are examples of experimental databases that can be used to illustrate the process by which classical statistical methodology can be applied in the correlated time series framework.

The first step in any time series investigation always involves careful scrutiny of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data. Before looking more closely at the particular statistical methods, it is appropriate to mention that two separate (but not necessarily mutually exclusive) approaches exist in time series analysis, commonly identified as the time domain approach and the frequency domain approach.

Time domain is a term used to describe the analysis of mathematical functions, or physical signals, with respect to time. In the time domain, the signal or function's value is known for all real numbers, for the case of continuous time, or at various separate instants in the case of discrete time. The time domain approach is generally motivated by the presumption that correlation between adjacent points in time is best explained in terms of a dependence of the current value on past values. The time

domain approach focuses on modeling some future value of a time series as a parametric function of the current and past values. In this scenario, we begin with linear regressions of the present value of a time series on its own past values and on the past values of other series. This modeling leads one to use the results of the time domain approach as a forecasting tool and is particularly popular with economists for this reason.

In electronics, control systems engineering, and statistics, frequency domain is a term used to describe the domain for analysis of mathematical functions or signals with respect to frequency, rather than time. A time-domain graph shows how a signal changes over time. Whereas a frequency-domain graph shows how much of the signal lies within each given frequency band over a range of frequencies. A frequency-domain representation can also include information on the phase shift that must be applied to each sinusoid in order to be able to recombine the frequency components to recover the original time signal.

A given function or signal can be converted between the time and frequency domains with a pair of mathematical operators called a transform. An example is the Fourier transform, which decomposes a function into the sum of a (potentially infinite) number of sine wave frequency components. The 'spectrum' of frequency components is the frequency domain representation of the signal. The inverse Fourier transform converts the frequency domain function back to a time function.

There are, obviously, numerous reasons to record and to analyze the data of a time series. Among these are the wishes to gain a better understanding of the data generating mechanism, the prediction of future values or the optimal control of a system. The characteristic property of a time series is the fact that the data are not generated independently, their dispersion varies in time, they are often governed by a trend and they have cyclic components. Statistical procedures that suppose independent and identically distributed data are, therefore, excluded from the analysis of time series. This requires proper methods that are summarized under time series analysis. Methods for time series analysis may be divided into two classes; frequency-domain methods and time-domain methods. The

former include auto-correlation, cross-correlation analysis, spectral analysis and recently wavelet analysis. Auto-correlation and cross-correlation analysis can also be completed in the time domain.

In most other statistics analysis, it is assumed that in time series analysis, the data consists of a systematic pattern (usually a set of identifiable components) and random noise (error) which usually makes the pattern difficult to identify. Most Time series analysis techniques involve some form of filtering out the noise in order to make the pattern more salient.

3.2 General Patterns of Time Series Analysis

Most Time Series patterns can be described in terms of two basic classes of components; **trend** and **seasonality**. The former represents a general systematic linear or (most often) nonlinear components that changes over time and does not repeat or at least does not repeat within the time range captured by our data. The later may have a formally similar nature. However, it repeats itself in systematic intervals over time. These two general classes of Time Series components may coexist in real life data. Other patterns that are seen in Time series analysis are **cyclical** and **irregular** patterns.

3.2.1 Trend Analysis

There are no proven "automatic" techniques to identify trend components in the time series data, however, as long as the trend is monotonous (consistently increasing or decreasing) that part of data analysis is typically not very difficult. If the time series data contain considerable error, then the first step in the process of trend identification is smoothing. The trend of a Time Series, such as registration of members of vehicles can be approximated by a straight line or a non-linear curve. A linear trend equation is used to represent a Time Series data that can be increasing or decreasing by equal amount from one period to another. The linear trend equation can be fitted to a Time Series using the least square method of fitting a straight line. However, if the there is a large number of time

periods, say 12 years and the magnitude of time figures is large, then it is computationally easier to fit the least squares line by using what is called the coded method.

For example, given Y= a + bx as the least square line equation, the estimate of '**a**' and '**b**' in the least square line can be computed and therefore, the forecasting for Y can be estimated. If the Time Series data tend to approximate a straight line trend, the equation develop by the least squares method can be used to predict (or forecast) figures for some future periods. This is done first by coding the year value for which a prediction is to be made. Then, by substituting the coded value to the least square equation, the predicted value can be computed.

3.2.1.1 Smoothing

Smoothing always involves some form of local averaging of data such that the nonsystematic components of individual observations cancel each other out. The most common technique is *moving average* smoothing which replaces each element of the series by either the simple or weighted average of *n* surrounding elements, where *n* is the width of the smoothing "window"

Medians can be used instead of means. The main advantage of median as compared to moving average smoothing is that its results are less biased by outliers (within the smoothing window). Thus, if there are outliers in the data (e.g., due to measurement errors), median smoothing typically produces smoother or at least more "reliable" curves than moving average based on the same window width. The main disadvantage of median smoothing is that in the absence of clear outliers it may produce more "jagged" curves than moving average and it does not allow for weighting.

In the relatively less common cases (in time series data), when the measurement error is very large, the *distance weighted least squares smoothing* or *negative exponentially weighted smoothing* techniques can be used. All those methods will filter out the noise and convert the data into a smooth curve that is relatively unbiased by outliers.

3.2.2 Seasonality analysis

Seasonal dependency (seasonality) is another general component of the time series pattern. It is formally defined as correlation dependency of order *k* between each *i*'th element of the series and the (*i*-*k*)'th element and measured by autocorrelation (i.e., a correlation between the two terms); *k* is usually called the *lag*. If the measurement error is not too large, seasonality can be visually identified in the series as a pattern that repeats every *k* elements.

Seasonality component exhibits a short-term pattern that recurs seasonally. The analysis of a seasonal variation over a period of time can also be useful in evaluating current figures. There are several ways of analyzing a time series in order to isolate the seasonal variation. The most popular one is the method of moving average. The moving average method is used in measuring the seasonal fluctuation of a time series. It can also be used to smooth out fluctuations by moving the mean values through the data.

3.2.3 Cyclical analysis

This type of component is very common with data on business and economic activities. It consists of a period of Prosperity followed by periods of recession, depression and recovery in that order. An important example of cyclical variation is what is called business cycles. In business and economic activities, if variations recur after yearly intervals, then they are considered cyclical.

3.2.4 Irregular analysis

Irregular component or variation in Time Series refers to the odd movements of a time series, which are due to chance. Events that may lead to such odd movements include industrial actions, earthquakes, floods, outbreak of epidemics and many more. Irregular variation is a combination of

episodic and residual (chance) variations. Episodic variation, though unpredictable, can be identified. For example, the effect of an earthquake on a national economy could easily be brought to bear, but earthquake itself could not have been predicted. Residual variation, on the other hand, is unpredictable and cannot be identified.

3.3. Autocorrelation Function (ACF)

Seasonal patterns of time series can be examined via correlograms. The correlogram (autocorrelogram) displays graphically and numerically the autocorrelation function (*ACF*), that is, serial correlation coefficients (and their standard errors) for consecutive lags in a specified range of lags (e.g., 1 through 30). Ranges of two standard errors for each lag are usually marked in correlograms but typically the size of auto correlation is of more interest than its reliability because we are usually interested only in very strong (and thus highly significant) autocorrelations.

3.4 Partial autocorrelations

Another useful method to examine serial dependencies is to examine the partial autocorrelation function (PACF). This is an extension of autocorrelation, where the dependence on the intermediate elements (those within the lag) is removed. In other words the partial autocorrelation is similar to autocorrelation, except that when calculating it, the (auto) correlations with all the elements within the lag are partialled out. If a lag of 1 is specified (i.e., there are no intermediate elements within the lag), then the partial autocorrelation is equivalent to auto correlation. In a sense, the partial autocorrelation provides a "cleaner" picture of serial dependencies for individual lags (not confounded by other serial dependencies).

3.4.1 Removing serial dependency

Serial dependency for a particular lag of *k* can be removed by differencing the series, that is converting each *i*'th element of the series into its difference from the (*i*-*k*)"th element. There are two major reasons for such transformations.

First, one can identify the hidden nature of seasonal dependencies in the series. As mentioned in the previous paragraph, autocorrelations for consecutive lags are interdependent. Therefore, removing some of the autocorrelations will change other autocorrelations, that is, it may eliminate them or it may make some other seasonalities more apparent.

The other reason for removing seasonal dependencies is to make the series stationary which is necessary for ARIMA and other techniques.

3.5 White Noise and IID Noise

White noise is a sequence $\{X_t\}$ of independent and identically distributed random variables such that the random variable has mean of zero and σ^2 variance.

I.e. E (X_t, $X_{t-j} = 0$ for all $j \neq 0$.

A sequence of independent random variable has $X_1,\ldots,X_n$ in which there is no trend or seasonal component and the random variables are identically distributed with zero (0) mean is said to be IID noise.

3.6 ARIMA Methodology

One approach, advocated in the landmark work of Box and Jenkins (1976) is the development of a systematic class of models called autoregressive integrated moving average (ARIMA) models to handle time-correlated modeling and forecasting. The approach includes a provision for treating more than one input series through multivariate ARIMA or through transfer function modeling. The defining feature of these models is that they are multiplicative models, meaning that the observed data are assumed to result from products of factors involving differential or difference equation operators responding to a white noise input. A more recent approach to the same problem uses additive models more familiar to statisticians. In this approach, the observed data are assumed to result from sums of series, each with a specified time series structure, for example, in economics, assume a series is generated as the sum of trend, a seasonal effect, and error. The state-space model that results is then treated by making judicious use of the celebrated Kalman filters and smoothers, developed originally for estimation and control in space applications. Specifically, the three types of parameters in the model are; the autoregressive parameters (p), the number of differencing passes (d) and the moving average parameters (q). The notation introduced by Box and Jenkins is summarized as ARIMA (p, d, q). For example, a model described as (0, 1, 2) means that it contains zero (0) autoregressive (p) parameters and 2 moving average (q) parameters which were computed for the series after it was differenced once.

3.7 Stationarity in Time Series.

In general, it is necessary for time series data to be stationary, so averaging lagged products over time will be a sensible thing to do. With time series data, it is the dependence between the values of the series that is important to measure. Stationarity in Time series has a constant mean, variance and autocorrelation through time (i.e. the distribution does not change over time). In Time series, a stationary series is called weakly stationary if both the mean $U_X(t) = U_X$ and the covariance function $\gamma_X(t+h,t) = \gamma_X(h)$ are independent of time t with h as the lag.

In Time series analysis, the process $\{Yt\}$ is said to be strictly stationary if the joint distribution of $Yt_1, Yt_2, \ldots, Yt_n$ is the same as the joint distribution of $Yt_1 - k, Yt_2 - k, \ldots, Yt_n - k$ for all choices of time points $t_1, t_2, \ldots, t_n$ and all choices of time lag k.

Thus, when $n = 1$, the (univariate) distribution of Yt is the same as that of $Yt - k$ for all t and k; in other words, the Y's are (marginally) identically distributed. It then follows that $E(Yt) = E(Yt - k)$ for all t and k so that the mean function is constant for all time. Additionally, $Var(Yt) = Var(Yt - k)$ for all t and k so that the variance is also constant over time. When $n = 2$ in the stationarity definition, the bivariate distribution of Yt and Ys must be the same as that of $Yt - k$ and $Ys - k$ from which it follows that $Cov(Yt, Ys) = Cov(Yt - k, Ys - k)$ for all t, s, and k.

3.8 Autoregressive model (AR)

In statistics and signal processing, an autoregressive (AR) model is a type of random process which is often used to model and predict various types of natural and social phenomena.

Autoregressive models are based on the idea that the current value of the series, $\boldsymbol{x_t}$, can be explained as a function of $\boldsymbol{p}$ past values, X_{t-1}, and $X_{t-2} \ldots X_{t-p}$, where $\boldsymbol{p}$ determines the number of steps into the past needed to forecast the current value.

Mathematically, a Time series autoregressive model is given by:

$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + Zt$, for any positive integer t where

$\{Zt\} \sim WN[0, \sigma^2]$. $\varphi_1, \varphi_2, \ldots, \varphi_p$ are the autoregressive model parameters and Z_t is the random error and $-1 < \varphi < 1$ for all p. Each observation is made up of a random error component $\{Zt\}$ and a linear combination of prior observations.

3.8.1 Calculation of the AR parameters

The AR (p) model is given by the equation

$$X_t = \sum_{i=1}^{P} \emptyset_i X_{t-i} + Z_t$$

where0_1,..., $\boldsymbol{0_p}$ are the parameters of the model and $\boldsymbol{Z_t}$ is white noise .There is a direct correspondence between these parameters and the covariance function of the process, and this correspondence can be inverted to determine the parameters from the autocorrelation function (which is itself obtained from the covariances). This is done using the Yule-Walker equations. Multiplying the AR (p) model by $\boldsymbol{X_{t-m}}$ results

$$X_t X_{t-m} = \sum_{i=1}^{P} \emptyset_i X_{t-i} X_{t-m} + Z_t X_{t-m}$$

Taking expectation through, we get

$$\gamma_m = E(X_t X_{t-m}) = \sum_{i=1}^{P} \emptyset_i E(X_{t-i} X_{t-m}) + E(Z_t X_{t-m})$$

$$\gamma_m = \sum_{i=1}^{P} \emptyset_i \gamma_{m-i} + \sigma^2 \delta_{m,o}$$

Thus the Yule-Walker equations.

Where m = 0... p, yielding p + 1 equations. γ_m is the autocorrelation function of $\{X_t\}$.

For m = 0,

$$\gamma_O = \sum_{i=1}^{q} \emptyset_i \gamma_{-i} + \sigma^2$$

For instance AR (1):

$$X_t = \emptyset_1 X_{t-1} + Z_t$$

$$\gamma_1 = \emptyset_1 \gamma_0 \text{, thus } \emptyset_1 = \gamma_1 / \gamma_0$$

3.9 Moving Average model (MA)

In time series analysis, the moving average (MA) model is a common approach for modeling univariate time series models. The notation MA (q) refers to the moving average model of order q. *A* moving average model of order q, abbreviated MA (q), is defined mathematically as:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots\cdots + \theta_q \varepsilon_{t-q}$$

This can also be simplify as:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where μ is the mean of the series, the θ1... θq are the parameters of the model and the ε_t, ε_{t-1}... are white noise error terms. The value of q is called the order of the MA model.

That is, a moving average model is conceptually a linear regression of the current value of the series against previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to come from the same distribution, typically a normal

distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with autoregressive models (AR models) because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models. Sometimes the autocorrelation function (ACF) and partial autocorrelation function (PACF) will suggest that a MA model would be a better model choice and sometimes both AR and MA terms should be used in the same model. The moving average model is essentially a finite impulse response filter with some additional interpretation placed on it.

3.9.1 The Backward Shift Operators

The backward shift operator B imposes a one-period time lag each time it is applied to a variable. lt can be applied either to the values of the time series $\{X_t\}$ or to the random errors of the time series. We write $B^1 X_t = X_t - i \forall i = 1 \ldots P$

Examples

Suppose $\{X_t\}$ is a time series for AR (p) is given by

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \ldots + \emptyset_p X_{t-p} + Z_t$$

$$Z_t = X_t - \emptyset_1 X_{t-1} - \emptyset_2 X_{t-2} - \ldots - \emptyset_p X_{t-p}$$

$$Z_t = X_t - \emptyset_1 B X_t - \emptyset_2 B^2 X_t - \ldots - \emptyset_p B^p X_t$$

$$Z_t = X_t(1 - \emptyset_1 B - \emptyset_2 B^2 - \ldots - \emptyset_p B^p)$$

$$Z_t = X_t \emptyset(B)$$

Where $\emptyset(B) = 1 - \emptyset_1 B - 0_2 B^2$; $\emptyset_p B^p$ is a polynomial in B of order p

Similarly for a time series $\{X_t\}$ is a MA (q) of order q such that

$X_t = Z_t + \theta_1 Z_{t\_x} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t\_q}$

$X_t = Z_t + \theta_1 B Z_t + \theta_2 B^2 Z_t + \ldots\ldots + \theta_q B^q Z_t$

$X_t = Z_t(1 + \theta_1 B + \theta_2 B^2 + \ldots\ldots + \theta_q B^q)$

$X_t = Z_t \theta(B)$

Where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots\ldots + \theta_q B^q$ is a polynomial in B of order q .Let's Consider an ARMA (p,q) process which is stationary and $X_t - \emptyset_1 X_{t-1} - \emptyset_2 X_{t-2} - \ldots -$

$\emptyset_p X_{t\_p} = Z_t + \theta_1 Z_{t\_1} + \theta_2 Z_{t\_2} + \ldots + \theta_q Z_{t\_q}$

where: $\{Z_t\} \sim WN[0,\sigma^2]$

using the Backshift operator defined as $B^i X_t = X_{t\_i}$,we have

$\emptyset(B) X_t = \theta(B) Z_t$ and the polynomials $\emptyset(B)$ and $\theta(B)$ have common factors.

3.10 Autoregressive Moving Average Model (ARMA)

In statistics and signal processing, autoregressive moving average (ARMA) models, sometimes called Box-Jenkins models after the iterative Box-Jenkins methodology usually used to estimate them, are typically applied to autocorrelated time series data. Given a time series of data Xt, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the ARMA (p, q) model where p is the order of the autoregressive part and q is the order of the moving average part.

The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving average terms. This model contains the AR (p) and MA (q) models.

An Auto-regressive Moving Average model of order (p, q) abbreviated as ARMA (p, q), is defined mathematically as:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \theta_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

The error terms ε_t are generally assumed to be independent identically-distributed random variables (i.i.d.) sampled from a normal distribution with zero mean- $\varepsilon_t \sim N(0, \sigma^2)$, Where σ^2 is the variance. These assumptions may be weakened but doing so will change the properties of the model. In particular, a change to the (i.i.d) assumption would make a rather fundamental difference.

3.10.1 Specification in terms of lag operator.

In some texts the ARMA (p, q) models will be specified in terms of the lag operator **L**. In these terms then, the AR (p) model is given by:

$$\varepsilon_t = (1 - \sum_{i=1}^{p} \theta_i L^i)\, x_t = \theta_i x_t$$

Where θi represents the polynomial

$$X_t = 1 - \sum_{i=1}^{p} \theta_i L^i$$

The MA (q) model is also given by:

$$X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\, \varepsilon_t = \theta \varepsilon_t$$

Where θ represents the polynomial

$$\theta = 1 + \sum_{i=1}^{q} \theta_i L^i.$$

Finally, the combined ARMA (p, q) model is given by

$$(1 - \sum_{i=1}^{p} \theta_i L^i)\, x_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\, \varepsilon_t$$

Some authors, including Box, Jenkins & Reinsel (1994) use a different convention for the auto-regression coefficients. This allows all the polynomials involving the lag operator to appear in a similar form throughout. Thus the ARMA model would be written as:

$$\theta X_t = q\varepsilon_t$$

ARMA models in general can, after choosing p and q, be fitted by least squares regression to find the values of the parameters which minimize the error term. It is generally considered good practice to find the smallest values of p and q which provide an acceptable fit to the data. For a pure AR model, the Yule-Walker equations may be used to provide a fit. Finding appropriate values of p and q in the ARMA (p, q) model can be facilitated by plotting the partial autocorrelation functions for an estimate of p, and likewise using the autocorrelation functions for an estimate of q. Further information can be gleaned by considering the same functions for the residuals of a model fitted with an initial selection of p and q.

The Yule-walker equation is defined as:

$$\gamma_m = \sum_{k=1}^{p} \theta_k \gamma_{m-k} + \sigma^2_\varepsilon \delta_{m,\,o}$$

Where m = 0... p, yielding p + 1 equations. γ_m is the autocorrelation function of X, is the standard deviation of the input noise process, and $\delta_{m,}0$ is the Kronecker delta function. Because the last part of the equation is non-zero only if m = 0, the equation is usually solved by representing it as a matrix for m > 0, thus getting equation:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \gamma_{-2} & \cdots \\ \gamma_1 & \gamma_0 & \gamma_{-1} & \cdots \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \end{pmatrix}$$

Solving all θ1 for m = 0, we have,

$$\gamma_0 = \sum_{k=1}^{p} \theta_k \gamma_{-k} + \sigma^2_\varepsilon$$

This allows us to solve $\sigma^2\varepsilon$.

The above equations (the Yule-Walker equations) provide one route to estimating the parameters of an AR (p) model, by replacing the theoretical covariances with estimated values. One way of specifying the estimated covariances is equivalent to a calculation using least squares regression of values Xt on the p previous values of the same series.

3.11 Invertibility Requirement

Without going into too much detail, there is a "duality" between the moving average process and the autoregressive process, that is, the moving average equation above can be rewritten (inverted) into an autoregressive form (of infinite order). However, analogous to the stationarity condition (|0| < 1, \ ***9***\ < 1), this can only be done if the moving average parameters follow certain conditions, that is, if the model is invertible. Otherwise, the series will not be stationary.

3.12 Model Identification

A series for ARIMA needs to be stationary, that is, it should have a constant mean, variance, and autocorrelation through time. Therefore, usually the series first needs to be differenced until it is stationary (this also often requires log transforming the data to stabilize the variance). The number of times the series needs to be differenced to achieve stationarity is reflected in the *d* parameter (ARIMA (p, d, q)). In order to determine the necessary level of differencing, you should examine the plot of the data and autocorrelogram. Significant changes in level (strong upward or downward changes) usually require first order non seasonal (lag=l) differencing; strong changes of slope usually require second order non seasonal differencing. Seasonal patterns require respective seasonal differencing. If the estimated autocorrelation coefficients decline slowly at longer lags, first order differencing is usually needed.

However, some time series may require little or no differencing, and that over differenced series produce less stable coefficient estimates.

3.13 Number of parameters to be estimated

Before the estimation can begin, we need to decide on (identify) the specific number and type of ARIMA parameters to be estimated. The major tools used in the identification phase are plots of the series, correlograms of auto correlation (ACF), and partial autocorrelation (PACF). The decision is not straightforward and in less typical cases requires not only experience but also a

good deal of experimentation with alternative models (as well as the technical parameters of ARIMA). However, a majority of empirical time series patterns can be sufficiently approximated using one of the 5 basic models that can be identified based on the shape of the autocorrelogram (ACF) and partial auto correlogram (PACF). The following brief summary is based on practical recommendations of Pankratz (1983); for additional practical advice McDowall, McCleary, Meidinger, and Hay (1980).

1. One autoregressive (p) parameter: ACF - exponential decay; PACF - spike at lag 1, no correlation for other lags.
2. Two autoregressive (p) parameters: ACF - a sine-wave shape pattern or a set of exponential decays; PACF - spikes at lags 1 and 2, no correlation for other lags.
3. One moving average (q) parameter: ACF - spike at lag 1, no correlation for other lags; PACF - damps out exponentially.
4. Two moving average (q) parameters: ACF - spikes at lags 1 and 2, no correlation for other lags; PACF - a sine-wave shape pattern or a set of exponential decays.
5. One autoregressive (p) and one moving average (q) parameter: ACF - exponential decay starting at lag 1; PACF - exponential decay starting at lag 1.

3.14 Estimation and Forecasting

At the next step *(Estimation)*, the parameters are estimated so that the sum of squared residuals is minimized. The estimates of the parameters are used in the last stage *(Forecasting)* to calculate new values of the series (beyond those included in the input data set) and confidence intervals for those predicted values. The estimation process is performed on transformed (differenced) data; before the forecasts are generated, the series needs to be *integrated* (integration is the inverse of differencing) so that the forecasts are expressed in values

compatible with the input data. This automatic integration feature is represented by the letter I in the name of the methodology (ARIMA = Auto-Regressive Integrated Moving Average)

Forecasting is Time series is the process of making statements about events whose actual outcomes have not yet been observed. Forecasting can also be describe as predicting what the will look like or what the future should look. A common place for example might be estimation for some variable of interest at some specific future date. Prediction is a similar, but more general term. Both might refer to formal statistical methods employing time series, cross-sectional or longitudinal data, or alternatively to less formal judgment methods. Usage can differ between areas of application for example in hydrology, the terms "forecast" and "forecasting" are some time reserved for estimates of values at certain specific future time, whiles the term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period. Risk and uncertainty are central to forecasting and prediction which are generally considered good practice to indicate the degree of uncertainty attaching to forecast. The process of climate change and increasing energy prices has lead to the usage of prediction. Forecasting is used in the practice of customer demand planning in every day business for manufacturing companies. The discipline of demand planning, also sometimes referred to as supply chain forecasting, embraces both statistical forecasting and a consensus process. An important, albeit often ignored aspect of forecasting, is the relationship it holds with planning.

3.15 The constant in ARIMA models

In addition to the standard autoregressive and moving average parameters, ARIMA models may also include a constant, as described above. The interpretation of a (statistically significant) constant depends on the model that is fit. Specifically, (1) if there

are no autoregressive parameters in the model, then the expected value of the constant is , the mean of the series; (2) if there are autoregressive parameters in the series, then the constant

represents the intercept. If the series is differenced, then the constant represents the mean or intercept of the differenced series; For example, if the series is differenced once, and there are no autoregressive parameters in the model, then the constant represents the mean of the differenced series, and therefore the *linear trend slope* of the un-differenced series.

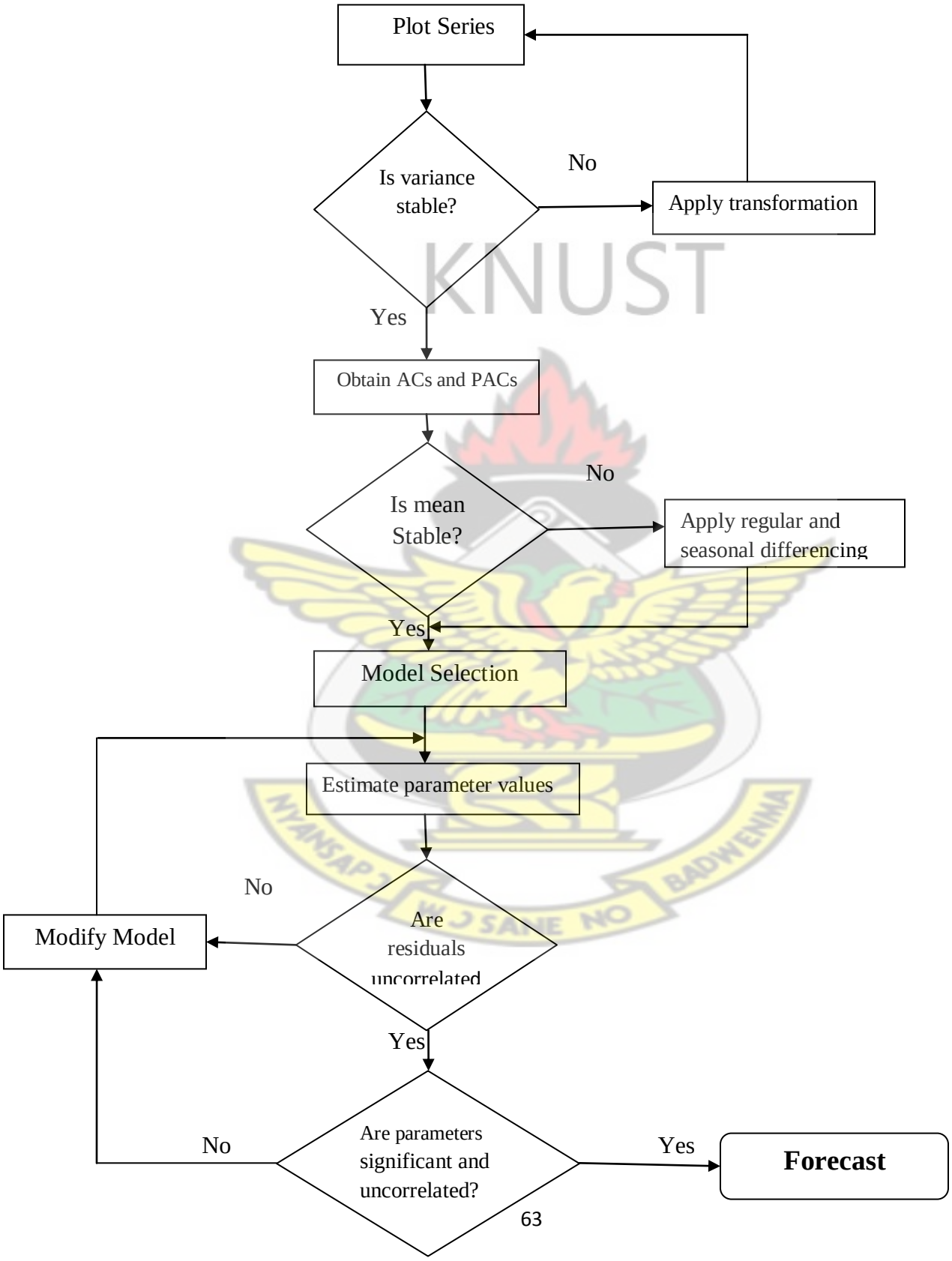
3.16 Box - Jenkins Method

The Box-Jenkins methodology is a statistically sophisticated way of analyzing and building a forecasting model which best represents a time series.

- First stage is the identification of the appropriate ARIMA model through the study of ACF and PACF. For example if the PACF cuts off after lag 1 and ACF decays then ARIMA is identified.
- The next stage is to estimate the parameters of the ARIMA model chosen.
- The third stage is the diagnostic checking of the model. The Q-statistic is used for the model adequacy check.

  If the model is not adequate the forecaster goes to stage 1 to identify an alternative model. This is then tested for adequacy an if adequate, the forecaster goes to the final stage of the methodology
- The final stage is where the forecaster uses the model chosen to forecast and the process end

3.17 BOX-JENKINS MODELING APPROACH



CHAPTER FOUR

DATA ANALYSIS

4.0 Introduction

This chapter looks at the analysis and the information that can be derived from the data collected and then makes inferences based on this information so as to be able to come out with solid conclusions and recommendations. The analysis employs Box-Jenkins method of analyzing Time Series data.

4.1 Data Presentation

The six years data comprising monthly enrolment of clients is purely secondary and was obtained from the office of Savelugu/Nanton District Mutual Health Insurance Scheme. The data spans the number of enrolment recorded from March 2005 to December 2010.

4.2 Preliminary Analysis

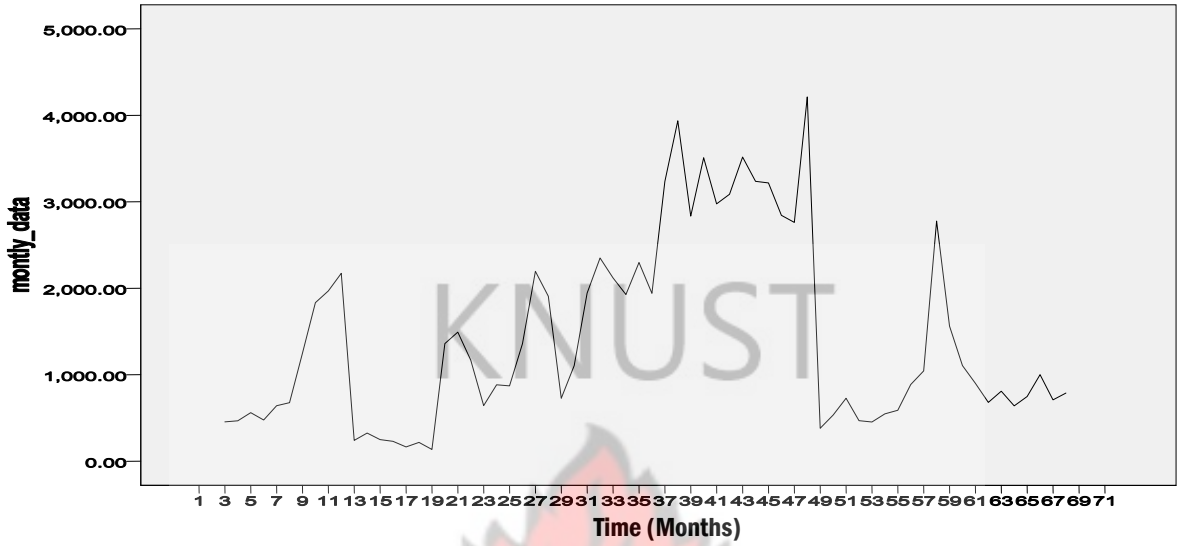


Figure 4.1 Time Series Plot of Monthly enrolment on NHIS data

The Time series plot in figure 4.1 above basically illustrates how the entire data set based on enrolment behaves within equally spaced time interval. It can be observed that the enrolment of individuals does not follow a specific systematic trend. It is also clear from the figure 4.1 that there is no form of seasonality nor periodicity but only an upward and downward moving trend in the number of enrolment. That is, it exhibits an irregular swinging pattern in the data. Therefore, there may be the need for differencing in order to attain stationarity.

From the figure 4.1 above, the Savelugu/Nanton Scheme started recording lowest level of enrolment of 455 clients in March 2005 and had a gradual increase over the months of April through November till it recorded 2,173 clients in December 2005.The number of registered individuals or enrollment decreased sharply by 241 clients in the month of January and increased moderately to July 2006. It then increased sharply from August, September and November all in 2006 until it experienced another sharp decrease in November and December 2006. The

enrolment pattern then continue to experience slightly upward and downward trend over the months of January through November of 2008 forming a zigzag diagram throughout the period under consideration until it registered a sharp increased of 4,213 in December of 2008 representing the highest enrolment figure from the year 2005 to the year 2008. The trend chased a very sharp decrease and begins to increase slowly forming a valley shape until in October where another slightly sharp increase of 2,777 experienced in the year 2009. There after, enrolment of clients continued to experienced slow and gentle slope over the period under study. However, the causes of the decrease were not known from the data.

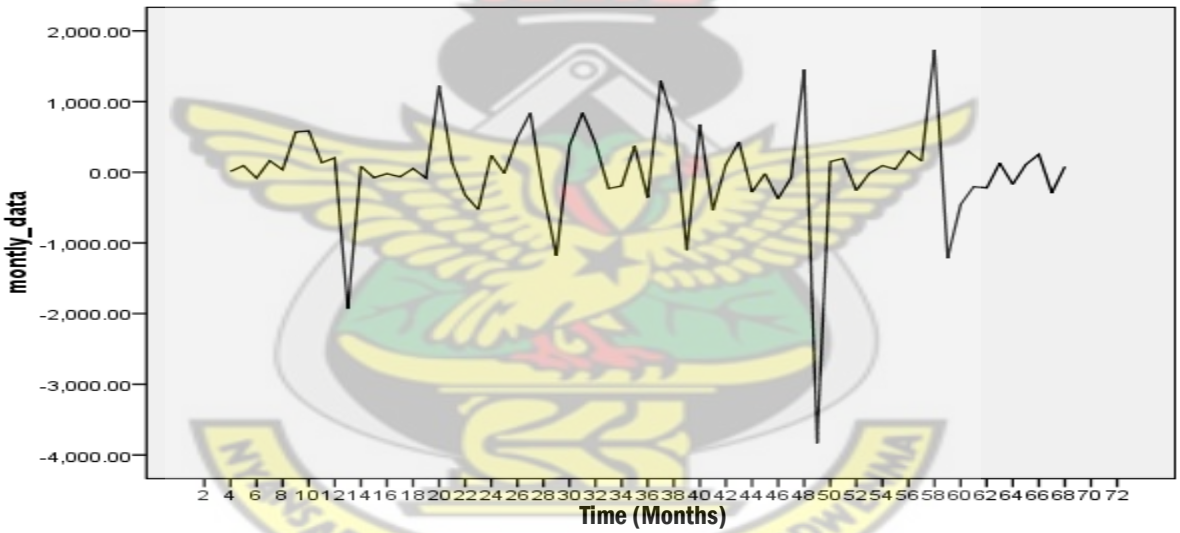


Figure 4.2 Time Series Plot of first order differenced data

The fig 4.2 above is the trajectory of the stationary differenced enrolment data set. It shows that there is no form of periodicity nor seasonality but only mimics the pattern of a random sequence. Here we observed an overall upward linear trend. Since the data was nonstationary and nonseasonal, the differencing has made it stationary and therefore further and precise analysis

can be carry out on the data set. The series moves around a fixed point and it shows that order one (1) differencing of the data sets make the monthly data set and its mean stationary.

After differencing the data, both the autocorrelation and partial autocorrelation functions were computed and is as shown in the table 4.1 and figure 4.3 and 4.5 below.

Table 4.1: ACF and PACF computed for the first sixteen lags.

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic			Partial Autocorrelation	Std. Error[a]
			Value	df	Sig.[b]		
1	-.225	.118	3.631	1	.057	-.225	.120
2	-.116	.117	4.616	2	.099	-.175	.120
3	.012	.116	4.628	3	.201	-.062	.120
4	-.099	.115	5.362	4	.252	-.143	.120
5	.105	.114	6.206	5	.287	.040	.120
6	-.012	.113	6.217	6	.399	-.010	.120
7	-.050	.112	6.415	7	.492	-.040	.120
8	.029	.112	6.484	8	.593	-.002	.120
9	-.392	.111	19.016	9	.025	-.428	.120
10	.418	.110	33.495	10	.000	.269	.120
11	.000	.109	33.495	11	.000	-.002	.120
12	-.097	.108	34.301	12	.001	.000	.120
13	.099	.107	35.165	13	.001	.054	.120
14	-.186	.106	38.253	14	.000	-.130	.120
15	.044	.105	38.430	15	.001	-.024	.120
16	.098	.104	39.323	16	.001	.011	.120

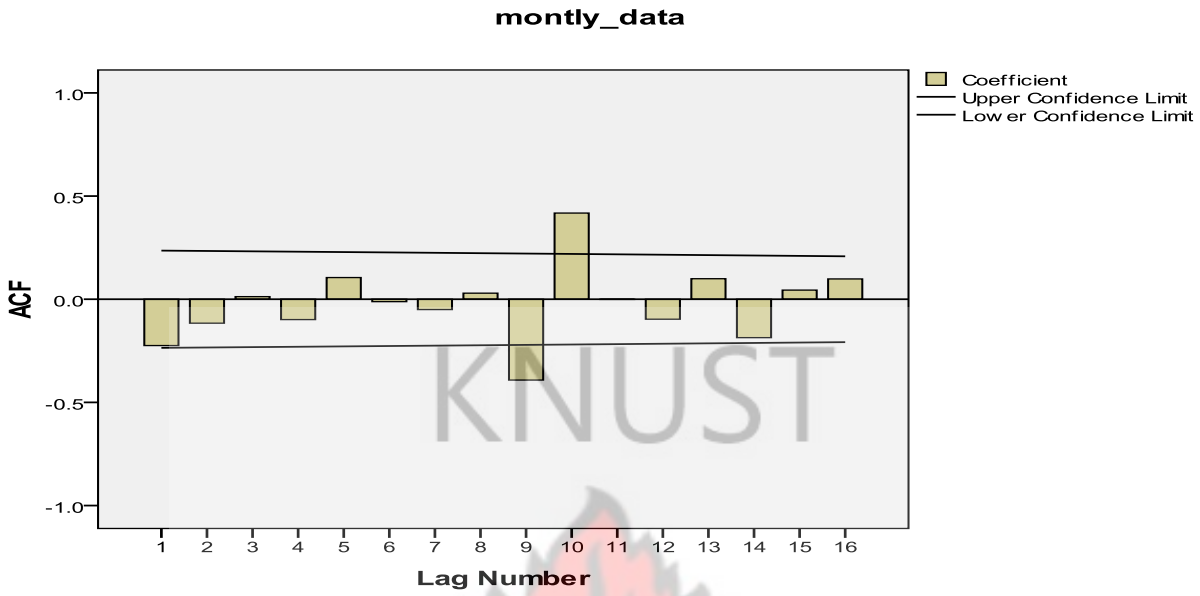


Figure 4.3: ACF Plot of differenced data set

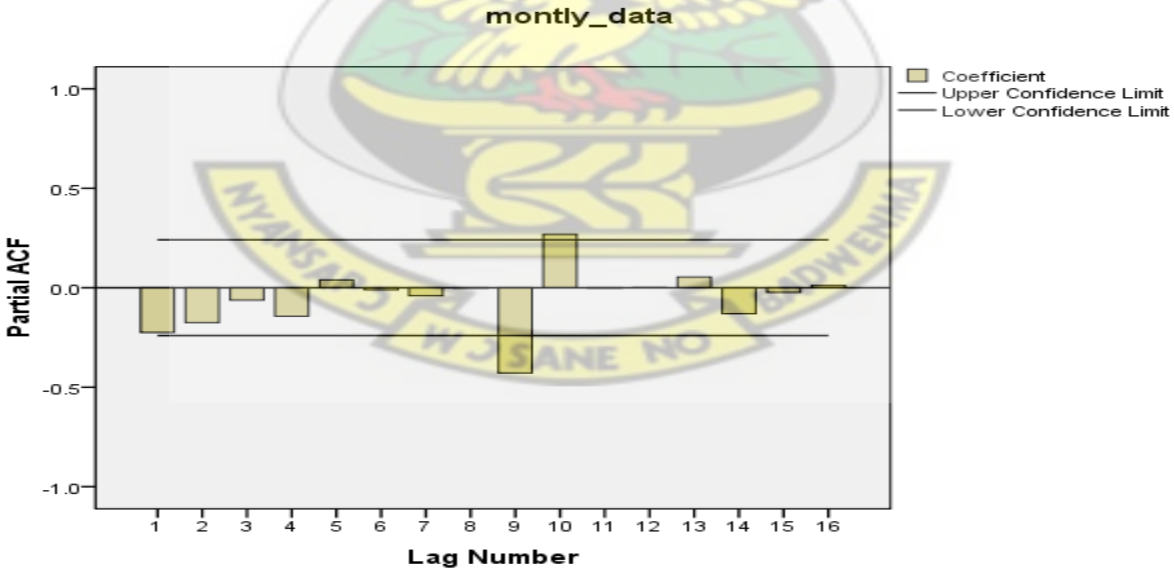


Figure 4.4: PACF Plot of differenced data set

It is observed from the figure 4.3 and 4.4 that, the Autocorrelation Function (ACF) of the first order differenced data set truncates after time lag of one (1) whilst the Partial Autocorrelation Functions (PACF) of the differenced data trailed to zero at lag 6 and continue to truncates at each time lag of one (1).

4.4 Model Selection

ARIMA models are usually estimated after transforming the variable under forecasting into a stationary series. A stationary series is one whose values vary over time only around a constant mean and variance.

Table 4.2: Selection of candidate model for prediction

No.	Model Type	AIC Value
1	ARIMA (0, 1, 1)	13.234
2	ARIMA (1, 1, 0)	13.260
3	ARIMA (1, 1, 1)	13.286

After inputting the data in R, the ***auto-***arima function in the forecast package was used. The ARIMA model (0, 1, 1) produced the least AIC value of 13.234 and hence the best model that fits the data set and can be used for forecasting the monthly enrolment data sets.

ARIMA (0, 1, 1) is therefore given as $y_t = \varepsilon_t - (1+\theta)\varepsilon_{t-1} + \theta\varepsilon_{t-2}$ where, θ is a parameter and ε_t is the residual term.

Table 4.3: ARIMA model parameters

		Estimate	Std Error	t	Sig.
Monthly data No transformation Difference					
Model_1	MA Lag 1	1.321	0.115	2.858	0.005

The table displayed the estimated model parameters that should be used in the model selected. Thus replacing the symbol with the estimated parameters gives $y_t = -1.321\varepsilon_{t-1} + 0.321\varepsilon_{t-2}$ as the appropriate model for predicting the enrolment values of the Savelugu/Nanton District Scheme.

Table 4.4: Testing for model adequacy

Model	Number of Predictors	Model Fit Statistics			Ljung Box-Q (18)			Number of outlier
		Stationary R-squared	R-squared	Normalized AIC	Statistics	DF	Sig	Number of outliers
Monthly_data model_1	0	0.075	0.532	13.234	25.668	17	0.081	0

H_O: Model is inadequate for prediction purposes

H_1: Model is adequate for prediction purposes

From the above, with a Ljung-Box Q Statistic (25.668) and a P-value (0.081), the null hypothesis Ho is rejected and we conclude that the model is statistically significant and adequate for prediction purposes. The co-efficient of determination R^2 = 53.2% means that 53.2% of the total variation in the enrolment values can be explained by the chosen mathematical model

Table 4.5: Predictions of enrollment of individuals

Month	**Predicted enrollment values**	
	2011	**2012**
January	877	738
February	678	669
March	1,170	985
April	850	874
May	989	925
June	651	630
July	637	720
August	995	932
September	1,328	964
October	1,245	988
November	1,799	1,540
December	2,112	929

After testing for the adequacy of the model, it was used to make predictions of enrollment of individuals for 2011 and 2012 as shown in table 4.5 above. It is observed from the predicted values that, enrollment of the individuals has been decreasing from month to month. However, high values of enrollment were recorded in the months of April and July of 2012 compared to those in 2011. The predicted values compared to the values of the previous enrollments of 2005 to 2010, exhibit a downward trend of enrollment of individuals of the scheme. This indicates that as the people get enrolled, the scheme will continue to experience a downward trend from year to year. Also, the model was used to make predictions from 2005 to 2010 and the differences of the predicted values from the actual values were computed as shown in the table at the appendix. It is observed that the differences of the predicted values from the actual values were not very great. This gave the indication that the selected model was good for predictions of the enrollment of the individuals to the scheme.

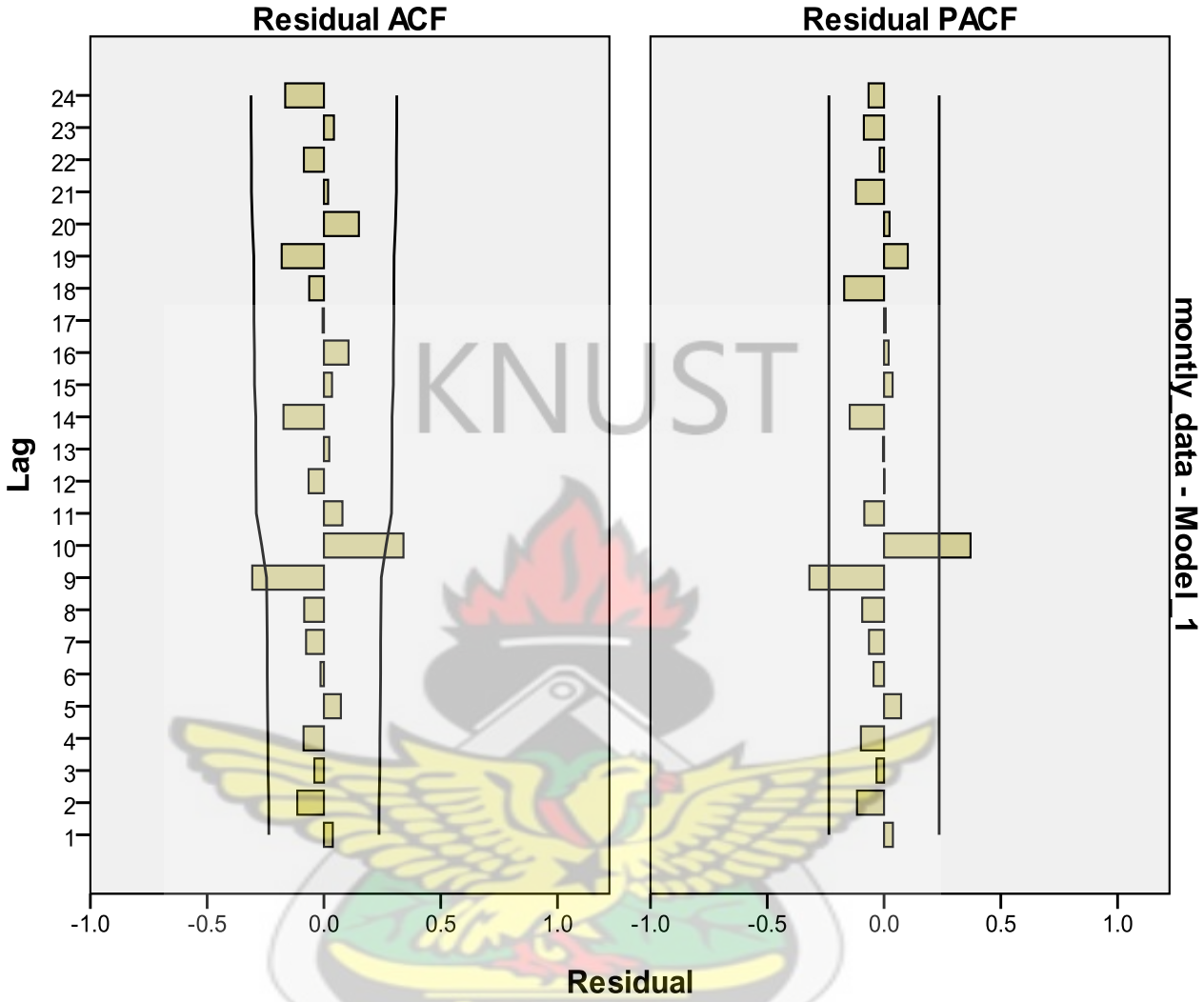


Figure 4.5: Residual Plots of ARIMA (0, 1, 1)

The Autocorrelation and Partial Autocorrelation functions of the residuals confirmed that there is no form of correlation amongst the residuals. Also, the spikes of lags 9 and 10 are due to the random nature of the series. This therefore shows that the selected model is good for prediction purposes.

CHAPTER FIVE

FINDINGS, SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

These chapter summaries the major findings based on the specific objectives of the study and also make recommendations based on the findings.

The core objective for the research was to explore the trend of enrolment over time for the Savelugu/Nanton District Mutual Health Insurance Scheme and get the best ARIMA model that can be used for predictions of enrolment values.

5.1 Findings

The following results were obtained from the research.

- Over the period understudy, it was observed that, since the inception of the Scheme in the year 2005, the highest enrolment (4,213) was record in December 2008.
- Most of the high enrolment values were recorded from August through December of each year.
- The enrolment of clients in the District has an increase trend over the period of inception to the year 2008 though other increase trend was obtained thereafter, this happened after a very sharp drop in December 2008.

- From the predicted values, it is observed that the difference between the enrollees kept reducing over time. This gives an indication that the enrolment of residents will gradually get to a point where majority if not all would have been registered under the scheme.

- The best ARIMA model for the Scheme was fitted and it is given by $y_t = -1.321\varepsilon_{t-1} + 0.321\varepsilon_{t-2}$

5.2 Summary

The core objective for the research as mentioned earlier, was to explore the trend of enrolment over time for the Savelugu/Nanton District Mutual Health Insurance and also to use Box-Jenkins methodology to model stochastic mechanism to assess the quantum of enrolment generated within a specific period and to predict or forecast future values of the series based on the history of the series.

The Box-Jenkins methodology comprises four-step iterative procedure, which includes model identification, model fitting, model diagnostic and forecasting. Since the enrolment figure is non-stationary series, the first difference was taken to make it stationary. The model identification process yielded an ARIMA (0, 1, 1) model for the enrolment series of the scheme. After fitting the model, the Ljung-Box Q Statistic was used to analyze the adequacy of the overall estimated model. The selected model was validated so as to ascertain its adequacy and its goodness of fit. This was done by observing the behavior of the residual plots. Since not all the plots contradict the stated assumptions of normality, the identitified model was considered to be statistically

preferable model to predict the enrolment of the scheme. Finally, the adequate model was used to forecast future time series values for the scheme.

5.3 Conclusion

The overall ARIMA model obtained was $y_t = -1.321\varepsilon_{t-1} + 0.321\varepsilon_{t-2}$. The model was used to make prediction for 2011 and 2012. The predicted values recorded were decreasing from month to month. Findings from the study also indicate that enrollment of individuals to the scheme had experienced an increase and a decrease linear trend from the year 2005 to 2010. The highest enrollment (4,213) from the inception of the scheme was recorded in December 2008. Thereafter, enrollments have been declining gradually from year to year. However, most of other few high enrollment values were recorded from August to December over the period under study and same was seen from the predicted values of 2011 and 2012.

5.4 Recommendations

From the above discussions, it is worth making some recommendations, the adoption of which will make the District Scheme and National Health Insurance Scheme as a whole, more attractive to the people of not the District, but all residents in the country.

- With the downwardly trend, there is the need for the expansion of public education as this can increase the enrollment of individuals in the Scheme.
- The office of the Savelugu/Nanton Scheme should engage themselves in a regular house-to-house registration of the individuals in order to increase their enrolments values.

- It is further recommended that the National Health Insurance Council (NHIC) in collaboration with the National health Insurance Authority (NHIA) should reduce the old age exemption for those in the informal sector to 60 years so that those in that group most of whom are not in active an gainful employment can be covered.

- More sources of funding should be established so that the scheme will be able to pay for the medical bills resulting from the registered clients.

- Everybody should see the National Health Insurance Scheme as a national policy aimed at providing quality healthcare and also as a way of poverty eradication strategy by the government of Ghana.

- Management of the district scheme should allow the people especially the very poor once to pay the premium by installments in order to encourage more residents to register.

- The district assembly in collaboration with the district NHIS office should work together in sensitizing the people through the use of community participation strategies to get greater number of people involve in the education of the importance of the National Health Insurance Scheme.

APPENDIX

PREDICTED ENROLLMENT VALUES FOR SAVELUGU/NANTON SCHEME.

Year/Month		Monthly enrolment values (Actual Values)	Predicted enrolment values.	Differenced data series
2005	March	455	455	0
	April	467	463.43	3.57
	May	561	529.2	31.8
	June	477	494.17	-17.17
	July	641	592.64	48.36
	August	677	649.21	27.79
	September	1,248.00	1,050.76	197.24
	October	1,833.00	1,575.33	257.67
	November	1,969.00	1,839.33	129.67
	December	2,173.00	2,063.09	109.91
2006	January	241	841.19	-600.19
	February	326	495.7	-169.7
	March	250	330.93	-80.93
	April	230	263.25	-33.25
	May	165	197.36	-32.36
	June	218	211.2	6.8
	July	136	160.77	-24.77
	August	1,360.00	964.98	395.02
	September	1,493.00	1,319.07	173.93
	October	1,169.00	1,218.43	-49.43

	November	643	832.55	-189.55
	December	884	867.05	16.95
2007	January	871	869.7	1.3
	February	1,358.00	1,197.16	160.84
	March	2,196.00	1,866.98	329.02
	April	1,907.00	1,893.82	13.18
	May	727	1,111.35	-384.35
	June	1,105.00	1,107.09	-2.09
	July	1,945.00	1,669.00	276
	August	2,350.00	2,125.68	224.32
	September	2,120.00	2,121.87	-1.87
	October	1,927.00	1,991.19	-64.19
	November	2,298.00	2,196.94	101.06
	December	1,941.00	2,025.31	-84.31
2008	January	3,235.00	2,836.53	398.47
	February	3,936.00	3,573.84	362.16
	March	2,834.00	3,077.70	-243.7
	April	3,510.00	3,367.60	142.4
	May	2,976.00	3,104.99	-128.99
	June	3,087.00	3,092.93	-5.93
	July	3,517.00	3,377.31	139.69
	August	3,237.00	3,283.22	-46.22
	September	3,218.00	3,239.48	-21.48
	October	2,843.00	2,973.60	-130.6
	November	2,762.00	2,831.70	-69.7
	December	4,213.00	3,758.00	455

2009	January	381	1,493.37	-1112.37
	February	533	849.34	-316.34
	March	727	767.3	-40.3
	April	469	567.26	-98.26
	May	453	490.64	-37.64
	June	547	528.43	18.57
	July	590	569.72	20.28
	August	886	781.82	104.18
	September	1,044.00	957.64	86.36
	October	2,777.00	2,177.71	599.29
	November	1,562.00	1,764.81	-202.81
	December	1,106.00	1,323.01	-217.01
2010	January	901	1,040.01	-139.01
	February	680	798.59	-118.59
	March	809	805.57	3.43
	April	640	694.54	-54.54
	May	748	730.39	17.61
	June	1,000.00	911.19	88.81
	July	709	775.6	0
	August	789		3.57
	September	734	722.3	31.8
	October	1724	725.21	-17.17
	November	1754	1713.64	48.36
	December	838	929.01	27.79

REFERENCES

1. Kuzin, J. and Howard B. (1992), *Institution features of Health Insurance and their effects on developing country's health systems- International Journal of health planning and management: volume 7, no. 1: pages 5-72.*

2. Arhin D.C (1995), *Health insurance in rural Africa, Lancet: volume 7, no. 3459: pages 44-45.*

3. National Health Insurance Authority (2004), *National Health Insurance policy framework for Ghana, Ghana Government.*

4. Ingilu, C.K. (2005), WHO world alliance for patient safety Conference: Opening speech, Nairobi, Kenya.

5. Ghana Health Service (2004), *Guidelines for design and implementation of Mutual Health Insurance Schemes in Ghana.*

6. Ghana Government (2003), The National Health Insurance Act, Act 650.

7. Venables, W.N. and B.D Ripley (1994), *Modern Applied Statistics with S-Plus, Springer-Verlag, New York, 2nd edition.*

8. Gebhard, K.J.W, (2007), *Introduction to modern time series analysis, Springer-Verlag, Berlin, Germany.*

9. Wei, W.W.S. (1990), *Time series Analysis, 2nd ed. Redwood City, CA, Adisson-wesley.*

10. Benony, K.G and Nathaniel K. H (2000), *Elements of Statistical analysis, City Printers, Accra, Ghana, 1st edition.*

11. WHO (2004), *Bulleting international of public Health: volume 82, No.12, pages 8991-970.*

12. Laloo R. et al (2004), South Africa Medical Journal-Access to healthcare in South Africa, the influence of race and class: volume 94, No. 8, pages 639-642.

13. Wei, W.W.S. (1990), *Time series Analysis, Univariate and Multivariate methods*

14. George Box P.E. and Gwilym Jenkins M. (1976), *Time series Analysis, Forecasting and control. Holden-Day, Oakland, California, USA, 2nd edition.*

15. Ruey-Chyn Tsaur and Ting-Chun Kuo, *An international Journal on "The adoptive fuzzy time series model with an application to Taiwan's tourism demand" Expert system with Applications: volume 38, No. 8*

16. Harri H. and Jukka R. (1988), *European Journal of operational Research on "An application of Time Series Methods to financial guarantee insurance" volume 37, No. 3, pages 398-408*

17. Peter J. B. and Richard A. D. (2001), *Introduction to Time Series and Forecasting, 2nd ed. Fort Collins, Colorado.*

18. Ott, R.L. (1993), *An introduction to Statistical Methods and Data Analysis, 4th ed. Duxbury press, Belmont, California, USA.*

19. Reiss, R.D. (1989), *Approximate Distributions of Order Statistics with applications to nonparametric Statistics, Springer Series in Statistics-New York.*

20. Adsera, A. (2004), *changing fertility rates in developed countries. The impact of labor market institutions. Journal of Population Economics: volume 17, pages 17-34.*

21. Salmen M. and Ploger P (2005), *An international Journal on Echo-state networks used for No. 6, volume 18, page 1953.*

22. Abdel-Aal, R.e and Mangoud A. M. (1998), *Modeling and forecasting monthly patient volume at a primary healthcare clinic using univariate Time Series Analysis.*

23. Box, G. E. P, Jenkins, G. M. and Reinsel, G.C. (1994), *Time Series Analysis: Forecasting and Control*, *Englewood Cliffs, NJ: Prentice Hall*

24. Choi, B. (1992), *ARMA Model Identification*, New York: Springer-Verlag.

25. Foeckens, E. W., Leeflang, P. S. H., and Wittink, D. R. (1999), "Varying Parameter Models to accommodate Dynamic Promotion Effects," *International Journal of Forecasting.*

26. http://www.wikipedia//the free encyclopedia (June 15, 2011)

27. http://www.nhis.gov.gh (June 15, 2011)

28. http://www.4shared.com (July 22, 2011)

29. http://www.sciencedirect.com (July 26, 2011)