CHAPTER 1

INTRODUCTION

1.1 Background to the Study

Malaria contributes substantially to the poor health situation in Africa. It is on record that, sub-Saharan Africa (SSA) accounts for 90% of the world's 300 to 500 million cases of malaria and 1.5 to 2.7 million deaths annually. About 90% of these deaths in Africa are of young children, suggesting some serious demographic consequences for the continent. Malaria is a great burden on the health system in Africa, as it is responsible for 20 to 40% of outpatient visits and 10 to15% of hospital admissions, according to the World Health Organisation (WHO, 1999). In sub-Saharan Africa (SSA), 10.8% of all disability-adjusted life years (DALYs) were lost to malaria in 1990. Again, among the 10 leading factors in DALYs in the world in 1998, malaria ranked eighth with a share of 2.8% of the global disease burden. In SSA however, malaria ranks second after HIV/AIDS, accounting for 10.6% of the disease burden. According to the World Bank, malaria accounted for an estimated 35 million DALYs lost in Africa in 1990 due to ill health and premature death (World Bank, 1993). The estimate was 39 million DALYs in 1998 and 36 million DALYs in 1999 (WHO, 1998, 1999, 2000). Furthermore, while malaria contributed 2.05% to total global deaths in 2000, it was responsible for 9.0% of all deaths in Africa (WHO, 2002).

The WHO also estimated that the total cost of malaria to Africa was US$1.8 billion in 1995 and US$2 billion in 1997 (WHO, 1997). Malaria is therefore a massive problem that affects all segments of society. While its effect on people of all ages is quite immense, the most serious impact of malaria is on pregnant women and children because they have less immunity. When a malaria infection is not properly treated in pregnant women, it can cause anaemia and also lead

to miscarriages, stillbirths, underweight babies and maternal deaths. Also, frequent cerebral malaria can lead to disabling neurological consequences. With regard to school children, malaria is a major cause of absenteeism in endemic countries. It is estimated that about 2% of children who recover from cerebral malaria suffer brain damage including epilepsy (WHO/UNICEF, 2003). Hence, among young children, frequent episodes of severe malaria may harm their learning abilities and educational attainment. This is a threat to human capital formation, which constitutes a key factor in economic development. The debilitating effects of malaria on adult victims are very disturbing. In addition to the time and money spent on preventing and treating malaria, it causes considerable pain and weakness among its victims, thereby affecting their ability to work. The adverse impact of the disease on household production and gross domestic product can be substantial. Malaria therefore is not only a public health problem but also a developmental problem.

At the national level, apart from the negative effect of lost productivity on the major sectors of the economy, malaria has negative effects on the growth of tourism, investment and trade, especially in endemic regions.

It constitutes a major socio-economic challenge to African countries since they are the region most affected by the disease. This challenge must be faced with resolve since good health is not only a basic human need but also a fundamental human right and a prerequisite for economic growth (Streeten, 1981).

The malaria burden is a challenge to human development. It is both a cause and consequence of under-development. As has been observed elsewhere, the disease is not homogeneous and uniform in Ghana. It is a localized problem with great differences from place to place. Prevalence shows diversities and variations to the extent that neighbouring communities can

display complete difference in transmission patterns. This is due to a combination of factors: meteorological and ecology of both human host and the vectors and proximity between human habitat and the breeding places. The peak period of malaria transmission occurs during the rainy season and often coincides with the peak period of agricultural activities such as planting and harvesting. Parasite strains appear to be including an increasing human tool with services on productivity. There is some reduction during the dry season.

1.2 Problem Statement

The malaria burden is a challenge to human development. It is both a cause and consequence of under-development. In Ghana, malaria is the number one cause of morbidity, accounting for 40 to 60% of outpatient visits. It is also the leading cause of mortality in children under 5 years of age, a significant cause of adult morbidity, and the leading cause of workdays lost to illness.

Despite the devastating effects of the disease, the importance of a malaria-free environment in promoting economic development and poverty reduction has not been fully appreciated in Ghana. Perhaps the impact of malaria has not been demonstrated in quantitative terms that might convince politicians, policy makers, programme managers and development partners to devote the needed attention and resources to combating this dreadful disease. This study is an attempt to fill this gap with appropriate information.

1.3 Objectives of the Study

The objectives of the study are to:

(i) provide us clues about the behaviour of the time series data

(ii) develop a statistical model that will aid in forecasting the incidence of the disease in the district.

1.3.1 Methodology

In the study district, the Government Hospital was contacted for data on morbidity of malaria. The data for monthly reported cases of malaria fever between the years 2001 to 2010 was obtained from the statistical Department of Effiduase Government Hospital. The KNUST library and the internet were other sources of vital information to this project. Data from January 2001 to December 2010 would be used for comparing the mean monthly reported cases. The malaria data from 2001 is taken as the based year and its mean reported cases would be compared with the mean reported cases of the other years (2002 to 2010) to see the trend of occurrence of the disease over the years under consideration. The two-sample t-test would be used for comparing the means of the reported cases. Minitab 16 is used to compare average monthly reported cases. Data from January 2001 to December 2010 would be used for developing a time series ARIMA model for predicting the number of reported cases in the district. Values for the months May and June 2010 would be reserved for post time series forecast. An SPSS version 16.0 is used in the formulation of the ARIMA model. Minitab is used to plot the graphs. Minitab software is used to determine the trend equation.

By the application of Box-Jenkins method, the data was analyzed and used to identify and select the best ARIMA model.

1.3.2 Scope

The study is confined to the morbidity of malaria cases reported at the Effiduase Government Hospital in the Sekyere-East District. The project is seeking information on monthly outpatient morbidity returns for ten (10) consecutive years (2001-2010) of the disease.

1.3.3 Justification of the Research

Several programmes have been initiated in this country to combat this disease of the poor people but its prevalent rate is still high and accounts for 40% out of the 70% communicable disease in Ghana in 2008. The National Malaria Control Programme (NMCP) is the mother agent for controlling the disease in this country. Other companies like Zoomlion Company Limited and AngloGold Ashanti have also joined in the fight against the disease.

Considering the concomitant loss of lives, cost in the medication and loss of productive hours, it requires renewed commitment from the government, non-governmental organizations and all and sundry to fight for complete eradication of the disease. Since the cost of treatment of the disease is directly proportional to the size of the potential benefits to be derived for the country, for a successful malaria control programme, this study will try to identify areas of high prevalence of malaria in the district. Indeed, very few research studies have been carried out on the incidence of malaria in the Ghanaian context to date. In particular, very little is known in the district concerning this very important subject. Based on available empirical evidence, it is necessary to furnish decision makers and other stakeholders with vital information regarding the incidence of malaria in the district for possible policy interventions. Additionally, it is important to contribute to knowledge on the incidence of malaria with a view to, among other things, stimulate further research.

1.3.4 Limitation of the Study

This project has successfully been accomplished not withstanding constraints encountered. Inadequate logistical arrangement to facilitate collection and storing information constituted a major drawback. Officers of the statistical department of the hospital had to search through bulk documents for the needed information. This made it a hard task for me in obtaining the required

data. Time allotted to carry out the study wasn't sufficient enough to cover a broader scope like the whole of Ashanti Region or country wide. The high cost of gathering and collection of data also prevented me form covering a wider scope to achieve a desirable and representative result.

1.3.5 Thesis Organization

Chapter 1 is the introduction which comprises the background to the problem, statement of the problem, objectives of the study, justification of the research, methodology and limitation.

Chapter 2 basically deals with the review of literature on malaria and time series modelling.

Chapter 3 deals with the methods used in the analysis. It comprises of the introduction, describing basic statistics, method and concept of time series and Box-Jenkins methodology.

The data used in the study and characteristics of the study area, data analysis and results are considered in Chapter 4.

The summary and conclusions including a discussion of the policy implications of the study are presented in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

2.1 MALARIA

A lot of people have done research on malaria. This chapter deals with the review of some literature works on malaria.

In humans, malaria is caused by four species of the plasmodium protozoa (single celled parasites) – plasmodium falciparum, plasmodium vivax, plasmodium ovale and plasmodium malariae. Of these species plasmodium falciparum accounts for the majority of infections and is the most lethal.

Several studies have been done on different aspects of the disease, from parasitology to finding a cure with drugs (chemotherapy) and to eradication of the disease by the use of insecticide treated net and insecticides.

Rashed, S. et al. (2000), conducted a study which was aimed at determining the effect of Permethrin insecticide treated nets (PITN) use on the incidence of febrile episodes and non household malaria expenses in Benin.

The study found out that, the use of PITNs decreased the risk of developing malaria by 34% in children in the rural areas; meanwhile, PITN use did not reduce prevention and treatment expenses. In a parasitology laboratory, malaria was found to be the major killer of paediatric illness and death in Kinshasa (Coene, 1991). In view of this, the treatment of fevers as malaria with chloroquine is no longer acceptable because the plasmodium falciparum had a resistance to chloroquine. According to the study, the differences in endemicity of malaria that existed between the various parts of town had to be taken into consideration alongside the ecological and socio-economic factors that underlie when planning for estimation of potential control methods.

The behavioural risk for malaria in the Machodinho resettlement area in the Amazonian forests of Brazil was examined (Castilla and Sawyer, 1993). Analysis of the study suggested that economic status and knowledge of the importance and behaviour of the mosquito in transmitting malaria are significant factors in determining prevalence risk irrespective of whether preventive precautions, for example, dichlorodiphenyl trichloroethane (DDT) spraying of houses and cleaning of vector breeding sites are to be undertaken in the endemic areas. However, the researchers found out that a higher economic status combined with better knowledge of the vector and DDT spraying of houses decreased the risk of infection. They suggested that a more positive implication is that control programmes must work harder and more intensively on behalf of poorer people especially migrants in order to diminish the disease burden for them. Sharma and colleagues (2001) carried out a study on the socioeconomic factors as well as on the human behaviour towards malaria on cross section of the Sundargarh district in India. They argued that poor socioeconomic status and socio-cultural factors play an important role in maintaining high degree of malaria transmission. They found that human behaviours such as location of hamlets, type of malaria transmitted, sleeping habits, and outdoor activities after dusk, poor knowledge about the disease and treatment seeking behaviour are of great significance as determinants of malaria transmission.

Malaria is also a major problem in Papua New Guinea as it accounts for a high proportion of sickness and death. This is because in addition to human suffering, it also put severe stress on the health facilities and directly hinders economic growth. It has been suggested that a malaria vaccine would be best, most cost effective and safe public health measure to reduce the burden of malaria (Reeder, 2001).

Whitty and Allan (2004) contend that the serious threat posed by the spread of drug-resistant malaria in Africa has been widely acknowledged. Chloroquine resistant malaria is now almost universal and resistant to successor drug, sulfadoxine-pyrimethamine (SP) is growing rapidly. If the question of cost of treatment is not successfully addressed this could lead to adverse result from the deployment of combination therapy as a first-line treatment. Adverse effect of costly treatment ranges from increase in delays in infected individuals presenting themselves to the health care facilities for treatment to exclusion of the poorest malaria sufferers from receiving treatment altogether.

Malaria has been one of the most prominent and ancient diseases which has been profiled and studied. It has been one of the greatest burdens to mankind, with a mortality rate that is unmatched by any other modern disease other than tuberculosis (Sudhakar et at., 2007). It remains the leading cause of death in children under five years in Africa (Houeto et al., 2007).

Malaria is one of the leading killer diseases in the tropical and subtropical countries. It therefore poses a serious health problem to these countries including Ghana. This disease is frequently called disease of the poor because its prevalent rate is very high in the poorest continent and in the poorest countries (Worral et al., 2003).

According to Nchinda, (2005) sub-Sahara Africa was never part of the global malaria eradication programme because the period coincided with colonial and immediate postcolonial period and so the indigenous had little or no power to initiate and sustain an eradication programme.

This review looks at the factors affecting the prevalent rate of this dreadful disease and some of the control measures which have already been initiated to control the disease. There are combination of factors which contribute to the resurgence of this disease in Ghana and Africa as a whole.

Poverty contributes to the prevalence of malaria in Africa. According to Pattanayak et al., (2003) many of the world's poorest people live in areas of high rates of malaria. These people do not have access to effective health care due to financial constraint. Worral et al,(2003) called malaria as a disease of poverty. The economic status of a vulnerable country plays another role in determining the equippedness and control measures in case of epidemics (Sudhakar et al., 2007).

A survey in Zambia found a substantially higher prevalence of malaria infections among the poorest population group. Poverty compels people to move from non-endemic areas to endemic areas in their quest to search for jobs. As with other diseases, malaria has unequal effect on different members of the population; pregnant women and children are most susceptible. Given that the intensity of malaria transmission and therefore the likelihood of control depends on the relative abundance of and contact patterns among susceptible, infected, infectious and immune individuals, it is essential to target mothers and children in treatment (Guerin et al., 2002). The immediate economic burden on households from losing mothers is devastating, whereas childhood malaria imposes future burdens (Pattanayak et al., 2003).

Another contributing factor to the high morbidity of malaria is altitude. Research shows that there is high prevalent rate of malaria in low altitude areas and the prevalent rate is low in high altitude areas. According to Wiwanitkit (2006), a previous research on altitude and malaria mosquito prevalence in Thailand indicated high prevalence of the disease in low altitude. However people living in high altitude areas may experience high prevalent rate of malaria if they create the enabling environment for the breeding of the vector. Environmental factors hinder efforts to control the disease. Inhabitants of houses surrounded by bushes or garbage heaps and swamps or stagnant water showed higher malaria prevalence and densities as compared with

those from cleaner surroundings (Nkuo-Akenji et al, 2006). Irrigation and deforestation have affected the transmission of the disease. Irrigation requires the construction of dams which serve as fertile breeding grounds for the parasite. In Sri Lanka, the construction of hydropower dams on Mahaweli River created pools with sandy and rocky nature, which are suitable for the breeding of anopheles culicifacies, the primary malaria vector for the country (Nkuo-Akenji et al., 2006). Deforestation changes the ecology of the vector and its option for the host (Pattanayak et al., 2003). Whereas the forest floor in primary growth tends to be heavily shaded and littered with a thick layer of organic matter that absorbs water and renders acidic, clear lands are generally more sunlit and prone to the formation of puddles with more neutral pH, which can favour specific anopheline larvae development (Patz et al., 2000).

According Lindsay et al., (2004), deforestation is one of the most potent factors at work in emerging and re-emerging of infectious diseases. Mining causes deforestation and environmental degradation. Mining pits dug during land dredging mining creates stagnant water pools serving as breeding ground for mosquitoes and other water-borne diseases (Wiwanitkit, 2009). In Kanchanaburi, Thailand the primary forest malaria vector, An. dirus increased mainly because breeding places were created by excavation work (Wiwanitkit, 2006). In sub-Sahara Africa, climate change has several features that could influence the prevalent rate of malaria. Rising temperature can extend the habitat of mosquitoes, shifting the boundaries of latitude and altitude for malaria transmission. Highland areas in Burundi, Kenya and Uganda which initially were malaria-free are now experiencing epidemics (Sulaiman, 2007). Floods and drought also have impact on the incidence of malaria. Drought leads to the formation of pool of stagnant water which creates a favourable habitat for the parasite. Relative humidity affects the transmission of malaria. It affects the survival of the vector.

The resistance of the parasite to anti-malaria drugs especially chloroquine is a major cause in the re-emerging of the disease. Resistance evolves through fundamental principles of natural selection and evolution, including diverse factors such as extent of treatment, nature and site of antibiotic action or genomic complexity of the parasite (Wilson, 2001). Perhaps the biggest threat to malaria control- be it prevention or treatment- is the increasing resistance to pesticides and drugs. Optimal control and treatment maximize the useful life span of insecticides and drugs. Resistance is more likely to emerge when background immunity is weak, parasite numbers in individuals are high, transmission is low, and insecticides and drug pressure is intense. Plasmodium falciparum has become variably resistant to all drug classes except the artemisinin derivatives. Multiple economic factors influence the inappropriate use of drugs and insecticides (Reed et al., 2002).

Since the discovery of the disease about 4,000 years ago, several control measures have been put in place to curb it but the incidence of the disease is still high in sub-Sahara Africa. The United States and some Europeans countries have been able to eradicate it through the use of insecticide and manipulation of the environment (Nkuo-Akenji et al., 2006). United States launched the National Malaria Eradication Program on 1st July 1947. Over 4,650,000 houses were sprayed by the end of 1949. In 1947, the malaria cases reported were 15,000 and reduced to 2,000 in 1950. The disease was considered eradicated in 1951.

The World Health Organization launched the global eradication of malaria in 1955. Unfortunately, this coincided with the struggle for independence in sub-Sahara Africa. Countries with temperate climates succeeded in eradicating the disease. Countries like India and Sri Lanka had sharp reduction in morbidity. However countries like Indonesia, Afghanistan, Haiti and Nicaragua made negligible progress (CDC, 2004).

The Roll Back Malaria (RBM) initiative launched in 1998 has the ambitious target of decreasing malaria mortality by 50% by the year 2010. Although several control and preventive measures will contribute to the achievement of this target, an essential contribution needs to come from a substantial reduction of the case-fatality rate for the disease (WHO, 2002). African leaders met in Abuja in 2000 to reaffirm their commitment to the RBM. The goals of the Abuja Declaration include ensuring that 60% of those with malaria have access to treatment within 24 hours of the onset of the symptoms; at least 60% of the at-risk pregnant women receive preventive drugs and at 18 least 60% of the at-risk sleep under bed nets (The African Summit on Roll Back Malaria, 2005). After nine years of implementing the Abuja Declaration, it appears Ghana is making a negligible success since the morbidity rate is still high and in Effiduase malaria is still the most common disease recorded daily at the health centres. The ability to prevent and treat the disease is a function of one's income so the issue of affordability in terms of treatment and the acquisition of the nets should not be downplayed if modest gain is to be achieved different from the others.

2.2 TIME SERIES AND MALARIA STUDIES

A mathematical model can help respond to the increasing threat of malaria in the district. The man who first discovered that malaria is transmitted though mosquitoes, Sir Roland Ross, developed the first mathematical model for malaria transmission in 1911. In presenting his model, Ross pointed out that "the mathematical method of treatment is really nothing but the application of careful reasoning to the problem at issue".

Autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models are some of the fundamental models for forecasting time series data. The MA model assumes that the series depends linearly on its

previous values and normally distributed error term. The autoregressive, model attempts to forecast the time series values based on historical data. The moving average component models the error terms which are correlated. Non-stationarity can be decreased by differencing the series with specific time lag. The main objective of a time series is to develop statistical model explaining the behaviour of a random variable changing over time which allows making future estimations of the said random variable (Suarez et al., 2009). Analysis of time series may serve a number of purposes. Often the main interest lies in the regressive model, for example relating infection incidence to staffing levels or antibiotic usage data. Other applications include forecasting and the development of the alert systems to detect periods or places where transmission exceeds some threshold (Brown et al., 2002).

Regarding planning in future needs in any system, traffic prediction accuracy is really important when defining required future capacity and planning any changes. A fairly accurate time series model could predict several years in the future, this being an advantageous skill when planning future requirements (Fillatre et al., 2003).

Lin et al., (2009) used time series analysis to investigate the relationship between the falciparum malaria in the endemic provinces and the imported malaria in the non-endemic provinces of China. An autoregressive integrated moving average model was first fit to the predictor variable. Of all the models tested, the seasonal ARIMA (1, 1, 1) and (0, 1, 1) model for malaria incidence fit the data best according to the according to AIC and goodness-of -fit criteria.

Briet et al., (2008) formulated a model for short term malaria prediction Sri Lanka. Exponentially moving average models, autoregressive integrated moving average models with seasonal components and seasonal multiplicative autoregressive integrated moving average (ARIMA) models were compared on monthly time series of district malaria cases for their ability

to predict the number of malaria cases one to four months ahead. The best model for forecasting and forecasting error varied strongly among the districts for instance, for the district of Ampara, for a one month forecasting horizon, the best model was an ARIMA (2, 1, 1) with seasonality through a harmonic with a period of one year and a harmonic with period of six months. For further forecasting horizons, the ARIMA (0, 1, 2) model with seasonality through a first order seasonal autoregressive and a first order seasonal moving average component was best for the district of Ampara. Contreras et al., (2003) developed a model for predicting the next-day electricity prices in mainland Spain and California markets using an ARIMA model. Their developed model was able to forecast the 24 market clearing prices of tomorrow. The ARIMA model is an effective tool for forecasting time series.

A good model is to be developed for forecasting the malaria cases. A model fitting quality is defined as the sum of the residuals‟ squares divided by the sample size. Its objective is to measure the model's capacity to produce the sample data (i.e. to verify how similar the modelled series and the actual series really are) (Guerrero, 2003).

CHAPTER 3

3.0 METHODOLOGY

Statistical forecasting methods fall under two major categories: Qualitative forecasting and Quantitative forecasting methods. The Qualitative methods are generally subjective in nature; they rely on opinions of experts for formulating the relationship. The Quantitative methods, on the other hand, involve statistical analysis of the historical data in an attempt to identify either the true mathematical relationship between the historical data or that relationship which is reasonably close to the true relationship. These methods can be further classified into two sub-types: Explanatory or Causal methods and Time series methods. Explanatory methods investigate the presence of other variables which affect the variable of interest. These other variables, called inputs, are then analyzed and a suitable relationship between the inputs and the variable of interest is formulated. The Time Series Methods are 'stand alone' methods. Time Series Methods investigate the historical pattern present in the variable of interest and, assuming that it will continue in future, use this association to predict the future values. The choice of the method depends on specific cases and the availability of data.

This chapter deals with the concept of time series, an overview of the Box-Jenkins ARIMA methodology, a sophisticated time series analysis technique, and time series methodologies. ARIMA methodology is the main tool used in identifying the forecasting models of chapter four. Software packages Minitab and SPSS are used as the platform to identify these models.

Section 3.1 deals with the concept of time series, 3.2. Introduces the basic terminology of Box-Jenkins methodology and the various classes of ARIMA models, followed by a discussion on the concept of 'Stationarity' of time series in section 3.3. In Section 3.4, the three-stage iterative method of fitting Box-Jenkins ARIMA models to a time series is explained, along with a discussion on the various statistical tests to measure the goodness of fit of such models. Section

3.5 looks at the procedure of forecasting using general ARIMA models. Seasonal ARIMA models and ARIMA transfer function models are covered in sections 3.6 and 3.7 respectively.

3.1 Concept of Time Series

Time series is the set of observations on a variable of interest that has been collected in time order that is, daily, weekly, monthly, etc. In other words it is a time dependent sequence. If the time series can be predicted exactly, it is said to be deterministic for example a person's salary may be determined according to the number of years worked but most time series are stochastic in nature in that the future values are determined based on the past values.

3.1.1 Components Of Time Series

Traditional time series are mainly concerned with decomposition. All time series contain at least one of the following four components: trend, cyclical, seasonal and irregular variations.

3.1.2 Trend

This refers to the general direction in which the graph of time series appears to be going over a long interval of time. In other words, it is a long-term growth or decay.

A deterministic trend model with a seasonal effect can take either an additive form,

$$X_t = m_t + s_t + Y_t, \qquad t = 0,1, \dots n, \tag{1.1}$$

or a multiplicative form, such as

$$X_t = m_t s_t Y_t, \qquad t = 0, 1, \dots, n \tag{1.2}$$

or a mixed form,

$$X_t = m_t s_t + Y_t, \qquad t = 0, 1, \dots, n \tag{1.3}$$

where $m_t = m(t)$ is a (usually slowly changing) function of time, so called 'trend

component', $s_t = s(t)$ is a periodical function of time and Y_t is a random noise component.

Model (1.2) can be easily transformed to the additive form by taking a logarithm of both sides.

Model (1.3) is often referred to as a multiplicative one.

3.1.3 Cyclic

This is the long-term oscillations or swing, about a trend line curves which may be periodic and may not be equal intervals of time. It is a wave-like fluctuation about trend. The length and the amplitude of the cycle are not constant as in the seasonal component but may vary from one to the next.

3.1.4 Seasonality

This refers to the identical or almost identical patterns which a time series appear to follow. It is a periodic change usually in year cycles. In other words, a regular recurring variation or fluctuation.

3.1.5 Irregular variations

The irregular component of a time series is the residual factor that accounts for the deviations of the actual time series value from what we would expect if the trend, cyclical and seasonal components completely explain the time series . It is caused by a short term unanticipated and non-recurring factors such as wars, earthquakes, floods and so that affect the time series. Since this component accounts for the random variability in the time series, it is unpredictable that is we cannot predict its impact on the time series in advance.

3. 2 Box-Jenkins ARIMA models: Definition and Terminology

Box-Jenkins ARIMA models use historical values of a single variable to forecast its future values; hence they are classified as univariate methods. The variable of interest must be

separated by equally spaced time intervals to apply Box-Jenkins methodology. Let's consider a discrete time series of n equally spaced observations in time:

$$y_t = y_1, y_2, y_3, y_4, \cdots, y_{n-1}, y_n$$

The basic essence of Box-Jenkins methodology is that it considers the observed time series, y_t, to be the outputs of an unobservable 'black box' process. The inputs to this black box are a series of independent random shocks e_t , as illustrated in Figure 3.1.

Figure 3-1: Box-Jenkins Black-Box Process

For statistical purposes, these random shocks are assumed to be normally distributed with zero mean and a constant variance. This sequence is typically referred to as 'white noise'. Thus, the Box-Jenkins approach views a time series as the result of transformation of a white noise process using a black box, which is nothing more than a linear filter.

In essence, the ARIMA model assumes that the outputs (observed time series values) may depend on:

1) The previous and current inputs (white noise or random shocks).

2) The previous output values of the time series under study, $y_{t-1}, y_{t-2}, \cdots$ in varying proportion.

How much each of these will determine the future output will depend on their associated coefficients.

Specifically, the Box-Jenkins approach proposes a simple linear form for the observed time series values:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \ldots - \theta_q e_{t-q} \quad (3.1)$$

or, $$\Phi(B)(1-B)^d y_t = \Theta(B) e_t \quad (3.2)$$

where $\Phi(B) = (1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$,

$\Theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$

$By_t = y_{t-1}$, B is the Backward shift operator ($By_3 = y_2, By_9 = y_8$ ..., etc)

d = order of differencing

Equation 3.1 shows the current output as a linear weighted sum of previous outputs and inputs. Note that only "*p*" nonzero output terms and "*q*" nonzero input terms are required to produce the current output. Thus only a finite number of recent inputs and outputs will have a statistically significant effect on the current output.

The general notation of ARIMA models is ARIMA (*p,d,q*), where " *p* " is the order of Autoregressive component, " *d* " is the order of differencing used and " *q* " is the order of Moving Average component in the model. The Autoregressive and Moving Average components are described below and the concept of differencing is described in the next section.

Depending on the above definition, the ARIMA models can be classified into:

1**) Autoregressive (AR) models:**

When the value of the current output y_t depends solely on p prior outputs and the current input (random shock) e_t , the Box-Jenkins model takes the form of

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t \quad (3.3)$$

or $\Phi(B) y_t = e_t$ (3.4)

and is called an Autoregressive model of order p, denoted by AR (p) or ARIMA (*p, 0, 0*).

2**) Moving Average (MA) models:**

When the current output y_t depends solely on the current input and q prior inputs, the Box-Jenkins model takes the form of

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (3.5)$$

or $y_t = \Theta(B) e_t \quad (3.6)$

and is called a Moving Average model of order q, denoted by MA (q) or ARIMA (*0, 0, q*).

3**) Mixed Autoregressive and Moving Average (ARMA) models:**

When the current output y_t depends on both the AR and MA processes, the Box-Jenkins model takes the form of equation (3.1) and is called an Autoregressive and Moving Average model, denoted by ARMA (p,q) or ARIMA *(p,0,q*).

3.3 Stationary Time Series

ARIMA time series models are designed for stationary time series. The Box-Jenkins methodology requires that the time series under analysis be "stationary" in both mean and variance.

In simplest non-statistical terms, the concept of stationarity can be explained as follows:

1) If the mean of the plotted series varies over time, the series is considered nonstationary in mean. If there is no evidence of a change in mean level over time, then the series is considered mean-stationary.

2) If the plotted series shows no obvious change in the variance over time, then the series is considered to be stationary in variance, otherwise it is considered to be non-stationary in variance. One of the other advantages of Box-Jenkins model is that it can be applied to nonstationary series after making them stationary using some sort of transformation. In order to induce mean stationarity in the (mean non-stationary) data, typically, a concept called 'differencing' is used. A difference of order one (or second or higher orders) is all that is required to achieve mean stationarity in the majority of cases. A difference of order one means that each value of the time series is subtracted from the immediate previous value:

$$w_t = \nabla y_t = y_t - y_{t-1} \quad (3.7)$$

It can be shown that $y_t - y_{t-1} = (1 - B)y_t$

Therefore,$\nabla = (1 - B)$, the difference operator. The result is a new time series w_t , having one less observation than the originaly_t series. A difference of order two implies that the first order differenced series is differenced again resulting in:

$$k_t = w_t - w_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2} \quad (3.8)$$

The result again is a new time series k_t , having two less observations than the original series y_t . This can be generalized to d^{th} order differencing, where *d* is the order of differencing required to achieve mean stationarity. After modelling the d*th* order differenced series with an appropriate ARMA model, to reclaim the modelled values corresponding to the original undifferenced series, it is necessary to reverse the differencing transformation and "integrate" *d* times. This is represented by "I" in the acronym ARIMA and the order of integration is same as the order of differencing ('d' in this case). Next, in order to make the series stationary in variance, if required, a different class of transformations can be carried out like the logarithmic transformation (taking log of the original data), square root, cubic root etc.

$w_t = In(y_t)$ (Natural logarithmic transformation)

$w_t = \sqrt{y_t}$ (Square root transformation), etc.

If the series is not positive throughout, it can be made positive by adding a suitable constant *c* to each observation of the series before the transformation is carried out. One of the best methods available to detect the proper transformation required to reduce heteroscadicity in data is using the Box-Cox transformation methodology, a general class of transformation which includes all

other transformations mentioned earlier as special cases. The Box-Cox transformation also makes the data more normal distribution-like.

Once the series of interest has achieved stationarity in both mean and variance using transformations, appropriate AR(p), MA(q), ARMA(p,q) or ARIMA(p,d,q) models can be fit to the series.

3.4 Box-Jenkins Methodology

The Box-Jenkins methodology is used in modelling the time series. The pioneers who popularized an approach which combines the moving average and autoregressive models were Box and Jenkins. Although both autoregressive and moving average approaches were known (and were originally investigated by Yule), the contribution of Box and Jenkins was in developing a systematic methodology for identifying and estimating models that could incorporate both approaches and this makes Box-Jenkins models a powerful class of models (Dobre et al., 2008).

There are four primary stages in building a Box-Jenkins time series model. These are model identification, estimation of the model parameters, diagnostic check of the residuals and model adequacy and forecasting.

3.4.1 Steps in Analyzing Data and Identifying ARIMA Models

Box and Jenkins recommend a three-stage iterative modelling strategy to fit ARIMA models to a time series of interest.

Step 1- Identification of the order of the ARIMA model:

At the identification stage, the historical data of the time series of interest is statistically analyzed and an appropriate subclass of models from the general ARIMA (p, d, q) family is selected.

The approach can be summarized as follows:

a) Suitably transform the time series y_t to remove the non-stationarity in variance (if present).

b) Difference the time series y_t as many times as is needed to produce mean stationarity (if required), hopefully reducing the process under study to the mixed Autoregressive Moving Average ARMA (*p*,*q*) process.

c) Identify the order of the ARMA model. That is, identify the autoregressive order ' p' and moving average order ' q' present in the transformed and differenced data.

The basic tools for model identification (steps (b) and (c)) are the graphs of estimated Sample Autocorrelation Function (ACF) and the estimated Sample Partial Autocorrelation Function (PACF) obtained from the series.

These graphs are used not only to help guess the form of the model, but also to obtain approximate estimates of the parameters (using Yule-Walker equations), which are useful at the estimation stage to provide starting values for iterative procedures employed during the estimation of final parameters.

For a time series$y_t, t \geq 1$, the autocorrelation coefficient at lag k is:

$$\rho_k = \frac{Cov(y_t, y_{t+k})}{var(y_t)} \tag{3.9}$$

The sample kth order autocorrelation is,

$$r_k = \frac{\sum_{t=1}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2} \tag{3.10}$$

The theoretical partial autocorrelation at lag k, r_{kk} , may be thought of as the autocorrelation between y_t andy_{t+k}, separated by a lag of k time intervals, with the effects of the intervening variables $y_{t+k}, y_{t+2}, \cdots, y_{t+k-1}$eliminated.

The sample partial autocorrelation coefficients can be computed as:

$r_{11} = r_1$,

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1}(r_{k-1,j} r_{k-j})}{1 - \sum_{j=1}^{k-1}(r_{k-1,j} r_j)}, \qquad (k = 2,3, \ldots) \tag{3.11}$$

Theoretically, it can be shown that an Autoregressive (AR) process of order p has an autocorrelation function of infinite extent, dominated by damped exponentials and sine waves, and a partial autocorrelation function that is zero after lag p. Conversely, the partial autocorrelation function of a Moving Average (MA) process of any order q is infinite in extent and its autocorrelation function is zero beyond lag q. For ARMA processes, the identification of the process order gets somewhat complicated by the fact that both the autocorrelation function and partial autocorrelation function are infinite in extent.

Table 3-1: Distinguishing characteristics of theoretical ACF and PACF

Process	ACF	PACF
AR(p)	Trails off towards zero (exponential decay or damped sine wave)	Cuts off zero after lag p
MA(q)	Cuts off to zero after lag q	Trails off towards zero (exponential decay or damped sine wave)
ARMA(p, q)	Trails off towards zero (exponential decay or damped sine wave)	Tails off towards zero (exponential decay or damped sine wave)

These opposite characteristics are used to identify the type and order of AR, MA or ARMA processes in the data. In addition, other patterns may also be present in the ACF and PACF plots which can help to identify the true orders of AR and MA coefficients. In practice the idealized procedure of significant spikes is confounded by sampling error in the estimated ACF and PACF and proper identification can become quite difficult depending on specific cases. Thus it requires some experience and judgment to identify a proper tentative form of the model.

Step 2- Estimation of the model parameters:

After a tentative form of the model is identified, the AR and MA parameters need to be estimated in the best possible manner.

There are fundamentally two ways of getting final estimates:

a) Trial and error – examine many different values of parameters and choose that value (or values, if more than one parameter is to be estimated) that minimizes the sum of squared residuals of fitting the model. The residual at each time step is the difference between the actual time series observation and the model output value at the same time step.

b) Iterative improvement – choose a preliminary estimate obtained from the identification procedure (Yule-Walker equations) and use an efficient nonlinear least-squares algorithm, called Marquardt algorithm, to refine the estimate iteratively.

Several methods exist to solve for the AR and MA coefficients using nonlinear square estimation. These methods are: the Maximum Likelihood Method, Unconditional Least Squares Method and the Conditional Least Squares Method.

In the present work, Conditional least squares method was employed since its computationally faster and under the assumption of normally distributed random shocks in the model, the Least squares parameter estimates are either exactly equal to or very nearly Maximum Likelihood estimates. In this work, the estimation of parameters was performed on Minitab and SPSS software package.

Step 3- Diagnostic Check of the residuals and model adequacy:

It is a common practice in ARIMA modelling to tentatively fit more than one model form to the data, estimate the parameters for each model and then perform a diagnostic check to test the validity of each model. The model which fits the best according to various statistical tests of fit is then selected for forecasting.

In particular, the following has to be performed:

1) A study of the residual series obtained after fitting the model to the data to see if any pattern remains accounted for. The ACF and PACF plots of the residual series help in detecting any unaccounted pattern.

2) A study of the sampling statistics of the current optimum solution to check if any further simplification of the model is possible. The residuals left over after fitting an ARIMA model should ideally be just random noise (white noise) with zero mean and constant variance. The following statistical tests for lack of fit were used in the present work to check for the randomness of the residuals:

1) **ACF and PACF plots of the residuals**: The ACF of the residuals obtained after fitting a proper model to the data must show no significant autocorrelations at any lag order. Similarly, the PACF plot of the residuals must show no significant spikes at any lag order. Absence of any significant spikes in the residual ACF and PACF plots demonstrate proper fitting. However, in practice, there maybe a few spikes which are close to significance. One might expect approximately 1 lag in every 20 lags to be statistically significant by chance alone for a 95% confidence limit test. Such spikes may not be a big concern; though their position of lag order also matters in deciding their importance and proper judgment should be used.

2) **Ljung-Box Chi-Square test**: Another measure of check for the randomness of residuals is using the Ljung-Box Chi-Square test. The null hypothesis is that the set of autocorrelations for residuals is white noise. This statistic measures the significance of residual autocorrelations as a set and points out if they are collectively significant:

H_0 : The data is random

H_1 : The data is not random

It is computed as:

$$\chi^2_m = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \qquad (3.12)$$

where n is the size of sample, r_k is the sample autocorrelation at lag k, and the m is the number of lags being tested. Each chi-square statistic is computed for all lags up to the indicated value and is not independent of the preceding chi-square values. If α is the significance level, the null hypothesis is rejected if:

$\chi^2_m > \chi^2_{1-\alpha,f}$, the α -quantile of the Chi-square distribution with f degrees of freedom. Apart from these tests to check residual randomness, more tests need to be carried out on the model itself to check its adequacy and best fit.

Two of the most important of such criteria are the Akaike's Information Criteria (AIC) and Schwarz's Bayesian Criteria (SBC). The AIC and SBC are used to compare competing models fit to the same series. The model with smaller AIC and SBC values is a statistically better fit.

1**) Akaike's Information Criteria (AIC):** It is a statistical tool for model selection and is grounded in the concept of entropy. It can be non-statistically described as a measure of trade-off between the precision and complexity of the model. The absolute value of AIC is not useful; the relative comparison of AIC values of different competing models can be used to infer the best model. The model with lowest AIC value is the best fit.

It is computed as:

$$AIC = -2\text{loglikelihood} + 2k, \tag{3.13}$$

where k = number for parameters in the model,

Loglikelihood = maximized value of log likelihood function for the estimated model.

Assuming the residuals to be normally and independently distributed, if the residual sum of squares is denoted by R, the AIC criterion becomes:

$$AIC = n\left[\ln\frac{2\pi R}{n} + 1\right] + 2k,\ R = \sum_{i=1}^{n} a_i^{-2} \tag{3.14}$$

The AIC criterion attempts to find the model that best explains the data with a minimum of free parameters. It imposes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over fitting in estimation and thus leads to selection of a parsimonious model.

2**) Schwarz's Bayesian Criteria (SBC)**: It is also called Bayesian Information Criteria (BIC). SBC is also a statistical tool for model selection, which penalizes over fitting of estimation. The model with lower SBC is generally the best fit.

It is computed as:

$$SBC = -2\ loglikelihood\ +\ k\ ln\ (n) \tag{3.15}$$

For normally and independently distributed residuals,

$$SBC = n\left[\ln\frac{R}{n}\right] + k\ln(n) \tag{3.16}$$

The SBC criterion penalizes free parameters more heavily than the AIC criterion. It must be noted that both AIC and SBC tests may not generally point to a common model as the best fit. In such cases, proper judgment is required in choosing the best fit for the model.

If, after performing the above checks of residual randomness and model adequacy, the model is found inadequate or if some significant autocorrelations are detected in the residual ACF plots,

the identification stage should be revisited and a new model reformulated by examining the ACF and PACF plots of the original series again and making a new interpretation. The knowledge of the left-over pattern in the residuals, as evidenced from the residual ACF and PACF series, may also be used in making a judgment to help identify a different form of tentative model. Thus, the three stages of the iterative process, viz. identification, estimation and diagnostic checking may have to be repeated multiple times until a satisfactory model is generated. Another aspect to keep in mind while fitting models is the 'principle of parsimony', which states that the best model for a given series is the very simplest model with least number of parameters which can account for the observed properties of the data. Thus, if two candidate models are finalized for the series under interest depending on the various tests outlined in this section and if they are comparable with respect to fitting adequacy and yielding white noise residuals, the model with minimum number of parameters must be preferred.

3.5 Forecasting Using ARIMA Models

Once an adequate and satisfactory model is fitted to the series of interest, forecasts can be generated using the model. Consider the general ARIMA model of

equation (3.1).

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \ldots - \theta_q e_{t-q} \quad (3.17)$$

The one-step ahead forecast for time $t + 1$ is given by,

$$y_{t+1} = \phi_1 y_t + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p+1} + e_{t+1} - \theta_1 e_t - \theta_2 e_{t-1} - \cdots - \theta_q e_{t-q+1} \quad (3.18)$$

Except e_{t+1} , the random shock at time t + 1, all other parameters are known.

Thus, setting e_{t+1} = 0, its true expected value, the one-step ahead forecasts can be

generated. Similarly, future forecasts can be 'bootstrapped' using these obtained forecasts y_t and setting the unrealized random shocks to 0 for each case. Also, the 95% confidence interval for the forecasts can be calculated.

If future inputs are available, they are used in the forecasting model; if they are not available, the inputs can be forecasted using their corresponding prewhitening ARIMA filters prior to forecasting the outputs. As in the case of univariate ARIMA models, the goal is to fit a parsimonious model with minimum of parameters which satisfy the various goodness of fit criteria. The autocorrelation and partial autocorrelation plots are studied along with the Ljung Box Chi-square test results to test the hypothesis for randomness of the residual series e_t . AIC and SBC criteria are used to select the best fit out of a set of competing models.

3.6 Seasonal ARIMA Models

Time series data may often display periodic (also called 'seasonal' in ARIMA context) behaviour. A periodic series has a pattern which repeats every's' time periods (s>1), where's' is also called the length of periodicity. ARIMA models for seasonal time series, popularly called SARIMA ('*S*' stands for seasonal), are built using the same three-stage iterative modelling procedure used for non-seasonal ARIMA models:

identification, estimation, and diagnostic checking. However, with seasonal data, attention must also be focused on the autocorrelation coefficients in the ACF and PACF plots occurring at the seasonal lags $s, 2s, 3s,\ldots$ etc. If seasonal non-stationarity is present in the data, as evidenced from the fact that the autocorrelation coefficients at the seasonal lags of ACF plot will not die out rapidly, proper order of seasonal differencing (denoted by '*D* ') may be required to make the data seasonal stationary. Secondly, the presence of seasonal autoregressive and moving average coefficients in the data needs to be determined on similar lines as was discussed for the non-

seasonal ARIMA model identification, but with using the autocorrelation coefficients of ACF and PACF plots at the seasonal lags. The general notation for seasonal ARIMA model is ARIMA (*P, D, Q*), where 'P' is the order of seasonal autoregressive component, '*Q*' is the order of seasonal moving average coefficient and '*D*' is the order of seasonal differencing used. In general, a time series often may contain both non-seasonal and seasonal components. Though the time series may be deseasonalized and a non-seasonal ARIMA model maybe fitted to the remainder, experience suggests that Box-Jenkins methodology provides good forecasts of periodic data series. Thus, it may be advisable to leave the seasonal component in the data and fit a general class of ARIMA model which accounts for both seasonality and non-seasonality. Such a general ARIMA model can be represented by the form ARIMA $(p, d, q)(P, D, Q)^s$. This is commonly referred to as a seasonal ARIMA multiplicative model and it is represented by:

$$\Phi(B^s)\Phi(B)(1-B^s)^D(1-B)^d y_t = \Theta(B^s)\Theta(B)e_t$$

The forecasts from seasonal ARIMA multiplicative models are generated in the similar fashion as with non-seasonal ARIMA models.

CHAPTER 4

4.0 DATA ANALYSIS AND RESULTS

Forecasting the future events has captivated the human imagination for centuries. Throughout history, predicting the future has been the major motivation behind the evolution of the science of astrology to astronomy to palmistry to tarot cards. With the advent of technology and statistical sciences, the realm of forecasting expanded rapidly to the scientific and technological fields. Today, the need for predicting the future events and the drivers of these events is overwhelming in many aspects of our daily life. Forecasting is finding its way into many different fields of applications than ever before.

4.1 Data Sources

The location and severity of malaria is mostly determined by climate and ecology (Gallup and Sachs, 2001). The area of potential transmission is influenced by climatic factors such as temperature, humidity and rainfall as well as the socio-economic condition of the population. These factors influence the development of both the vector and the parasite. Thus, based on the agro-ecological zones in Ghana, the Sekyere-East district was selected for this study. The study was conducted at the Effiduase Government Hospital.

4.2 Data Collection

In the study district, data for monthly reported cases of malaria fever between the years 2001 to 2010 was obtained from the statistical Department of Effiduase Government Hospital.

4.3 Characteristics of the Sekyere East District

The Sekyere East district lies in the north-eastern part of the Ashanti Region, in the forest zone. Almost 70% of the land area of the district is in the Greater Afram Plains to the north and is covered with Guinea savannah woodland while the vegetation of the southern portion is moist, Semi-deciduous forest. The district experiences two peak rainfall periods in a year but the rains are heavier in the southern parts. As in the rest of the forest zone, mean annual rainfall ranges between 125 cm and 200 cm. The mean monthly temperature is 260° C, with mean monthly humidity of 70 to 80%. Sekyere East has a total population of 157,396, representing 4.4% of the total population of the Ashanti Region (*2000 Population and Housing Census*, 2002). The urban population of the district is 33.7% of the total, with a high overall concentration of people in the southern portion, where population density is 72.8 people per square kilometre while that of the northern parts is 7.8 people per square kilometre. The principal economic activities are agriculture and commerce and the major crops include cocoa, kola nuts, plantain, cassava and cocoyam. The favourable climate and soil conditions enable crop production throughout the year. The district has one public hospital, one private hospital and one mission hospital as well as more than eight clinics and health posts. Malaria is the leading cause of morbidity, accounting for over 60% of outpatient consultations. It is also the first among the major causes of inpatient admissions in the district, with an annual average of 1,666 cases (58% of the total) between 2000 and 2003. Cerebral malaria and malaria with severe anaemia were the second most frequent causes of recorded deaths during the same period.

In this chapter, a suitable model is identified to forecast the incidence of malaria in Effiduase in the Sekyere East District.

4.3.1 Data Analysis and Results

The inputs used in the model are the historical data for each month's morbidity from Jan 1, 2001 through December 31, 2010.

In section 4.4.1, the various inputs to the model are identified. In section 4.5, the actual model is identified to forecast these incidences. The model is statistical in nature and its functioning is validated in section 4.6 with various tests of significance. Sections 4.7 and 4.8 present the results of fitting and forecasting, respectively. Section 4.8 also presents the conclusions of the chapter.

4.3.2 Identifying the Forecasting Model

Box-Jenkins ARIMA methodology was employed to fit the final models. A detailed discussion on Box-Jenkins ARIMA methodology can be found in Chapter 3. The identification of the ARIMA model will be carried out according to the following steps:

4.3.3 Transformation of the Inputs and Output Series

For illustration purpose, plots of the two input series to the model are shown in Figures 4.1 and 4.2 for the months of Jan 1, 2001- Dec 31, 2010. Also shown in Figure 4.3 is the plot of actual morbidity for the same period. The preliminary requirement of ARIMA methodology is that the series should be stationary in both mean and variance (Section 3.2, chapter 3).

Analysis of the autocorrelation plots revealed that each of the series is mean non-stationary in non-seasonal orders (Section 3.2, Chapter 3). Thus, suitable differencing was required for each of the series to render mean-stationarity. First order non-seasonal differencing was found to be sufficient for the series (Table 4.5).

Figures 4.8-4.10 present the autocorrelation and partial autocorrelation plots of the series obtained after differencing with the optimal non-seasonal orders. From the figures, it can be deduced that each of the differenced series is mean-stationary.

4.3.4 Descriptive Statistics of Monthly Reported Cases

The monthly reported cases from January 2001 to December 2010 were fed into a SPSS Spreadsheet and analysed to obtain the descriptive statistics.

Table 4.1 Reported cases of malaria (2001-2010)

Month/Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	655	**640**	**637**	**741**	**860**	**897**	**475**	**561**	**572**	**700**	**709**	**1161**
2002	**966**	**451**	**743**	**864**	**1001**	**1182**	**1202**	**1185**	**1192**	**1247**	**1281**	**1391**
2003	**992**	**1037**	**1059**	**998**	**1011**	**1037**	**1019**	**1201**	**1224**	**1249**	**1359**	**1377**
2004	**927**	**967**	**997**	**976**	**379**	**574**	**624**	**655**	**917**	**842**	**1187**	**1402**
2005	**1392**	**1455**	**1721**	**1929**	**1876**	**2023**	**2032**	**2038**	**1993**	**2009**	**2014**	**2029**
2006	**2047**	**2020**	**2063**	**2069**	**1997**	**2079**	**2097**	**3009**	**3017**	**2929**	**2957**	**2987**
2007	**2524**	**2493**	**2176**	**2105**	**2100**	**2577**	**2859**	**1641**	**1594**	**2158**	**1861**	**1288**
2008	**1558**	**2080**	**2045**	**2634**	**4628**	**2553**	**2059**	**1895**	**1812**	**2427**	**2416**	**1940**
2009	**2070**	**2086**	**2563**	**2517**	**2461**	**2482**	**3602**	**3820**	**2988**	**3232**	**3191**	**2712**
2010	**2956**	**2529**	**2627**	**2675**	**2459**	**2606**	**2835**	**2747**	**2427**	**2263**	**2748**	**2886**

Source: Effiduase Government Hospital-Ashanti Region

The table below shows the descriptive statistics of the reported cases.

4.3.5 DESCRIPTIVE STATISTICS OF MONTHLY REPORTED CASES

The monthly reported cases from January 2001 to January 2010 were fed into a Minitab spread sheet and analysed to obtain the descriptive statistics. Let $\mu_1, \mu_2, \ldots, \mu_{10}$ denote the monthly reported cases in 2001, 2002,...,2010 respectively.

The table below shows the descriptive statistics of the reported cases.

Table 4.2 Descriptive Statistics of Reported Cases

Descriptives

Morbidity

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
2001	12	717.33	183.653	53.016	600.65	834.02	475	1161
2002	12	1058.75	266.893	77.045	889.17	1228.33	451	1391
2003	12	1130.25	143.583	41.449	1039.02	1221.48	992	1377
2004	12	870.58	279.929	80.809	692.72	1048.44	379	1402
2005	12	1875.92	230.484	66.535	1729.47	2022.36	1392	2038
2006	12	2439.25	478.363	138.091	2135.31	2743.19	1997	3017
2007	12	2114.67	460.000	132.791	1822.40	2406.94	1288	2859
2008	12	2337.25	789.817	228.001	1835.42	2839.08	1558	4628
2009	12	2810.33	560.837	161.900	2453.99	3166.67	2070	3820
2010	12	2646.50	203.634	58.784	2517.12	2775.88	2263	2956
Total	120	1800.08	846.799	77.302	1647.02	1953.15	379	4628

In 2001, the average monthly reported case was 717.3; the least number was recorded as 475 in July and the highest number was recorded in December as 1161. In 2002, the average monthly reported case was 1058; the least number was recorded in February as 451. In 2003, 2004,2005, 2006, 2007, 2008, 2009 and 2010, the average monthly reported cases were 1130, 870, 1875.9, 2439, 2115, 2337, 2810 and 2646.5 respectively. Their respective highest reported cases were 1377 (in December), 1402 (in December), 2038(in August), 3017(in September), 2859(in July), 4628(in May), 3820(in August) and 2956 (in January).

4.3.6 TWO-SAMPLE T-TEST AND CONFIDENCE INTERVAL

The two-sample t-test was used to compare the sample means of reported cases of malaria between the year 2001 (the base year of the corrected data) to the tenth year, 2010. Minitab package was used in this analysis and the results are displayed in appendix C.

4.3.7 Two-Sample T-Test and Confidence Interval for Monthly Reported Cases in 2001 and 2002

Assuming equal variances, the null hypothesis and the alternative hypotheses stated are as follows;

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

From the Minitab output in appendix C, the difference between the average number of reported cases in 2001 and 2002 was -341.5. The difference was then tested to find out whether it was significant or not. The T-value (calculated value) was found to be -3.65 and the p-value=0.002. Since the p-value was less than 0.05 level of significance, the null hypothesis was then rejected and concluded that the difference in the average monthly reported cases between 2001 and 2002 was statistically not the same at 5% level of significance. This indicated an increase in the number of reported cases. Also from the output, the 95% confidence interval of the difference between the average monthly reported cases in 2001 and 2002 was between -537.3 and -145.7.

4.3.8 Two-Sample T-Test and Confidence Interval for Monthly Reported Cases in 2001 and 2003

The null and alternative hypotheses for comparing the annual reported cases have been stated below.

$H_0: \mu_1 - \mu_3 = 0$

$H_1: \mu_1 - \mu_3 \neq 0$

From the Minitab output in appendix C, the difference between the average number of reported cases in 2001 and 2003 is -413. The difference was then tested to find out whether it was significant or not. The T-value (calculated value) was found to be -6.14 and the p-value=0.000.

For smaller p-value, the null hypothesis is rejected. So the null hypothesis was rejected at 5% and this indicated that there was significant increase in the number of malaria reported cases. The 95% confidence interval for the means was between -553.4 and -272.6.

4.3.9 Two-Sample T-Test and Confidence Interval for Monthly Reported Cases in 2001 and 2004

The null and alternative hypotheses for comparing the annual reported cases have been stated below.

$H_0: \mu_1 - \mu_4 = 0$

$H_1: \mu_1 - \mu_4 \neq 0$

From the Minitab output in appendix C, the difference between the average number of reported cases in 2001 and 2004 is -153.3. The difference was then tested to determine whether it was significant or not. The T-value (calculated value) was found to be -6.14 and the p-value=0.000 at 20 degrees of freedom. For smaller p-value, the null hypothesis is rejected. So the null hypothesis was rejected at 5% and this indicated that there was significant increase in the number of malaria reported cases in 2001 and 2004. The 95% confidence interval of the means was between -356.3 and 49.7.

4.3.10 Two-Sample T-Test and Confidence Interval for Monthly Reported Cases in 2001 and 2005

The null and alternative hypotheses for comparing the annual reported cases have been stated below.

$H_0: \mu_1 - \mu_5 = 0$

$H_1: \mu_1 - \mu_5 \neq 0$

From the Minitab output in appendix C, the difference between the means was -1158.6 indicating an increase in the number of reported cases. The difference was then tested to determine whether it was significant or not. The T-value (calculated value) was found to be -13.62 and the p-value=0.000 at 20 degrees of freedom. For smaller p-value, the null hypothesis is rejected. So the null hypothesis was rejected at 5% and this indicated that there was significant decrease in the malaria reported cases comparing the number in 2001 with that of 2005. The 95% confidence interval of the difference between the monthly reported cases in 2005 and 2008 was between -1336.1 and -981.1.

4.3.11 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2002 and 2003

The monthly reported cases of the disease in 2002 and 2003 were compared. The null and alternative hypotheses were

$H_0: \mu_2 - \mu_3 = 0$

$H_1: \mu_2 - \mu_3 \neq 0$

The Minitab output in appendix C indicated that difference between the means is –71.5. This implied that the reported cases increased between 2002 and 2003. The test statistic value, T-value calculated is -0.82 and the p-value=0.426. Since the p-value is greater than 0.05 level of significance, the null hypothesis is accepted and concluded that the difference in the increase in the monthly reported cases of malaria between 2002 and 2003. The 95% confidence interval of the difference between the average monthly reported cases in 2003 and 2003 was between -257 to 114.

4.3.12 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2002 and 2004

The mean monthly reported cases of the disease in 2002 and 2004 were compared. The null and alternative hypotheses were

$H_0: \mu_2 - \mu_4 = 0$

$H_1: \mu_2 - \mu_4 \neq 0$

The Minitab output indicated there was a drop in the number by 188. It was the tested to find whether is significant or not. The T-value (calculated value) was 1.69 and the p-value at 21 degrees of freedom was found to be 0.107. Since the p-value was greater than the 0.05 level of significance, the null hypothesis was accepted and concluded that there was no significant decrease in the monthly reported cases of malaria between 2002 and 2004. The 95% confidence interval of the difference between the average monthly reported cases in 2002 and 2004 was between -44 and 420.

4.3.13 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2002 and 2005

The monthly reported cases of the disease in 2002 and 2005 were compared. The null and alternative hypotheses were

$H_0: \mu_2 - \mu_5 = 0$

$H_1: \mu_2 - \mu_5 \neq 0$

The Minitab output indicated an increase in the number of the reported cases between 2002 and 2005 by -817. The T-value was found (calculated value) to be -8.03 and the p-value at 21 degrees of freedom was found to be 0.000. Since the p-value was less than the 0.05 level of significance, the null hypothesis was rejected and concluded that there was significant decrease

in the monthly reported cases of malaria between 2002 and 2005. The 95% confidence interval of the difference between the average monthly reported cases in 2002 and 2005 was between -1029 and -605.

4.3.14 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2003 and 2004

The monthly reported cases of the disease in 2003 and 2004 were compared. The null and alternative hypotheses were

$H_0: \mu_3 - \mu_4 = 0$

$H_1: \mu_3 - \mu_4 \neq 0$

The Minitab output indicated a drop in the number of reported cases between 2003 and 2004 by 259.7. The T-value was found (calculated value) to be 2.86 and the p-value at 16 degrees of freedom was found to be 0.011. Since the p-value was less than the 0.05 level of significance, the null hypothesis was rejected and concluded that there was significant decrease in the monthly reported cases of malaria between 2003 and 2004. The 95% confidence interval of the difference between the average monthly reported cases in 2003 and 2004 was between 67.2 to 452.2.

4.3.15 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2003 and 2005

The monthly reported cases of the disease in 2003 and 2005 were compared. The null and alternative hypotheses were

$H_0: \mu_3 - \mu_5 = 0$

$H_1: \mu_3 - \mu_5 \neq 0$

The Minitab output indicated an increase in the number of reported cases by -745.6. The T-value calculated was -9.51 and the p-value at 18 degrees of freedom was found to be 0.000. Since the p-value was less than the 0.05 level of significance, the null hypothesis was rejected and concluded that there was significant decrease in the monthly reported cases of malaria between 2003 and 2005. The 95% confidence interval of the difference between the average monthly reported cases in 2003 and 2005 was between -910.3 and -580.9.

4.3.16 Two-Sample T-Test and Confidence Interval for the Monthly Reported Cases in 2004 and 2005

The monthly reported cases of the disease in 2004 and 2005 were compared. The null and alternative hypotheses were

$H_0: \mu_4 - \mu_5 = 0$

$H_1: \mu_4 - \mu_5 \neq 0$

The Minitab output indicated an increased number of reported cases by -1005. The T-value calculated was -9.60 and the p-value at 21 degrees of freedom was found to be 0.000. Since the p-value was less than the 0.05 level of significance, the null hypothesis was rejected and concluded that there was significant difference in the monthly reported cases of malaria between 2004 and 2005. The 95% confidence interval of the difference between the average monthly reported cases in 2004 and 2005 was between -1223 and -788.

For the rest of the analysis, refer to appendix C

4.4 Trend Analysis

The series by inspection exhibits some upward trend with no hint of seasonal variations. There might be individual series with seasonality, but it appears that seasonality is not a prominent feature of the data in general.

From table 4.1 and figure, the time series has a linear trend model given by,$y_t = 572.8 + 20.3 * t$ where y_t is the number of monthly reported cases and t is the given month.

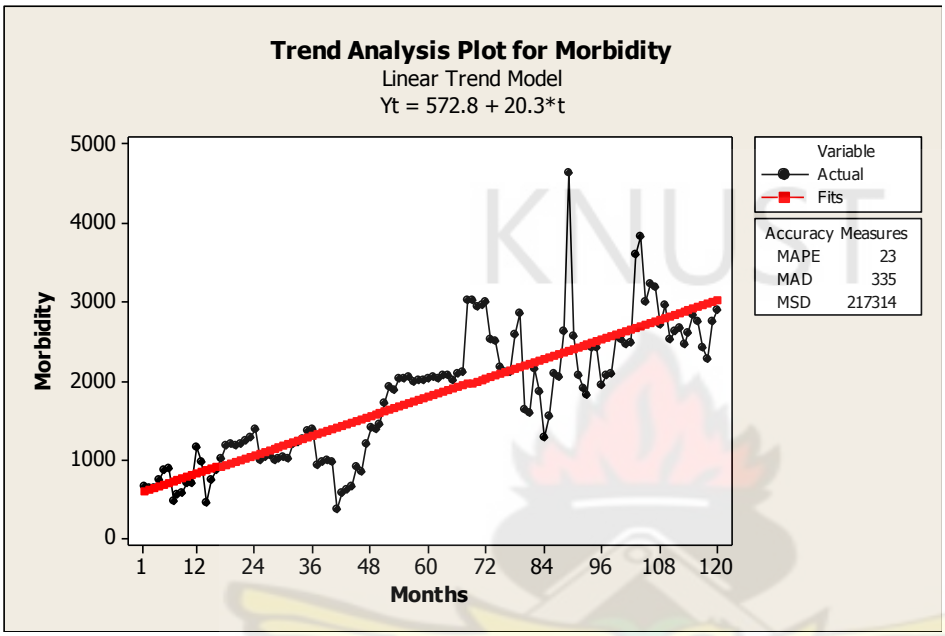
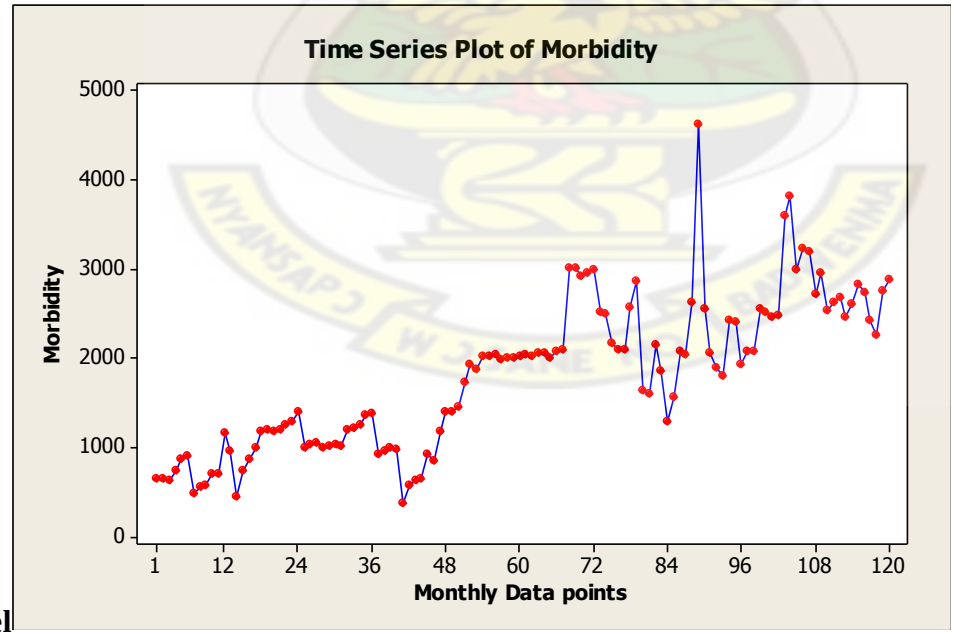


Figure 4.1: Linear Trend



model

Figure 4.2: Time Series Plot of Malaria cases

4.4.1 Preliminary Test for Stationarity

A stationary series has constant mean, constant variance and constant autocorrelation structure. The time series plot of the monthly reported cases from January 2001 to December 2010 is shown in figure 4.3. The plot shows that the time series is non-stationary. The values had irregular swings and hence had irregular variability. But the fact the there were irregular fluctuations in the plot was not a clear indication that the series was non-stationary. This called for verification to confirm the non-stationarity as indicated by the plotted graph below. The ACF and PACF were used to verify the non-stationarity of the series. Tables 4.3 and 4.4 below show the results of the test.

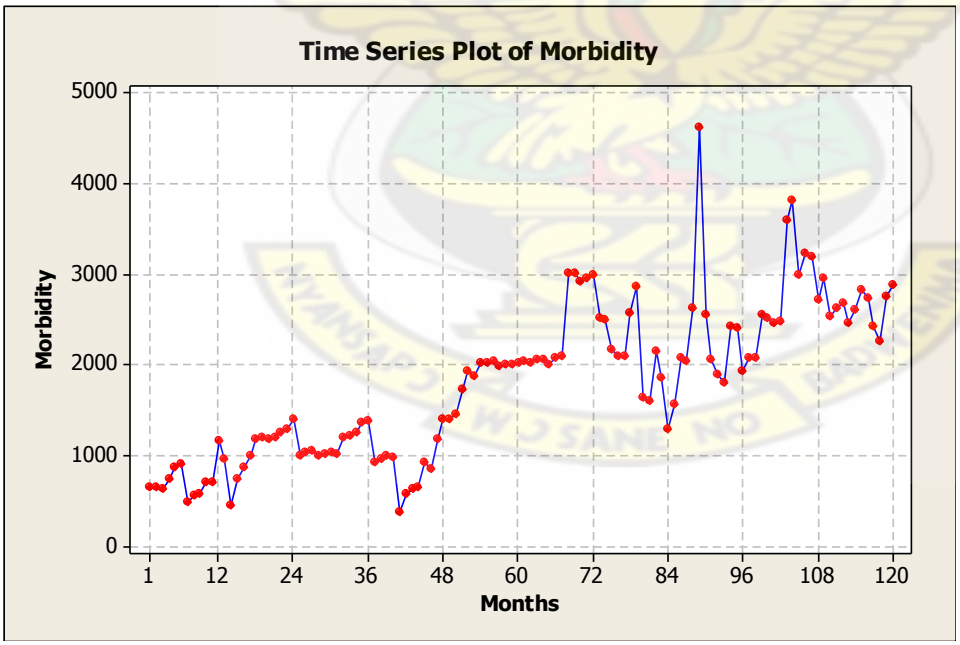


Figure 4.3 Monthly reported cases of Malaria

Table 4.3 Autocorrelation Function: Morbidity

Lag	ACF	T	LBQ
1	0.874254	9.58	94.03
2	0.781047	5.38	169.72
3	0.755431	4.27	241.12
4	0.710198	3.52	304.78
5	0.672070	3.03	362.28
6	0.647370	2.72	416.10
7	0.611573	2.42	464.56
8	0.576170	2.18	507.95
9	0.555532	2.02	548.66
10	0.566931	2.00	591.43
11	0.536145	1.83	630.04
12	0.502304	1.67	664.24
13	0.480706	1.56	695.86
14	0.482265	1.53	727.98
15	0.485007	1.51	760.78
16	0.460606	1.41	790.64
17	0.444800	1.34	818.77
18	0.404501	1.20	842.25
19	0.356913	1.05	860.72
20	0.352891	1.03	878.95
21	0.325739	0.94	894.64
22	0.267419	0.77	905.32
23	0.243937	0.69	914.30
24	0.241948	0.69	923.23
25	0.250018	0.71	932.86
26	0.238542	0.67	941.72
27	0.219290	0.61	949.29
28	0.203503	0.57	955.88
29	0.199395	0.56	962.28
30	0.222026	0.62	970.30
31	0.225546	0.62	978.67
32	0.204697	0.57	985.64
33	0.208506	0.57	992.95
34	0.214844	0.59	1000.81
35	0.210893	0.58	1008.47
36	0.200623	0.55	1015.49

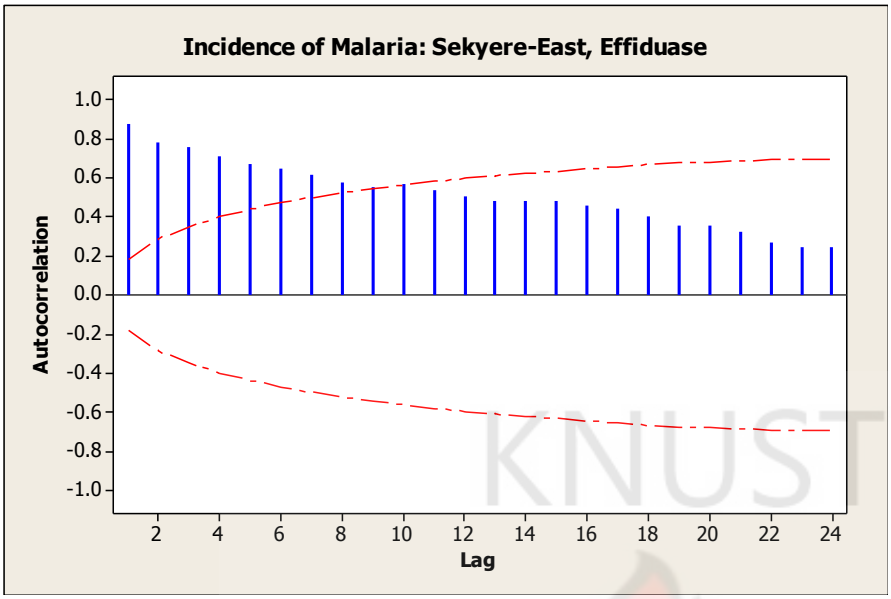


Figure 4.4 Autocorrelation coefficient for the times: incidence of malaria in Effiduase

Table 4.4 Partial Autocorrelation function: Morbidity

Lag	PACF	T
1	0.874254	9.58
2	0.070970	0.78
3	0.251658	2.76
4	-0.025845	-0.28
5	0.076137	0.83
6	0.034372	0.38
7	-0.012940	-0.14
8	0.000136	0.00
9	0.036673	0.40
10	0.162414	1.78
11	-0.114332	-1.25
12	0.017385	0.19
13	-0.044947	-0.49
14	0.134953	1.48
15	0.035271	0.39
16	-0.067693	-0.74
17	0.031921	0.35
18	-0.155214	-1.70
19	-0.033346	-0.37
20	0.035399	0.39
21	-0.066586	-0.73
22	-0.113662	-1.25
23	0.054724	0.60
24	0.046952	0.51

25	0.098134	1.08
26	-0.037346	-0.41
27	-0.044055	-0.48
28	0.020913	0.23
29	0.054520	0.60
30	0.106803	1.17
31	-0.051703	-0.57
32	-0.025280	-0.28
33	0.100582	1.10
34	0.032179	0.35
35	-0.049724	-0.54
36	-0.016779	-0.18

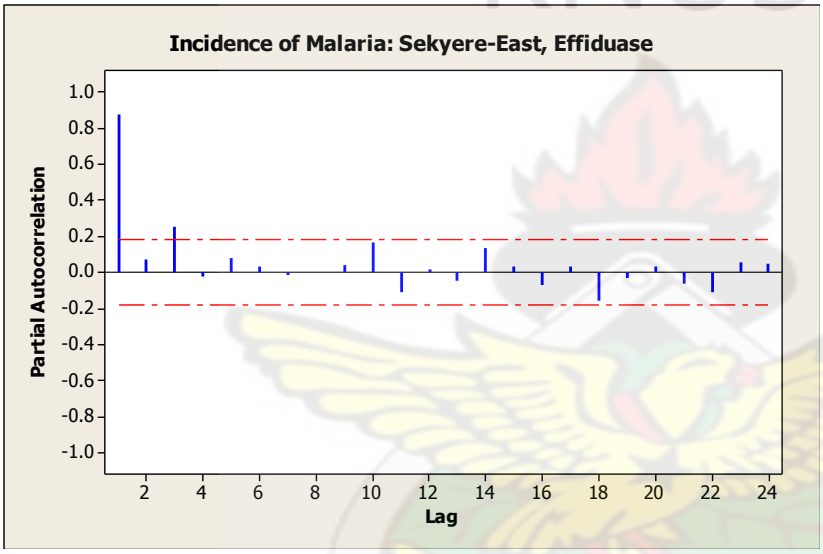


Figure 4.5 Partial autocorrelation coefficients for the time series: incidence of malaria Effiduase

4.5 Model Identification

In this section, we implement the Box-Jenkins approach to the identification and modelling of time series. This was done using Minitab and SPSS software.

Stage 1: Identification

The first step in identifying a preliminary model was to examine the autocorrelations and partial autocorrelations for the raw data. These values are shown in figures 4.3 and 4.4 respectively. The autocorrelations, which were very large at first, did not trail off towards zero quickly. They

appeared to be forming a sine wave pattern, but because the damping process was so slow, it was concluded that the process was non-stationary. It was also noted that the first 36 coefficients were outside the 2σ limits. Since there are 120 observations the 2σ confidence limits were found to be $-2\frac{1}{\sqrt{120}} < r < 2\frac{1}{\sqrt{120}}$

$$-0.183 < r < +0.183$$

Figure 4.5 contains the plot of partial autocorrelations. The coefficient of order 1 was very large. This indicates that the time series often occurs when the autocorrelations do not rapidly die down. Differencing was necessary at this point to identify the ARIMA model property.

Upon differencing the ARIMA process to achieve a stationary ARMA process for forecasting, new diagrams of autocorrelation coefficients and partial autocorrelation coefficients were drawn. Figure 4.6 contains the autocorrelation of the first differences. These autocorrelations rapidly trail off towards zero after lag 2. The partial autocorrelations in figure 4.7 cut off after lag 2. Both these patterns indicate an ARIMA (2, 1, 0) model. However, this is only a tentative choice.

Table 4.5: Autocorrelation coefficients for first-differenced time series (computer-generated SPSS output)

Autocorrelations

Series:Morbidity

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	Df	Sig.[b]
1	-.142	.091	2.453	1	.117
2	-.279	.090	12.014	2	.002
3	.084	.090	12.881	3	.005
4	-.045	.089	13.136	4	.011
5	-.065	.089	13.676	5	.018

6	.035	.089	13.836	6	.032
7	.010	.088	13.848	7	.054
8	-.081	.088	14.689	8	.065
9	-.136	.087	17.111	9	.047
10	.194	.087	22.101	10	.015
11	.015	.087	22.129	11	.023
12	-.055	.086	22.536	12	.032
13	-.166	.086	26.278	13	.016
14	.007	.085	26.286	14	.024
15	.141	.085	29.050	15	.016
16	-.081	.085	29.973	16	.018
17	.131	.084	32.402	17	.013
18	.097	.084	33.736	18	.014
19	-.198	.083	39.372	19	.004
20	.100	.083	40.821	20	.004
21	.140	.082	43.681	21	.003
22	-.129	.082	46.145	22	.002
23	-.090	.082	47.371	23	.002
24	-.063	.081	47.981	24	.003
25	.062	.081	48.564	25	.003
26	.034	.080	48.742	26	.004
27	.015	.080	48.776	27	.006
28	-.058	.079	49.308	28	.008
29	-.130	.079	52.009	29	.005
30	.055	.079	52.493	30	.007
31	-.002	.078	52.493	31	.009
32	.003	.078	52.494	32	.013
33	.023	.077	52.579	33	.017
34	.050	.077	52.997	34	.020
35	.059	.076	53.588	35	.023
36	-.001	.076	53.588	36	.030

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

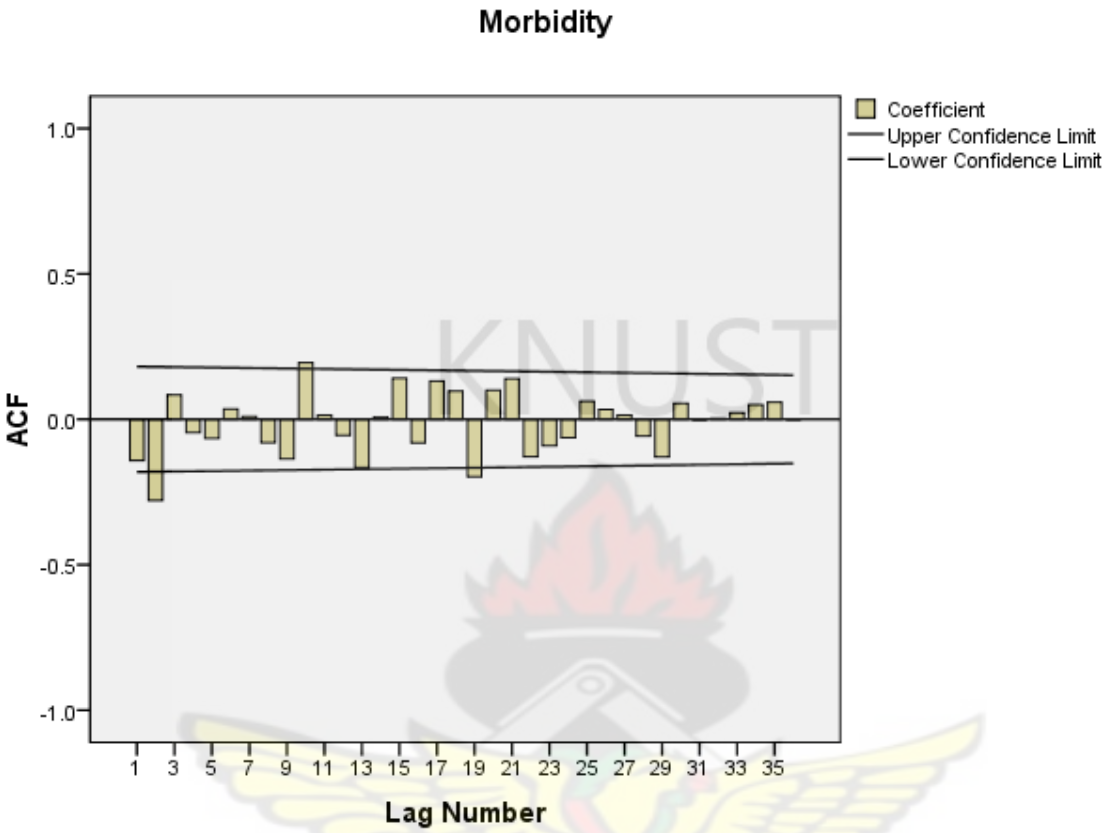


Figure 4.6 Autocorrelation coefficients for first-differenced time series (Note: computer – generated SPSS output)

Table 4.6: Partial autocorrelation coefficients of first-differenced series (computer generated SPSS output)

Partial Autocorrelations

Series:Morbidity

Lag	Partial Autocorrelation	Std. Error
1	-.142	.092
2	-.305	.092
3	-.013	.092
4	-.131	.092
5	-.085	.092

6	-.047	.092
7	-.033	.092
8	-.103	.092
9	-.217	.092
10	.079	.092
11	-.049	.092
12	.015	.092
13	-.274	.092
14	-.105	.092
15	-.004	.092
16	-.132	.092
17	.103	.092
18	.042	.092
19	-.050	.092
20	.061	.092
21	.105	.092
22	-.055	.092
23	-.011	.092
24	-.139	.092
25	-.009	.092
26	.015	.092
27	-.023	.092
28	-.088	.092
29	-.166	.092
30	-.021	.092
31	-.147	.092
32	-.124	.092
33	-.128	.092
34	.082	.092
35	-.058	.092
36	-.043	.092

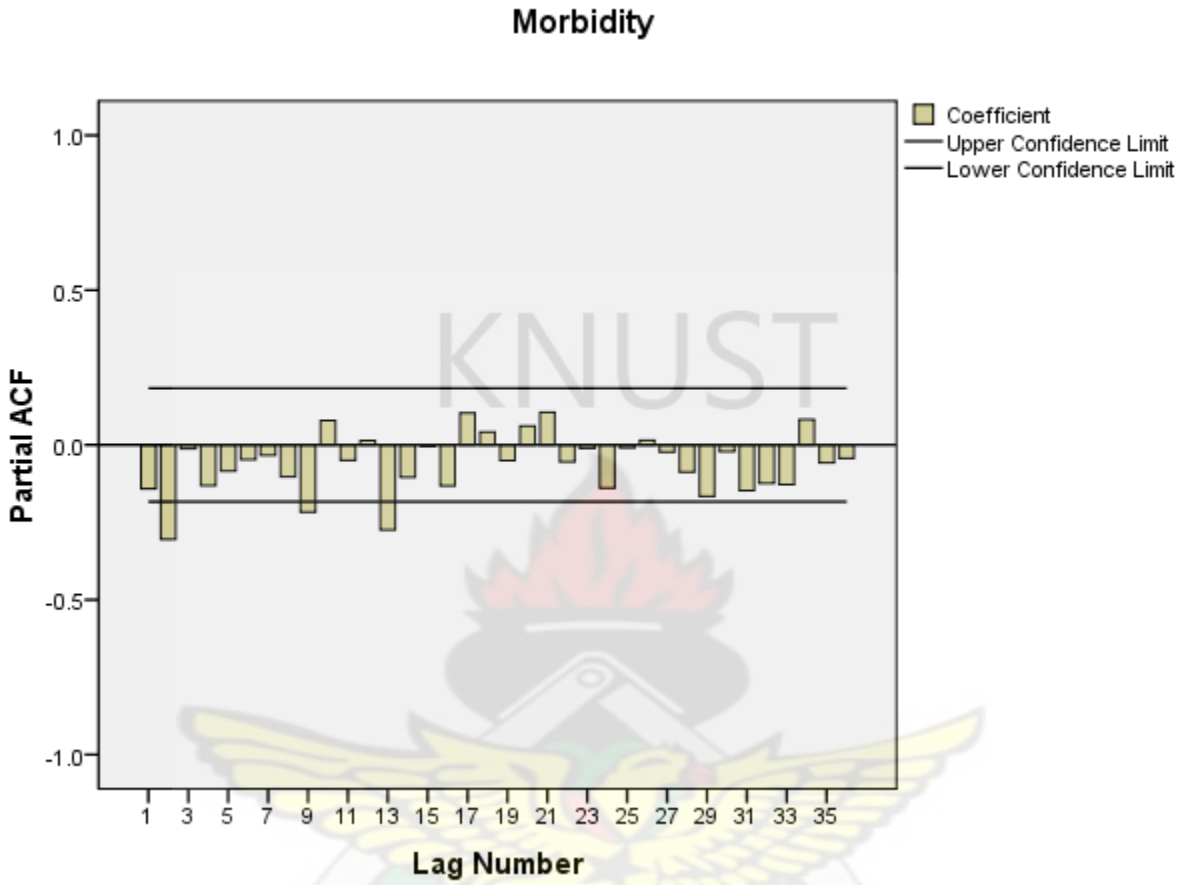


Figure 4.7 Partial autocorrelation coefficients of first-differenced series (computer-generated SPSS output)

Table 4.7

Model Statistics

Model	Number of Predictors	Model Fit statistics			Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	MAPE	Normalized BIC	Statistics	DF	Sig.	
Morbidity-Model_1	0	.053	14.358	11.929	11.198	17	.846	0

Table 4.8

ARIMA Model Parameters

		Estimate	SE	t	Sig.
Morbidity-Model_1 Morbidity	Natural Log	-.2331	.090	-2.593	.011
	AR Lag 2				
	Difference	1			

From the statistics of the model, we have it that using Expert modeller of SPSS, the number of predictors to compete with the model identified is zero and outliers to show that there is any element of trend of seasonality is also zero. We therefore conclude that the data does not exhibit any seasonality.

4.6 Stage 2: Estimation

Once the preliminary model is chosen, the estimation stage begins. The estimate is of the model:

$$(1-B)(1-\phi_1 B)Y_t = e_t \qquad (4.1)$$

where the $(1-B)Y_t$ are the first differences of the original values of Table 4.1 expressed in terms of deviations. The purpose of estimation is to find the parameter estimates that minimize the mean square error (MSE). The process is iterative, and the final value of the parameter estimates may be significantly different from the initialized values of the estimation procedure. However, the estimates will usually converge on an optimal value for the parameters with a small number of iterations. If the algorithm fails to converge after some specified number of iterations, an examination of the trail values will indicate the direction of changes needed and new initial estimates can be made. An AR (2) model is tentatively chosen for the sample data on the morbidity. The program converges after five iterations, producing the tentative model AR (2) where $\hat{\phi}_1 = -0.1868 \ and \ \hat{\phi}_2 = -0.3096$. The final estimate of the model is shown in figure

4.11. The estimate from this model is -0.233, the standard deviation is 0.0884 and 0.0890 and the t ratio (equal to the estimate divided by the standard deviation) is -2.593. The MSE is 381.540. Finally, the computer output produces forecasts with 95 confidence limits for periods121-125 using the Minitab software.

Estimates at each iteration
Iteration SSE Parameters
0 19653622 0.100 16.963
1 18706667 -0.050 19.579
2 18547333 -0.137 21.195
3 18546932 -0.142 21.300
4 18546931 -0.142 21.306
5 18546931 -0.142 21.306
Relative change in each estimate less than 0.0010
Final Estimates of Parameters
Type Coef SE Coef T P
AR 1 -0.1419 0.0915 -1.55 0.124
Constant 21.31 36.50 0.58 0.561
Differencing: 1 regular difference
Number of observations: Original series 120, after differencing 119
Residuals: SS = 18546908 (backforecasts excluded)
MS = 158521 DF = 117
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag 12 24 36 48
Chi-Square 21.3 46.6 53.3 56.7
DF 10 22 34 46
P-Value 0.019 0.002 0.019 0.135
Forecasts from period 120

		95% Limits	
Period	Forecast	Lower	Upper	Actual
121	2887.72	2107.20	3668.25	
122	2908.79	1880.29	3937.28	
123	2927.10	1691.10	4163.10	
124	2945.81	1533.53	4358.09	
125	2964.46	1395.45	4533.47	

Figure 4.8 Box-Jenkins estimation of ARIMA (1, 1, 0) model (Note: computer-generated Minitab output)

4.7 Stage 3: Diagnostic Testing

To check the adequacy of the model, we estimated and plotted the autocorrelations of the residuals to determine whether they are significantly different from zero (Table 4.9). The limits of 2σ confidence intervals are

$$-2\frac{1}{\sqrt{119}} \leq r \leq 2\frac{1}{\sqrt{119}} \quad (4.2)$$

$$-0.183 \leq r \leq +0.183$$

Table 4.9 Autocorrelation coefficients for the residuals of the fitted ARIMA (2, 1, 0) model

Autocorrelations

Series:Morbidity

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	Df	Sig.[b]
1	-.142	.091	2.453	1	.117
2	-.279	.090	12.014	2	.002
3	.084	.090	12.881	3	.005
4	-.045	.089	13.136	4	.011
5	-.065	.089	13.676	5	.018
6	.035	.089	13.836	6	.032
7	.010	.088	13.848	7	.054
8	-.081	.088	14.689	8	.065

9	-.136	.087	17.111	9	.047
10	.194	.087	22.101	10	.015
11	.015	.087	22.129	11	.023
12	-.055	.086	22.536	12	.032
13	-.166	.086	26.278	13	.016
14	.007	.085	26.286	14	.024
15	.141	.085	29.050	15	.016
16	-.081	.085	29.973	16	.018
17	.131	.084	32.402	17	.013
18	.097	.084	33.736	18	.014
19	-.198	.083	39.372	19	.004
20	.100	.083	40.821	20	.004
21	.140	.082	43.681	21	.003
22	-.129	.082	46.145	22	.002
23	-.090	.082	47.371	23	.002
24	-.063	.081	47.981	24	.003
25	.062	.081	48.564	25	.003
26	.034	.080	48.742	26	.004
27	.015	.080	48.776	27	.006
28	-.058	.079	49.308	28	.008
29	-.130	.079	52.009	29	.005
30	.055	.079	52.493	30	.007
31	-.002	.078	52.493	31	.009
32	.003	.078	52.494	32	.013
33	.023	.077	52.579	33	.017
34	.050	.077	52.997	34	.020
35	.059	.076	53.588	35	.023
36	-.001	.076	53.588	36	.030

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

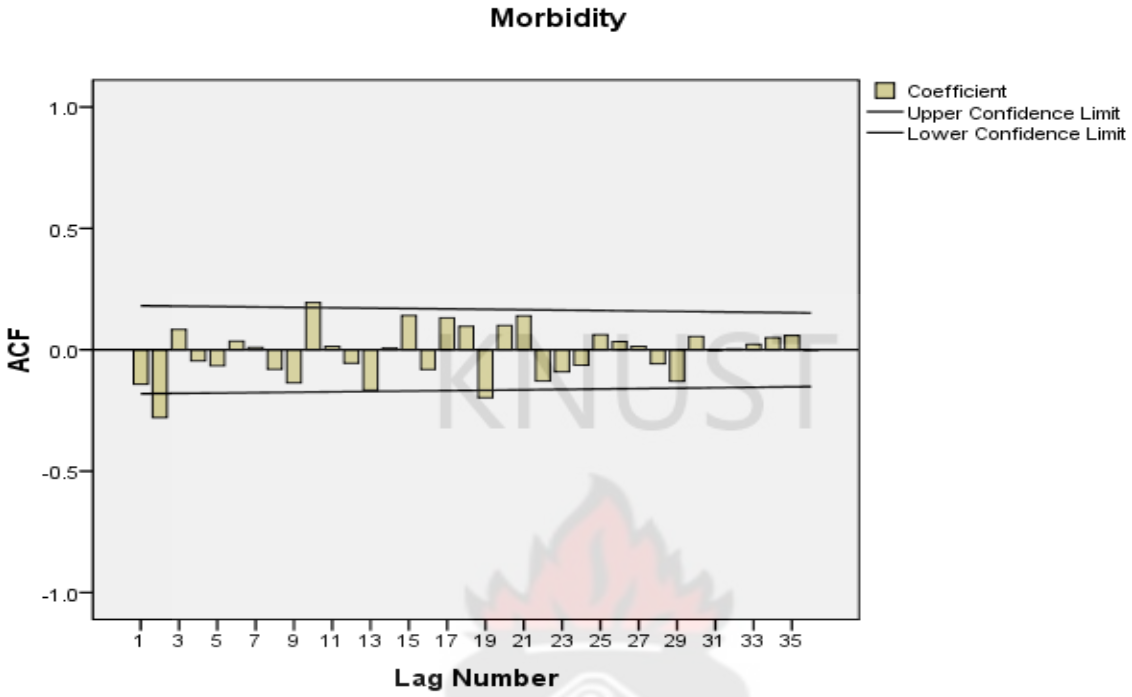


Figure 4.9 Autocorrelation coefficients for the residuals of the fitted ARIMA (2, 1, 0) model (Note: computer- generated SPSS output)

4.8 Stage 4: Forecasting

Figures4.1- 4.3 show that the time series data is not stationary in mean value, we then corrected through appropriate differencing of the data. In this case, we applied ARIMA (2, 1, 0) model.

Model parameters are shown as following Table 4.10

Table 4.10 Final Estimates of parameters

Final Estimates of Parameters				
Type	Coef	SE Coef	T	P
AR 1	-0.1868	0.0884	-2.11	0.037
AR 2	-0.3096	0.0890	-3.48	0.001
Constant	26.53	34.88	0.76	0.449

Estimated model parameters of malaria case model

We obtained the model in the form

$$Y_t = Y_{t-1} - 0.1868\,(Y_{t-1} - Y_{t-2}) - 0.3096(Y_{t-2} - Y_{t-3}) \quad (4.3)$$

with the MAPE = 14.358%.

Figure 4. 6, proof that the selected ARIMA (2, 1, 0) is an appropriate model.

We forecast for five steps ahead using the above ARIMA (2, 1, 0) model with observation number 121 as the starting value. See Appendix A for details.

The computer output shown in figure 4.10 indicates that the nonseasonal parameter estimates are significantly different from zero at low significance level. Their t ratios are -2.11 and -3.48 respectively. Three iterations were necessary to achieve the results. To indicate whether this model is adequate for forecasting, two test for randomness of the residuals were performed.

First, the ACFs for the residuals from the model ARIMA (2, 1, 0) are presented in figure 4.9. Note that all the autocorrelation coefficients of the residuals lie within 2σ limits of $\pm$ 0.183. Hence, ACF is white noise or random. Hence none of the autocorrelations are significantly different from zero (at 2σ limits), indicating that the model is adequate. Finally, the range of these autocorrelations is from-0.142 to+0.082 and shows no pattern.

The second test of the adequacy of the model is the Ljung-Box (Q)test. When the first $k = 24$ autocorrelations are used for the test, the null and alternative hypotheses are as follows:

H_0: the model is adequate

H_1: the model is not adequate

The number of degrees of freedom$(d.f.)$ is

$$d.f = k - p - q = 24 - 1 = 23$$

From the tabulated value of χ^2 we find $\chi^2 = 35.1725$ at a significance level of 0.05 for $d.f. = 23$.

The test statistic is the Q - statistic

$$Q = n(n+2)\sum_{k=1}^{m}\left(\frac{r_k^2(e)}{n-k}\right) \sim \chi^2_{m-r}; \quad (4.4)$$

where $r_k(e)$ = the residual autocorrelation at lag k

n = the number of residuals

m = the number of time lags included in the test

The calculated value is

$$Q = 120\text{x}122\left[\frac{(-0.142)^2}{119} + \frac{(-0.279)^2}{118} + \frac{(0.084)^2}{117}\right] \quad (4.5)$$

$= 120\text{x}122\text{x}0.000889 = 13.0211$

Since the calculated value is less than the tabulated value, we can accept the null hypothesis at the 0.05 level and conclude that the model is adequate.

The selection of the appropriate model depended on the values of Normalized BIC and the ACF together with the PACF. The graphs of the ACF and PACF are shown above in figures 4.6 and 4.7 respectively. Two tentative models were entertained and the model with the minimum Normalized BIC was chosen. The models and their corresponding Normalized BIC values have been illustrated below in table 4.11.

Table 4.11**: Tentative ARIMA Models and their corresponding normalized BIC.**

Model	**Normalized BIC**
ARIMA(2, 1,0)	0.713
ARIMA(1, 1, 0)	1.045

$$Y_t = Y_{t-1} - 0.1868\,(Y_{t-1} - Y_{t-2}) - 0.3096(Y_{t-2} - Y_{t-3}) \quad (4.6)$$

4.1

ARIMA Model: Morbidity

Estimates at each iteration				
Iteration	SSE	Parameters		
0	20840874	0.100	0.100	15.078
1	18419264	-0.005	-0.050	18.926
2	17083045	-0.110	-0.200	23.162
3	16793854	-0.183	-0.304	26.309
4	16793124	-0.187	-0.309	26.514
5	16793122	-0.187	-0.310	26.526
Relative change in each estimate less than 0.0010				
Final Estimates of Parameters				
Type	Coef	SE Coef	T	P
AR 1	-0.1868	0.0884	-2.11	0.037
AR 2	-0.3096	0.0890	-3.48	0.001
Constant	26.53	34.88	0.76	0.449
Differencing: 1 regular difference				
Number of observations: Original series 120, after differencing 119				
Residuals: SS = 16792900 (backforecasts excluded)				
MS = 144766 DF = 116				
Modified Box-Pierce (Ljung-Box) Chi-Square statistic				
Lag	12	24	36	48
Chi-Square	10.7	32.7	41.5	44.8

DF	9	21	33	45
P-Value	0.299	0.050	0.148	0.480

Forecasts from period 120

		95% Limits		
Period	Forecast	Lower	Upper	Actual
121	2736.60	1990.70	3482.49	
122	2748.30	1786.89	3709.71	
123	2818.90	1776.95	3860.85	
124	2828.61	1680.13	3977.10	
125	2831.47	1566.04	4096.90	

Figure 4.10 Forecast from period 120 using the tentative model (2, 1, 0)

The result of these forecasts for the first period is a function of the last period's actual observation and the last period's error. All future forecast (periods121-125) are based only on predicted values of Y_t since the future values of e_t are unknown. We can now calculate the forecast error and its variance for the ARIMA (1, 1, 0) model to obtain a forecast confidence interval. As we shall see, the forecast confidence interval Y_t is related to the forecast confidence interval for the differenced series BY_t.

Estimate of the forecasts and forecast errors by extrapolation using ARIMA (1, 1, 0)

$$Y_t = Y_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \mu + e_t \tag{4.7}$$

To begin, we restate the fitted forecast model:

$$(1 - B)(1 - \phi_1 B)Y_t = e_t \tag{4.8}$$

Note the use of the backshift operator to describe the first difference $(1 - B)$ and the AR portion of the model $(1 - \phi_1 B)$. The terms can be multiplied out and rearranged as follows:

$$(1 - \phi_1 B - B + \phi_1 B^2) Y_t = e_t$$

$$Y_t = Y_{t-1} + \phi_1 (Y_{t-1} - Y_{t-2}) + e_t \quad (4.9)$$

For details see Appendix A

Table 4.12 Comparison of two AR (1) and AR (2) models

		ARIMA(1,1,0)	ARIMA(2,1,0)
		1 degree of differencing	*1 degree of differencing*
Ljung-Box(Q)		46.6	32.7
d.f		23	23
Forecast errors:			
Period	*Period Ahead*		
121	1	−136.551	175.9344
122	2	362.627	9.8603
123	3	358.819	−56.3112
124	4	355.0514	7.4662
125	5	351.3234	−329.376
Variance of residuals (MSE)		0.0984	0.0882
AR parameter		-0.1395	-0.3068
t value		-1.53	-3.46

We also compared the ARIMA (2, 1, 0) model estimated above with an ARIMA (1, 1, 0) model (figure 4.9). The purpose is to determine which model is most adequate.

The results of the AR (1) and AR (2) models are compared in table 4.12. The value of Q is smaller for the (2, 1, 0) differenced model than for the model (1, 1, 0) with differencing. The model (2, 1, 0) with differencing appears more suitable as judged by this measure. If the forecast

errors for 1-5 periods into the future are checked, the ARIMA (1, 1, 0) model has slightly closer estimates. However, the variance in the residuals (MSE) is smaller for the AR (2, 1, 0) model which is 0.0882. Most important is the coefficient for the ARIMA (1, 1, 0) model, which is 0.0984 and hence the model AR (1, 1, 0) with differencing is not acceptable for forecasting. Finally, the autocorrelation function (ACF) for the residuals from the (1, 1, 0) model indicates the r_1 and r_2 are outside the 2σ limits. Thus the model does not yield evidence of stability in the forecasts.

To begin, consider the forecast error for the one-period-ahead forecast $Y_1(1)$:

$$e_t(1) = Y_{t+1} - \hat{Y}_t(1) \tag{4.10}$$

where the circumflex indicates the estimated value. In turn, if we assign $Z_t = Y_t - Y_{t-1}$, then

$$e_t(1) = Y_t + Z_{t+1} - Y_t - \hat{Z}_t(1) = e_{t+1} \tag{4.11}$$

which has a variance σ_e^2. The two-period ahead forecast is given by

$$e_t(2) = Y_{t+2} - \hat{Y}_t(2) \tag{4.12}$$

$$= Y_t + Z_{t+1} + Z_{t+2} - Y_t - \hat{Z}_t(1) - \hat{Z}_t(2)$$

$$= \left[Z_{t+1} - \hat{Z}_t(1)\right] + \left[Z_{t+2} - \hat{Z}_t(2)\right]$$

$$= (1 + \phi_1)e_{t+1} + e_{t+2} \tag{4.13}$$

This has a variance

$$E[\sigma_t^2(2)] = \sigma_e^2[(1 + \phi_1)^2 + 1] \tag{4.14}$$

The forecast error is equal to the two-period error for $\hat{Z}_t(2)$ in addition to the one-period error for$\hat{Z}_t(1)$. Thus the error in$\hat{Y}_t(2)$ is an accumulation of the forecast errors in$\hat{Z}_t(1)$ and $\hat{Z}_t(2)$:

$$e_t(n) = \sum_{i=1}^{n} e_{t+1} \sum_{j=0}^{n-i} \phi_1^j \quad (4.15)$$

And this has a variance

$$E[\sigma_t^2(n)] = \sigma_e^2 \sum_{i=1}^{n} \left(\sum_{j=0}^{n=1} \phi_1^j\right)^2 \quad (4.16)$$

The above forecast values show that malaria cases in five years ahead will be in the range given in the above table and if care is not taken, escalating cases will be witnessed in the coming years.

In conclusion, although we examined properties of only a simple ARIMA model, some of our conclusions apply to more complicated (higher-order) ARIMA models. In particular, an MA model of order q has a memory of only q periods, since there are only q error terms.

Finally, practicing forecasters find it very useful to compare forecasts of several competing fitted time series models as we did in table 4.11. This is particularly helpful if the forecaster is having difficulty in choosing among a variety of possible models.

At least 50 observations are usually required for Box-Jenkins estimation. Even more observations are recommended of a seasonal model.

CHAPTER 5

5.0 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Discussion of Results

The study shows that in the first quarter of 2001, the month January recorded the highest prevalence of malaria cases (655). In the second quarter of that same year, the highest reported case was in June with a total number of 897. The highest number of reported cases in third quarter was in September with a total number of 572. In the last quarter of that same year, the highest number of reported cases was in December. In all, the highest number of reported cases in the whole year was in December (1161) and the least number was in the month of July with a total number of 475. The number of reported cases per month was 717.3.

In the year 2002, the highest number of reported cases in the first quarter was in the month of January with a number of 966. June had the highest number of cases in the second quarter with a total number of 1182. In the third quarter, the highest number was 1202 and it is recorded in July. In the fourth quarter, the highest number was 1391 in December. In all, the highest number of reported cases in 2002 was in the month of December. The least number of reported cases was 451 and it was recorded in February. The average monthly number of reported cases in the year was 1058.8. In the year 2003, the study indicates that the highest number of reported cases in the first quarter was in the month of March with a total of 1059 cases. The highest number in the second quarter was 1037 cases in the month of June. The third quarter recorded the highest number in September with a total of 1224 cases. The highest number of cases in the fourth quarter was in December with a total of 1377. The highest number of reported cases in the whole

year was in December and the least number of reported cases was 992 which occurred in January. The average reported a case per month of the year was 1130.3.

The study also shows that the highest number of reported cases in 2004 for the first quarter was in the month of March with a total of 997 cases. The second quarter had the highest value in April with a total of 976 cases. In the third quarter, the highest number of reported cases was 917 in September. The fourth quarter had a total of 1402 cases in December. The highest number of reported cases for the whole year was in January with a total of 1402 cases. The least number of reported cases in the whole year was 379 in the month of May. The average monthly reported cases were 870.6.

In 2005, the highest value in the first quarter was in the month of March with a total of 1721 cases. The second quarter had its highest value in the month of June with a total of 2023 cases. The highest number of reported cases for the third quarter was in the month of June with a total of 2023 cases. The fourth quarter had a record of 2029 cases in the month of December as the highest number of reported cases. The highest number of reported cases for the whole year of 2005 was 2038 in the month of August. The number of reported cases per month in that year was 1875.9. The rest of the analysis can be referenced from Tables 4.1 and 4.2 respectively.

Comparatively, the number of reported cases per month (717) in 2001 indicates the lowest among all the years. This figure indicates that on the average, 717 people had malaria in each month in 2001. The average monthly reported case in 2002 is 1059.This implies that 1059 had malaria every month in 2002. This indicates an increase in 342 the average reported cases. This probably could be attributed to the increase in the population of the people in that area. In 2003, the reported cases per month further increased to 1130. In 2004 the number of reported cases per month was 871. This figure however indicates a drop in the number of reported cases. In 2005,

an average of 1876 cases was recorded. Comparing the number of reported cases per month in 2005 with 2004, there is a further increase in the number of reported cases. The least number of reported cases per month is in 2001. Observing the trend of the reported cases from 2001 to 2004, it is most likely that the number of reported cases of the disease will increase in 2010 with the implementation of the National Health Insurance scheme.

Comparing the number of reported cases per month in 2001 with that of 2002, the result indicates that there is a significant difference in the number of reported cases. The number of reported cases in 2001 is lower than that of 2002. One is 95% confident that the difference in the means of 2001 and 2002 is between -672.76 and -10.07. The interval does not contain zero and this confirms the fact that the number of reported cases per month between these years are significantly different from zero. Comparing the number of reported cases in 2001 with that of 2003 indicates that the number of reported cases in 2001 is significantly different from that of 2003. One can say there is 95% confidence that the difference in the number of reported cases per month in 2001 and 2003 is between-744.26 and -81.57. Also comparing the number of reported cases in 2005 with 2008 and 2009 indicate the average number of monthly reported cases in 2005 is significantly different from those of 2008 and 2009. The number of reported cases per month in 2005 is lower than those of 2008 and 2009. There is 95% confidence that the difference in the number of reported cases per month in 2005 and 2008 is between -792.66 and -129.99. There is 95% confidence that the difference in the number of reported cases per month in 2005 and 2009 is between -1265.76 and 603.07.

Comparing the number of monthly reported cases per month in 2006 and 2001 to 2004 it is realized that there is no significant difference in their means. There is 95% confidence that the average number of reported cases per month is between 1390.57and 2053.26, 1049.15 and

1711.85, 977.65 and 1640.35, 1237.32 and 1900.35 respectively. However, considering the number of reported cases per month in 2006 and 2007, there is statistical evidence that the average number of reported cases in 2006 is different from that of 2007. The average number of reported cases in 2006 is 2439 and that of 2007 is 2114.67. So in 2006, the average number of reported cases is higher than that of 2007. There is 95% confidence that the difference in means is between -6.76 and 655.93. Considering the number of reported cases per month in 2006 and 2008, there is statistical evidence that the average number of reported cases in 2006 is different from that of 2008. The average number of reported cases in 2006 is 2439 and that of 2008 is 2337.25. So in 2006, the average number of reported cases are higher than that of 2008. There is 95% confidence that the difference in means is between -229.35 and 433.35. There is evidence that the difference in the average number of reported cases in 2006 and 2009 is significantly different from zero. The average number of reported cases in 2009 is 2810. There is 95% confidence that the difference in the average number of cases in 2006 and 2009 is between-702.43 and -39.74. There is the evidence that the difference in the number of reported cases per month in 2007 and 2008 is different from zero. There is 95% confidence that the difference in the number of reported cases per month is between -553.93 to 108.76. In 2007 and 2009, there is evidence that the difference in the number of reported cases per month is significantly different from zero. There is 95% confidence that the difference in the number of reported cases in the years 2007 and 2009 is between -1027.01 and -364.32. There is statistical evidence that the number of reported cases per month in 2008 and 2009 is significantly different from zero. There is 95% confidence that the difference in the number of reported cases per month in 2008 and 2009 is between -804.43 and -141.74.

The time series model developed for predicting the number of reported cases of malaria in the Sekyere-East District is ARIMA (2, 1, 0). This model can be used by researchers for forecasting malaria reported cases in the district. However, it should be updated from time to time with the incorporation of current data.

5.2 Findings

The study reveals that the introduction of the National Health Insurance Scheme is having positive impact on the malaria reported cases in the District. Since its inception in 2002, the reported cases per month in subsequent years from 2001 to 2010 have increased significantly. Despite the fact that people have been complaining about the low-cost drugs administer at the health centre under health insurance scheme, they still breathe a sigh of relief for the fact that many people are able to access health care with a minimum cost.

In general, Malaria is not only a health problem but also a developmental problem in Ghana. It imposes significant financial hardships on households and the national economy. The burden of malaria is, therefore, a challenge to human development manifesting itself as a cause and consequence of under-development. Malaria's impact on households and society can be assessed in at least three important dimensions namely, health, social and economic. The impact of malaria in all the dimensions is to a large extent little appreciated, especially with the emergence of the HIV/AIDS pandemic.

5.3 Summary of Results

Comparing the average monthly reported cases of 2001 with each year's reported cases, it is realized that the number of reported cases has been increasing significantly over the years. The number of reported cases from 2007 to 2010 is higher than that of preceding years, the factor attributable to increase in population over the years and also for the fact that the community has

seen the usefulness of the National Health Insurance scheme and now access the health facilities more than ever before. For details check from Appendix B

The ARIMA model developed for predicting the monthly reported malaria cases is ARIMA (2, 1, 0):$(Y_t - Y_{t-1}) = \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + e_t$

$$(Y_t) = Y_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + e_t$$

$$(Y_t) = Y_{t-1} - 0.1868(Y_{t-1} - Y_{t-2}) - 0.3096(Y_{t-2} - Y_{t-3}) + e_t \quad (5.1)$$

The model was used to predict a five-month lead period of the reported cases.

See Appendix A for details.

5.4 Conclusion

The model was essentially a stand-alone model since no relevant inputs for the model were available. Nevertheless, reasonable fitting accuracy (R square = 0.053, MAPE =14.358) was achieved for 120 months of historical data and the generated forecasts were adequately accurate. Based on the accuracy of forecasts obtained from the various models built in this thesis, it was demonstrated that Box and Jenkins ARIMA model can be successfully employed for the purpose of forecasting time dependent series.

The results of this work demonstrates the usefulness and motivates the need of employing the statistical technique of ARIMA methodology in forecasting time series applications, either independently or in conjunction with the traditional methods, to result in a less computationally and data intensive method. Nevertheless, some work still needs to be done to validate further the use of such techniques in various other scenarios faced by time series forecasting.

It can be concluded that the Sekyere-East District has conditions that favour the breeding of mosquitoes, the vector that causes malaria. The prevalence of malaria is more pronounced among certain population subgroups. Malaria presents significant costs to the affected

households since it is possible to experience multiple and repeated attacks in a year. The district, which has the two rainy seasons, is the hardest hit by these vectors because of the weather. In this case, the district must be given priority attention in annual budgets to enable it combat the disease. In particular, there is the need for a strong collaboration among major stakeholders including the Government, District Assemblies, Non-Governmental Organisations and the community to devise holistic, effective, and cost-saving methods for prevention, control and treatment of the disease. Though the use of insecticides for example coils, sprays are identified as the major method of protection due to their availability and affordability for many households, the efficacy of some of these numerous brands on the market may be questionable. In the short-term, the efficacy of these products needs to be assessed by concerned authorities in order not to endanger the health of the people.

5.5 RECOMMENDATIONS

While advocating continuation of education on the use of the ITNs, it is recommended that efforts must be seriously made by the major players in the health sector to make the net readily available in the communities at low prices to enable the ordinary Ghanaian to purchase it.

The decision to seek medical care from a health provider is influenced by several factors but the perceived quality of the provider and the proximity of the health facility are major determinants of health seeking behaviours. The proximity of the facility affects the cost of transportation and more importantly the cost of time. In order to improve timeliness of treatment, the service consequently would have to be closer to patients especially those in the remote and malarious endemic areas like the Asukorkor, Apemso, Naama, Ahensan etc in the district. The mobile outreach programme of the Ghana Health Service must be well equipped so that difficulty could be minimised at the service. Malaria reduction strategies should be incorporated into Ghana's

Poverty Reduction Strategy. It is anticipated that with a considerable reduction in poverty levels, households and communities would become increasingly responsible for the improvement of their health status and quality of life.

REFERENCES

1. Briet O. J.T., Vounatsou P., Amerasinghe, P. H., (2008). "Malaria seasonality and rainfall seasonality in Sri Lanka are correlated in space", Geospatial Health 2(2), pp. 183-190

2. Brown, V., Issak, M.A., Rossi, M., Barboza, P. and Paugam, A. (2002). "Epidemic of malaria in North Eastern Kenya", Lancet 352, 1356–1357.

3. Center for Disease Control and Prevention (CDC), National Center for infectious Disease, Malaria Epideminology Branch, (2004). "Global Malaria Prevention and Control Programme". http//www. cdc.gov/ncidod/dpd/parasites/malaria. Accessed

4. Contreras, J., Espínola, R., Nogales, F. J. andAntonio J. Conejo, A. J., (2003). "IEEE Transactions on Power Systems", Vol. 18, No. 3, pp.1014-1020.

5. Fillatre, L., Marakov, D., Vaton, S. (2003). "Forecasting seasonal trafficflows", Computer Science Department, ENST Bretagne, Brest, Paris.

6. Gallup and Sachs, (2001); The Economic Burden of malaria. Centre for International Development at Harvard University.

7. __Ghana Statistical Service (2000); Poverty Trends in Ghana in the 1990s. GSS, Accra.
   _,(2000); Population and Housing Census in Ghana. GSS, Accra.

8. Glenn *et al.* (1996); Economic cost of Health; Effects of occupational exposure to Hazardous substance. University of Sidney.

9. Guerin, P. J., Olliaro, P., Nosten, F., Druilhe, P., Laxminarayan, R., Binka, F., Kilama, W.L., Ford,

10. Guerrero, V. M., (2003). "Análisis estadístico de series de tiempo económicas", Segunda edición, México.

11. Houeto, D., D‟Hoore, W., Ouendo, E. M., Charlie, D. and Deccache, A., (2007). "Malaria control among children under five in sub-Saharan Africa: the role of empowerment and parents‟ participation besides the clinical strategies", ARHEN. Kendall, M. G., and A. Stuart. 1966. "The advanced theory of statistics: Design and Analysis and Time-Series", Charles Griffin & Co. Ltd., London, United Kingdom, Vol. 3.

12. Lin, H., Lu, L., Tian, L., Zhou, S., Wu, H., B, Y., Ho, S. C. and Liu, Q., (2009). "Spatial and temporal distribution of falciparum malaria in China", Malaria Journal, 8:130 doi:10.1186/1475-2875-8-130

13. Lindsay W. S and Birley M., (2004). "Rural Development and Malaria Control in Sub-Sahara Africa", EcoHealth, Springer New York, Volume 1 Number 2. pp 129-137.

14. McKenzie D, 2003. "Measuring Inequality with Asset Indicators". Bureau for Research in Economic Analysis of Development Working Paper No. 042: Cambridge, MA: Center for International Development, Harvard University.

15. Nchinda, T.C. (2005). "Malaria: A Reemerging Disease in Africa", World Health Organization, Geneva.

16. Nkuo-Akenji, T., Ntonifor, N. N, Ndukum M. B, Kimbi H. K, Abongwa E. L, Nkwescheu A, Anong D. N, Songmbe, M, Boyo M. G, Ndamukong, K. N. and Titanji, V. P. K. (2006).

17. Pattanayak, S.K., Mercer D.E., Sills, E. and Yang, J. C. (2003). "Taking Stock of the Agroforestry Adoption Studies", Agroforestry Systems 57 (3): 173-186.

18. Rashed, S. et al. (2000), "Economic Impact of Febrile Morbidity and Use of

Permethrin- Treated Bed Nets in a Malarious Area II" in Social Science and Medicine, pp49, Toronto, Canada

19. Reed, S.D., Laxminarayan, R., Black, D.J., and Sullivan, S.D., (2002). "Economic issues and antibiotic resistance in the community." The Annals of Pharmacotherapy, 36(1): 148-54

20. Streeten, Paul (1981); First thing first; meeting basic human needs in developing countries. The World Bank. Washington.

21. Suarez, C.A.H, Parra, O. J. S. and Diaz, A. E. (2009). Revista Ingenieria Investigacion, Volume 29 Number 2. pp 65-67.

22. Sudhakar, P., and Subramani, P., (2007). "Literature Review: Insights into Formulating a Protective Malaria Medicine, Journal of Young Investigators", Volume 16, Issue 1. Sulaiman, S., (2007). How is climate change shifting Africa‟s malaria map? http//www.scidev.net.com. Accessed on Friday, 9 October 2009.

23. The African Summit on Roll Back Malaria (2005) Abuja, Nigeria, April 25 2000, World Health Organization, Geneva (document WHO/CDS/RBM/2000.17). 79

24. UN (2003); Millennium Indicators; Combat HIV/AIDS, Malaria and other Diseases. United Nations Statistical Division. *http://millenniumindicators.un.org/*unsd/mi/mi_goals.asp

25. UNDP (2002); 'Science, Technology and Development'. Ghana Human Development

Report 2000.

26. WHO (2005a) The Roll Back Malaria strategy for improving access to treatment through home management of malaria WHO, Geneva (document WHO/HTM/MAL/2005.1101)

27. Wilson, M.L., (200)1. "Ecology and Infectious Disease." In Ecosystem Change and Public Health, The Johns Hopkins University Press: Baltimore, MD: 285-291.

28. Wiwanitkit V. (2006). Hellenic Society of Haematology, "Correlation between Prevalence of Malaria and Altitude, a Case Study in a Rural Endemic area of Thailand", HAEMA 2006; 9(1). pp 56-58.

29. Worrall, E., Basu, S. and Hanson, K., (2003). **"**The relationship between socio-economic status and malaria", London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.

30. WHO (1990); World Malaria Situation. World Health Statistics, Quarterly 43.
___, (1992); Health Dimensions of Economic Reform. Geneva.
__, (1997); World Malaria Situation in 1994, part 1. WHO weekly Epidemiological Record 36: 269-274.
___, (1999); World Health Report, 1999; *making a Difference*. *http:/www.who.int/whr/1999/*
___, (2002); "Macroeconomics and Health Initiative in Accra". Press Release No. 013/02. *http:/www.who.int/macrohealth/infocentre/press/bulletin/en/html*

31. WHO /AFRO, (2001); A framework for Estimating the Economic Burden of Malaria in the African Region. WHO, Harare, Zimbabwe.

32. WHO/UNICEF, (2003); Africa Malaria Report. WHO/CDS/MAL/2003.1093.2003. Geneva.

5.8 APPENDICES

Appendix A
Estimate of the forecast by extrapolation using ARIMA (1, 1, 0)

Forecast of period 121,

We can now substitute in the equation $\phi_1 = 0.9895$ and Y_{t-1}andY_{t-2}. To forecast for period 121,

$$\hat{Y}_{121} = Y_{120} + \phi_1(Y_{120} - Y_{119})$$

with e_t set equal to zero. By substitution we have,

$$\hat{Y}_{121} = 2886 + 0.9895(2886 - 2748)$$

$$\hat{Y}_{121} = 3022.551$$

Forecast error of period 121

$$e_t(1) = Y_{120} - \hat{Y}_{121}$$

$$e_t(1) = 2886 - 3022.551$$

$$e_t(1) = -136.551$$

Forecast of period 122,

$$\hat{Y}_{122} = Y_{121} + \phi_1(Y_{121} - Y_{120})$$

$$\hat{Y}_{122} = 3022.551 + 0.9895(3022.551 - 2886)$$

$$\hat{Y}_{122} = 3157.668$$

$$= 2659.924$$

Forecast error of period 122,

$$e_t(2) = Y_{121} - \hat{Y}_{122}$$

$$e_t = 3022.551 - 2659.924$$

$$= 362.627$$

Forecast of period 123,

$$\hat{Y}_{123} = Y_{122} + \phi_1(Y_{122} - Y_{121})$$

$$\hat{Y}_{123} = 2659.924 + 0.9895(2659.924 - 3022.551)$$

$\hat{Y}_{123} = 2301.105$

Forecast error of period 123

$e_t(3) = Y_{122} - \hat{Y}_{123}$

$e_t(3) = 2659.924 - 2301.105$

$e_t(3) = 358.819$

Forecast of period 124,

$\hat{Y}_{124} = Y_{123} + \phi_1(Y_{123} - Y_{122})$

$\hat{Y}_{124} = 2301.105 + 0.9895(2301.105 - 2659.924)$

$\hat{Y}_{124} = 1946.0536$

Forecast error of period 124

$e_t(4) = Y_{123} - \hat{Y}_{124}$

$e_t(4) = 2301.105 - 1946.0536$

$e_t(4) = 355.0514$

Forecast of period 125,

$\hat{Y}_{125} = Y_{124} + \phi_1(Y_{124} - Y_{123})$

$\hat{Y}_{125} = 1946.0536 + 0.9895(1946.0536 - 2301.105)$

$\hat{Y}_{125} = 1594.7302$

Forecast error of period 125,

$e_t(5) = Y_{124} - \hat{Y}_{125}$

$e_t(5) = 1946.0536 - 1594.7302$

$e_t(5) = 351.3234$

$= 2886 - 3227.9075$

$= -341.9075$

In a similar manner, we can forecast for more periods ahead.

Estimate of the forecast by extrapolation using ARIMA (2, 1, 0)

$$Y_t = Y_{t-1} - 0.1868\,(Y_{t-1} - Y_{t-2}) - 0.3096(Y_{t-2} - Y_{t-3}) \quad (4.4)$$

To use this equation to forecast one period ahead, Y_{t+1}, = the subscripts are increased by 1 throughout, as in the following equation:

$$\hat{Y}_{t+1} = Y_t - 0.1868(Y_t - Y_{t-1}) - 0.3096(Y_{t-1} - Y_{t-2})$$

Forecast of the period 121

$$\hat{Y}_{121} = Y_{120} - 0.1868(Y_{120} - Y_{119}) - 0.3096(Y_{119} - Y_{118})$$

$$\hat{Y}_{121} = 2886 - 0.1868(2886 - 2748) - 0.3096(2748 - 2263)$$

$$\hat{Y}_{121} = 2710.0656$$

Forecast error of the period 121

$$e_t(1) = Y_{120} - \hat{Y}_{121}$$

$$e_t(1) = 2886 - 2710.0656$$

$$e_t(1) = 175.9344$$

Forecast of the period 122,

$$\hat{Y}_{122} = Y_{121} - 0.1868(Y_{121} - Y_{120}) - 0.3096(Y_{120} - Y_{119})$$

$$\hat{Y}_{122} = 2710.0656 - 0.1868(2710.0656 - 2886) - 0.3096(2886 - 2748)$$

$$\hat{Y}_{122} = 2700.2053$$

Forecast error of the period 122

$$e_t(2) = Y_{121} - \hat{Y}_{122}$$

$$e_t(2) = 2710.0656 - 2700.2053$$

$$e_t(2) = 9.8603$$

Forecast of the period 123

$$\hat{Y}_{123} = Y_{122} - 0.1868(Y_{122} - Y_{121}) - 0.3096(Y_{121} - Y_{120})$$

$$\hat{Y}_{123} = 2700.2053 - 0.1868(2700.2053 - 2710.0656) - 0.3096(2710.0656 - 2886)$$

$$\hat{Y}_{123} = 2756.5165$$

Forecast error of the period 123

$$e_t(3) = Y_{122} - \hat{Y}_{123}$$

$$e_t(3) = 2700.2053 - 2756.5165$$

$$e_t(3) = -56.3112$$

Forecast of the period 124

$$\hat{Y}_{124} = Y_{123} - 0.1868(Y_{123} - Y_{122}) - 0.3096(Y_{122} - Y_{121})$$

$$\hat{Y}_{124} = 2756.5165 - 0.1868(2756.5165 - 2700.2053) - 0.3096(2700.2053 - 2710.0656)$$

$$\hat{Y}_{124} = 2749.0503$$

Forecast error of the period 124

$$e_t(4) = Y_{123} - \hat{Y}_{124}$$

$$e_t(4) = 2756.5165 - 2749.0503$$

$$e_t(4) = 7.4662$$

Forecast of the period 125

$$\hat{Y}_{125} = Y_{124} - 0.1868(Y_{124} - Y_{123}) - 0.3096(Y_{123} - Y_{122})$$

$$\hat{Y}_{125} = 2749.0503 - 0.1868(2749.0503 + 56.3112) - 0.3096(-56.3112 - 2700.2053)$$

$$\hat{Y}_{125} = 3078.4263$$

Forecast error of the period 125

$$e_t(5) = Y_{124} - \hat{Y}_{125}$$

$$e_t(5) = 2749.0503 - 3078.4263$$

$$e_t(5) = -329.376$$

Appendix B

ANOVA

Morbidity						
		Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)	6.688E7	9	7431188.370	44.304	.000

	Linear term	Contrast	5.805E7	1	5.805E7	346.082	.000
		Deviation	8832134.096	8	1104016.762	6.582	.000
Within Groups			1.845E7	110	167730.580		
Total			8.533E7	119			

Post Hoc Tests

Multiple Comparisons

Morbidity

LSD

(I) YEAR, not periodic	(J) YEAR, not periodic	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
	2002	-341.417*	167.198	.044	-672.76	-10.07
	2003	-412.917*	167.198	.015	-744.26	-81.57
	2004	-153.250	167.198	.361	-484.60	178.10
	2005	-1158.583*	167.198	.000	-1489.93	-827.24
	2006	-1721.917*	167.198	.000	-2053.26	-1390.57
	2007	-1397.333*	167.198	.000	-1728.68	-1065.99
	2008	-1619.917*	167.198	.000	-1951.26	-1288.57
	2009	-2093.000*	167.198	.000	-2424.35	-1761.65
	2010	-1929.167*	167.198	.000	-2260.51	-1597.82
2002	2001	341.417*	167.198	.044	10.07	672.76
	2003	-71.500	167.198	.670	-402.85	259.85
	2004	188.167	167.198	.263	-143.18	519.51
	2005	-817.167*	167.198	.000	-1148.51	-485.82
	2006	-1380.500*	167.198	.000	-1711.85	-1049.15
	2007	-1055.917*	167.198	.000	-1387.26	-724.57
	2008	-1278.500*	167.198	.000	-1609.85	-947.15
	2009	-1751.583*	167.198	.000	-2082.93	-1420.24
	2010	-1587.750*	167.198	.000	-1919.10	-1256.40
2003	2001	412.917*	167.198	.015	81.57	744.26
	2002	71.500	167.198	.670	-259.85	402.85
	2004	259.667	167.198	.123	-71.68	591.01

	2005	-745.667*	167.198	.000	-1077.01	-414.32
	2006	-1309.000*	167.198	.000	-1640.35	-977.65
	2007	-984.417*	167.198	.000	-1315.76	-653.07
	2008	-1207.000*	167.198	.000	-1538.35	-875.65
	2009	-1680.083*	167.198	.000	-2011.43	-1348.74
	2010	-1516.250*	167.198	.000	-1847.60	-1184.90
2004	2001	153.250	167.198	.361	-178.10	484.60
	2002	-188.167	167.198	.263	-519.51	143.18
	2003	-259.667	167.198	.123	-591.01	71.68
	2005	-1005.333*	167.198	.000	-1336.68	-673.99
	2006	-1568.667*	167.198	.000	-1900.01	-1237.32
	2007	-1244.083*	167.198	.000	-1575.43	-912.74
	2008	-1466.667*	167.198	.000	-1798.01	-1135.32
	2009	-1939.750*	167.198	.000	-2271.10	-1608.40
	2010	-1775.917*	167.198	.000	-2107.26	-1444.57
2005	2001	1158.583*	167.198	.000	827.24	1489.93
	2002	817.167*	167.198	.000	485.82	1148.51
	2003	745.667*	167.198	.000	414.32	1077.01
	2004	1005.333*	167.198	.000	673.99	1336.68
	2006	-563.333*	167.198	.001	-894.68	-231.99
	2007	-238.750	167.198	.156	-570.10	92.60
	2008	-461.333*	167.198	.007	-792.68	-129.99
	2009	-934.417*	167.198	.000	-1265.76	-603.07
	2010	-770.583*	167.198	.000	-1101.93	-439.24
2006	2001	1721.917*	167.198	.000	1390.57	2053.26
	2002	1380.500*	167.198	.000	1049.15	1711.85
	2003	1309.000*	167.198	.000	977.65	1640.35
	2004	1568.667*	167.198	.000	1237.32	1900.01
	2005	563.333*	167.198	.001	231.99	894.68
	2007	324.583	167.198	.055	-6.76	655.93
	2008	102.000	167.198	.543	-229.35	433.35
	2009	-371.083*	167.198	.029	-702.43	-39.74

	2010	-207.250	167.198	.218	-538.60	124.10
2007	2001	1397.333*	167.198	.000	1065.99	1728.68
	2002	1055.917*	167.198	.000	724.57	1387.26
	2003	984.417*	167.198	.000	653.07	1315.76
	2004	1244.083*	167.198	.000	912.74	1575.43
	2005	238.750	167.198	.156	-92.60	570.10
	2006	-324.583	167.198	.055	-655.93	6.76
	2008	-222.583	167.198	.186	-553.93	108.76
	2009	-695.667*	167.198	.000	-1027.01	-364.32
	2010	-531.833*	167.198	.002	-863.18	-200.49
2008	2001	1619.917*	167.198	.000	1288.57	1951.26
	2002	1278.500*	167.198	.000	947.15	1609.85
	2003	1207.000*	167.198	.000	875.65	1538.35
	2004	1466.667*	167.198	.000	1135.32	1798.01
	2005	461.333*	167.198	.007	129.99	792.68
	2006	-102.000	167.198	.543	-433.35	229.35
	2007	222.583	167.198	.186	-108.76	553.93
	2009	-473.083*	167.198	.006	-804.43	-141.74
	2010	-309.250	167.198	.067	-640.60	22.10
2009	2001	2093.000*	167.198	.000	1761.65	2424.35
	2002	1751.583*	167.198	.000	1420.24	2082.93
	2003	1680.083*	167.198	.000	1348.74	2011.43
	2004	1939.750*	167.198	.000	1608.40	2271.10
	2005	934.417*	167.198	.000	603.07	1265.76
	2006	371.083*	167.198	.029	39.74	702.43
	2007	695.667*	167.198	.000	364.32	1027.01
	2008	473.083*	167.198	.006	141.74	804.43
	2010	163.833	167.198	.329	-167.51	495.18
2010	2001	1929.167*	167.198	.000	1597.82	2260.51
	2002	1587.750*	167.198	.000	1256.40	1919.10
	2003	1516.250*	167.198	.000	1184.90	1847.60
	2004	1775.917*	167.198	.000	1444.57	2107.26

	2005	770.583*	167.198	.000	439.24	1101.93
	2006	207.250	167.198	.218	-124.10	538.60
	2007	531.833*	167.198	.002	200.49	863.18
	2008	309.250	167.198	.067	-22.10	640.60
	2009	-163.833	167.198	.329	-495.18	167.51

*. The mean difference is significant at the 0.05 level.

APPENDIX C

Minitab output of the comparison of the average monthly reported cases.

Two-Sample T-Test and CI: 2001, 2002

Sample N Mean StDev SEMean
1 12 717 184 53
2 12 1059 267 77
Difference = mu (1) - mu (2)
Estimate for difference: -341.5
95% CI for difference: (-537.3, -145.7)
T-Test of difference = 0 (vs not =): T-Value = -3.65 P-Value = 0.002 DF = 19

Two-Sample T-Test and CI: 2001, 2003

Sample N Mean StDev SEMean
1 12 717 184 53
2 12 1130 144 41
Difference = mu (1) - mu (2)
Estimate for difference: -413.0
95% CI for difference: (-553.4, -272.6)
T-Test of difference = 0 (vs not =): T-Value = -6.14 P-Value = 0.000 DF = 20

Two-Sample T-Test and CI: 2001, 2004

Sample N Mean StDev SEMean
1 12 717 184 53
2 12 871 280 81
Difference = mu (1) - mu (2)
Estimate for difference: -153.3
95% CI for difference: (-356.3, 49.7)
T-Test of difference = 0 (vs not =): T-Value = -1.59 P-Value = 0.130 DF = 18

Two-Sample T-Test and CI: 2001, 2005

Sample N Mean StDev SE Mean
1 12 717 184 53
2 12 1876 231 67
Difference = mu (1) - mu (2)
Estimate for difference: -1158.6
95% CI for difference: (-1336.1, -981.1)

T-Test of difference = 0 (vs not =): T-Value = -13.62 P-Value = 0.000 DF = 20

Two-Sample T-Test and CI: 2001, 2006

Sample	N	Mean	StDev	SE Mean
1	12	717	184	53
2	12	2439	478	138

Difference = mu (1) - mu (2)
Estimate for difference: -1722
95% CI for difference: (-2039, -1405)
T-Test of difference = 0 (vs not =): T-Value = -11.65 P-Value = 0.000 DF = 14

Two-Sample T-Test and CI: 2001, 2007

Sample	N	Mean	StDev	SE Mean
1	12	717	184	53
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: -1398
95% CI for difference: (-1704, -1091)
T-Test of difference = 0 (vs not =): T-Value = -9.77 P-Value = 0.000 DF = 14

Two-Sample T-Test and CI: 2001, 2008

Sample	N	Mean	StDev	SE Mean
1	12	717	184	53
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -1620
95% CI for difference: (-2130, -1110)
T-Test of difference = 0 (vs not =): T-Value = -6.92 P-Value = 0.000 DF = 12

Two-Sample T-Test and CI: 2001, 2009

Sample	N	Mean	StDev	SE Mean
1	12	717	184	53
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -2093
95% CI for difference: (-2461, -1725)
T-Test of difference = 0 (vs not =): T-Value = -12.28 P-Value = 0.000 DF = 13

Two-Sample T-Test and CI 2001, 2010

Sample	N	Mean	StDev	SEMean
1	12	717	184	53
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -1929.2
95% CI for difference: (-2093.8, -1764.6)
T-Test of difference = 0 (vs not =): T-Value = -24.37 P-Value = 0.000 DF = 21

Two-Sample T-Test and CI: 2002, 2003

Sample	N	Mean	StDev	SEMean
1	12	1059	267	77
2	12	1130	144	41

Difference = mu (1) - mu (2)
Estimate for difference: -71.5
95% CI for difference: (-257.0, 114.0)
T-Test of difference = 0 (vs not =): T-Value = -0.82 P-Value = 0.426 DF = 16

Two-Sample T-Test and CI: 2002, 2004

Sample	N	Mean	StDev	SEMean
1	12	1059	267	77
2	12	871	280	81

Difference = mu (1) - mu (2)
Estimate for difference: 188
95% CI for difference: (-44, 420)
T-Test of difference = 0 (vs not =): T-Value = 1.69 P-Value = 0.107 DF = 21

Two-Sample T-Test and CI :2002, 2005

Sample	N	Mean	StDev	SEMean
1	12	1059	267	77
2	12	1876	231	67

Difference = mu (1) - mu (2)
Estimate for difference: -817
95% CI for difference: (-1029, -605)
T-Test of difference = 0 (vs not =): T-Value = -8.03 P-Value = 0.000 DF = 21

Two-Sample T-Test and CI : 2002, 2006

Sample	N	Mean	StDev	SE Mean
1	12	1059	267	77
2	12	2439	478	138

Difference = mu (1) - mu (2)
Estimate for difference: -1380
95% CI for difference: (-1714, -1047)
T-Test of difference = 0 (vs not =): T-Value = -8.73 P-Value = 0.000 DF = 17

Two-Sample T-Test and CI: 2002, 2007

Sample	N	Mean	StDev	SE Mean
1	12	1059	267	77
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: -1056
95% CI for difference: (-1380, -732)
T-Test of difference = 0 (vs not =): T-Value = -6.88 P-Value = 0.000 DF = 17

Two-Sample T-Test and CI: 2002, 2008

Sample	N	Mean	StDev	SE Mean
1	12	1059	267	77
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -1278
95% CI for difference: (-1798, -758)
T-Test of difference = 0 (vs not =): T-Value = -5.31 P-Value = 0.000 DF = 13

Two-Sample T-Test and CI: 2002, 2009

Sample	N	Mean	StDev	SE Mean
1	12	1059	267	77
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -1751
95% CI for difference: (-2133, -1369)
T-Test of difference = 0 (vs not =): T-Value = -9.76 P-Value = 0.000 DF = 15

Two-Sample T-Test and CI: 2002, 2010

Sample	N	Mean	StDev	SEMean
1	12	1059	267	77
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -1587.7
95% CI for difference: (-1789.8, -1385.6)
T-Test of difference = 0 (vs not =): T-Value = -16.38 P-Value = 0.000 DF = 20

Two-Sample T-Test and CI: 2003, 2004

Sample	N	Mean	StDev	SEMean
1	12	1130	144	41
2	12	871	280	81

Difference = mu (1) - mu (2)
Estimate for difference: 259.7
95% CI for difference: (67.2, 452.2)
T-Test of difference = 0 (vs not =): T-Value = 2.86 P-Value = 0.011 DF = 16

Two-Sample T-Test and CI: 2003, 2005

Sample	N	Mean	StDev	SEMean
1	12	1130	144	41
2	12	1876	231	67

Difference = mu (1) - mu (2)
Estimate for difference: -745.6
95% CI for difference: (-910.3, -580.9)
T-Test of difference = 0 (vs not =): T-Value = -9.51 P-Value = 0.000 DF = 18

Two-Sample T-Test and CI: 2003, 2006

Sample	N	Mean	StDev	SE Mean
1	12	1130	144	41
2	12	2439	478	138

Difference = mu (1) - mu (2)
Estimate for difference: -1309
95% CI for difference: (-1623, -995)
T-Test of difference = 0 (vs not =): T-Value = -9.08 P-Value = 0.000 DF = 12

Two-Sample T-Test and CI: 2003, 2007

Sample	N	Mean	StDev	SE Mean
1	12	1130	144	41
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: -985
95% CI for difference: (-1285, -684)
T-Test of difference = 0 (vs not =): T-Value = -7.08 P-Value = 0.000 DF = 13

Two-Sample T-Test and CI: 2003, 2008

Sample	N	Mean	StDev	SE Mean
1	12	1130	144	41
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -1207
95% CI for difference: (-1717, -697)
T-Test of difference = 0 (vs not =): T-Value = -5.21 P-Value = 0.000 DF = 11

Two-Sample T-Test and CI: 2003, 2009

Sample	N	Mean	StDev	SE Mean
1	12	1130	144	41
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -1680
95% CI for difference: (-2044, -1315)
T-Test of difference = 0 (vs not =): T-Value = -10.05 P-Value = 0.000 DF = 12

Two-Sample T-Test and CI: 2003, 2010

Sample	N	Mean	StDev	SE Mean
1	12	1130	144	41
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -1516.2
95% CI for difference: (-1666.7, -1365.7)
T-Test of difference = 0 (vs not =): T-Value = -21.08 P-Value = 0.000 DF = 19

Two-Sample T-Test and CI: 2004, 2005

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	1876	231	67

Difference = mu (1) - mu (2)
Estimate for difference: -1005
95% CI for difference: (-1223, -788)
T-Test of difference = 0 (vs not =): T-Value = -9.60 P-Value = 0.000 DF = 21

Two-Sample T-Test and CI: 2004, 2006

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	2439	478	138

Difference = mu (1) - mu (2)
Estimate for difference: -1568
95% CI for difference: (-1906, -1231)
T-Test of difference = 0 (vs not =): T-Value = -9.81 P-Value = 0.000 DF = 17

Two-Sample T-Test and CI: 2004, 2007

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: -1244
95% CI for difference: (-1571, -918)
T-Test of difference = 0 (vs not =): T-Value = -8.01 P-Value = 0.000 DF = 18

Two-Sample T-Test and CI: 2004, 2008

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -1466
95% CI for difference: (-1989, -944)
T-Test of difference = 0 (vs not =): T-Value = -6.06 P-Value = 0.000 DF = 13

Two-Sample T-Test and CI: 2004, 2009

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -1939
95% CI for difference: (-2323, -1556)
T-Test of difference = 0 (vs not =): T-Value = -10.72 P-Value = 0.000 DF = 16

Two-Sample T-Test and CI: 2004, 2010

Sample	N	Mean	StDev	SE Mean
1	12	871	280	81
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -1775.9
95% CI for difference: (-1984.3, -1567.5)
T-Test of difference = 0 (vs not =): T-Value = -17.77 P-Value = 0.000 DF = 20

Two-Sample T-Test and CI: 2005, 2006

Sample	N	Mean	StDev	SE Mean
1	12	1876	231	67
2	12	2439	478	138

Difference = mu (1) - mu (2)
Estimate for difference: -563
95% CI for difference: (-890, -237)
T-Test of difference = 0 (vs not =): T-Value = -3.68 P-Value = 0.002 DF = 15

Two-Sample T-Test and CI : 2005, 2007

Sample	N	Mean	StDev	SE Mean
1	12	1876	231	67
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: -239
95% CI for difference: (-554, 76)
T-Test of difference = 0 (vs not =): T-Value = -1.61 P-Value = 0.127 DF = 16

Two-Sample T-Test and CI: 2005, 2008

Sample	N	Mean	StDev	SE Mean
1	12	1876	231	67
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -461
95% CI for difference: (-979, 57)
T-Test of difference = 0 (vs not =): T-Value = -1.94 P-Value = 0.076 DF = 12

Two-Sample T-Test and CI : 2005, 2009

SampleN	Mean	StDev	SE Mean
1 12	1876	231	67
2 12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -934
95% CI for difference: (-1310, -559)
T-Test of difference = 0 (vs not =): T-Value = -5.34 P-Value = 0.000 DF = 14

Two-Sample T-Test and CI: 2005, 2010

Sample	N	Mean	StDev	SE Mean
1	12	1876	231	67
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -770.6
95% CI for difference: (-955.2, -586.0)
T-Test of difference = 0 (vs not =): T-Value = -8.68 P-Value = 0.000 DF = 21

Two-Sample T-Test and CI: 2006, 2007

Sample	N	Mean	StDev	SE Mean
1	12	2439	478	138
2	12	2115	460	133

Difference = mu (1) - mu (2)
Estimate for difference: 324
95% CI for difference: (-74, 722)
T-Test of difference = 0 (vs not =): T-Value = 1.69 P-Value = 0.105 DF = 21

Two-Sample T-Test and CI: 2006, 2008

Sample	N	Mean	StDev	SE Mean
1	12	2439	478	138
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: 102
95% CI for difference: (-458, 662)
T-Test of difference = 0 (vs not =): T-Value = 0.38 P-Value = 0.706 DF = 18

Two-Sample T-Test and CI: 2006, 2009

Sample	N	Mean	StDev	SE Mean
1	12	2439	478	138
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -371
95% CI for difference: (-813, 71)
T-Test of difference = 0 (vs not =): T-Value = -1.74 P-Value = 0.096 DF = 21

Two-Sample T-Test and CI : 2006, 2010

Sample	N	Mean	StDev	SE Mean
1	12	2439	478	138
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -208
95% CI for difference: (-529, 114)
T-Test of difference = 0 (vs not =): T-Value = -1.38 P-Value = 0.188 DF = 14

Two-Sample T-Test and CI: 2007, 2008

Sample	N	Mean	StDev	SE Mean
1	12	2115	460	133
2	12	2337	790	228

Difference = mu (1) - mu (2)
Estimate for difference: -222
95% CI for difference: (-779, 335)
T-Test of difference = 0 (vs not =): T-Value = -0.84 P-Value = 0.412 DF = 17

Two-Sample T-Test and CI: 2007, 2009

Sample	N	Mean	StDev	SE Mean
1	12	2115	460	133
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -695
95% CI for difference: (-1131, -259)
T-Test of difference = 0 (vs not =): T-Value = -3.32 P-Value = 0.003 DF = 21

Two-Sample T-Test and CI : 2007, 2010

Sample	N	Mean	StDev	SE Mean
1	12	2115	460	133
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -532
95% CI for difference: (-841, -222)
T-Test of difference = 0 (vs not =): T-Value = -3.66 P-Value = 0.002 DF = 15

Two-Sample T-Test and CI : 2008, 2009

Sample	N	Mean	StDev	SE Mean
1	12	2337	790	228
2	12	2810	561	162

Difference = mu (1) - mu (2)
Estimate for difference: -473
95% CI for difference: (-1058, 112)
T-Test of difference = 0 (vs not =): T-Value = -1.69 P-Value = 0.107 DF = 19

Two-Sample T-Test and CI: 2008, 2010

Sample	N	Mean	StDev	SE Mean
1	12	2337	790	228
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: -310
95% CI for difference: (-823, 204)
T-Test of difference = 0 (vs not =): T-Value = -1.31 P-Value = 0.213 DF = 12

Two-Sample T-Test and CI: 2009, 2010

Sample	N	Mean	StDev	SE Mean
1	12	2810	561	162
2	12	2647	204	59

Difference = mu (1) - mu (2)
Estimate for difference: 164
95% CI for difference: (-209, 536)
T-Test of difference = 0 (vs not =): T-Value = 0.95 P-Value = 0.360 DF = 13