Sudan University of Science and Technology

College of Graduate Studies

Predicting Malaria Cases in Sudan Based on Time Series analysis

التنبؤ بحالات الملاريا في السودان اعتماداً على تحليل السلاسل الزمنية

This Thesis Submitted in Partial Fulfillment for the Requirement of the Degree of M.Sc. in Biomedical Engineering

Student:

Sofia Ali Mohamed Ali

Supervisor:

Dr. Fragoon Mohamed Ahmed

November 2018

آية

"قَالُواْ سُبْحَانَكَ لاَ عِلْمَ لَنَا إِلاَّ مَا عَلَّمْتَنَا إِنَّكَ أَنتَ الْعَلِيمُ الْحَكِيم"

سورة البقرة (32)

Dedication

***To my Father**, who encourage, enthusiasm and support me to be the best.*
***To my Mother, my beloved family and my friends** for theirs love.*
***To Alaa Yahia** for supporting, helping, and standing by me.*

Abstract

Malaria is one of the most prevalent and debilitating diseases afflicting humans in Africa, and it is responsible for the fifth greatest number of deaths due to infectious diseases in the world and it is from the most ten infectious diseases that causing deaths In Sudan. That's why the study of infectious diseases represents one of the richest areas in mathematical biology. The infectious diseases need fast responses, and appropriately modeling, also predicting the outcome of disease spread over time and across space is a critical step toward informed development of effective strategies for public health intervention and decision making. Using mathematical representations for infectious diseases let the essential elements grasped quickly, and captured well. A lot of difficulties face the responsible organization in the process of collecting infectious diseases data by traditional ways in addition to the presence of more than 1500 health center in Sudan, as well; there is no applied method in Sudan to predict the spread of infectious diseases. Determine the best and most efficient mathematical model for predicting the new cases of infectious diseases in Khartoum, Al-Gadaref and Sennar based on the previous (history) data and visualize diseases distribution development (ArcMap) is important thing to save lives.

Using time series analysis (Auto Regressive model, Moving Average model, Mixed Model, and Exponential Smoothing model) found that the simple models (AR and MA) represented the data better in Khartoum, Al-Gadaref and Sennar states in the seasonal data and Sennar in the non-seasonal data. While the Exponential Smoothing and Mixed model (ARIMA) are better in representing Khartoum and Al-Gadaref non-seasonal data. Which prove that; not all data can be representing using the same forecasting model.

المستخلص

تعد الملاريا من أكثر الأمراض التي تصيب البشر في إفريقيا، وهي مسؤولة عن خامس أكبر عدد من الوفيات بسبب الأمراض المعدية في العالم وهي من أكثر عشر أمراض معدية تسبب الوفاة في السودان. هذا هو السبب في أن دراسة الأمراض المعدية تمثل واحدة من أغنى المجالات في علم الأحياء الرياضي. تحتاج الأمراض المعدية إلى استجابات سريعة ، ونمذجة مناسبة، والتنبؤ بنتيجة انتشار المرض مع مرور الوقت هي خطوة حاسمة نحو تطوير للاستراتيجيات الفعالة لتَدّخُل الصحة العامة وصناعة القرار. استخدام التمثيلات الرياضية للأمراض المعدية تسمح للعناصر الأساسية بإدراكها بسرعة والتقاطها بشكل جيد. تواجه الجهات المسؤولة الكثير من الصعوبات في عملية جمع بيانات الأمراض المعدية بالطرق التقليدية بالإضافة إلى وجود أكثر من 1500 مركز صحي في السودان، كذلك لا توجد طريقة مُطبّقة في السودان للتنبؤ بانتشار الأمراض المعدية. تحديد أفضل نموذج رياضي وأكثرها كفاءة للتنبؤ بالحالات الجديدة للأمراض المعدية في الخرطوم والقضارف وسنار استناداً إلى البيانات السابقة وتصور تطوير توزيع الأمراض (ArcMap).

استخدام تحليل السلاسل الزمنية (نموذج الانحدار التلقائي ، نموذج المتوسط المتحرك ، النموذج المختلط ، نموذج التنعيم الأسّي) وجد أن النماذج البسيطة (AR و MA) مثلت البيانات بشكل أفضل في ولايات الخرطوم والقضارف وسنار في البيانات الموسمية وسنار في البيانات غير الموسمية. في حين أن النموذج الأسّي والمختلط (ARIMA) أفضل في تمثيل بيانات الخرطوم والقضارف غير الموسمية. مما يثبت أنه لا يمكن أن تُمَثل جميع البيانات باستخدام نموذج التنبؤ نفسه.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

1.1 General Review

Surveillance of emerging infectious diseases is vital for the early identification of public health threats and the infectious disease spread is a major threat to public health and economy. Based on the statistics of the World Health Organization (WHO), 25% of human death is caused by infectious diseases. The spread of an infectious disease involves characteristics of the agent such as virus and bacteria, the host, and the environment in which transmissions take place. That's why the study of infectious diseases represents one of the richest areas in mathematical biology. In the study of infectious diseases, the essential elements are quickly grasped, and well-captured within mathematical representations and we need this because the infectious diseases can be classified as sporadic (occurs occasionally), endemic (constantly present in a population), epidemic (many cases in a region in short time), and pandemic (worldwide epidemic) which need fast responses, and appropriately modeling and actually predicting the outcome of disease spread over time and across space is a critical step toward informed development of effective strategies for public health intervention and decision making. And web-based surveillance tools and epidemic intelligence methods, used by all major public health institutions, are intended to facilitate risk assessment and timely outbreak detection [1-3].

1.2 Problem Statement

A lot of difficulties face the responsible organization in the process of collecting infectious diseases data by using the traditional ways in addition to existence of more than 1500 health centers in Sudan. As well, there is no applied method in Sudan to predict the spread of infectious diseases which is important thing to save lives, by planning the significant procedures and determine the best time to take the vaccinations.

1.3 Objectives

There are two categories of the research objectives:

1.3.1 General Objective

Determine the best and most efficient mathematical model as proposed model for predicting the new Malaria cases in Sudan based on the previous (history) data.

1.3.2 Specific Objectives

1. To compare between the mathematical models to find the suitable one for each state.
2. To predict the new cases of Malaria using the efficient model (AR, MA, ARIMA, or Exponential Smoothing).
3. To visualize Malaria distribution in Sudan.

1.4 Research Methodology

This research is a qualitative and quantitative research which used the in-depth interviews with the staff of the epidemiological administration in the Ministry of Federal Health to study the procedures of collecting data process and identify the key challenges and difficulties of conducting this process. The analysis mainly focused on identifying the major challenges to proposed forecasting model to predict the new cases.

The goals of this research are mainly achieved by dividing the research methodology into two stages as shown in figure 1.1 and figure 1.2:

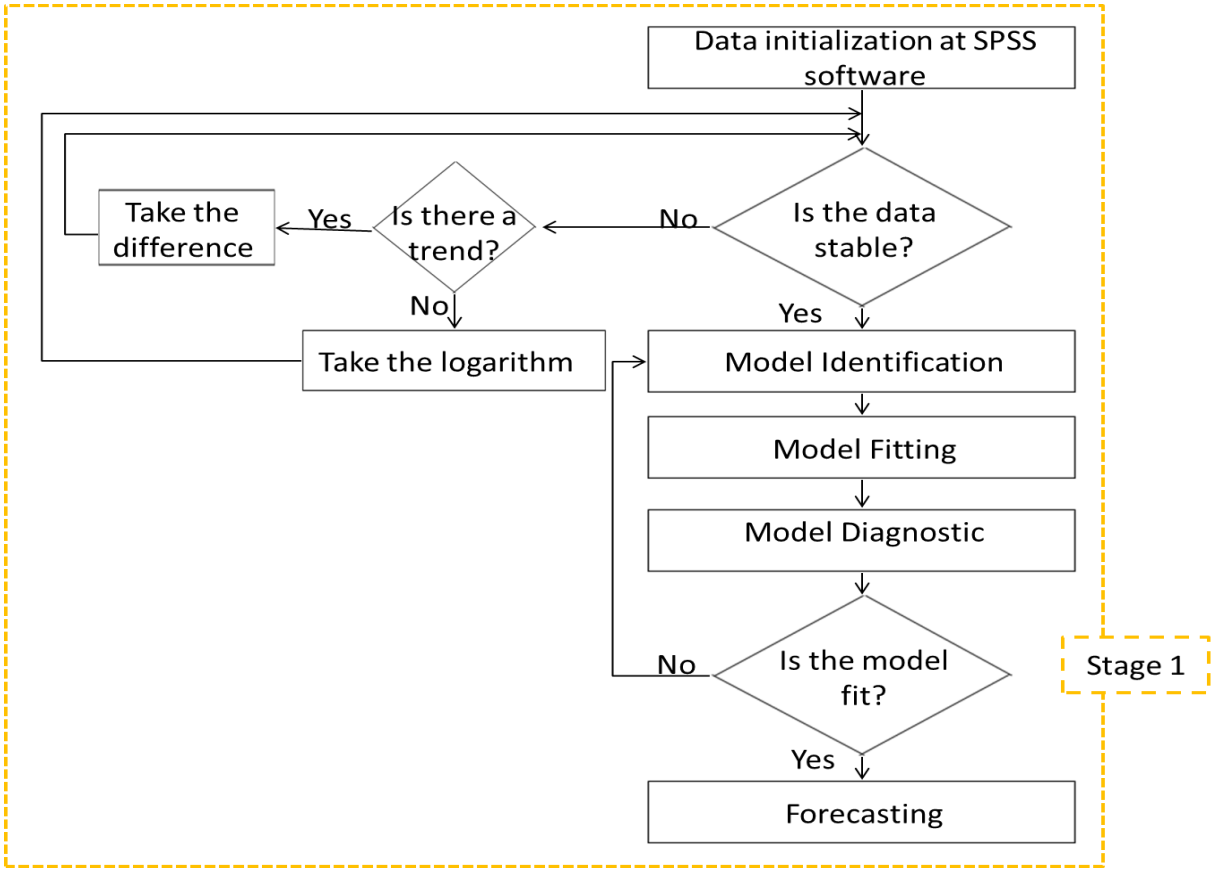

Figure 1.1 Stage 1 of the Malaria Model Predictor

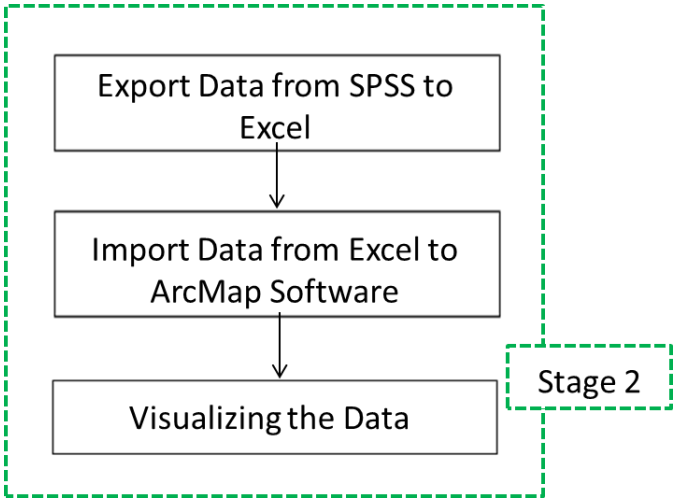

Figure 1.2 Stage 2 of the Malaria Model Predictor

1.5 Thesis Organization

The thesis consists of seven chapters. The sticky chapter discusses the problem definition, justification for carrying out the research and objectives.

Chapter two reviews some of the previous studies related to infectious diseases mapping, distribution, spread and prediction .

Introduction to infectious disease, how we classified it, the principle of disease transmission, how to control infection, the importance of predicting infectious diseases, spatial analysis and statistical modeling of infectious diseases, have been discussed in chapter three.

Chapter four described a detailed description of the research methodology which includes theories, models, materials, methods used, the data collection and analysis techniques.

The model design, have been viewed in chapter five. Whilst the main results and discussions have been included in chapter six viewed. However, chapter seven included the conclusions and recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The first application of mathematical modeling in the area of infectious diseases appears to have emerged by Daniel Bernoulli during the 18th century. Bernoulli's formula for the endemic prevalence of susceptible has so far escaped attention. It involves the lifetime risk of the infection, the force of infection and the life expectancy at birth. A new formula for the basic reproduction number is derived which involves the average force of infection, the average case fatality and the life expectancy at the time of infection. One can use this estimate to assess the gain in life expectancy if only a fraction of the population is immunized [4].

Therefore, the following sections seek to review recent literature related to infectious diseases modeling and predicting methods, as well time series forecasting techniques.

2.2 Visualization and Predicting using Geographical Information System (GIS)

Several studies concentrated on using GIS, the research that done by Norstrøm and Madelaine used GIS in the field of surveillance and monitoring of animal diseases by recording and reporting disease information and visualized on a map, and in the case of an outbreak they used GIS for identifying the location of the case farm and all farms at risk within a specified area of the outbreak and used programs packages as @Risk integrated with the GIS for simulation and find risk factors for the spread of the modeled disease [5].

Other research studies, concerned on using GIS software in combination with remote sensing images/data to detect potential breeding ponds by using the image from SPOT 5 and converted into proper format, then, they used ArcView software and calculated mosquito density and evaluated cross-potential risks [6]. While [7] the surveillance study using Geographical information system and remote sensing was initiated in Jodhpur Cantonment area for mapping the distribution of malaria vectors. Using Survey of India toposheet 1 : 50,000 base map has been prepared. Using hand-held GPS, the area was surveyed and important landmarks which include major roads, drains, water bodies, major mosquito breeding sites, etc. were geo-referenced. Contour layers have also been digitized. The indoor and outdoor densities of adult mosquitoes were recorded from various GPS registered localities

using manual aspirators (per man hour density) and CDC light traps. The larval density of mosquitoes from different water bodies was recorded by using ladle. Different species of mosquitoes collected from each sector were identified for species composition. Village wise map of District Dhar had been digitized using GIS software ARC/View 3.2 and attribute data were collected from state health authorities, NVDA and Survey of India had been attached. Thematic maps of altitude, soil, rainfall, forest cover, temperature, etc. have also been prepared. Trend analysis of epidemiological data from 2002– 2004 had been done. Submerged villages under Indira Sagar Dam have been mapped. The data on various entomological and parasitological parameters are being collected through periodic surveys and are regularly entered in GIS-based framework to view the impact of the construction of dams in space and time.

2.3 Mathematical Modeling

Other researchers focused on describing mathematical model for a susceptible-infectious epidemic that simulates spatial and temporal patterns of disease spread. The parameterized model was implemented in a GIS to simulate disease spread [8]. While [9] discussed and presented main mathematical approaches of modeling used for the surveillance and predicting infectious disease outbreaks which allow rapid assessment, and those methods are namely, statistical methods for surveillance of outbreaks and identification of spatial patterns in real epidemics, mathematical models within the context of dynamical systems (also called state-space models) used to prediction the evolution of an on-going epidemic spread, and machine learning/ expert methods for the prediction too. A latter research performed areal co-kriging interpolation in ArcGIS 10.1 Geostatistical Analyst, predictive values and their associated standard errors were calculated for all areas within and between input polygons. It works by estimating values based on data collected in polygons and predicts values for a new set of polygons in the same data domain that differ in size and shape from the original [10].

2.4 Internet

In the systematic review that done by Christakiab and Eirini focused on finding new methods for regional and global infectious disease surveillance and in epidemic modeling to predict and prevent future infectious diseases. The review

discuss plenty manners of surveillance and prediction, namely event-based surveillance, agent-based models, structured metapopulation models, mobile phone for movement track, social media for data collection and remote sensing for diseases prediction. Some of these methods used risk factors, environmental change and population movements. The review was found that, many papers discuss surveillance and prediction, but the main challenge is sophisticated software for data acquisition and analysis and the high cost of using some methods [3].

The researchers used different ways to test validation and evaluate model performance in the GIS by examining the correspondence between predicted patterns of disease spread and over 1000 geo-located field observations of disease presence. Additionally, they examined the nature of prediction errors by eco-region, vegetation composition, and climate [8].

Upon validation, the areal co-kriging model produced root-mean-square standardized errors close to 1, indicating valid prediction errors. Mean standardized errors (the average difference between the measured and predicted values) were close to 0, and root-mean-square error (indicating how closely the model predicts the measured value) and average standard errors (average of the prediction standard errors) were low, with little difference between them [10].

CHAPTER THREE

THEORATICAL BACKGROUND

3.1 Introduction

Infectious disease spread is a major threat to public health and economy. Based on the statistics of the World Health Organization (WHO), 25% of human death is caused by infectious diseases. The spread of an infectious disease involves characteristics of the agent such as virus and bacteria, the host, and the environment in which transmissions take place. The study of the distribution of determinants of disease and injury in human populations call epidemiology, it's a discipline that includes both infectious and noninfectious diseases. Most epidemiologic studies of infectious diseases have concentrated on the factors that influence acquisition and spread, because this knowledge is essential for developing methods of prevention and control. Historically, epidemiologic studies and the application of the knowledge gained from them have been central to the control of the great epidemic diseases, such as cholera, plague, smallpox, yellow fever, and typhus. An understanding of the principles of epidemiology and the spread of disease is essential to all medical personnel, whether their work is with the individual patient or with the community. Appropriately modeling and actually predicting the outcome of disease spread over time and across space is a critical step toward informed development of effective strategies for public health intervention. Given the ongoing risk of infectious diseases worldwide, it is important to develop appropriate analysis methods, models, and tools to assess and predict the disease spread and evaluate the disease risk [2] & [11].

3.2 Infection

Infection is the invasion and multiplication in/on body tissue of microorganisms that produce signs and symptoms along with an immune response. Such reproduction injures the host either by causing cellular damage from microorganism-produced toxins or intracellular multiplication or by competing with host metabolism. The host's own immune response may increase tissue damage, which may be localized (as in infected pressure ulcers) or systemic. The very young and the very old are most susceptible to infections. Microorganisms that cause infectious diseases are difficult to overcome for many reasons:

- Some bacteria develop a resistance to antibiotics.

- Some microorganisms, such as human immunodeficiency virus (HIV), include many different strains, and a single vaccine can't provide protection against them all.
- Most viruses resist antiviral drugs.
- Some microorganisms localize in areas that make treatment difficult, such as the central nervous system and bone.
- New infectious agents, such as HIV and severe acute respiratory syndrome–coronavirus, occasionally arise.
- Opportunistic microorganisms can cause infections in immune compromised patients.
- Much of the world's ever-growing population has not received immunizations.
- Increased air travel by the world's population can speed a virulent microorganism to a heavily populated urban area within hours.
- Biological warfare and bioterrorism with organisms such as anthrax, plague, and smallpox are an increasing threat to public health and safety throughout the world.
- Invasive procedures and the expanded use of immunosuppressive drugs increase the risk of infection for many [12].

3.3 Routes of disease transmission

Various transmissible infections may be acquired from others by direct contact, by aerosol transmission of infectious secretions, or indirectly through contaminated inanimate objects or materials. Some, such as malaria, involve an animate insect vector. These routes of spread are often referred to as horizontal transmission, in contrast to vertical transmission from mother to fetus.

3.4 Infectious diseases, their transmission and research needs

Infectious diseases are also known as transmissible diseases or communicable diseases. The illness of infectious diseases is caused by the infection, presence, and growth of pathogenic biological agents (known as pathogens) in an individual host organism. Pathogen is the microorganism (or microbe) that causes illness. Infectious pathogens include viruses, bacteria, fungi, protozoa, multicellular parasites, and aberrant proteins known as prions. These pathogens are the cause of

disease epidemics, in the sense that without the pathogen, no infectious epidemic occurs. The organism that a pathogen infects is called the host. In the human host, a pathogen causes illness by either disrupting a vital body process or stimulating the immune system to mount a defensive reaction [11].

3.5 Stages of infection

The infection divided into four stages:

3.5.1 Stage I: Incubation

The duration of this stage can range from instantaneous to several years. Pathogen is replicating, and the infected person becomes contagious, thus capable of transmitting the disease.

3.5.2 Stage II: Prodromal stage

In this stage the host makes vague complaints of feeling unwell and is still contagious.

3.5.3 Stage III: Acute illness

Microorganisms actively destroy host cells and affect specific host systems in this stage. As well, patient recognizes which area of the body is affected.

3.5.4 Stage IV: Convalescence

This stage begins when the body's defense mechanisms have contained the microorganisms, and when damaged tissue is healing [12].

Based on the frequency of occurrence, infectious diseases can be classified as sporadic (occurs occasionally), endemic (constantly present in a population), epidemic (many cases in a region in short period), and pandemic (worldwide epidemic). An infectious disease is termed contagious if it is easily transmitted from one person to another. The transmission mechanisms of infectious diseases can be categorized as contact transmission, vehicle transmission, and vector transmission. Contact transmission can occur by direct contact (person-to-person) between the source of the disease and a susceptible host, indirect contact through inanimate objects (such as contaminated soils), or droplet contact via mucus

droplets in coughing, sneezing, laughing or talking. Vehicle transmission involves a media. Based on the media type in transmission, the infectious diseases can be categorized as airborne (diseases transmitted through the air such as influenza, anthrax, measles), foodborne (diseases transmitted through the foods such as Hepatitis A and E), and waterborne (diseases transmitted through the water such as Cholera) [11 & 12].

3.6 Malaria

Malaria is one of the most prevalent and debilitating diseases afflicting humans. It's responsible for the fifth greatest number of deaths due to infectious diseases. It is endemic in Central and most of South America, most of Africa, the Middle East, India, and Southeast Asia, including southern China in which most areas including uplands report malaria transmission throughout the year, though it increases during and soon after the rainy season. It's considered the most important vector-borne disease, causing an estimated 190 –311 million clinical episodes, and 708,000 - 1,003,000 deaths in 2008 worldwide (CDC). And Sudan is highly affected by malaria with 7.5 million cases and 35,000 deaths every year. The entire population is at risk of malaria epidemics with a very high burden on government and population. The usefulness of forecasting methods in predicting the number of future incidences is needed to motivate the development of a system that can predict future incidences. Immunologically vulnerable populations are generally afflicted by malaria epidemics and people of all age groups remain susceptible to the full range of its clinical effects. Its spread in a community poses unique intervention strategies. Prophylaxis Malaria is a parasitic infection caused by Plasmodium spp. characterized by fever, headache, and hemolytic anemia. It is transmitted by the Anopheles mosquito. There are four species: P. malariae, P. vivax, P. ovale, and P. falciparum. The most lethal form of malaria is caused by P. falciparum. Most P. falciparum are resistant to the tradition prophylaxis agent, chloroquine. Generally P. malariae, P. vivax, and P. ovale are sensitive to chloroquine. There are, though, increasing reports of chloroquine-resistant P. vivax in Southeast Asia [13 & 14].

The best prevention against malaria is avoidance of mosquito bites. Travelers to malarious regions should be advised that the Anopheles sp. mosquito bites from dusk to dawn, and nighttime outdoor activity should be limited. Medical

prophylaxis against malaria depends on the region of the world, length of stay in the endemic area, and activities of the traveler. Prophylactic agents are taken before travel to ensure adequate blood levels during travel and post travel to eradicate protozoa potentially in the bloodstream. Since prophylaxis is not completely effective and only approximately 50% of travelers fully adhere to their prophylactic regimens, travelers should be educated about the symptoms of malaria (unexplained fever with or without headache, chills, weakness, vomiting, and diarrhea). All travelers to endemic areas need to recognize that malaria can be fatal unless treated early and prompt medical evaluation is necessary for symptoms suggestive of malaria. Areas of drug-resistant P. falciparum malaria are rapidly changing. Throughout most of the world, the traditional prophylactic agent, chloroquine sulfate, can no longer be used to prevent malaria [13].

3.7 Epidemics

The characterization of epidemics and their recognition in a community involve several quantitative measures and some specific epidemiologic definitions. Infectivity, in epidemiologic terms, equates to attack rate and is measured as the frequency with which an infection is transmitted when there is contact between the agent and a susceptible individual. The disease index of an infection can be expressed as the number of persons who develop the disease divided by the total number infected. The virulence of an agent can be estimated as the number of fatal or severe cases per the total number of cases. Incidence, the number of new cases of a disease within a specified period, is described as a rate in which the number of cases is the numerator and the number of people in the population under surveillance is the denominator. This is usually normalized to reflect a percentage of the population that is affected. Prevalence, which can also be described as a rate, is primarily used to indicate the total number of cases existing in a population at risk at a point in time.

3.8 Control of epidemics

The first principle of control is to be aware of the existence of an epidemic. This awareness is sometimes immediate because of the high incidence of disease, but often the evidence is obtained from routine disease reports to health departments and records of school and work absenteeism. The causative factors must be identified and studies to determine the route of transmission (eg, food poisoning) must be initiated. Measures must then be adopted to control the spread and development of further infection [15].

3.9 Infection control and treatment

3.9.1 Control

The best way to control infections is to break the weakest link in the chain of infection (usually the mode of transmission). Many strategies exist to prevent or control the transmission of infectious agents, and they fall into four general categories:

- Control or elimination of infectious agents by appropriate sanitation, disinfection, and sterilization
- Control of transmission through proper hand hygiene, effective ventilation, and aseptic technique
- Reservoir control. In health care settings, a number of interventions are directed at controlling or destroying infectious reservoirs:
  - Using disposable equipment and supplies whenever possible.
  - Disinfecting or sterilizing equipment as soon as possible after use.
  - Using appropriate equipment for each patient.
  - Handling and disposing of patient secretions, excretions, and exudates properly.
  - Helping to identify and treat persons who are infection carriers. To help reduce the number of reservoirs in both the community and the health care setting, patients should be encouraged to obtain active and passive immunizations, to practice positive health behaviors, to avoid high-risk behavior, and to maintain first-line defenses.

- Isolating infected patients, according to Centers for Disease Control and Prevention (CDC) recommendations, to limit the chance that they will transmit the infection.

3.9.2 Treatment

Treatment for infections can vary widely. Vaccines may be administered to induce a primary immune response under conditions that won't cause disease. If infection occurs, treatment is tailored to the specific microorganism causing the infection. Drug therapy should only be used when appropriate. Supportive therapy can play an important role in fighting infections.

- Antibiotics work in a variety of ways, depending on the class of drug used. Antibiotic action is either bactericidal or bacteriostatic. Antibiotics may inhibit cell wall synthesis, protein synthesis, bacterial metabolism, or nucleic acid synthesis or activity, or they may increase cell membrane permeability.
- Antifungal drugs destroy the invading microorganism by increasing cell membrane permeability. The antifungal binds sterols in the cell membrane, resulting in leakage of intracellular contents, such as potassium, sodium, and nutrients.
- Antiviral drugs stop viral replication by interfering with DNA synthesis.

Almost any infectious disease, under certain circumstances, may be considered to be a true emergency [12].

3.10 Surveillance of emerging infectious diseases

Surveillance of emerging infectious diseases is vital for the early identification of public health threats. Emergence of novel infections is linked to human factors such as population density, travel and trade and ecological factors like climate change and agricultural practices. A wealth of new technologies is becoming increasingly available for the rapid molecular identification of pathogens but also for the more accurate monitoring of infectious disease activity. Web-based surveillance tools and epidemic intelligence methods, used by all major public health institutions, are intended to facilitate risk assessment and timely outbreak detection [3].

3.11 Time Series

A time series is a sequence of measurements over time. Time series is anything which is observed sequentially over the time at regular interval like hourly, daily, weekly, monthly, quarterly etc. Time series data is important when you are predicting something which is changing over the time using past data. In time series analysis the goal is to estimate the future value using the behaviors in the past data.

3.11.1 Components of a Time Series

•Secular Trend

The long-term trends of sales, employment, stock prices, and other business and economic series follow various patterns. Some move steadily upward, others decline, and still others stay the same over time.

• Cyclical Variation

A typical business cycle consists of a period of prosperity followed by periods of recession, depression, and then recovery with no fixed duration of the cycle. There are sizable fluctuations unfolding over more than one year in time above and below the secular trend.

• Seasonal Variation

The patterns of change within a year, typically repeating themselves.

• Residual Variation

The residual fluctuations, often called chance fluctuations, are unpredictable, and they cannot be identified.

3.11.2 Time series analysis

Time series analysis is done by computer, not by hand. Many of the computations are not difficult but can be extremely tedious if done by hand. Consider calculating full and partial autocorrelations between pairs of scores at 25 to 30 different lags. Several researchers are devoted to time- series analysis, some of them highly mathematical. The primary method for ARIMA models is Box, et al. (1994). The methods that demonstrate at least some of the equations with numbers are Glass, Wilson, and Gottman (1975) and McDowall, et al. (1980). A few less mathematical, more computer- oriented sources are Cryer (1986); McCleary and Hay (1980); and McCain and McCleary (1979). [16]

3.12 Forecasting Process

Forecast is to predict or estimate a future activity level. A forecast is dependent on the analysis of historic and/or current data to produce these estimates.

3.12.1 Approaches to forecasting

The process of forecasting can be broadly categorized into two approaches: objective or quantitative forecasts and subjective or qualitative forecasts.

3.12.1.1 Subjective forecasts

Subjective or qualitative forecasts rely to a large extent on an in-depth knowledge of the activity being forecast by those responsible for producing the forecast. The forecast might be created by reading reports and by consulting experts for information and then using this information in a relatively unspecified or unstructured way to predict a required activity. The main problem with this approach is that there is no clear methodology which can be analyzed to test how a forecast may be improved in order that past mistakes are avoided. As a subjective, or qualitative, the forecast is very dependent on the individuals involved, it is prone to problems when the key players responsible for the forecasting process change. This method of forecasting does not usually require much mathematical input and therefore a spreadsheet will play an accompanying role as opposed to a central role.

3.12.1.2 Objective forecasts

An objective or quantitative approach to forecasting requires a model to be developed which represents the relationships deduced from the observation of one or more different numeric variables. This is generally achieved by first recording historic data and then using these historical facts to hypothesize a relationship between the items to be forecast and the factors believed to be affecting it. The spreadsheet is clearly an ideal tool for this type of analysis and thus can play a central role in the production of such forecasts.

Objective forecasting methods are sometimes considered to be more dependable than subjective methods because they are less affected by what the forecasters would like the result to be. Furthermore, forecasting models can incorporate means

of assessing the accuracy of the forecast by comparing what actually happened with what were forecast and adjusting the data to produce more accurate figures in the future, and the objective forecasting methods:

3.12.1.2.1 Simple Moving Average (SMA)

A simple moving average (SMA) is the simplest type of technique of forecasting. Basically, a simple moving average is calculated by adding up the last 'n' period's values and then dividing that number by 'n'. So the moving average value is considering as the forecast for next period. Moving averages can be used to quickly identify whether selling is moving in an uptrend or a downtrend depending on the pattern captured by the moving average.

3.12.1.2.2 Exponential Smoothing (SES)

This is the second well known method to produce a smoothed Time Series. Exponential Smoothing assigns exponentially decreasing weights as the observation get older. Exponential smoothing is usually a way of "smoothing" out the data by removing much of the "noise" (random effect) from the data by giving a better forecast.

3.12.1.2.3 Types of Exponential Smoothing Methods

3.12.1.2.3.1 Simple Exponential Smoothing

If you have a time series that can be described using an additive model with constant level and no seasonality, you can use simple exponential smoothing to make short-term forecast.

3.12.1.2.3.2 Holt's Exponential Smoothing

Holt's exponential smoothing can be used to make short-term forecasts if the time series can be described using an additive model with increasing or decreasing trend and no seasonality.

3.12.1.2.3.3 Winters' Three Parameter Linear and Seasonal Exponential Smoothing

Holt-Winters exponential smoothing can be used to make short-term forecasts if the time series can be described using an additive model with increasing or decreasing the trend and seasonality.

3.12.1.2.4 Autoregressive Integration Moving Average (ARIMA)

A statistical technique uses time series data to predict future. The parameters used in the ARIMA is (P, d, q) which refers to the autoregressive, integrated and moving average parts of the data set, respectively. ARIMA modeling will take care of trends, seasonality, cycles, errors and non-stationary aspects of a data set when making forecasts.

ARIMA is mainly used to project future values using historical time series data. Its main application is in short forecasting with minimum 38-40 historical data points with minimum number of outliers.

3.12.1.2.5 Neural Network (NN)

Artificial neural network (ANN) is basically machine learning approach that models human brain and consists of a number of artificial neurons. Their ability to learn by example makes them very flexible and powerful.

Neural networks, has its own strength to derive meaning from complicated or imprecise data, and most of the time can be used to detect the pattern and trend in the data, which cannot be detectable easily from human eye or any computer techniques. We also have some of the advantage of NN like Adaptive learning, self-organization, real-time operation, fault tolerance. [17-20]

3.13 GIS for Early Detection and Response to Infectious Disease

Successful understanding and response to infectious disease outbreaks depend greatly on the ability to consider the surrounding context. Disease spreads geographically, and interventions occur in relation to human, institutional, climatic, and other kinds of landscapes. Because GIS technology relates many kinds of data to geographic location, it excels in tracking not only disease spread but also laboratory specimen and medical supply whereabouts, hospital bed availability, testing facility proximity, vulnerable population locations, and medical personnel distribution. Built-in GIS analysis tools provide effective early warning systems and preparedness programs that generate meaningful information that public health

officials need to make effective decisions—at the community, national, and global levels. During an outbreak, GIS provides tools that speed the collection of accurate field data. Complex statistical and other analyses applied with GIS technology provide relevant information to support sound decisions [21].

CHAPTER FOUR

METHODOLOGY & DATA ANALYSIS

4.1 Methodology

The methodology that is follow in the research is as following:

4.1.1 Study design

This research is a qualitative and quantitative research which studies the procedures of infectious diseases data collection process and identifies the key challenges and difficulties of conducting this process. Besides, propose a forecasting model to predict the new cases.

4.1.2 Study area

This study was conducted in Khartoum. It is the capital and largest city of Sudan

4.1.3 Study setting

The study concentrated on the epidemiological administration in the Ministry of Federal Health.

4.1.4 Study population

The staff of the epidemiological administration in the Ministry of Federal Health are the study population.

4.1.5 Methods

In the qualitative approach the In-depth interviews was used as the data collection method to study the procedures of collecting data process and identify the key challenges and difficulties of conducting this process.

While the quantitative approach is conducted by the three following stages:

i. The first stage concern with gathering of data of the infectious diseases in Sudan from epidemiological administration in the Ministry of Federal Health and to use Excel and SPSS softwares to find the suitable mathematical model to predict the new cases.
ii. The second stage focused on using the suitable mathematical model and maps it using Arcmap software to visualize the development and spread of the diseases.
iii. The third stage clarifies the integration and the linkage between Arcmap software and an online website (ESRI) for online and direct new cases entering in addition to entering the information that has relation with the disease.

4.2 Data analysis

The study was conducted to verify the existence of the research problem and to demonstrate the importance of the research objective (designing a forecasting program). The analysis mainly focused on identifying major challenges in collecting and processing the data of the infectious diseases.

4.2.1 Analysis of the in-depth interviews

Through the results of the analysis of the in depth interviews that were conducted with the personnel those responsible for collecting and processing the data of the infectious diseases found that the procedures of this process and the key challenges and difficulties of conducting it are as follows:

4.2.1.1 Data collection methods

- The data is collecting from about 1,500 centers out of a total of more than 5000 centers using traditional ways such as telephone and e-mails.
- The data is collecting daily or weekly depending on the classification of the disease.
- The role of the State Ministry of Health in each state in Sudan with regard to the data collection process of infectious diseases from its affiliated health facilities is that it works to collect data in its daily and annual profiles and then send these data to the Federal Ministry of Health and the figure (4.1) below shows the sequence of the data collection process:

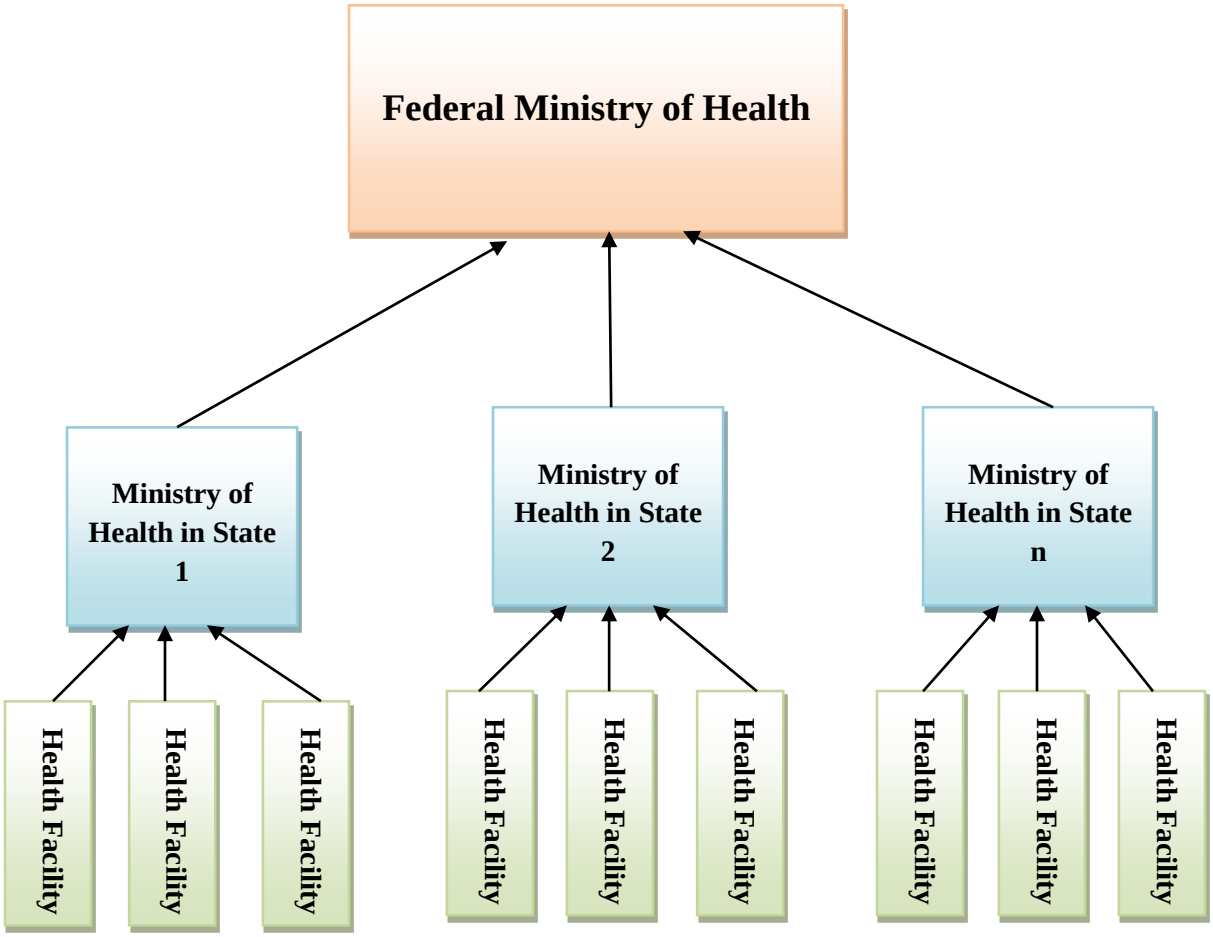

Figure 4.1 The sequence of the data collection process

4.2.1.2 Data processing and presentation methods

The Data that has been collected in the daily and weekly forms it coordinated in monthly form and displayed on maps using the Arcmap program and in the form of tables containing all the states of Sudan and all diseases and their statistics.

4.2.1.3 Difficulties and problems facing them:

The key challenges and difficulties that facing these institutions in the process of obtaining the information that enable them to meet the requirements of the management of the infectious diseases are as follows:

1- The number of health centers is large and geographically distant and scattered.

2- The methods used to collect data are traditional, therefore can’t cover all health centers.

3- The data do not represent all health centers in Sudan.

4- The actions taken on the basis of this data are inaccurate.

5- The paper-based methods of data collection are ineffective because they need large storage spaces relative to the large number of data and can be lost or damaged.

4.3 Final Result

The analysis found that, in addition to the difficulty of the collection process, predictive methods are not used to obtain future values for the spread of the disease.

In fact, the prediction process is relying on the experience, where the existence of the previous data and other variables is taken advantage of. Therefore, there is a real need for a program predicting the new cases, because through this program there is:

- The possibility of entering data from the health centers to the program directly without the need for Traditional ways of communication.
- The feature of analyzing the data by program to obtain future values based on the previous data.

CHAPTER FIVE

MODEL DESIGN

5.1 Introduction

Finding the suitable model for a time series is a difficult task that needs a lot of search and experience. The next steps show how to choose, test and validate the certain model to use it in the forecasting for Malaria cases in Khartoum, Al-Gadaref and Sennar states.

5.2 Non Seasonal Model – Yearly

Choosing the suitable model for the three states; using the data for the years (1996 up to 2016).

5.2.1 Khartoum

5.2.1.1 Time Series Stability

The ARIMA (auto-regressive, integrated, moving average) model of a time series is defined by three terms (p, d, q). The middle element in the model; d, is investigated before p and q. The goal is to determine if the process is stationary and, if not, to make it stationary before determining the values of p and q. Recall that a stationary process has a constant mean and variance over the time period of the study.

The first step in the analysis is to plot the time series; Figure 5.1 shows cases of the Malaria in Khartoum State. The two relevant features of the plot are central tendency and dispersion.

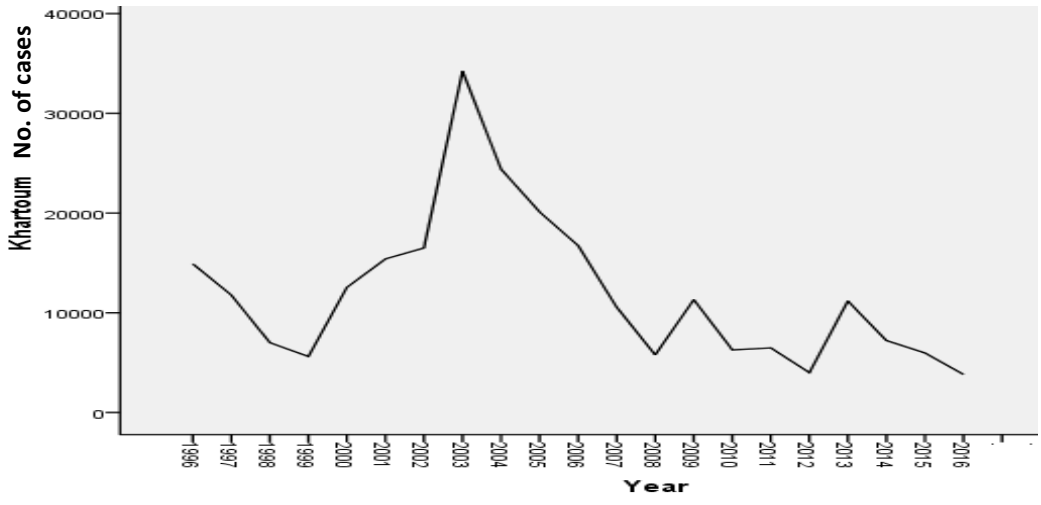

Figure 5.1 Yearly cases of the Malaria in Khartoum State

It is obvious that there is a shift in both the trend and the dispersion over time for the series. If the mean is changing, the trend is removed by differencing once or twice. If the variability is changing, the process can be made stationary by logarithmic transformation.

Differencing is the easiest way to make a non-stationary mean stationary (flat). The number of times you have to difference to make the process stationary determines the value of d. If d = 0, the model is already stationary and has no trend. When the series is differenced once, d=1 and linear trend is removed. When the difference is then differenced, d=2 and both linear and quadratic trend are removed. For non-stationary series, d values of 1 or 2 are usually adequate to make the mean stationary.

To see if the process is stationary after linear trend is removed, the first difference of the cases plotted against years, as seen in Figure 5.2.

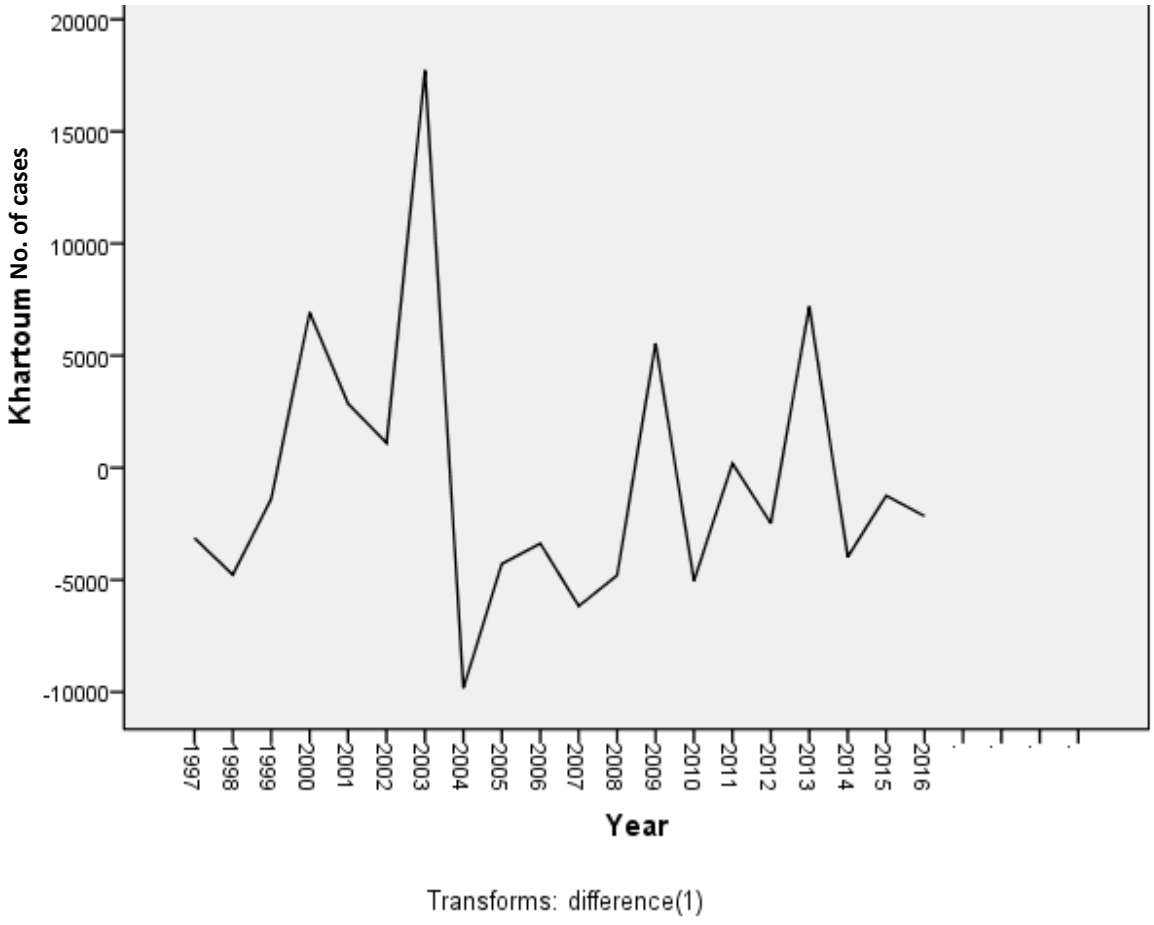

Figure 5.2 the first difference of Khartoum State yearly cases

It is clear that the series is still has the shift in the mean, so second differencing appear necessary which is shown in Figure 5.3

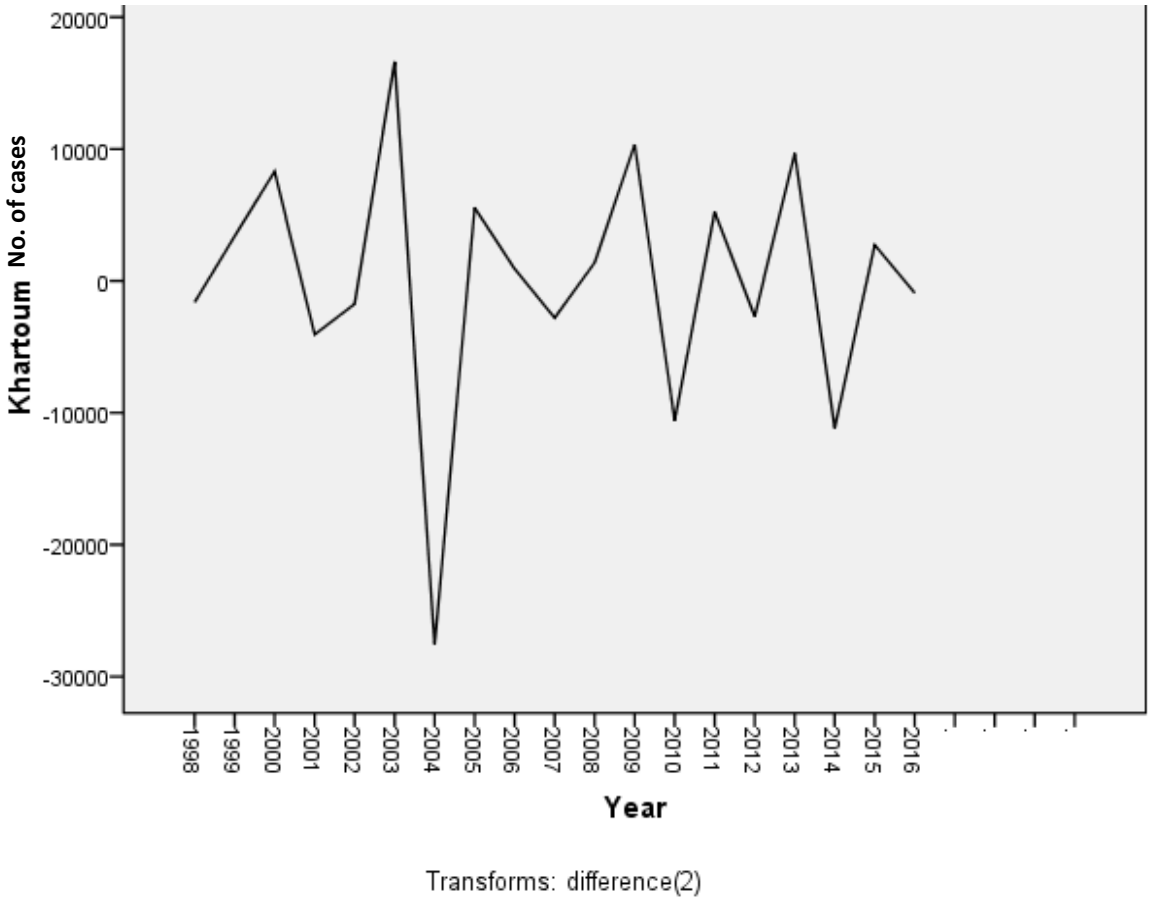

Figure 5.3 the second difference of Khartoum State yearly cases

The series now appears stationary with respect to central tendency, however, the variability seems to be increasing and decreasing over time. As we have mentioned earlier, the changing in variability can be made stationary by logarithmic transformation. The transformed difference is plotted in Figure 5.4.

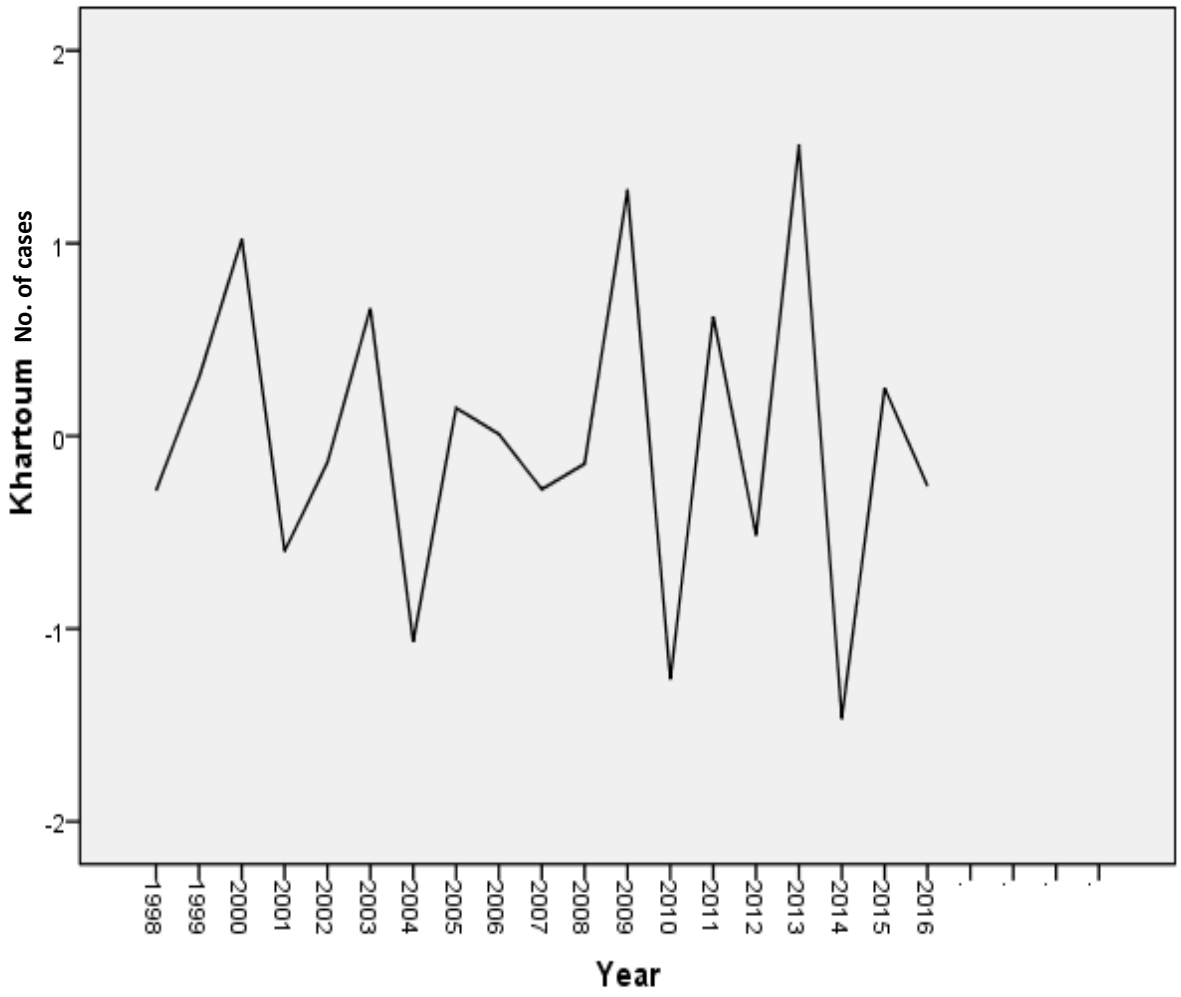

Figure 5.4 the natural logarithm and first difference of Khartoum State yearly cases

5.2.1.2 Model Identification

This stage is one of the most important and most difficult stages of the model building. The difficulty lies in determining the variables that must be included in the model or should be excluded from it.

As it is mentioned before, the ARIMA (auto-regressive, integrated, moving average) model of a time series is defined by three terms (p, d, q). Identification of a time series is the process of finding integer, usually very small (0, 1, or 2), values

of p, d, and q that model the patterns in the data. When the value is 0, the element is not needed in the model.

5.2.1.2.1 ACFs and PACFs

Autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) are examined to see which of the three patterns (AR, MA & ARIMA) are present in the data. Autocorrelations are self- correlations of the series with itself, removed one or more periods in time; partial autocorrelations are self- correlations with intermediate autocorrelations partialed out. Various auto- regressive and moving average patterns leave distinctive footprints on the autocorrelation and partial autocorrelation functions.

Both autocorrelations and partial autocorrelations are computed for sequential lags in the series. The first lag has an autocorrelation between Y_{t-1} and Y_t, the second lag has both an autocorrelation and partial autocorrelation between Y_{t-2} and Y_t, and so on. ACFs and PACFs are the functions across all the lags.

Examine the ACF and PACF of the logarithmic, first difference series as shown in Figure 5.5 and Figure 5.6 respectively.

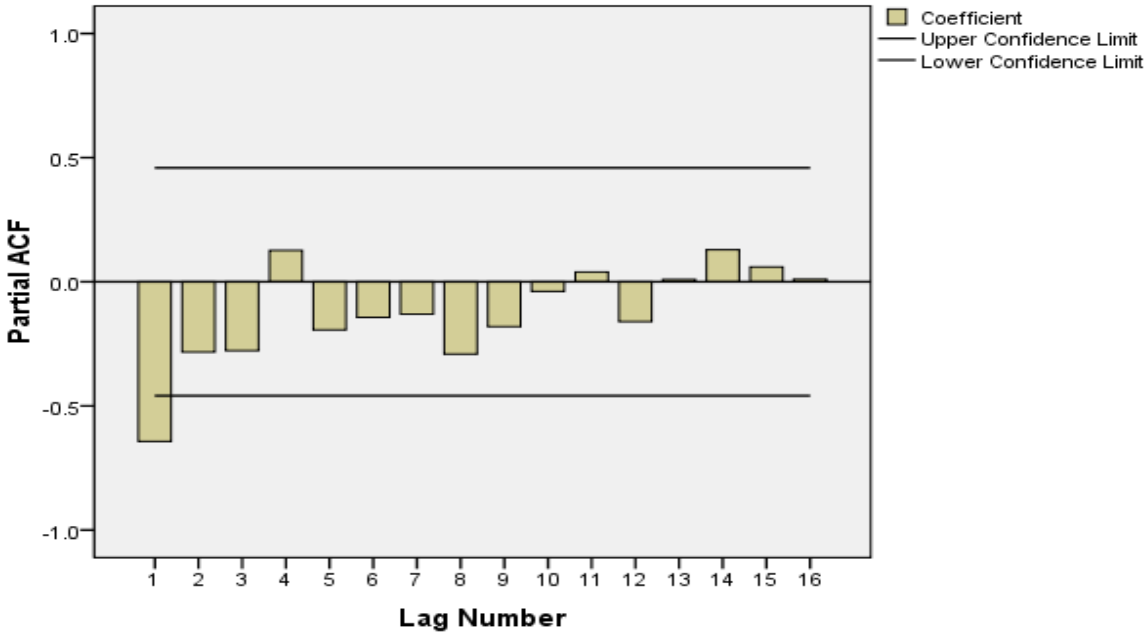

Figure 5.5 the ACF of the natural logarithm and second difference of Khartoum State yearly cases

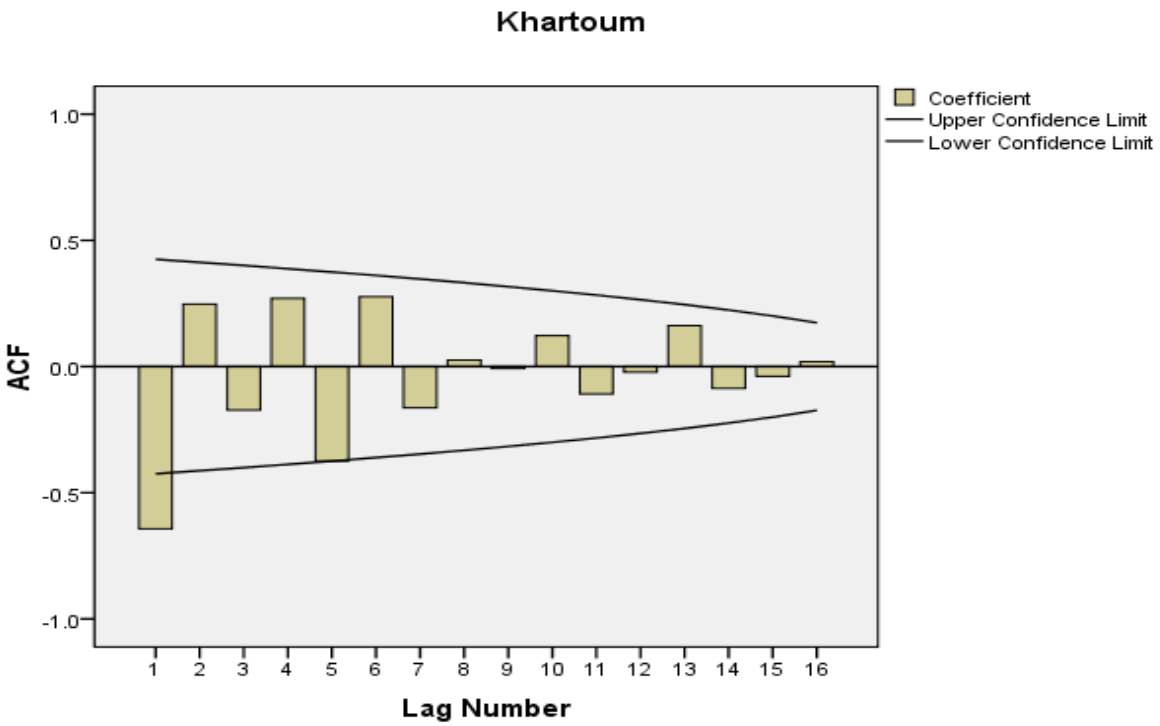

Figure 5.6 the PACF of the natural logarithm and second difference of Khartoum State yearly cases

5.2.1.2.2 Auto- Regressive Components (AR)

The auto- regressive components represent the memory of the process for preceding observations. The value of p is the number of auto- regressive components in an ARIMA (p, d, q) model. The value of p is 0 if there is no relationship between adjacent observations. When the value of p is 1, there is a relationship between observations at lag 1 and the correlation coefficient Φ1 is the magnitude of the relationship. When the value of p is 2, there is a relationship between observations at lag 2 and the correlation coefficient Φ2 is the magnitude of the relationship. Thus, p is the number of correlations you need to model the relationship and the equation is shown in Eq. 5.1.

$$\hat{Y}_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + a_t \qquad \text{Eq. 5.1}$$

When the PACF of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive that mean adding an AR term to the model is needed.

The lag at which the PACF cuts off is the indicated number of AR terms. Thus, examine the PACF in Figure 5.5 lead us to propose model AR (1), which define by Eq. 5.2,

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{\emptyset_1 Y_{t-1}} + \boldsymbol{a_t} \quad \text{Eq. 5.2}$$

5.2.1.2.3 Moving Average Components (MA)

The moving average components represent the memory of the process for the preceding random shocks. The value q indicates the number of moving average components in an ARIMA (p, d, q) model. When q is zero, there are no moving average components. When q is 1, there is a relationship between the current score and the random shock at lag 1, and the correlation coefficient Θ1 represents the magnitude of the relationship. When q is 2, there is a relationship between the current score and the random shock at lag 2, and the correlation coefficient Θ2 represents the magnitude of the relationship and the equation is shown in Eq. 5.3.

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad \text{Eq. 5.3}$$

When the ACF of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative that mean adding an MA term to the model is needed. The lag at which the ACF cuts off is the indicated number of MA terms. Thus, examine the ACF in Figure 5.5 lead us to propose model MA (1), which define by Eq. 5.4.

$$\boldsymbol{X_t = a_t - \theta_1 a_{t-1}} \quad \text{Eq. 5.4}$$

5.2.1.2.4 Mixed Model (ARIMA)

Sometimes a series has both auto- regressive and moving average components so both types of correlations are required to model the patterns and the equation is shown in Eq. 5.5.

$$\hat{Y}_t = a_t + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad \text{Eq. 5.5}$$

And if we look again at the two Figures 5.5, we propose a model ARIMA (1,2,1) because p is 1, q is 2, and we have applied the first difference to the series, which define by the equation shown in 5.6.

$$\widehat{Y}_t = \emptyset_1 Y_{t-1} - \theta_1 a_{t-1} + a_t \quad \text{Eq. 5.6}$$

5.2.1.2.5 Exponential Smoothing (SES)

Simple Exponential Smoothing is given by equation 5.7.

$$S_{t+1} = \alpha y_t + (1-\alpha) S_t \qquad 0 < \alpha \leq 1,\ t > 0 \quad \text{Eq. 5.7}$$

Where,

S_{t+1}: Next point Forecasting.

y_t: Actual value for period t.

St: Forecast value for period t.

5.2.1.3 Model Estimation (Fitting)

Estimating the values of the parameters in models consists of estimating the ∅ parameters from an auto- regressive model or the θ parameters from a moving average model, and the following rules apply:

- Parameters must differ significantly from zero and all significant parameters must be included in the model.
- All auto- regressive parameters, ∅ must be between -1 and 1.

If there are two such parameters (p=2) they must also meet the following requirements:

$$\emptyset_1 + \emptyset_2 < 1 \text{ and}$$

$$\emptyset_2 - \emptyset_1 < 1$$

These are called the bounds of stationarity for the auto- regressive parameters.

- All moving average parameters, θ must be between -1 and 1. If there are two such parameters (q = 2), they must also meet the following requirements:

$$\theta_1 + \theta_2 < 1 \text{ and}$$

$$\theta_2 - \theta_1 < 1$$

These are called the bounds of invertibility for the moving average parameters.

5.2.1.3.1 Estimating the model AR (1) with the second series differencing

Estimating the model AR (1) with the second series differencing gives the model ARIMA (1,2,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.1

Table 5.1 the results of the AR (1) estimation of Khartoum State yearly cases

			Estimate	SE	t	Sig.
Khartoum-Model_1	Khartoum	**Constant**	103.705	1195.824	.087	.932
		AR **Lag 1**	-.562	.196	-2.864	.011
		Difference	2			

From Table 5.1 it is clear that AR coefficient is between -1 and 1, and *p* value of 0.011 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.1 form the equation shown in Eq. 5.8.

$$\widehat{Y}_t = a_t - 0.562\, Y_{t-1} \quad \text{Eq. 5.8}$$

5.2.1.3.2 Estimating the model MA (1) with the second series differencing

Estimating the model MA (1) with the second series differencing gives the model ARIMA (0,2,1) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.2.

Table 5.2 the results of the MA (1) estimation of Khartoum State yearly cases

		Estimate	SE	t	Sig.
Khartoum-Model_1 Khartoum	**Constant**	-5.347	318.560	-.017	.987
	Difference	2			
	MA Lag 1	1.000	218.666	.005	.996

From Table 5.2 the θ =1 and *p* value of 0.996 > 0.05 indicates that the correlation isn't significantly different from zero.

5.2.1.3.3 Estimating the model ARIMA (1,2,1)

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.3.

Table 5.3 the results of the ARIMA (1,2,1) estimation of Khartoum State yearly cases

			Estimate	SE	t	Sig.
Khartoum-Model_1	Khartoum	**Constant**	-71.345	262.224	-.272	.789
		AR **Lag 1**	-.102	.288	-.354	.728
		Difference	2			
		MA **Lag 1**	.986	3.215	.307	.763

From Table 5.3 the AR coefficient -0.102 is between -1 and 1, the MA coefficient 0.986 is between -1 and 1 and their *p* values 0.728 > 0.05 and 0.763 > 0.05 respectively indicate that the correlation isn't significantly different from zero.

5.2.1.3.4 Estimating SES model

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.4.

Table 5.4 The result of estimating SES for Khartoum State yearly cases

Model		Estimate	SE	t	Sig.
Khartoum-Model_1	**Alpha (Level)**	.856	.219	3.902	.001

From Table 5.4 it is clear that alpha 0.856 ≤ 1, and *p* value of 0.001 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.1 form the equation shown in Eq. 5.9.

$$S_{t+1} = .856\, y_t + (1-.856)\, S_t \quad \text{Eq. 5.9}$$

5.2.1.4 Model Diagnosis

Ljung-Box test is the series correlation test applied to ensure that the proposed model is appropriate, and since the ARIMA (0,2,1) and ARIMA (1,2,1) aren't significantly different from zero we'll not make the test for them.

5.2.1.4.1 ARIMA (1,2,0) model

Table 5.5 shows the result of Ljung-Box test for ARIMA (1,2,0) using the software SPSS 17.0.

Table 5.5 the result of Ljung-Box test of ARIMA (1,2,0) for Khartoum State yearly cases

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
Khartoum-Model_1	14.902	17	.603

The Ljung-Box test as shown in the Table 5.5 indicates that there is no self-correlation between the errors (sig. 0.603 > 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.7.

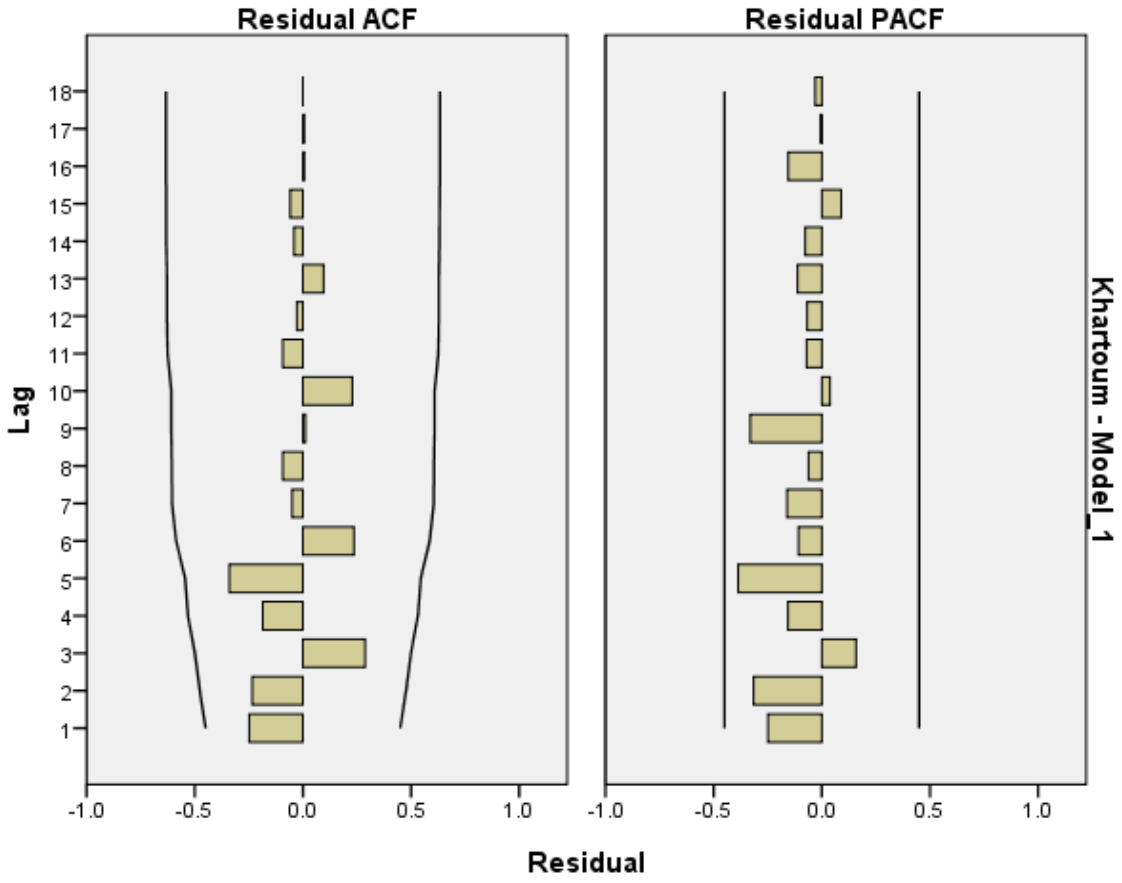

Figure 5.7 the ACF and PACF residuals of ARIMA (1,2,0) for Khartoum State yearly cases

5.2.1.4.2 Simple Exponential Smoothing (SES) model

Table 5.6 shows the result of Ljung-Box test for Simple Exponential Smoothing using the software SPSS 17.0.

Table 5.6 the result of Ljung-Box test of SES for Khartoum State yearly cases

Model	Ljung-Box Q(18)

	Statistics	DF	Sig.
Khartoum-Model_1	9.833	17	.910

The Ljung-Box test as shown in the Table 5.4 indicates that there is no self-correlation between the errors (sig. 0.910 > 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.8.

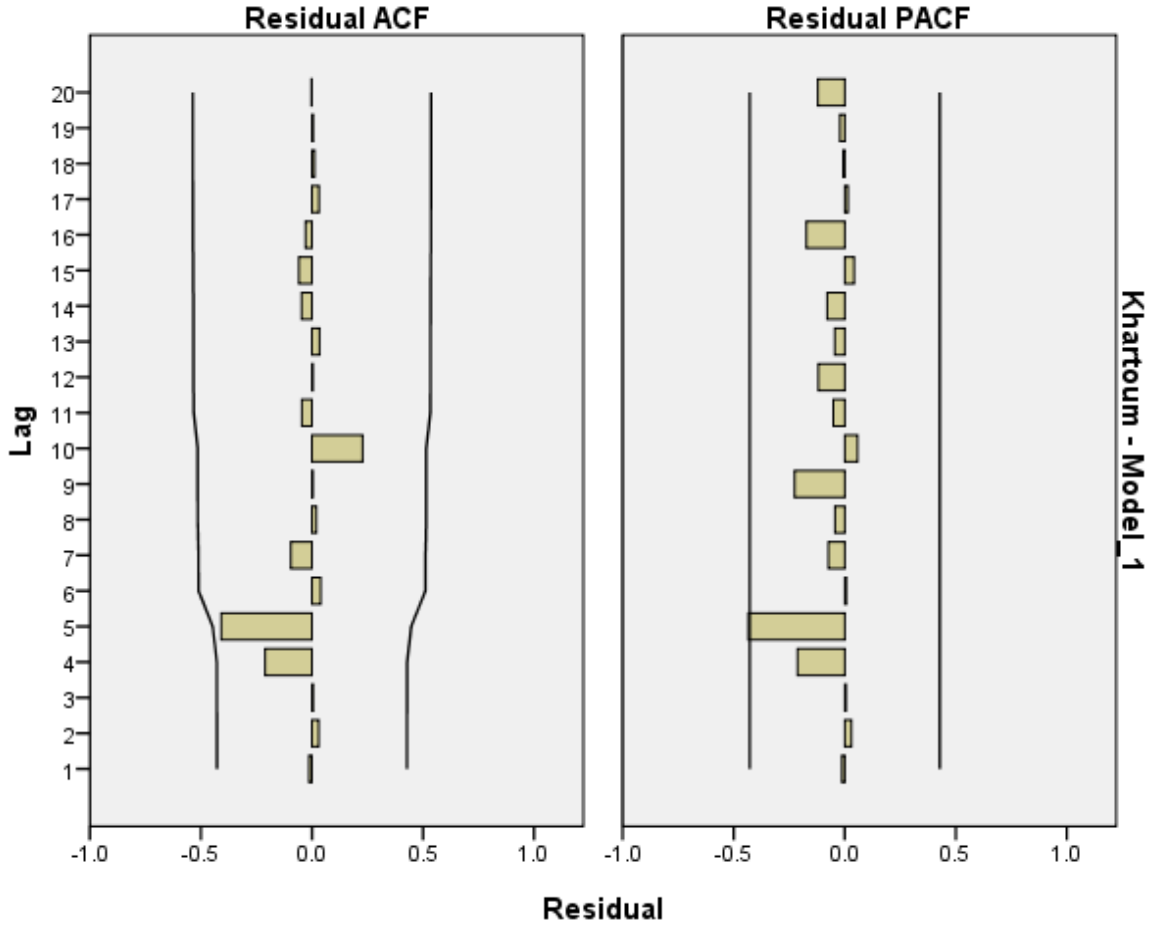

Figure 5.8 the ACF and PACF residuals of SES for Khartoum State yearly cases

5.2.2 Al-Gadaref

5.2.2.1 Time Series Stability

The first step in the analysis is to plot the time series; Figure 5.9 shows cases of the Malaria in Al-Gadaref State.

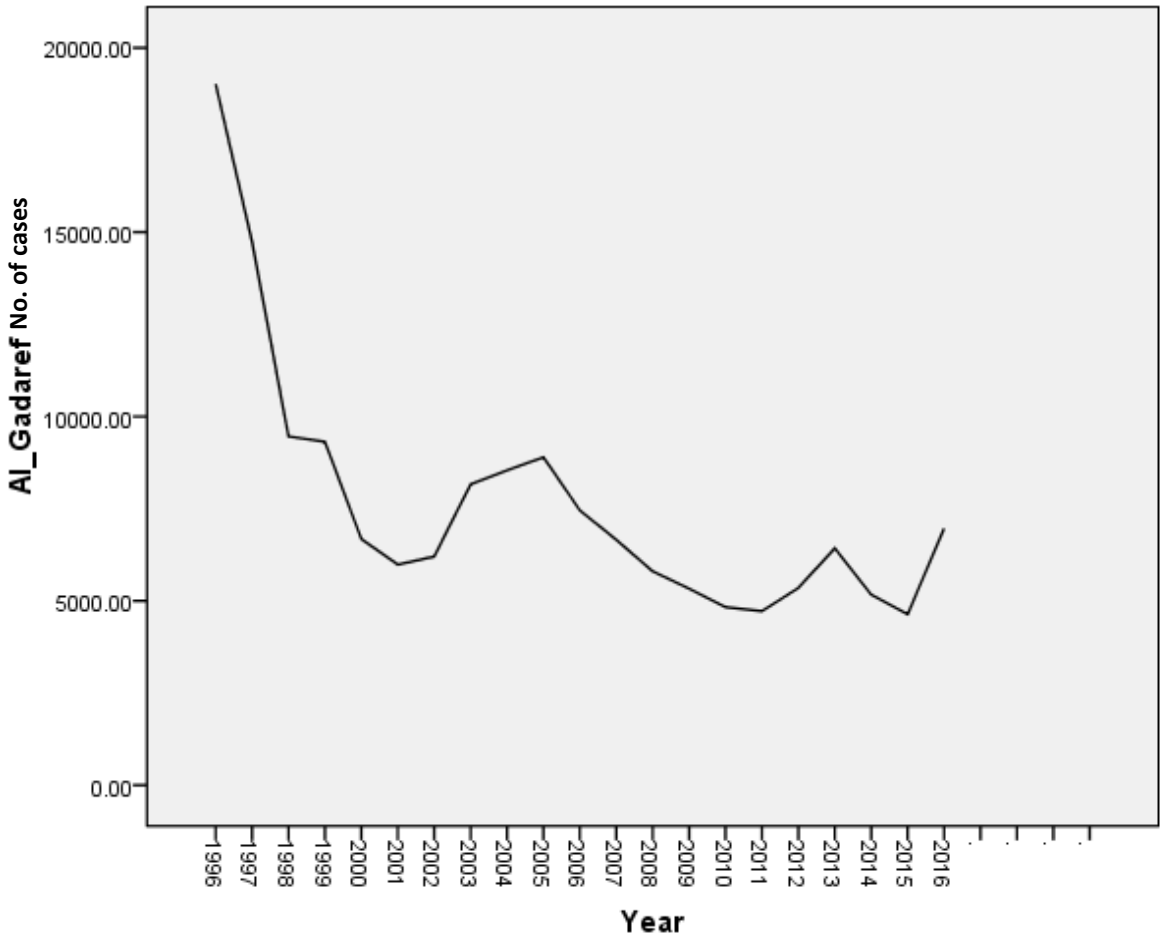

Figure 5.9 Yearly cases of the Malaria in Al-Gadaref State

It is obvious that there is a shift in both the trend and the dispersion over time for the series.

To see if the process is stationary after linear trend is removed, the first difference of the cases plotted against years, as seen in Figure 5.10.

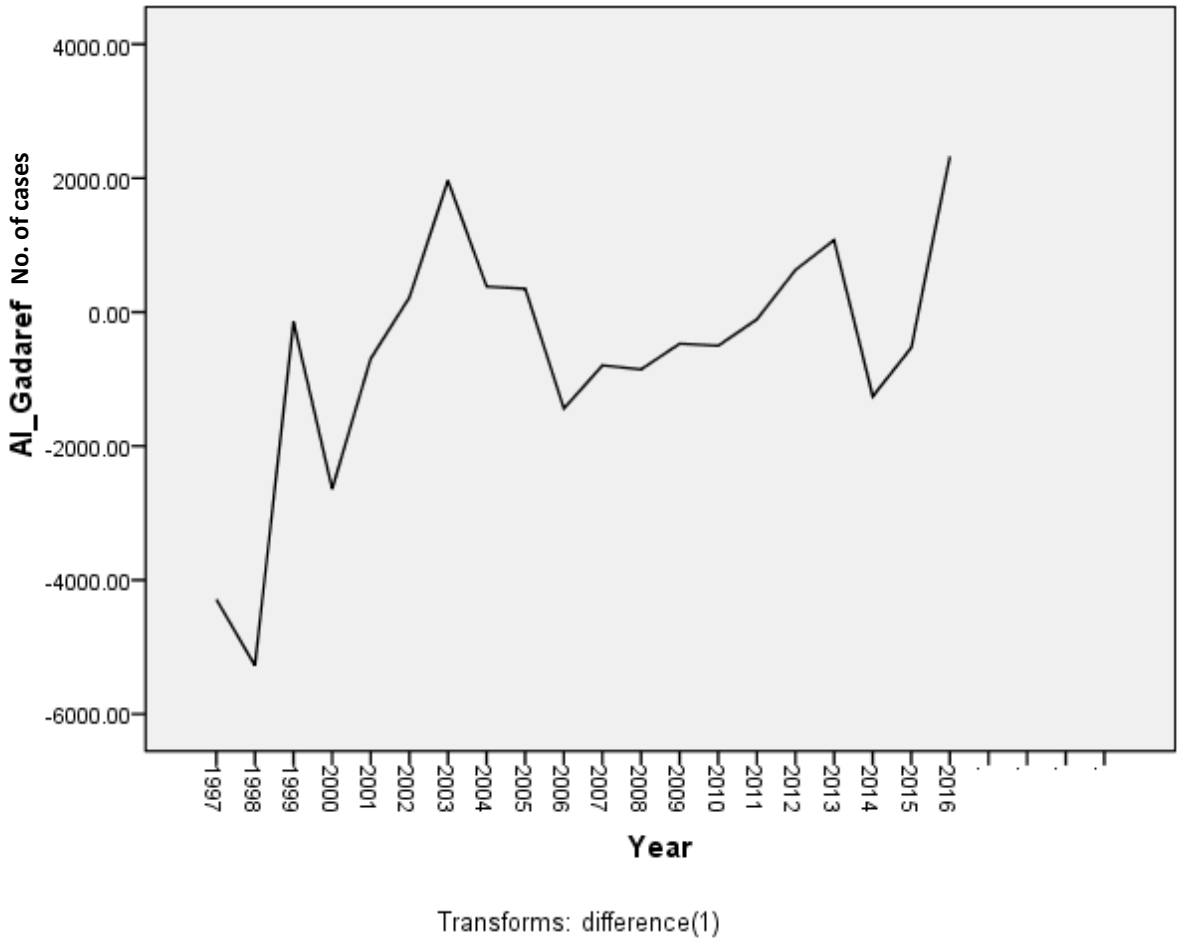

Figure 5.10 the first difference of Al-Gadaref State yearly cases

It is clear that the series is still has the shift in the mean, so second differencing appear necessary which is shown in Figure 5.11

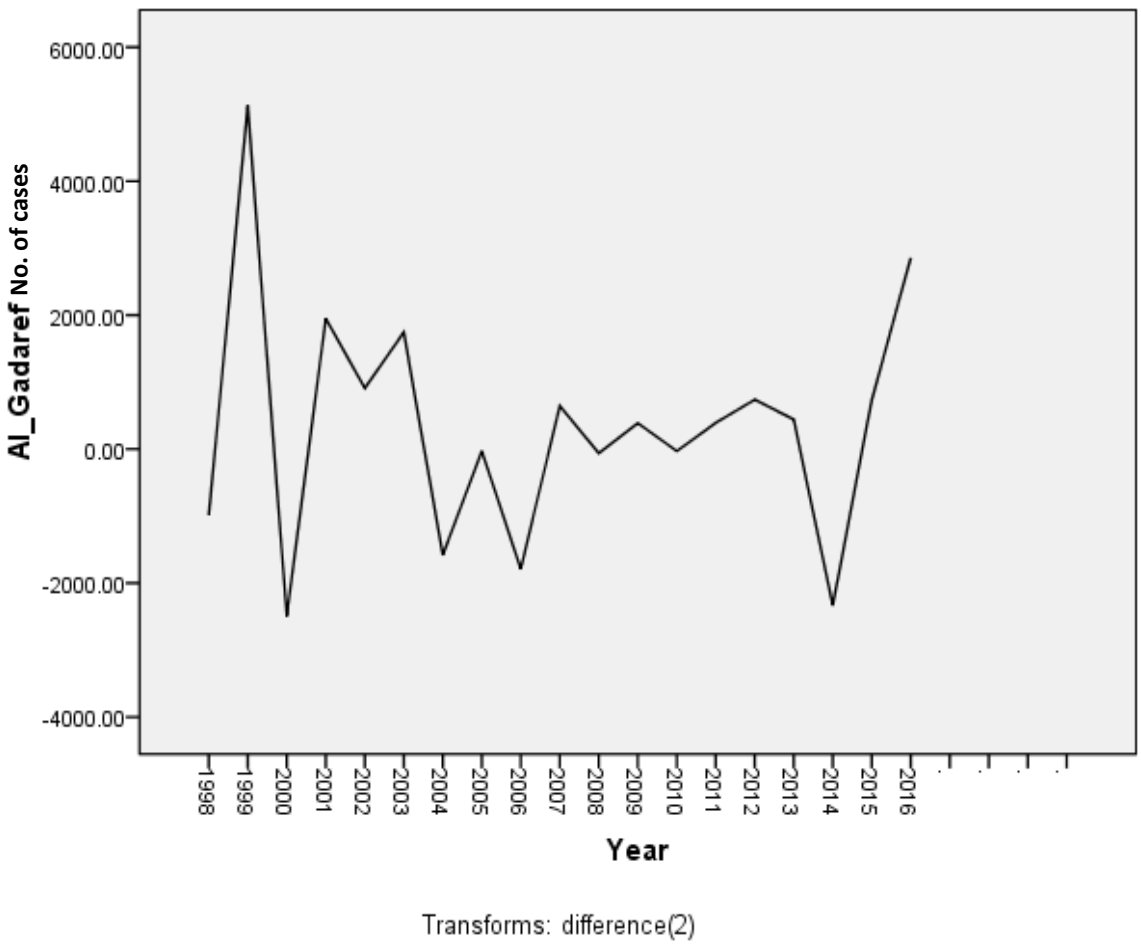

Figure 5.11 the second difference of Al-Gadaref State yearly cases

The series now appears stationary with respect to central tendency; however, the logarithmic transformation is plotted in Figure 5.12.

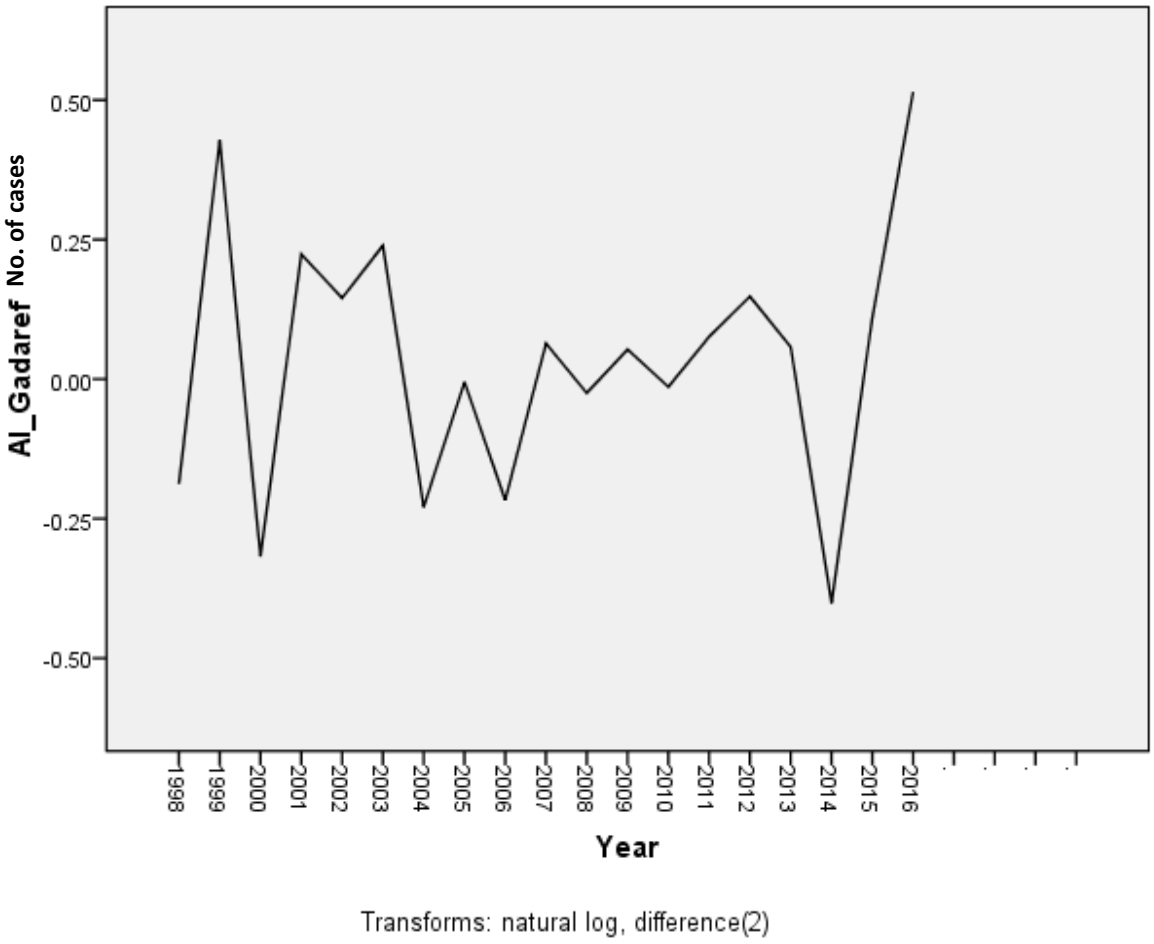

Figure 5.12 the natural logarithm and second difference of Al-Gadaref State yearly cases

5.2.2.2 Model Identification

For determine AR, MA, and ARIMA components.

5.2.2.2.1 ACFs and PACFs

Examine the ACF and PACF of the second difference series as shown in Figure 5.13 and Figure 5.14 respectively.

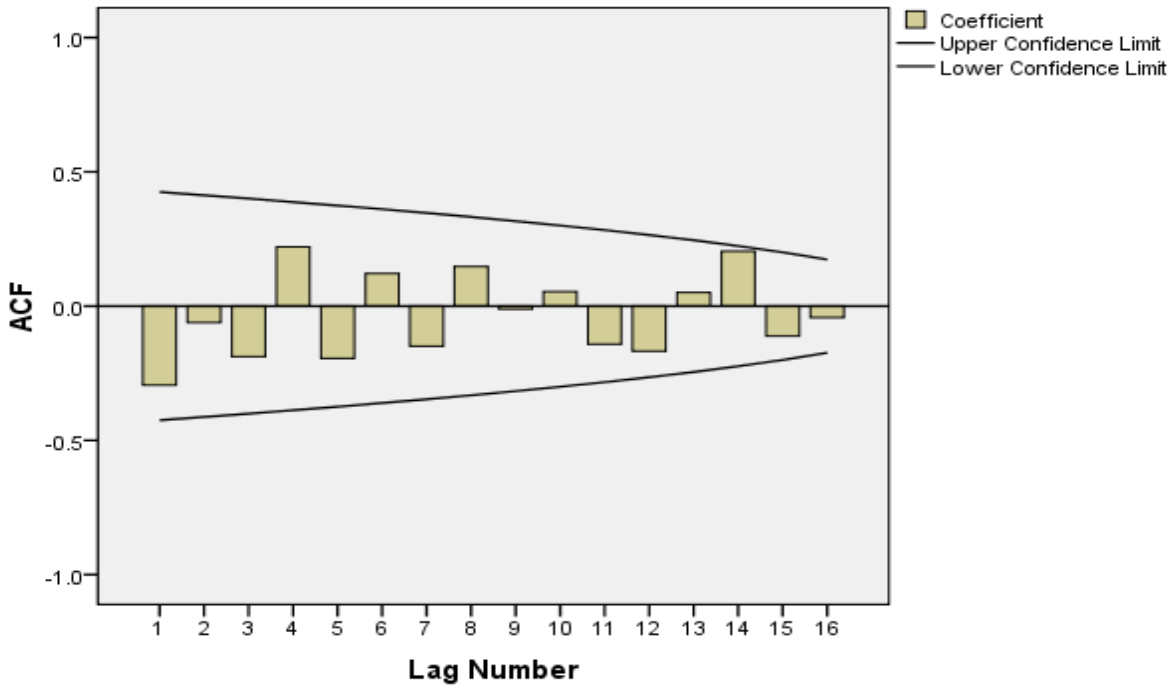

Figure 5.13 the ACF of the natural logarithm and second difference of Al-Gadaref State yearly Cases

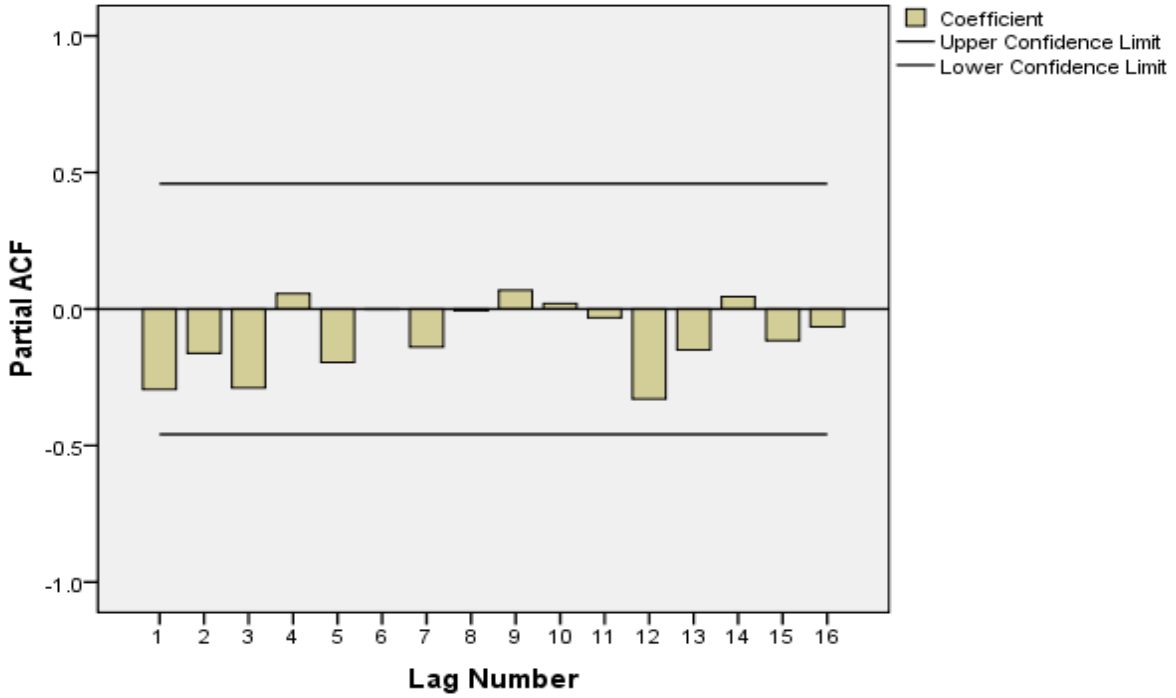

Figure 5.14 the and PACF of the natural logarithm and second difference of Al-Gadaref State yearly Cases

5.2.2.2.2 Auto- Regressive Components (AR)

Examine the PACF in Figure 5.14 lead us to propose model AR (0).

5.2.2.2.3 Moving Average Components (MA)

Examine the ACF in Figure 5.13 lead us to propose model MA (0).

5.2.2.2.4 Mixed Model (ARIMA)

While there is no AR and MA component the model will be ARIMA (0,2,0).

5.2.2.2.5 Brown Exponential Smoothing (BES)

The BES is representing by equation 5.10.

Level equation $\ell_t=\alpha y_t+(1-\alpha)(\ell_{t-1}+b_{t-1})$

Trend equation $b_t=\beta*(\ell_t-\ell_{t-1})+(1-\beta*)b_{t-1}$

Forecast equation $\hat{y}_{t+h|t}=\ell_t+hb_t$ Eq. 5.10

Where,

ℓ_t denotes an estimate of the level of the series at time t.

b_t denotes an estimate of the trend (slope) of the series at time t.

α is the smoothing parameter for the level, $0\leq\alpha\leq1$.

$\beta*$ is the smoothing parameter for the trend, $0\leq\beta*\leq1$.

5.2.2.3 Model Estimation (Fitting)

While AR and MA components are 0 then we'll not make the estimation for them.

5.2.2.3.1 Estimating the model ARIMA (0,2,0)

The results of the ARIMA (0,2,0) model estimation are shown using the statistical software SPSS 17.0 in the Table 5.7.

Table 5.7 the results of the ARIMA (0,2,0) model estimation of Al-Gadaref State yearly cases

		Estimate	SE	t	Sig.
Al_Gadaref-Model_1 Al_Gadaref	**Constant**	348.000	419.414	.830	.418
	Difference	2			

5.2.2.3.2 Estimating BES model

The results of the BES estimation are shown using the statistical software SPSS 17.0 in the Table 5.8.

Table 5.8 the results of the BES model estimation of Al-Gadaref State yearly cases

Model		Estimate	SE	t	Sig.
Al_Gadaref-Model_1	**Alpha (Level and Trend)**	.875	.088	9.959	.000

From Table 5.8 it is clear that alpha $0.875 \leq 1$, and p value of $0.000 < 0.05$ indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.1 form the equation shown in Eq. 5.11.

$$S_{t+1} = .875\, y_t + (1-.875)\, S_t \qquad \text{Eq. 5.11}$$

5.2.2.4 Model Diagnosis

5.2.2.4.1 ARIMA (0,2,0) model

Table 5.9 shows the result of Ljung-Box test for ARIMA (0,2,0) using the software SPSS 17.0.

Table 5.9 the result of Ljung-Box test of ARIMA (0,2,0) for Al-Gadaref State yearly cases

	Ljung-Box Q(18)		
Model	**Statistics**	**DF**	**Sig.**
Al_Gadaref-Model_1	25.346	18	.116

The Ljung-Box test as shown in the Table 5.9 indicates that there is no self-correlation between the errors (sig. $0.116 > 0.05$), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.15.

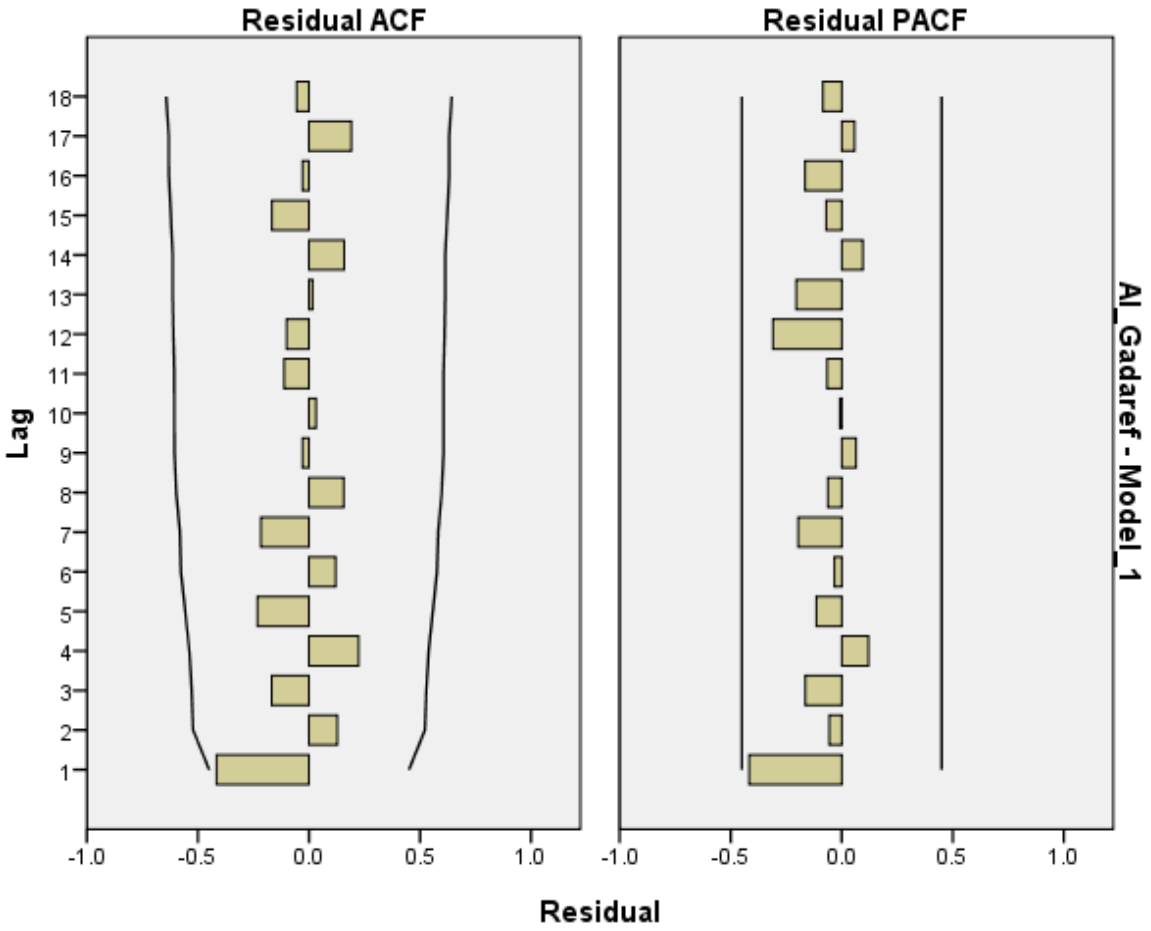

Figure 5.15 the ACF and PACF residuals of ARIMA (0,2,0) for Al-Gadaref State yearly cases

5.2.2.4.2 Brown Exponential Smoothing BES model

Table 5.10 shows the result of Ljung-Box test for Brown Exponential Smoothing using the software SPSS 17.0.

Table 5.10 the result of Ljung-Box test of BES for Al-Gadaref State yearly cases

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
Al_Gadaref-Model_1	21.036	17	.225

The Ljung-Box test as shown in the Table 5.10 indicates that there is no self-correlation between the errors (sig. 0.225 > 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.16.

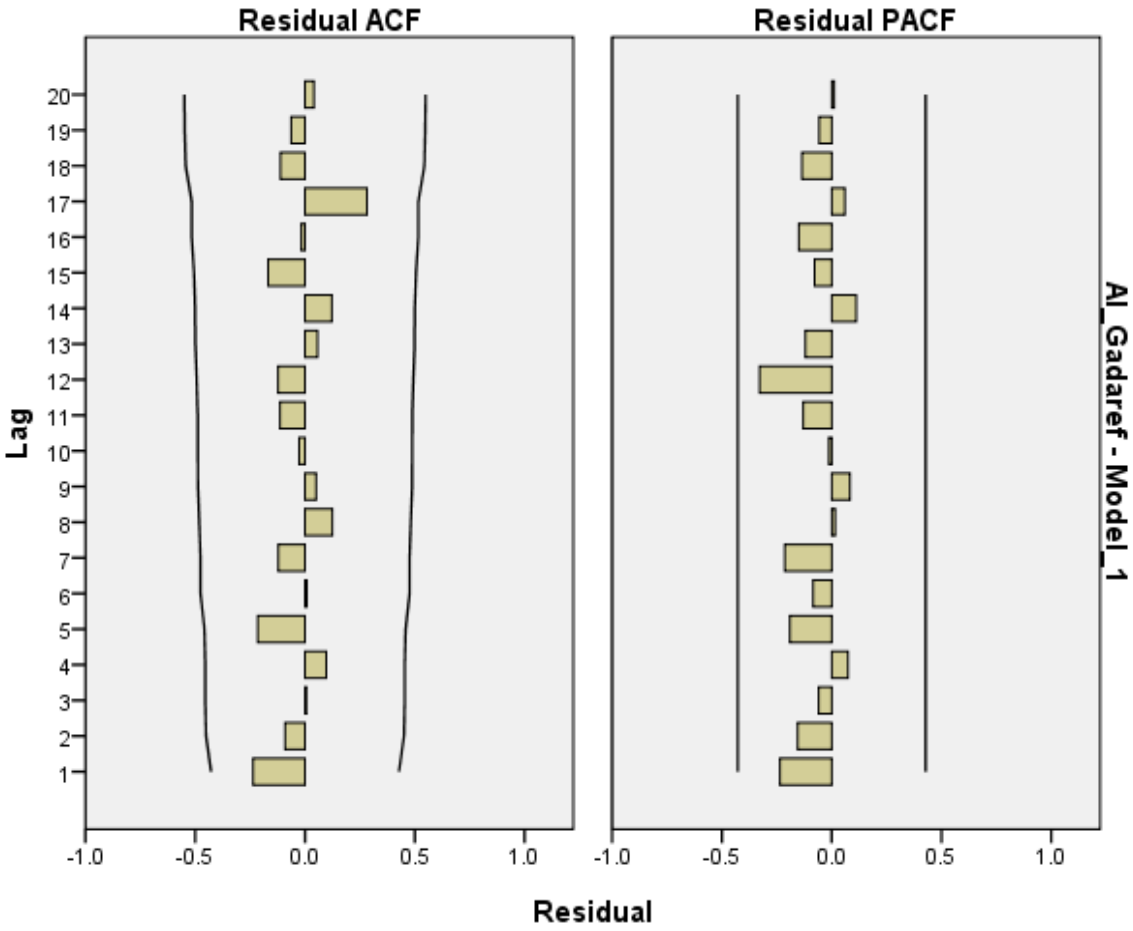

Figure 5.16 the ACF and PACF residuals of BES for Al-Gadaref State yearly cases

5.2.3 Sennar

5.2.3.1 Time Series Stability

The first step in the analysis is to plot the time series; Figure 5.17 shows cases of the Malaria in Sennar State.

Figure 5.17 Yearly cases of the Malaria in Sennar State

It is obvious that there is a shift in both the trend and the dispersion over time for the series. The first difference of the cases plotted against years, as seen in Figure 5.18.

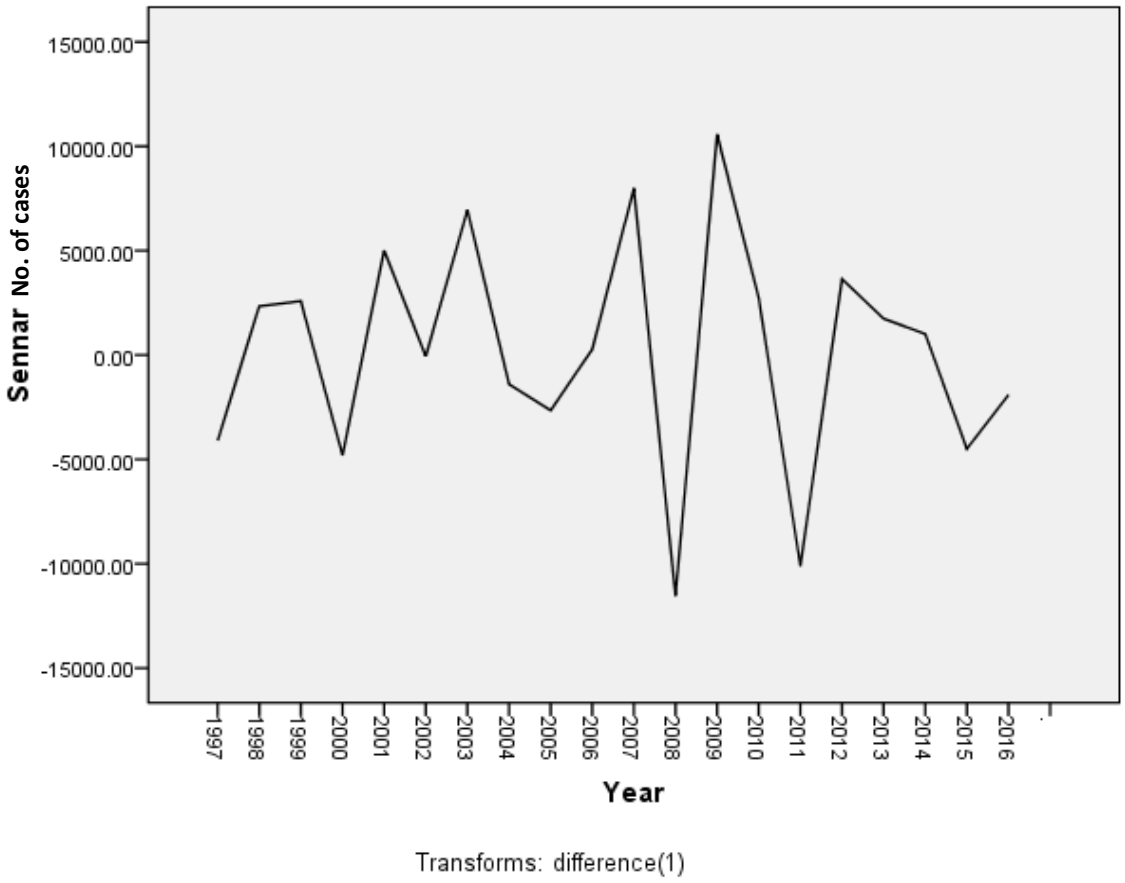

Figure 5.18 the first difference of Sennar State's Cases

The series now appears stationary with respect to central tendency, so second differencing does not appear necessary. However, the variability seems to be increasing and decreasing over time.

The transformed difference is plotted to see if both mean and variance are now stabilized, as seen in Figure 5.19.

Figure 5.19 the natural logarithm and first difference of Sennar State yearly cases

5.2.3.2 Model Identification

To determine the components of AR, MA, and ARIMA models.

5.2.3.2.1 ACFs and PACFs

Examine the ACF and PACF of the first difference series as shown in Figure 5.20 and Figure 5.21 respectively.

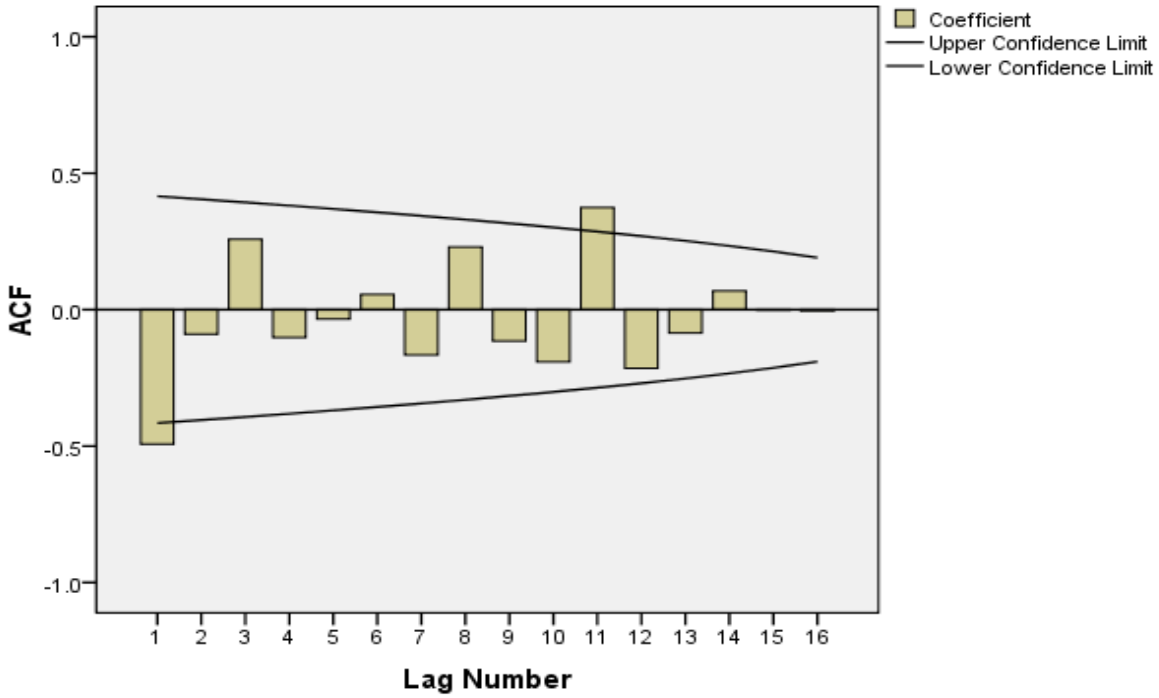

Figure 5.20 the ACF of the natural logarithm and first difference of Sennar State yearly cases

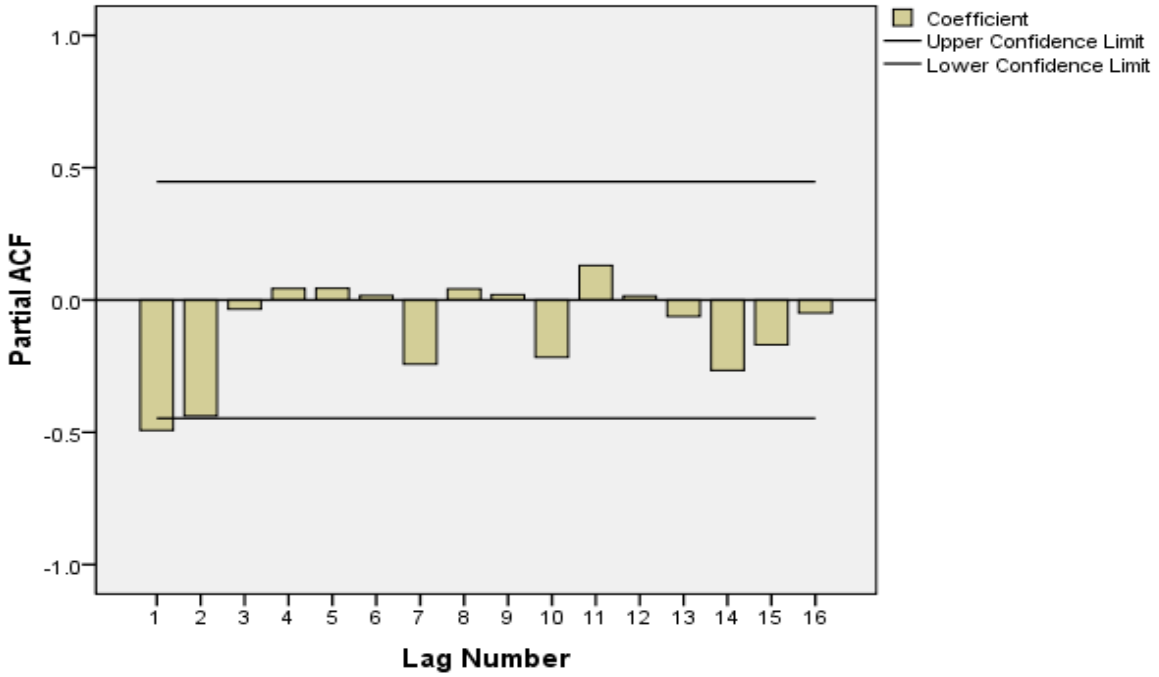

Figure 5.21 the PACF of the natural logarithm and first difference of Sennar State yearly cases

5.2.3.2.2 Auto- Regressive Components (AR)

Examine the PACF in Figure 5.16 lead us to propose model AR (1), which define by Eq. 5.12.

$$\widehat{Y}_t = \emptyset_1 Y_{t-1} + a_t \quad \text{Eq. 5.12}$$

5.2.3.2.3 Moving Average Components (MA)

Examine the ACF in Figure 5.16 lead us to propose model MA (1), which define by Eq. 5.13.

$$X_t = a_t - \theta_1 a_{t-1} \quad \text{Eq. 5.13}$$

5.2.3.2.4 Mixed Model (ARIMA)

And if we look at the two Figures 5.16, we propose a model ARIMA (1,1,1) because p is 1, q is 1, and we have applied the first difference to the series, which define by the equation shown in 5.14.

$$\widehat{Y}_t = \emptyset_1 Y_{t-1} - \theta_1 a_{t-1} + a_t \quad \text{Eq. 5.14}$$

5.2.3.3 Model Estimation (Fitting)

Insure that; the models coefficients are significantly differ from zero.

5.2.3.3.1 Estimating the model AR (1) with the first series differencing

Estimating the model AR (1) with the first series differencing gives the model ARIMA (1,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.11.

Table 5.11 the results of the ARIMA (1,1,0) estimation of Sennar State yearly Cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	.025	.049	.521	.608
		AR **Lag 1**	-.519	.199	-2.607	.018
		Difference	1			

From Table 5.11 it is clear that AR coefficient is between -1 and 1, and *p* value of 0.018 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.10 form the equation shown in Eq. 5.15.

$$\widehat{Y}_t = -0.519\, Y_{t-1} + a_t \qquad \text{Eq. 5.15}$$

5.2.3.3.2 Estimating the model MA (1) with the first series differencing

Estimating the model MA (1) with the first series differencing gives the model ARIMA (0,1,1) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.11.

Table 5.12 the results of the ARIMA (1,1,0) estimation of Sennar State yearly Cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	.038	.021	1.862	.079
		Difference	1			
		MA Lag 1	.808	.254	3.173	.005

From Table 5.12 it is clear that MA coefficient is between -1 and 1, and *p* value of 0.005 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.11 form the equation shown in Eq. 5.16.

$$X_t = a_t - 0.808\, a_{t-1} \qquad \text{Eq. 5.16}$$

5.2.3.3.3 Estimating the model ARIMA (1,1,1)

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.13.

Table 5.13 the results of the ARIMA (1,1,1) estimation of Sennar State yearly cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	.037	.024	1.549	.140
		AR **Lag 1**	-.206	.342	-.602	.555
		Difference	1			
		MA **Lag 1**	.648	.332	1.953	.067

From Table 5.13 the AR coefficient -0.206 is between -1 and 1, the MA coefficient 0.648 is between -1 and 1 and their *p* values 0.555 > 0.05 and 0.067 > 0.05 respectively indicate that the correlation isn't significantly different from zero.

5.2.3.3.4 Estimating HES model

The results of the HES estimation are shown using the statistical software SPSS 17.0 in the Table 5.14.

Table 5.14 the results of the BES model estimation of Sennar State yearly cases

Model		**Estimate**	**SE**	**t**	**Sig.**
Sennar-Model_1	**Alpha (Level)**	.099	.099	1.001	.329
	Gamma (Trend)	1.744E-6	.043	4.061E-5	1.000

From Table 5.14 it is clear that alpha 0.099 ≤ 1, and *p* value of 0.329 > 0.05 and *p* value of gamma 1 > 0.05, indicate that the correlation isn't significantly different from zero.

5.2.3.4 Model Diagnosis

Insure that; there is no self-correlation between the errors.

5.2.3.4.1 ARIMA (1,1,0)

Table 5.15 shows the result of Ljung-Box test.

Table 5.15 the result of Ljung-Box test of ARIMA (1,1,0)

Ljung-Box Q(18)		
Statistics	**DF**	**Sig.**
23.189	17	.143

The Ljung-Box test as shown in the Table 5.15 indicates that there is no self-correlation (sig. 0.143 > 0.05) between the errors, and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.22.

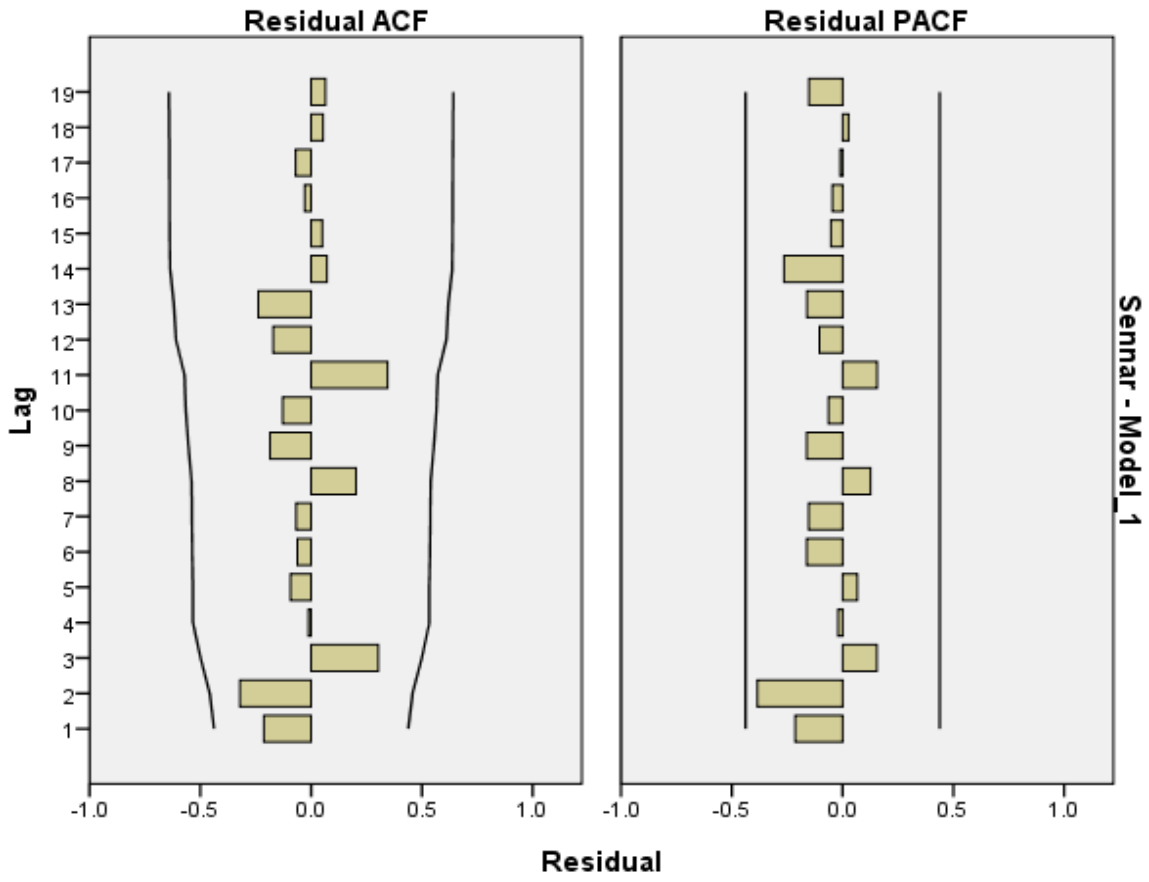

Figure 5.22 the ACF and PACF residuals of ARIMA (1,1,0) for Sennar State yearly cases

5.2.3.4.2 ARIMA (0,1,1)

Table 5.16 shows the result of Ljung-Box test.

Table 5.16 the result of Ljung-Box test of ARIMA (0,1,1)

Ljung-Box Q(18)		
Statistics	**DF**	**Sig.**
18.431	17	.362

The Ljung-Box test as shown in the Table 5.16 indicates that there is no self-correlation (sig. 0.362 > 0.05) between the errors, and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.23.

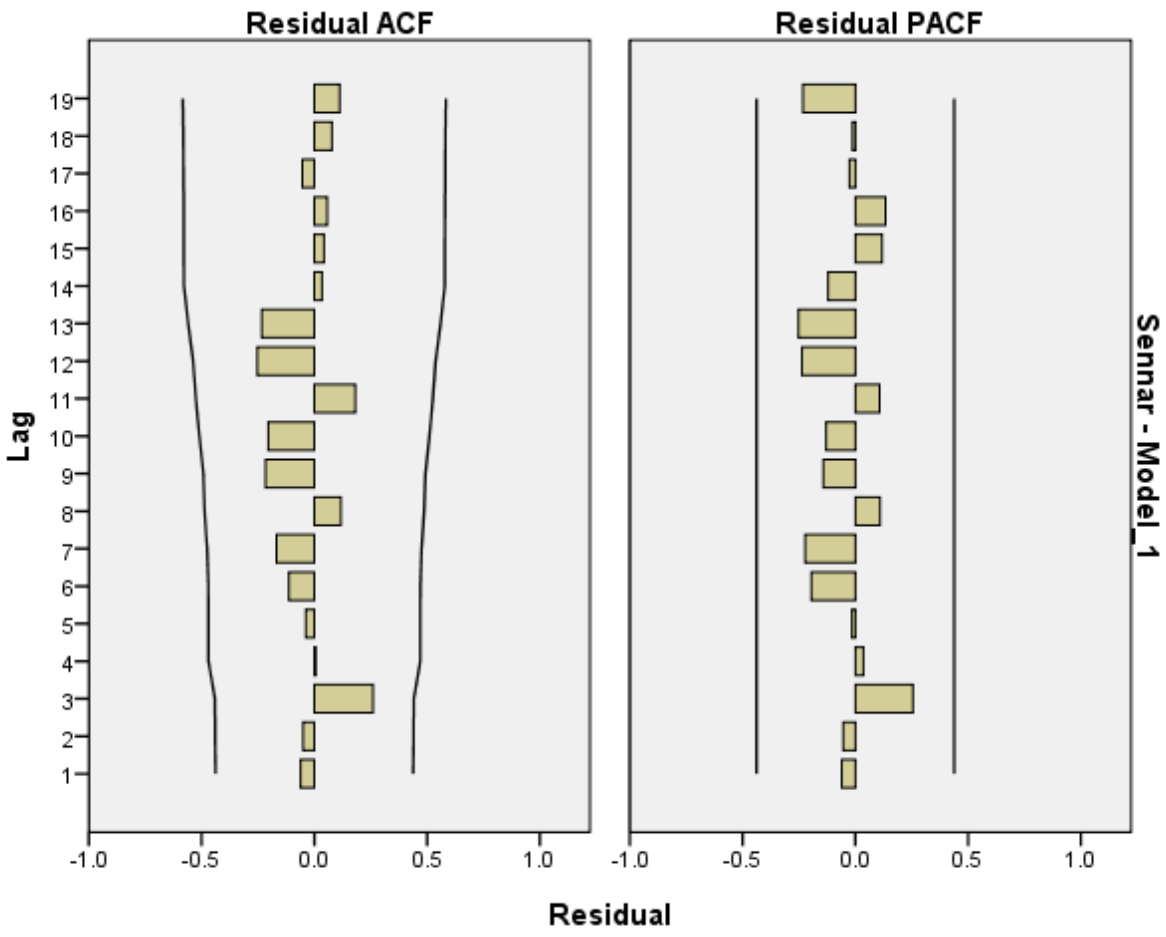

Figure 5.23 the ACF and PACF residuals of ARIMA (1,1,0) for Sennar State yearly cases

5.3 Seasonal Model – Monthly

Choosing the suitable model for the three states; using the monthly data for the years (2015 up to 2018).

5.3.1 Khartoum

5.3.1.1 Time Series Stability

The ARIMA (auto-regressive, integrated, moving average) model of a seasonal time series is defined by terms (p, d, q)(P,D,Q). The capital elements in the model; investigated the seasonal part of the model.

The first step in the analysis is to plot the seasonal time series; Figure 5.24 shows monthly cases of the Malaria in Khartoum State.

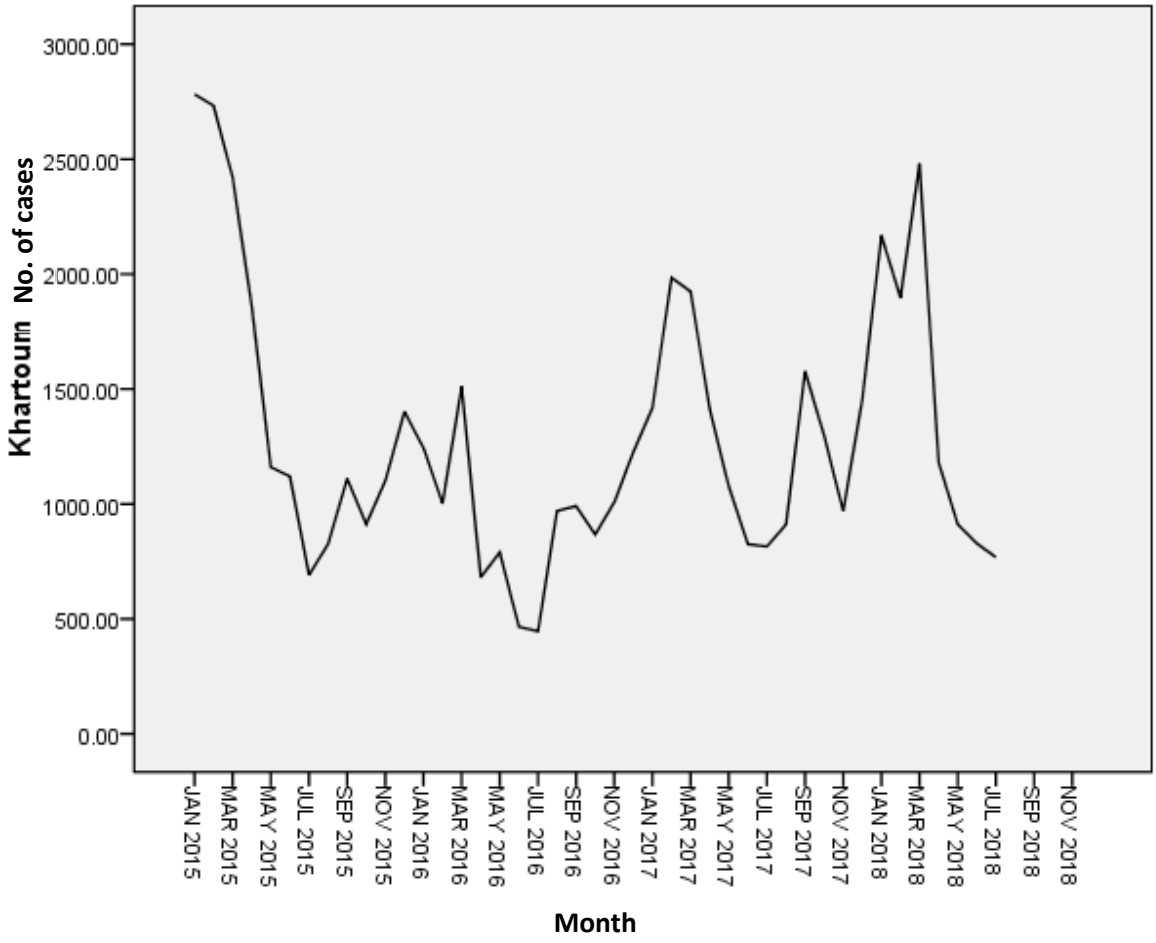

Figure 5.24 Monthly cases of the Malaria in Khartoum State

It is obvious that there is a shift in both the trend and the dispersion over time for the series.

To see if the process is stationary after linear trend is removed, the first difference of the non-seasonal part plotted against months, as seen in Figure 5.25.

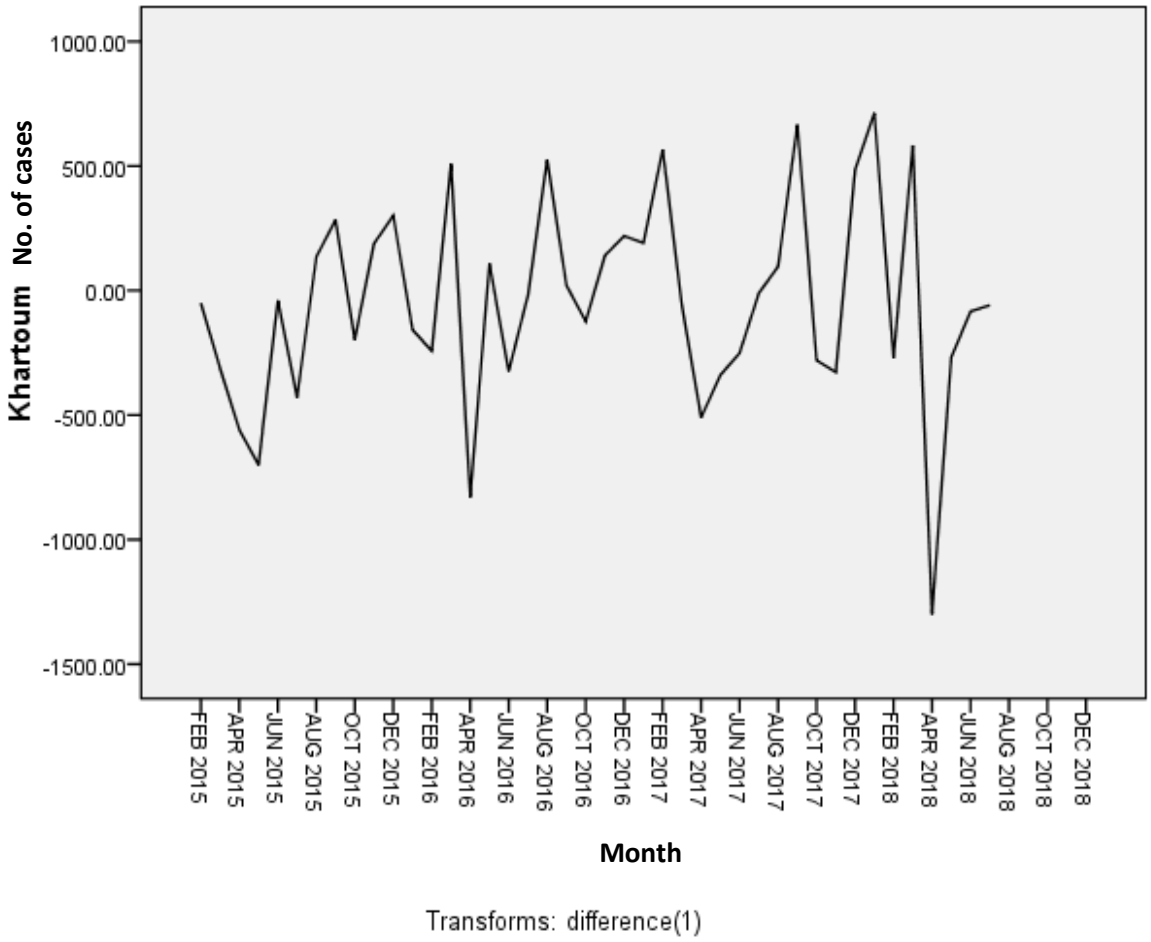

Figure 5.25 the first difference of Khartoum State monthly non-seasonal part

It is clear that the series is still has the shift in the mean in seasonal (it is clear in every April in the last figure) part, so first differencing appear necessary for seasonal part which is shown in Figure 5.26.

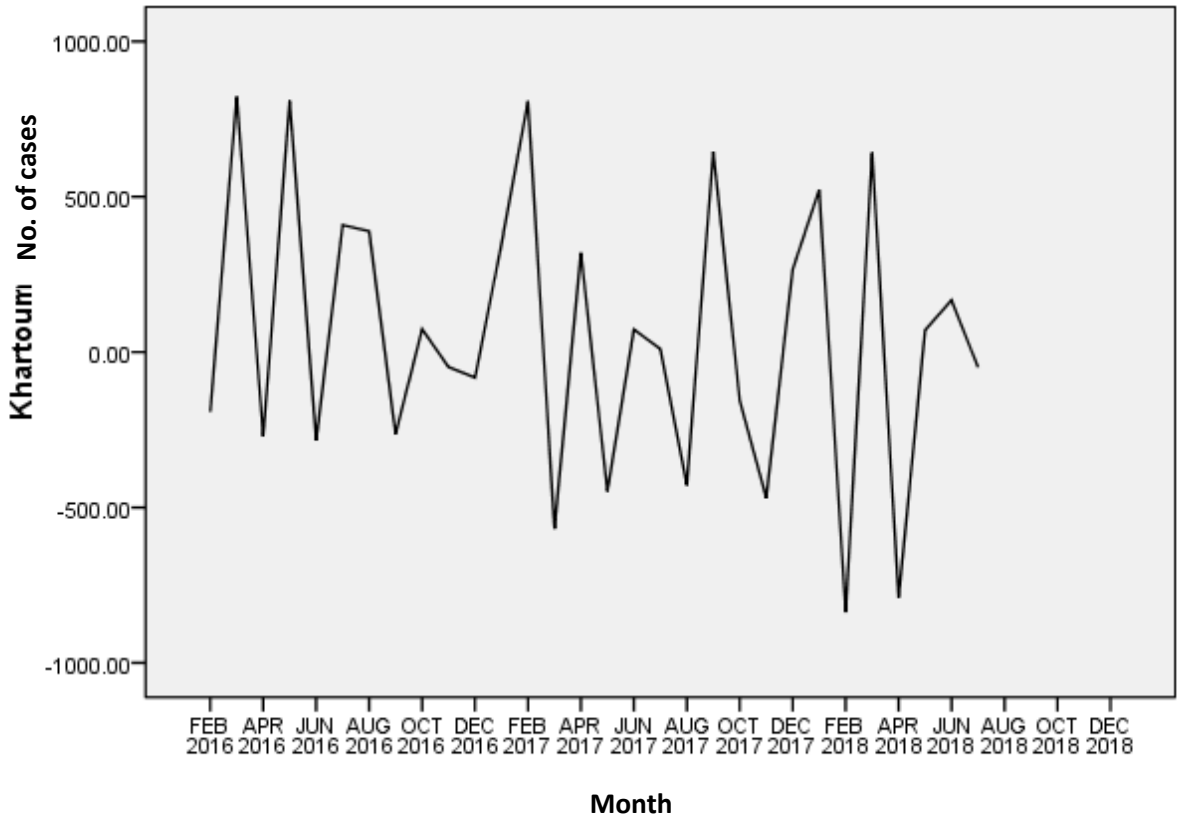

Figure 5.26 the first difference of Khartoum State monthly seasonal and non-seasonal parts

The series now appears stationary with respect to central tendency, however, the variability seems to be increasing and decreasing over time. As we have mentioned earlier, the changing in variability can be made stationary by logarithmic transformation. The transformed difference is plotted in Figure 5.27.

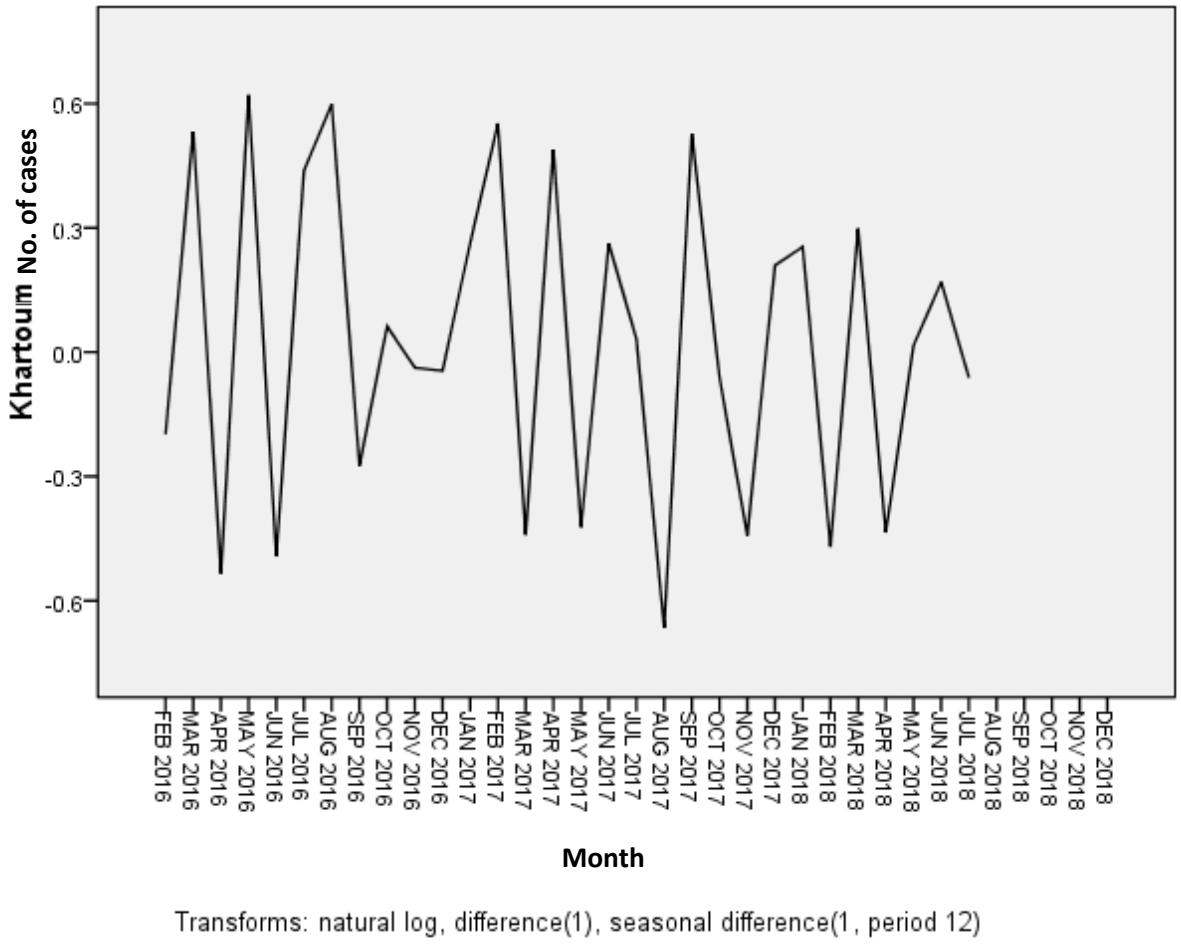

Figure 5.27 the natural logarithm and first difference of Khartoum State monthly Cases

5.3.1.2 Model Identification

Determine the components of the models (AR, MA, and ARIMA).

5.3.1.2.1 ACFs and PACFs

Examine the ACF and PACF of the logarithmic, first difference series as shown in Figure 5.28 and Figure 5.29 respectively.

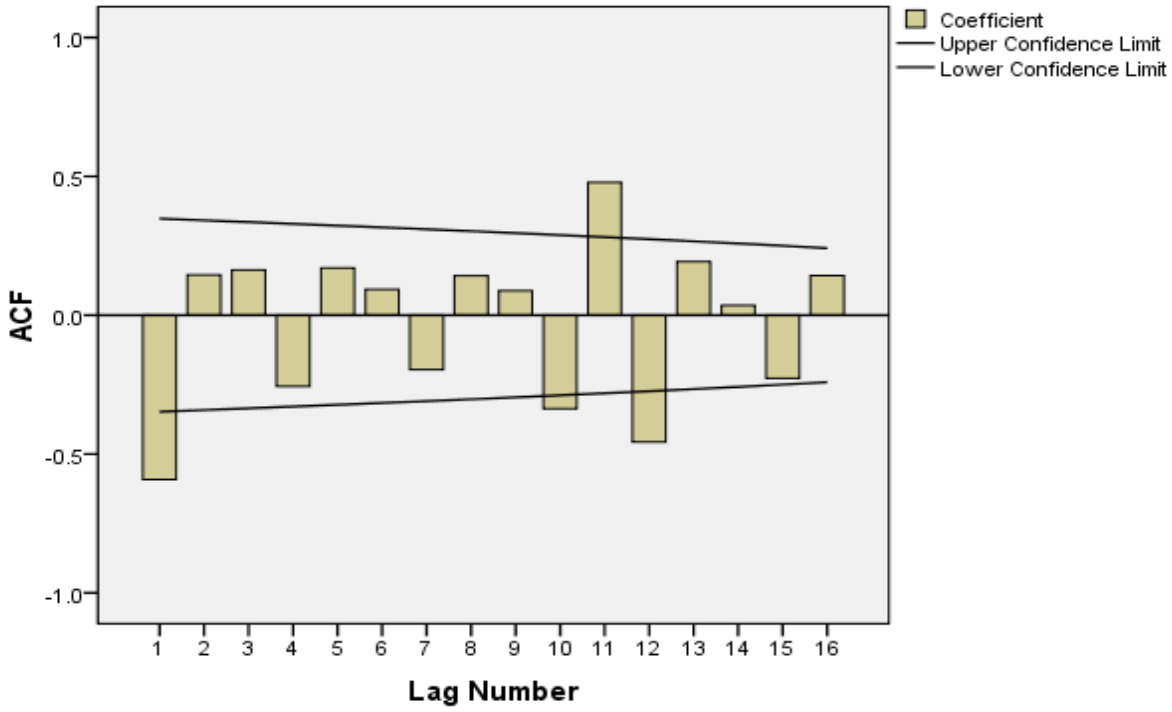

Figure 5.28 the ACF of the natural logarithm and first difference of Khartoum State monthly cases

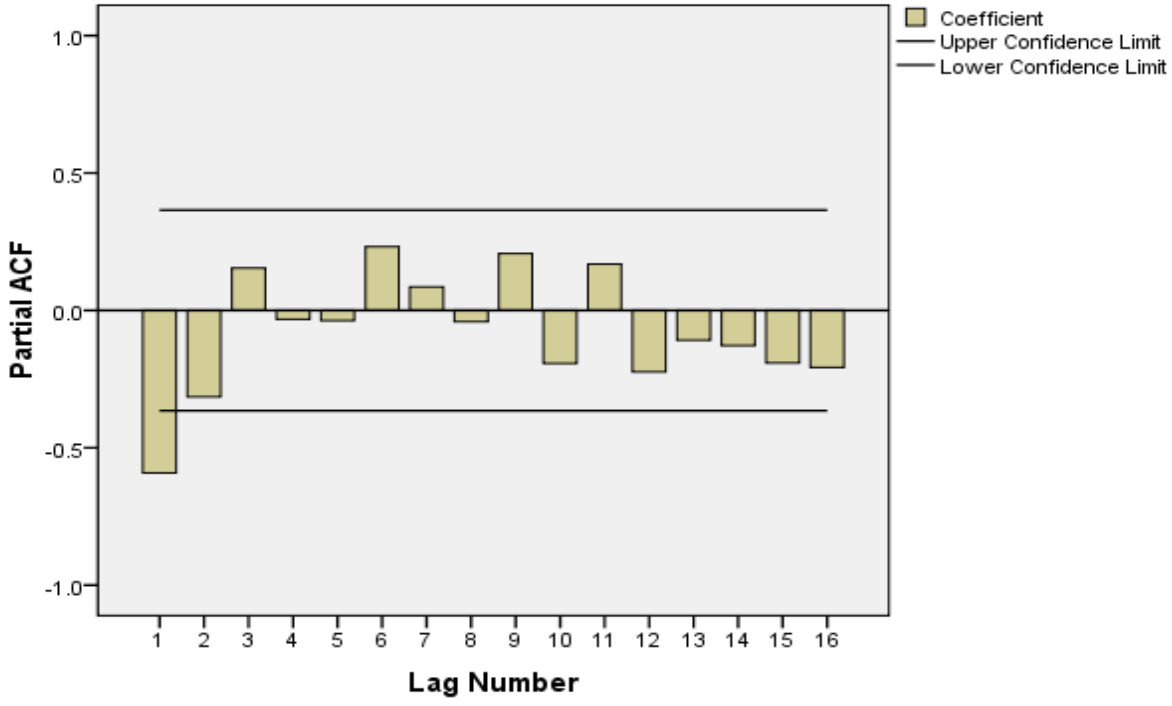

Figure 5.29 the PACF of the natural logarithm and first difference of Khartoum State monthly cases

5.3.1.2.2 Auto- Regressive Components (AR)

Examine the PACF in Figure 5.23 lead us to propose model AR (1), which define by Eq. 5.17.

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \emptyset_{\boldsymbol{1}}\boldsymbol{Y}_{\boldsymbol{t-1}} + \boldsymbol{a}_{\boldsymbol{t}} \quad \text{Eq. 5.17}$$

5.3.1.2.3 Moving Average Components (MA)

Examine the ACF in Figure 5.23 lead us to propose model MA (1), which define by Eq. 5.18

$$\boldsymbol{X}_{\boldsymbol{t}} = \boldsymbol{a}_{\boldsymbol{t}} - \boldsymbol{\theta}_{\boldsymbol{1}}\boldsymbol{a}_{\boldsymbol{t-1}} \quad \text{Eq. 5.18}$$

5.3.1.2.4 Mixed Model (ARIMA)

And if we look again at the two Figures 5.23, we propose a model ARIMA (1,1,1)(0,1,0) because p is 1, q is 1, and we have applied the first difference to the seasonal and non-seasonal series, which define by the equation shown in 5.19.

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \emptyset_{\boldsymbol{1}}\boldsymbol{Y}_{\boldsymbol{t-1}} - \boldsymbol{\theta}_{\boldsymbol{1}}\boldsymbol{a}_{\boldsymbol{t-1}} + \boldsymbol{a}_{\boldsymbol{t}} \quad \text{Eq. 5.19}$$

5.3.1.2.5 Winters' Multiplicative

It is extend for Holt's method to capture seasonality. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations; one for the level ℓ_t, one for the trend b_t, and one for the seasonal component s_t, with corresponding smoothing parameters α, $\beta*$ and γ as shown in Eq. 5.20.

$$\begin{aligned}
\hat{y}_{t+h|t} &= (\ell_t + hb_t)s_{t+h-m(k+1)} \\
\ell_t &= \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}
\end{aligned} \quad \text{Eq. 5.20}$$

5.3.1.3 Model Estimation (Fitting)

Insure that; the models coefficients significantly differ from zero.

5.3.1.3.1 Estimating the model AR (1) with the first series differencing

Estimating the model AR (1) with the first series differencing gives the model ARIMA (1,1,0)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.17.

Table 5.17 the results of the AR (1) estimation of Khartoum State monthly cases

		Estimate	SE	t	Sig.
Khartoum-Model_1 Khartoum	**Constant**	53.774	47.466	1.133	.267
	AR Lag 1	-.534	.158	-3.380	.002
	Difference	1			
	Seasonal Difference	1			

From Table 5.17 it is clear that AR coefficient is between -1 and 1, and *p* value of 0.002 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.18 form the equation shown in Eq. 5.21.

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \boldsymbol{a}_{\boldsymbol{t}} - \mathbf{0.534}\, \boldsymbol{Y}_{\boldsymbol{t-1}} \quad \text{Eq. 5.21}$$

5.3.1.3.2 Estimating the model MA (1) with the first series differencing

Estimating the model MA (1) with the first series differencing gives the model ARIMA (0,1,1)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.18.

Table 5.18 the results of the MA (1) estimation of Khartoum State monthly cases

		Estimate	SE	t	Sig.
Khartoum-Model_1 Khartoum	**Constant**	49.687	41.009	1.212	.236
	Difference	1			
	MA Lag 1	.465	.171	2.714	.011
	Seasonal Difference	1			

From Table 5.19 MA coefficient is between -1 and 1, and *p* value of 0.011 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.18 form the equation shown in Eq. 5.22.

$$\boldsymbol{X}_{\boldsymbol{t}} = \boldsymbol{a}_{\boldsymbol{t}} - \mathbf{0.465}\, \boldsymbol{a}_{\boldsymbol{t-1}} \quad \text{Eq. 5.22}$$

5.3.1.3.3 Estimating the model ARIMA (1,1,1)(0,1,0)

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.19.

Table 5.19 the results of the ARIMA (1,1,1)(0,1,0) estimation of Khartoum State monthly cases

			Estimate	SE	t	Sig.
Khartoum-Model_1	Khartoum	**Constant**	52.959	44.768	1.183	.247
		AR Lag 1	-.445	.315	-1.413	.169
		Difference	1			
		MA Lag 1	.127	.354	.359	.722
		Seasonal Difference	1			

From Table 5.19 the AR coefficient -0.445 is between -1 and 1, the MA coefficient 0.127 is between -1 and 1 and their *p* values 0.169 > 0.05 and 0.722 > 0.05 respectively indicate that the correlation isn't significantly different from zero.

5.3.1.3.4 Estimating Winters' Multiplicative model

The results of the winters' Multiplicative estimation are shown using the statistical software SPSS 17.0 in the Table 5.20.

Table 5.20 the results of the winter's Multiplicative model estimation of Khartoum State monthly cases

Model		Estimate	SE	t	Sig.
Khartoum-Model_1	**Alpha (Level)**	.101	.025	4.061	.000
	Gamma (Trend)	.926	.318	2.911	.006
	Delta (Season)	.229	.057	4.033	.000

From Table 5.20 it is clear that alpha, gamma and delta (0.101, 0.926 and 0.229 respectively) ≤ 1, and the *p* values (0.000, 0.006 and 0.000 respectively) < 0.05 which indicates that the correlation is significantly different from zero.

5.3.1.4 Model Diagnosis

Since the ARIMA (1,1,1)(0,1,0) isn't significantly different from zero we'll not make the test for it.

5.3.1.4.1 ARIMA (1,1,0)(0,1,0)

Table 5.21 shows the result of Ljung-Box test for ARIMA (1,1,0)(0,1,0) using the software SPSS 17.0.

Table 5.21 the result of Ljung-Box test of ARIMA (1,1,0)(0,1,0) for Khartoum State monthly cases

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
Khartoum-Model_1	20.399	17	.254

The Ljung-Box test as shown in the Table 5.21 indicates that there is no self-correlation between the errors (sig. 0.254 > 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.30.

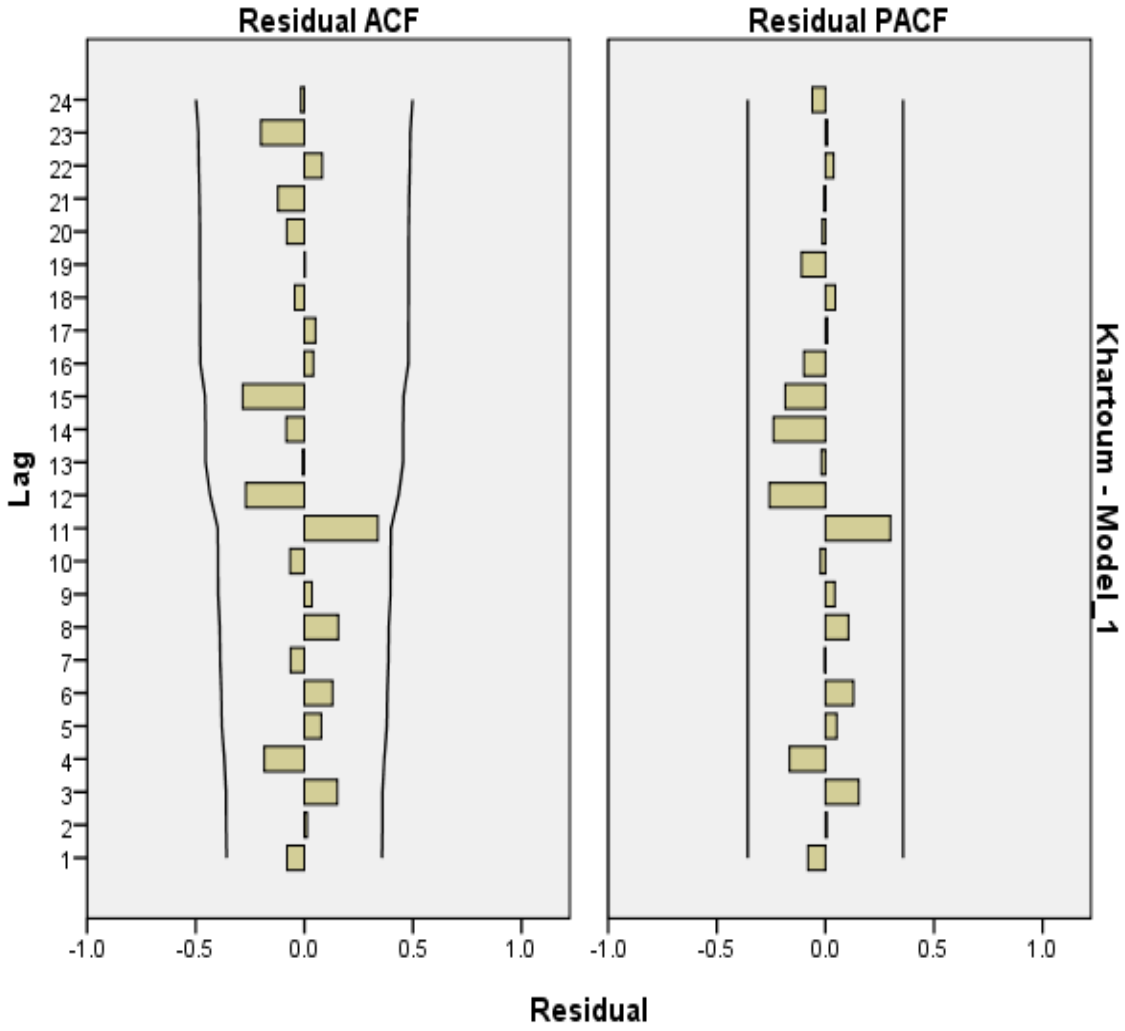

Figure 5.30 the ACF and PACF residuals of ARIMA (1,1,0)(0,1,0) for Khartoum State monthly cases

5.3.1.4.2 ARIMA (0,1,1)(0,1,0)

Table 5.22 shows the result of Ljung-Box test for ARIMA (0,1,1)(0,1,0) using the software SPSS 17.0.

Table 5.22 the result of Ljung-Box test of ARIMA (0,1,1)(0,1,0) for Khartoum State monthly cases

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
Khartoum-Model_1	22.934	17	.151

The Ljung-Box test as shown in the Table 5.22 indicates that there is no self-correlation between the errors (sig. 0.151 < 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.31.

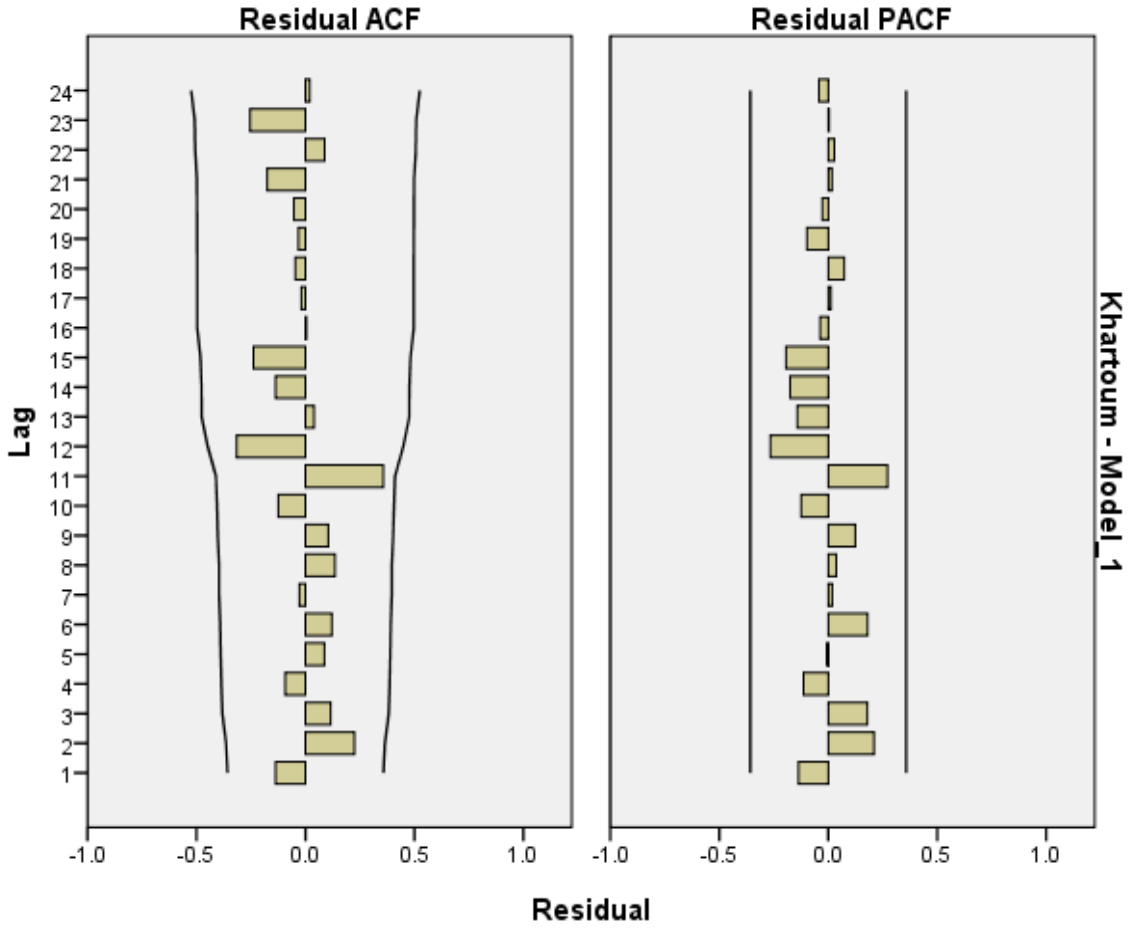

Figure 5.31 the ACF and PACF residuals of ARIMA (0,1,1)(0,1,0) for Khartoum State monthly cases

5.3.1.4.3 Winters' Multiplicative

Table 5.23 shows the result of Ljung-Box test for winters' Multiplicative using the software SPSS 17.0.

Table 5.23 the result of Ljung-Box test of winters' Multiplicative for Khartoum State monthly cases

Model	Ljung-Box Q(18)		
	Statistics	DF	Sig.
Khartoum-Model_1	20.805	15	.143

The Ljung-Box test as shown in the Table 5.23 indicates that there is no self-correlation between the errors (sig. 0.143 < 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.32.

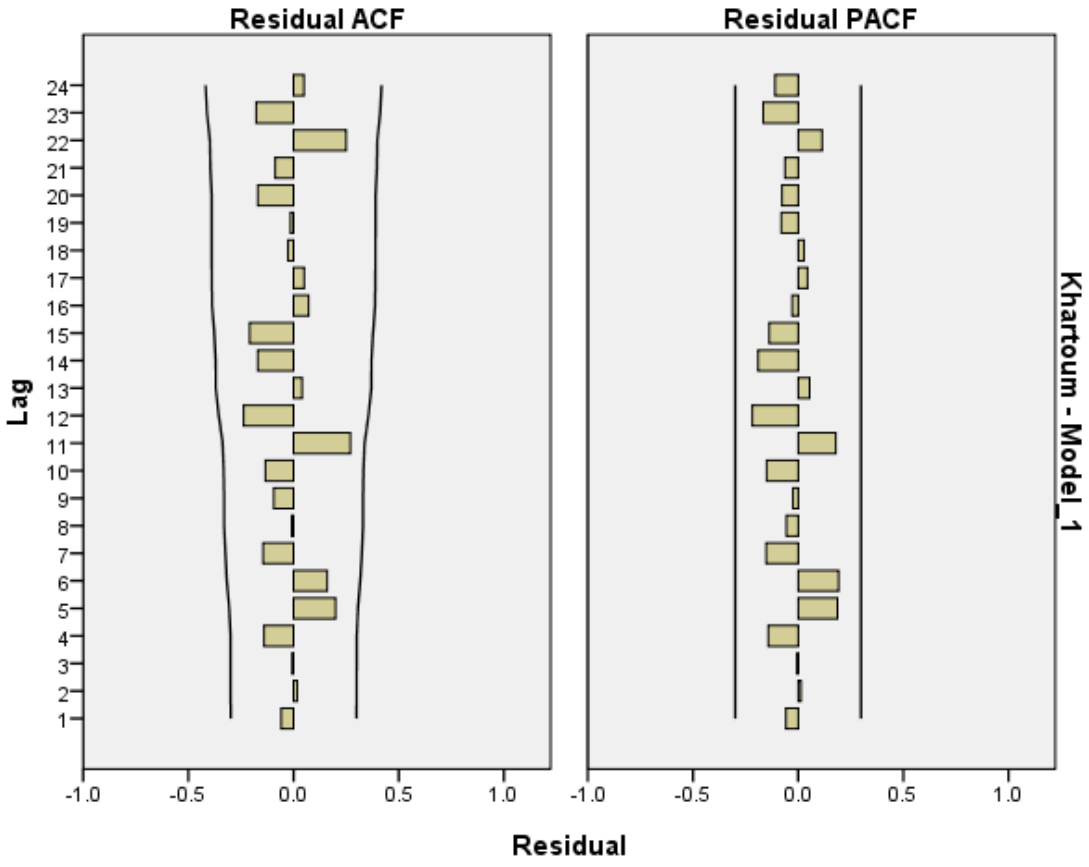

Figure 5.32 the ACF and PACF residuals of winters' Multiplicative for Khartoum State monthly cases

5.3.2 Al-Gadaref

5.3.2.1 Time Series Stability

The first step in the analysis is to plot the seasonal time series; Figure 5.33 shows monthly cases of the Malaria in Al-Gadaref State.

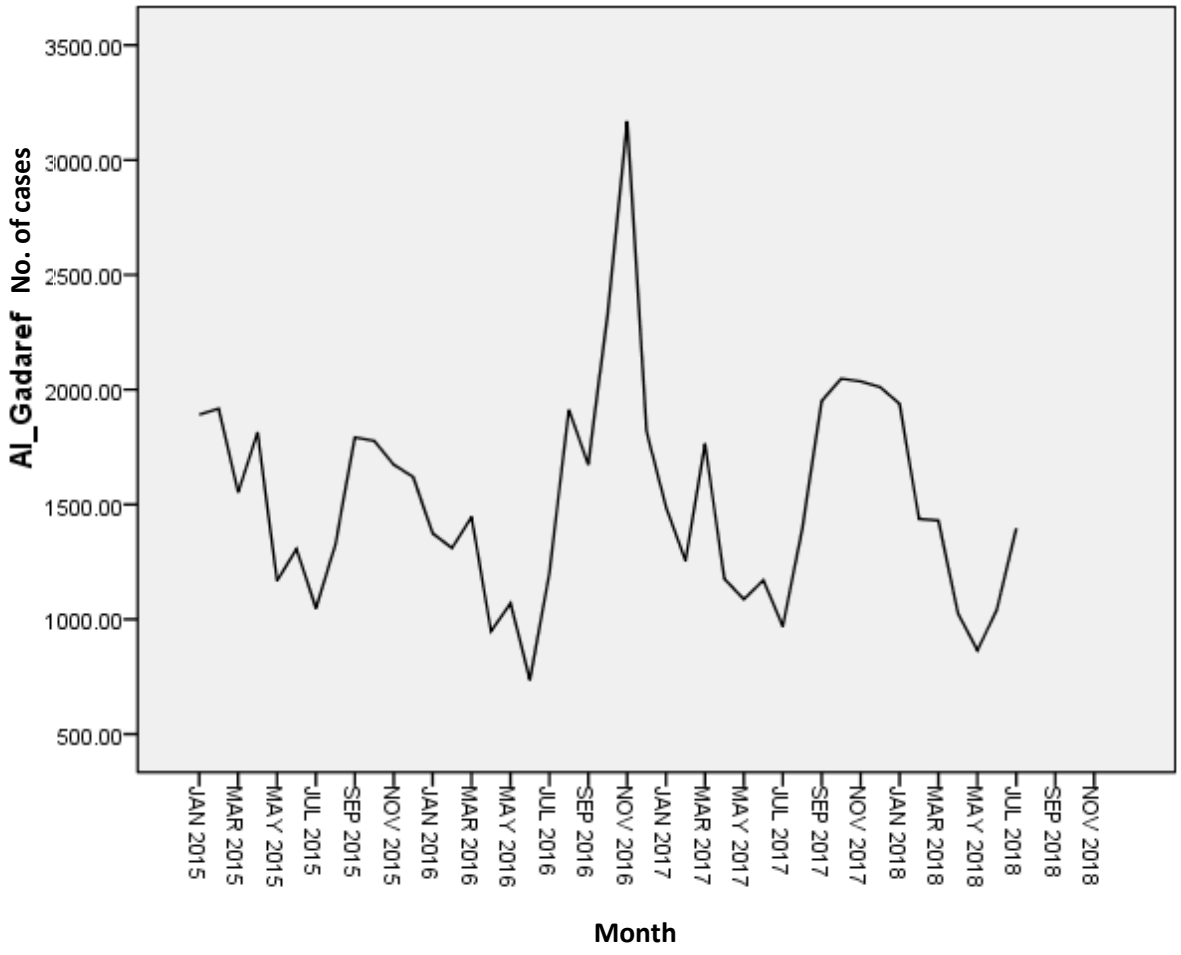

Figure 5.33 Monthly cases of the Malaria in Al-Gadaref State

It is obvious that there is a shift in both the trend and the dispersion over time for the series.

To see if the process is stationary after linear trend is removed, the first difference of the non-seasonal part plotted against months, as seen in Figure 5.34.

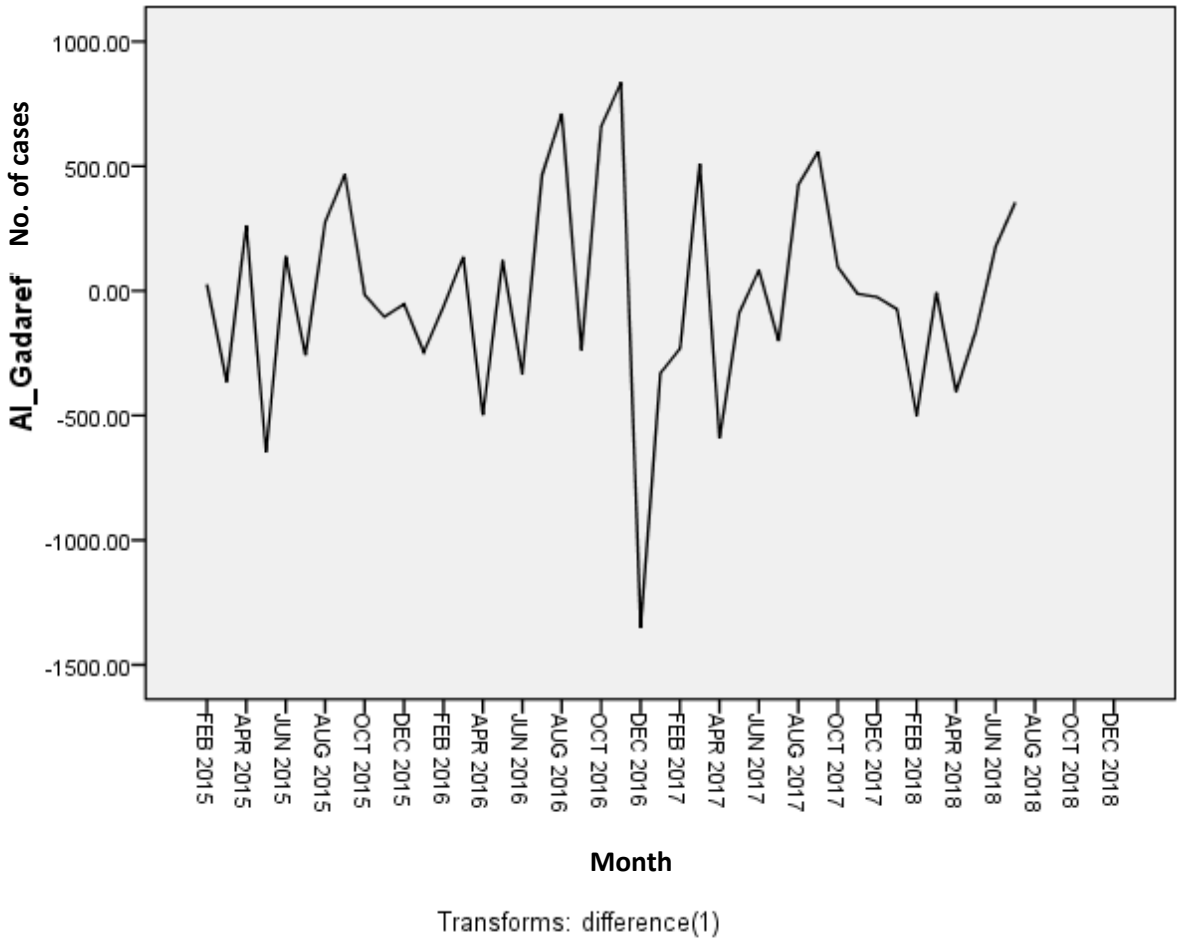

Figure 5.34 the first difference of Al-Gadaref State monthly non-seasonal part

It is clear that the series is still has the shift in the mean in seasonal (it is clear in every June in the last figure) part, so first differencing appear necessary for seasonal part which is shown in Figure 5.35.

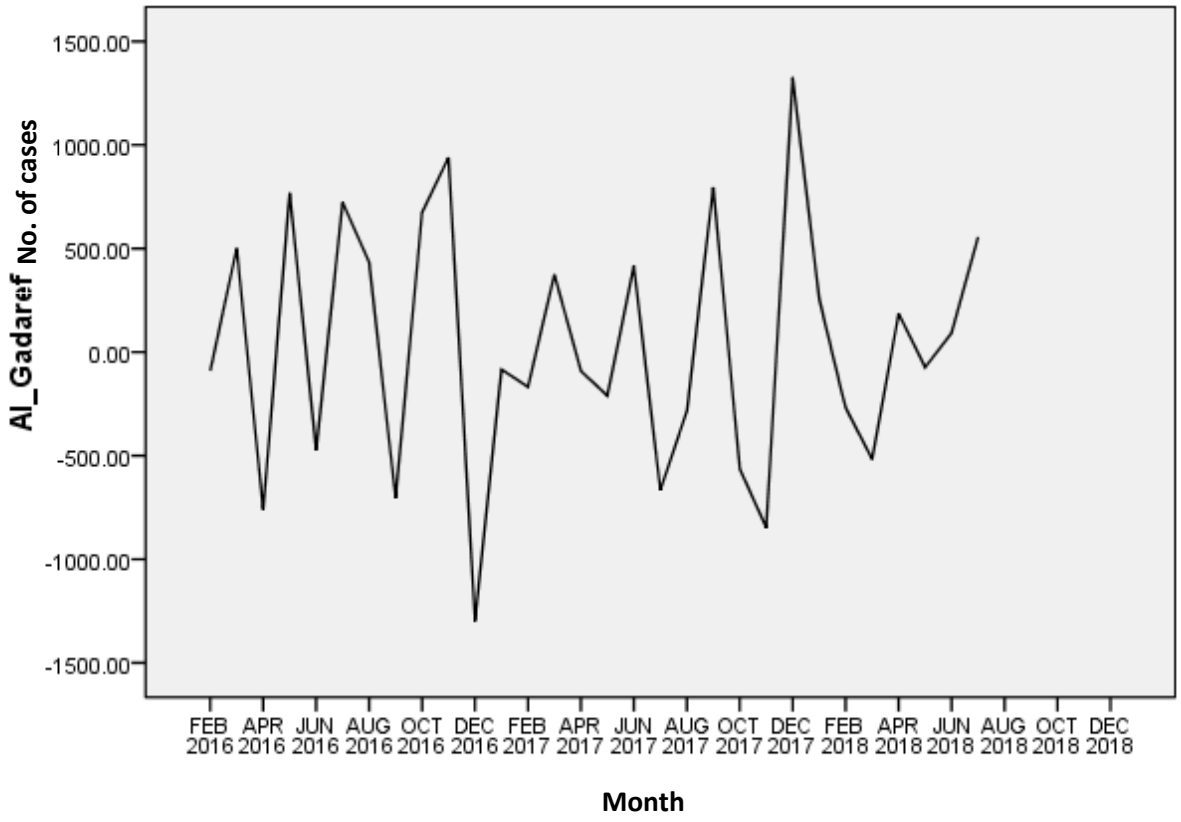

Figure 5.35 the first difference of Al-Gadaref State monthly seasonal and non-seasonal parts

The series now appears stationary with respect to central tendency, however, the variability seems to be increasing and decreasing over time. As we have mentioned earlier, the changing in variability can be made stationary by logarithmic transformation. The transformed difference is plotted in Figure 5.36.

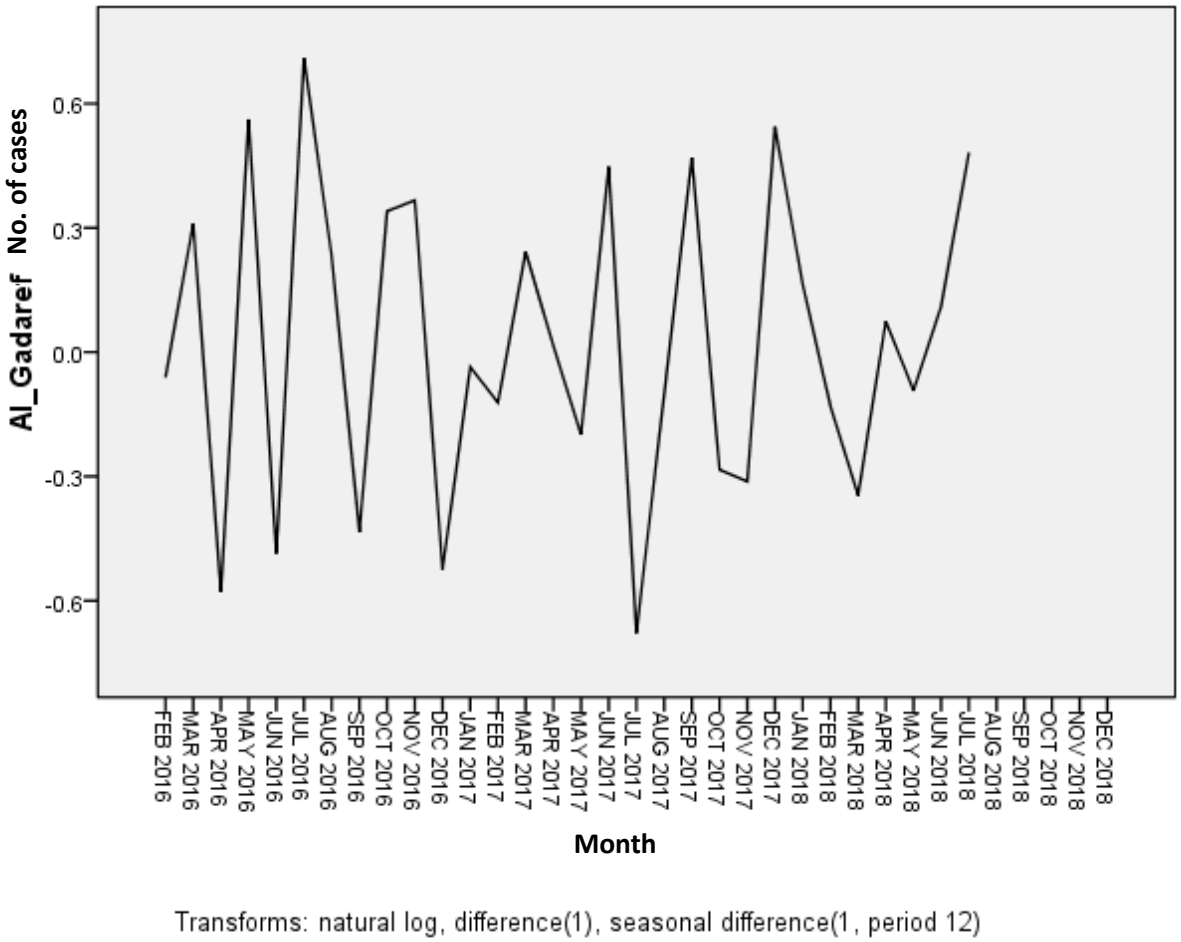

Figure 5.36 the natural logarithm and first difference of Al-Gadaref State monthly Cases

5.3.2.2 Model Identification

Determine the components of the models (AR, MA, and ARIMA).

5.3.2.2.1 ACFs and PACFs

Examine the ACF and PACF of the logarithmic, first difference series as shown in Figure 5.37 and Figure 5.38 respectively.

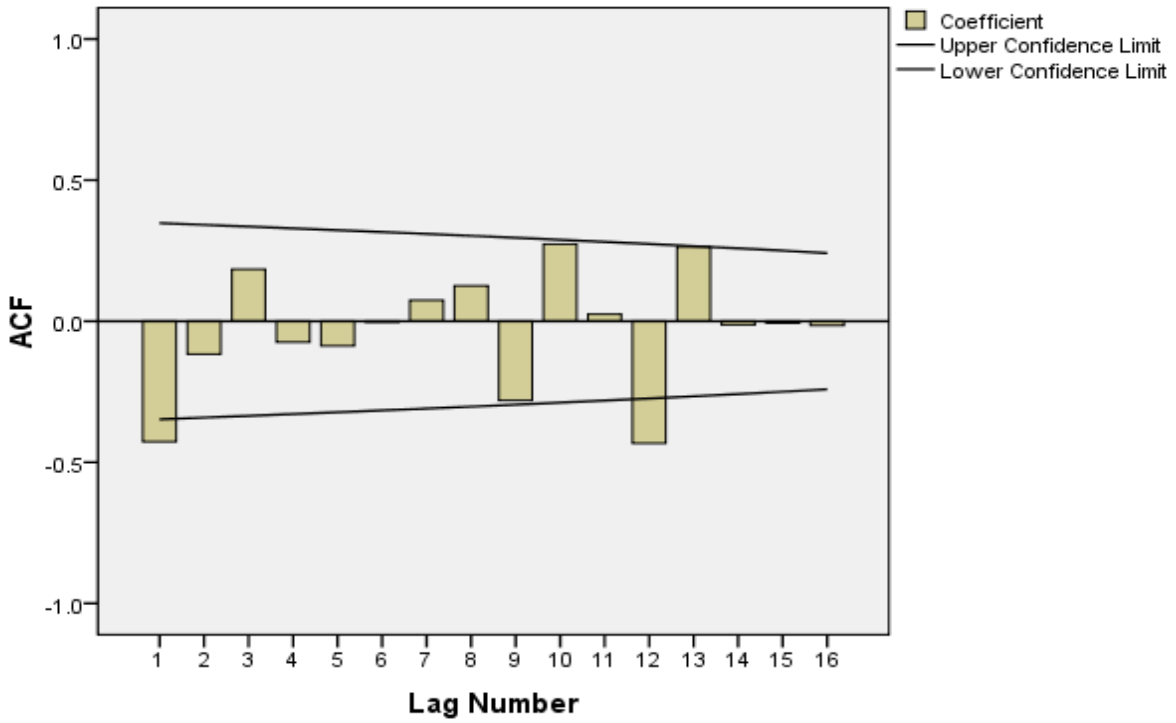

Figure 5.37 the ACF of the natural logarithm and first difference of Al-Gadaref State monthly cases

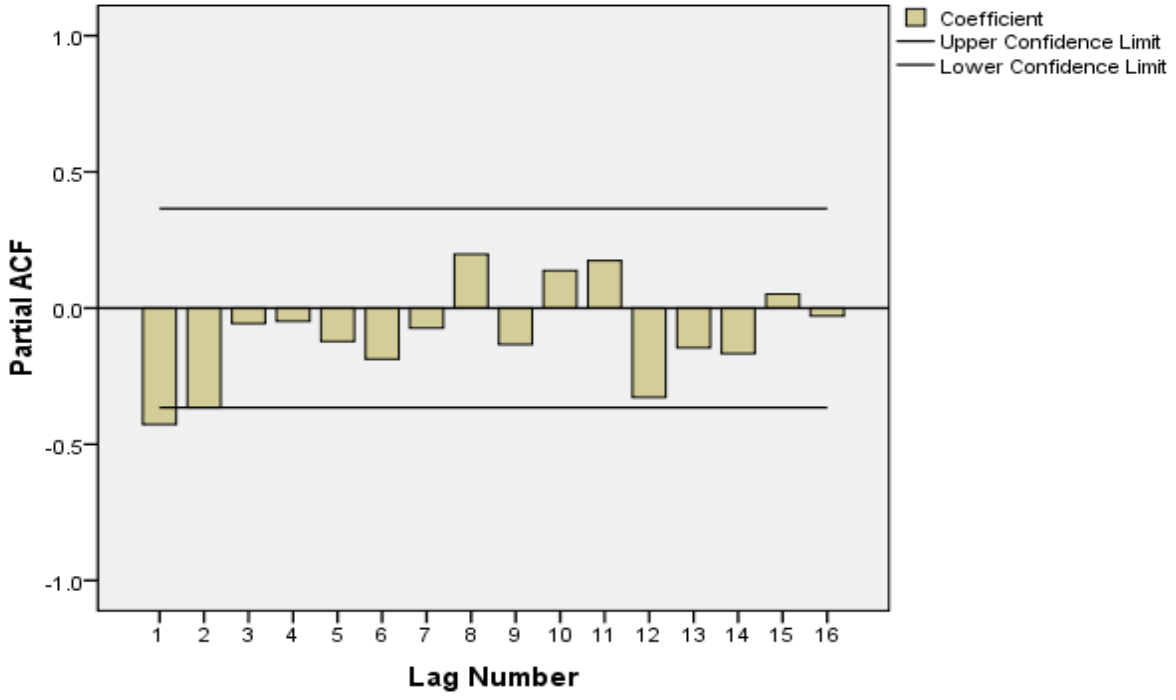

Figure 5.38 the PACF of the natural logarithm and first difference of Al-Gadaref State monthly cases

5.3.2.2.2 Auto- Regressive Components (AR)

Examine the PACF in Figure 5.23 lead us to propose model AR (1), which define by Eq. 5.23

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \emptyset_{\mathbf{1}}\boldsymbol{Y}_{\boldsymbol{t-1}} + \boldsymbol{a}_{\boldsymbol{t}} \qquad \text{Eq. 5.23}$$

5.3.2.2.3 Moving Average Components (MA)

Examine the ACF in Figure 5.23 lead us to propose model MA (1), which define by Eq. 5.16,

$$\boldsymbol{X}_{\boldsymbol{t}} = \boldsymbol{a}_{\boldsymbol{t}} - \boldsymbol{\theta}_{\mathbf{1}}\boldsymbol{a}_{\boldsymbol{t-1}} \qquad \text{Eq. 5.24}$$

5.3.2.2.4 Mixed Model (ARIMA)

And if we look again at the two Figures 5.23, we propose a model ARIMA (1,1,1)(0,1,0) because p is 1, q is 1, and we have applied the first difference to the seasonal and non-seasonal series, which define by the equation shown in 5.17.

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{t}} = \emptyset_{\mathbf{1}}\boldsymbol{Y}_{\boldsymbol{t-1}} - \boldsymbol{\theta}_{\mathbf{1}}\boldsymbol{a}_{\boldsymbol{t-1}} + \boldsymbol{a}_{\boldsymbol{t}} \qquad \text{Eq. 5.24}$$

5.3.2.2.5 Simple Seasonal Exponential Smoothing (SSES)

It is exponential smoothing with level and seasonality as shown in equations 5.25.

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)L_{t-1}$$
$$S_t = \delta(Y_t - L_t) + (1-\delta)S_{t-s}$$
$$Y_t = L_t + S_t \qquad \text{Eq. 5.25}$$

5.3.2.3 Model Estimation (Fitting)

Insure that; there is no self-correlation between the errors.

5.3.2.3.1 Estimating the model AR (1) with the first series differencing

Estimating the model AR (1) with the first series differencing gives the model ARIMA (1,1,0)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.24.

Table 5.24 the results of the AR (1) estimation of Al-Gadaref State monthly Cases

			Estimate	SE	t	Sig.
Al_Gadaref-Model_1	Al_Gadaref	**Constant**	.019	.044	.427	.673
		AR **Lag 1**	-.436	.174	-2.497	.019
		Difference	1			
		Seasonal Difference	1			

From Table 5.24 it is clear that AR coefficient is between -1 and 1, and *p* value of 0.019 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.18 form the equation shown in Eq. 5.26.

$$\widehat{Y}_t = a_t - 0.436\, Y_{t-1} \quad \text{Eq. 5.26}$$

5.3.2.3.2 Estimating the model MA (1) with the first series differencing

Estimating the model MA (1) with the first series differencing gives the model ARIMA (0,1,1)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.25.

Table 5.25 the results of the MA (1) estimation of Al-Gadaref State monthly Cases

		Estimate	SE	t	Sig.
Al_Gadaref-Model_1	Al_Gadaref **Constant**	.011	.017	.645	.524
	Difference	1			
	MA Lag 1	.724	.138	5.238	.000
	Seasonal Difference	1			

From Table 5.25 MA coefficient is between -1 and 1, and *p* value of 0.000 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.19 form the equation shown in Eq. 5.27.

$$X_t = a_t - 0.724\, a_{t-1} \quad \text{Eq. 5.27}$$

5.3.2.3.3 Estimating the model ARIMA (1,1,1)(0,1,0)

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.26.

Table 5.26 the results of the ARIMA (1,1,1)(0,1,0) estimation of Al-Gadaref State monthly cases

	Estimate	SE	t	Sig.

Al_Gadaref-Model_1	Al_Gadaref	**Constant**	.011	.017	.635	.531
		AR **Lag 1**	.023	.278	.084	.934
		Difference	1			
		MA **Lag 1**	.739	.192	3.844	.001
		Seasonal Difference	1			

From Table 5.26 the AR coefficient -0.023 is between -1 and 1, the MA coefficient 0.739 is between -1 and 1 and their *p* values 0.934 > 0.05 and 0.001 < 0.05 respectively indicate that the correlation isn't significantly different from zero.

5.3.2.3.4 Estimating SSES model

The results of the SSES estimation are shown using the statistical software SPSS 17.0 in the Table 5.27.

Table 5.27 the results of the SSES model estimation of Khartoum State monthly cases

Model		**Estimate**	**SE**	**t**	**Sig.**
Al_Gadaref-Model_1	**Alpha (Level)**	.300	.113	2.656	.011
	Delta (Season)	2.428E-6	.267	9.094E-6	1.000

From Table 5.27 it is clear that alpha 0.300 ≤ 1 and the *p* value 0.011 < 0.05 but delta > 1 and the *p* value =1 which indicates that the correlation isn't significantly different from zero.

5.3.2.4 Model Diagnosis

Since the ARIMA (1,1,1)(0,1,0) isn't significantly different from zero we'll not make the test for it.

5.3.2.4.1 ARIMA (1,1,0)(0,1,0)

Table 5.28 shows the result of Ljung-Box test for ARIMA (1,1,0)(0,1,0) using the software SPSS 17.0.

Table 5.28 the result of Ljung-Box test of ARIMA (1,1,0)(0,1,0) for Al-Gadaref State monthly cases

Ljung-Box Q(18)		
Statistics	**DF**	**Sig.**

Ljung-Box Q(18)		
Statistics	**DF**	**Sig.**
26.055	17	.073

The Ljung-Box test as shown in the Table 5.28 indicates that there is no self-correlation between the errors (sig. 0.73 > 0.05), and this is confirmed by the auto-correlations and partial auto-correlations of the residuals as in Figure 5.39.

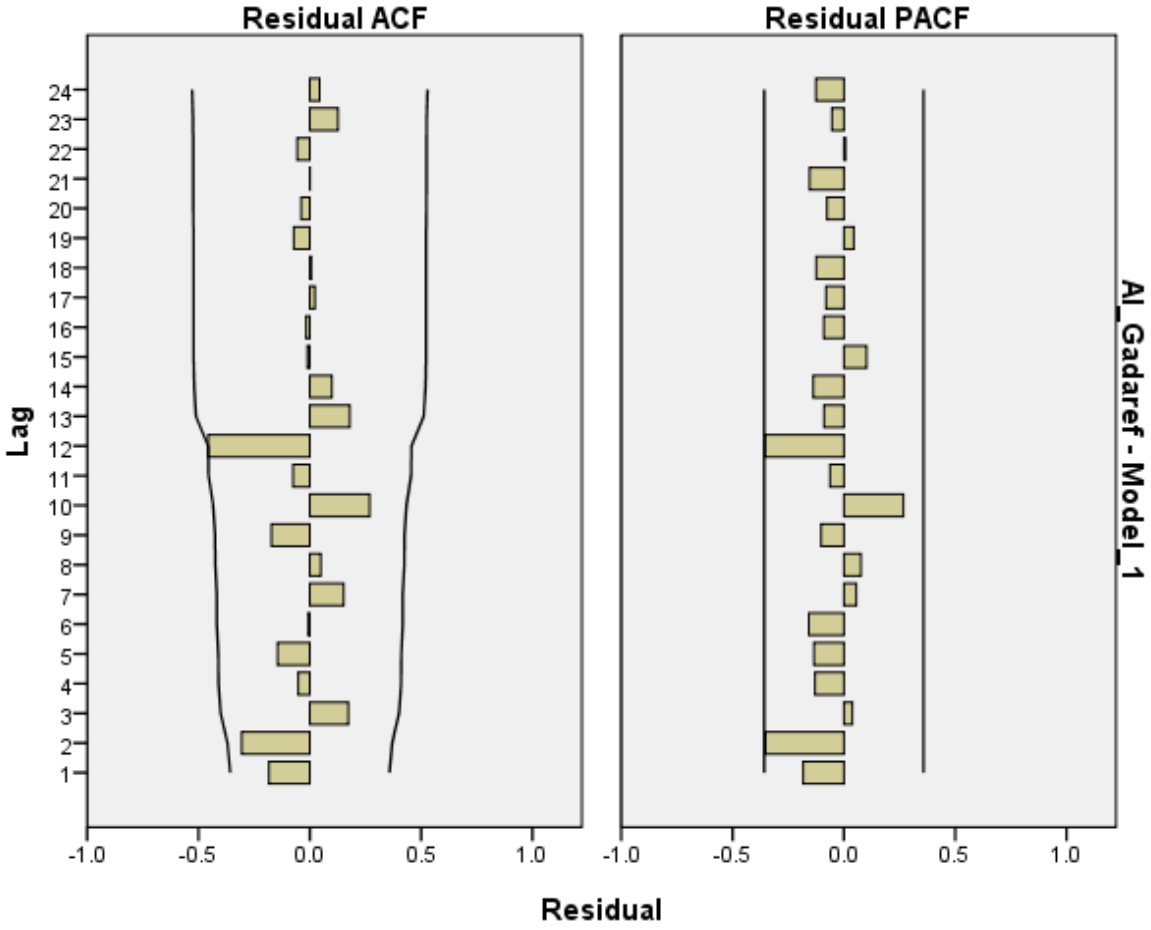

Figure 5.39 the ACF and PACF residuals of ARIMA (1,1,0)(0,1,0) for Al-Gadaref State monthly cases

5.3.2.4.2 ARIMA (0,1,1)(0,1,0)

Table 5.29 shows the result of Ljung-Box test for ARIMA (0,1,1)(0,1,0) using the software SPSS 17.0.

Table 5.29 the result of Ljung-Box test of ARIMA (0,1,1)(0,1,0) for Khartoum State monthly cases

Ljung-Box Q(18)		
Statistics	**DF**	**Sig.**
15.858	17	.534

The Ljung-Box test as shown in the Table 5.22 indicates that there is no self-correlation between the errors (sig. 0.534 > 0.05).

5.3.3 Sennar

5.3.3.1 Time Series Stability

The first step in the analysis is to plot the seasonal time series; Figure 5.40 shows monthly cases of the Malaria in Sennar State.

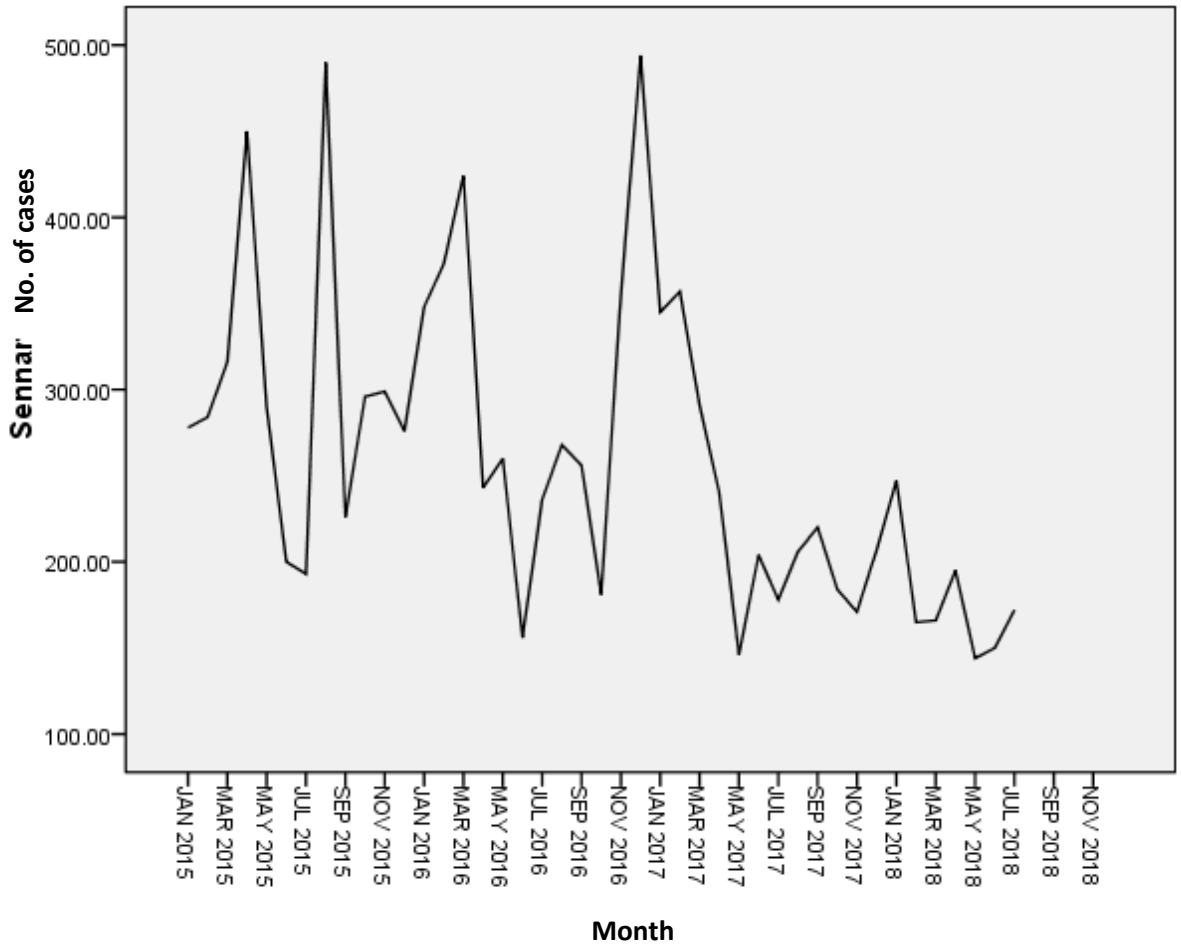

Figure 5.40 Monthly cases of the Malaria in Sennar State

It is clear that the series is still has the shift in the mean in seasonal (it is clear in every June in the last figure) part, so first differencing appear necessary for seasonal part which is shown in Figure 5.41.

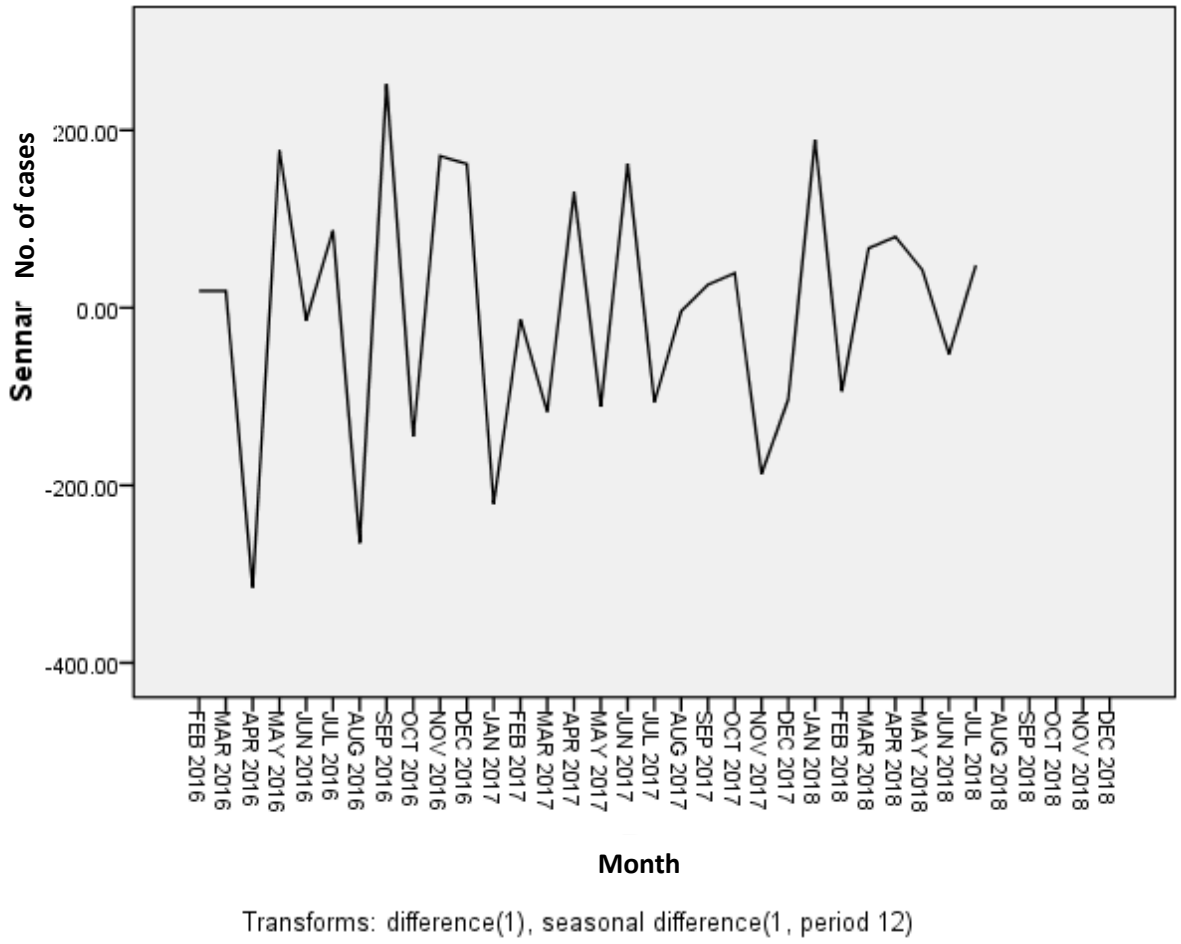

Figure 5.41 the first difference of Sennar State monthly seasonal and non-seasonal parts

The series now appears stationary with respect to central tendency, however, the variability seems to be increasing and decreasing over time. As we have mentioned earlier, the changing in variability can be made stationary by logarithmic transformation. The transformed difference is plotted in Figure 5.41.

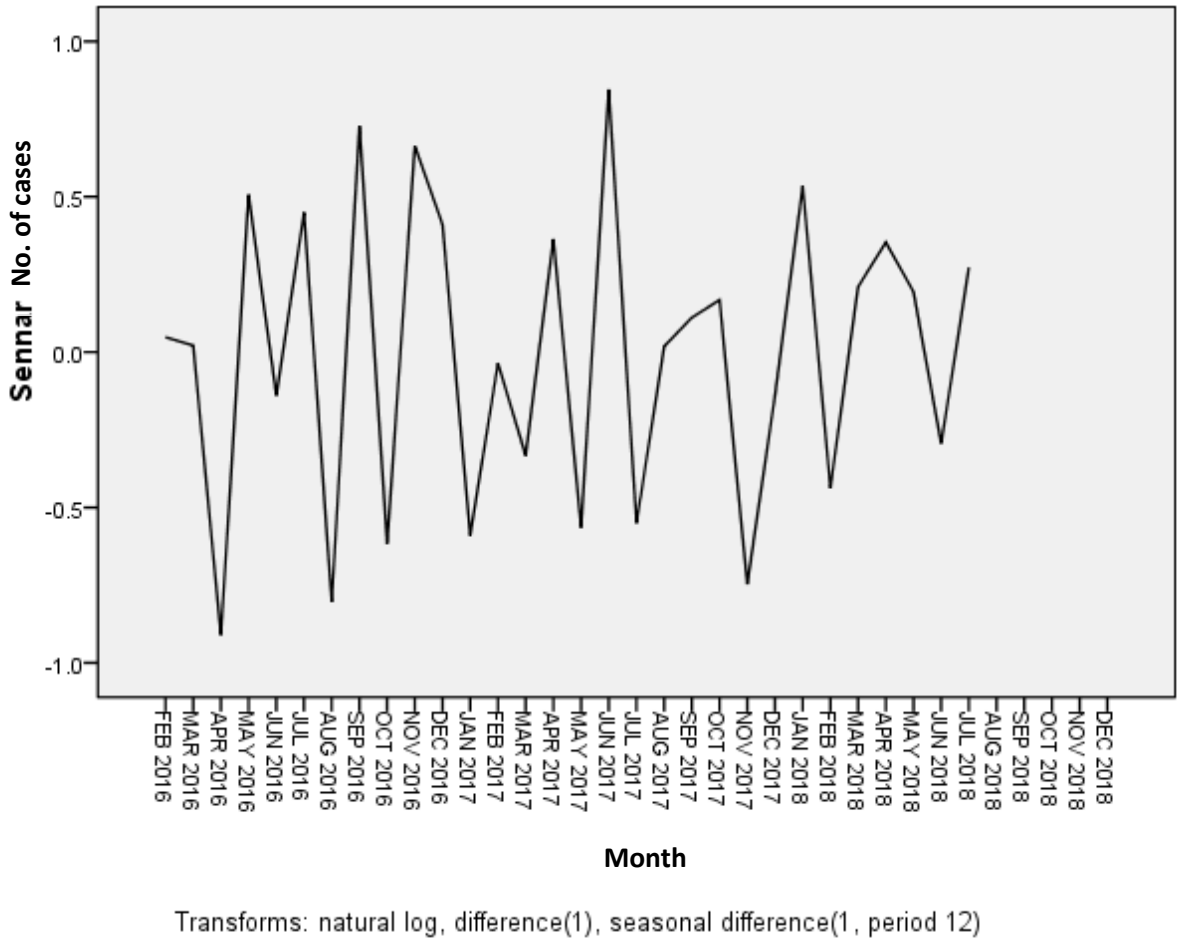

Figure 5.42 the natural log and first difference of Sennar State monthly non-seasonal part

5.3.3.2 Model Identification

Determine the models (AR, MA, and ARIMA) components.

5.3.3.2.1 ACFs and PACFs

Examine the ACF and PACF of the logarithmic, first difference series as shown in Figure 5.43 and Figure 5.44 respectively.

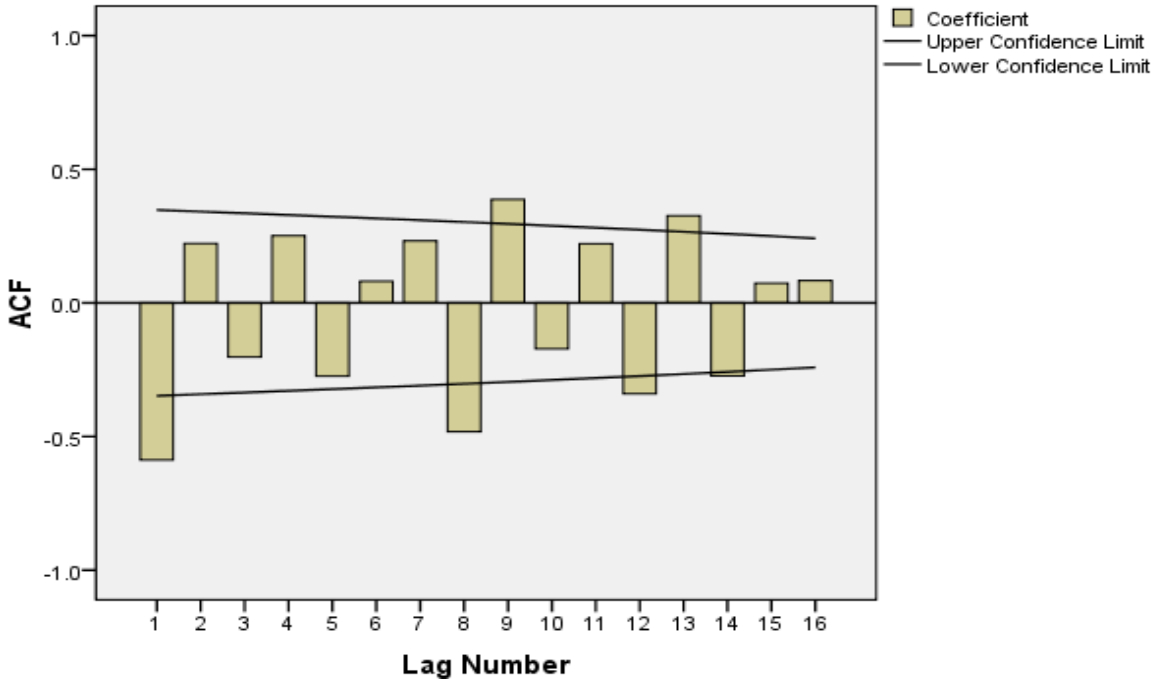

Figure 5.43 the ACF of the natural logarithm and first difference of Sennar State monthly cases

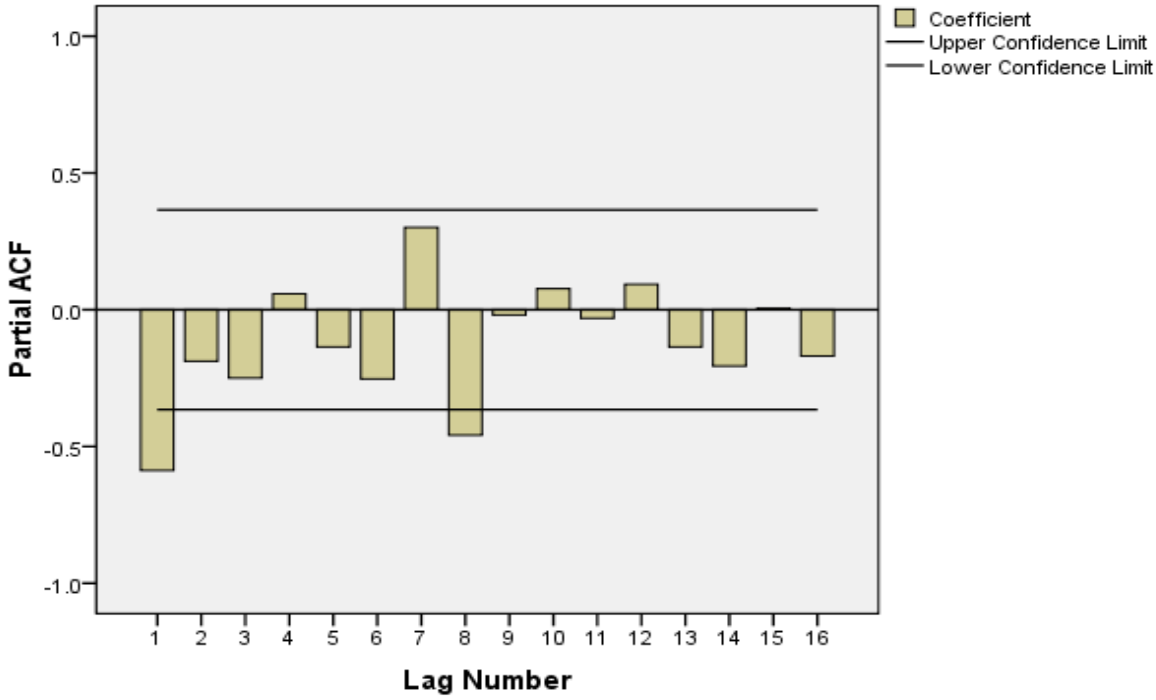

Figure 5.44 the ACF and PACF of the natural logarithm and first difference of Sennar State monthly cases

5.3.3.2.2 Auto- Regressive Components (AR)

Examine the PACF in Figure 5.23 lead us to propose model AR (1), which define by Eq. 5.28.

$$\widehat{Y}_t = \emptyset_1 Y_{t-1} + a_t \quad \text{Eq. 5.28}$$

5.3.3.2.3 Moving Average Components (MA)

Examine the ACF in Figure 5.23 lead us to propose model MA (1), which define by Eq. 5.29.

$$X_t = a_t - \theta_1 a_{t-1} \quad \text{Eq. 5.29}$$

5.3.3.2.4 Mixed Model (ARIMA)

And if we look again at the two Figures 5.23, we propose a model ARIMA (1,1,1)(0,1,0) because p is 3, q is 1, and we have applied the first difference to the seasonal and non-seasonal series, which define by the equation shown in 5.30.

$$\widehat{Y}_t = \emptyset_1 Y_{t-1} - \theta_1 a_{t-1} + a_t \quad \text{Eq. 5.30}$$

5.3.3.2.5 Winters' Multiplicative

5.3.3.3 Model Estimation (Fitting)

Insure that; there no self-correlation between the errors.

5.3.3.3.1 Estimating the model AR (1) with the first series differencing

Estimating the model AR (1) with the first series differencing gives the model ARIMA (1,1,0)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.30.

Table 5.30 the results of the AR (1) estimation of Sennar State monthly Cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	-.013	.047	-.274	.786
		AR **Lag 1**	-.576	.153	-3.758	.001
		Difference	1			
		Seasonal Difference	1			

From Table 5.30 it is clear that AR coefficient is between -1 and 1, and *p* value of 0.001 < 0.05 indicates that the correlation is significantly different from zero. Thus, the coefficients in the Table 5.18 form the equation shown in Eq. 5.31.

$$\widehat{Y}_t = a_t - 0.576\, Y_{t-1} \quad \text{Eq. 5.31}$$

5.3.3.3.2 Estimating the model MA (1) with the first series differencing

Estimating the model MA (1) with the first series differencing gives the model ARIMA (0,1,1)(0,1,0) where the results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.31.

Table 5.31 the results of the MA (1) estimation of Sennar State monthly Cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	-.015	.008	-1.960	.060
		Difference	1			
		MA Lag 1	.999	12.611	.079	.937
		Seasonal Difference	1			

From Table 5.31 MA coefficient is between -1 and 1, and *p* value of 0.937 > 0.05 indicates that the correlation isn't significantly different from zero.

5.3.3.3.3 Estimating the model ARIMA (1,1,1)(0,1,0)

The results of the estimation are shown using the statistical software SPSS 17.0 in the Table 5.32.

Table 5.32 the results of ARIMA (1,1,1)(0,1,0) estimation of Sennar monthly cases

			Estimate	SE	t	Sig.
Sennar-Model_1	Sennar	**Constant**	-.015	.008	-1.953	.061
		AR **Lag 1**	.004	.217	.017	.986
		Difference	1			
		MA **Lag 1**	.993	1.656	.599	.554
		Seasonal Difference	1			

From Table 5.32 the AR coefficient 0.004 is between -1 and 1, the MA coefficient 0.993 is between -1 and 1 and their *p* values (0.986 and 0.554 respectively) > 0.05 indicate that the correlation isn't significantly different from zero.

5.3.3.3.4 Estimating SSES model

The results of the SSES estimation are shown using the statistical software SPSS 17.0 in the Table 5.33.

Table 5.33 the results of the SSES model estimation of Khartoum State monthly cases

Model		Estimate	SE	t	Sig.
Sennar-Model_1	**Alpha (Level)**	.300	.123	2.435	.019
	Delta (Season)	2.645E-6	.260	1.019E-5	1.000

From Table 5.33 it is clear that alpha $0.300 \leq 1$ and the *p* value $0.019 < 0.05$ but delta > 1 and the *p* value =1 which indicates that the correlation isn't significantly different from zero.

5.3.3.4 Model Diagnosis

Since the ARIMA (1,1,1)(0,1,0) isn't significantly different from zero we'll not make the test for it.

5.3.3.4.1 ARIMA (1,1,0)(0,1,0)

Table 5.34 shows the result of Ljung-Box test for ARIMA (1,1,0)(0,1,0) using the software SPSS 17.0.

Table 5.34 the result of Ljung-Box test of ARIMA (1,1,0)(0,1,0) for Al-Gadaref State monthly cases

	Ljung-Box Q(18)		
Model	**Statistics**	**DF**	**Sig.**
Sennar-Model_1	26.894	17	.060

The Ljung-Box test as shown in the Table 5.34 indicates that there is no self-correlation between the errors (sig. 0.60 > 0.05).

CHAPTER SIX

RESULTS & DISCUSSION

6.1 Results

6.1.1 Non Seasonal Model - Yearly

6.1.1.1 Khartoum

6.1.1.1.1 Comparison of the predictability of the four models

Table 6.1 the result of the comparison between the four models for Khartoum yearly cases

Model	**ARIMA (1,2,0)**	**ARIMA (0,2,1)**	**ARIMA (1,2,1)**	**Simple Exponential Smoothing α=.856**
MAPE(%)	42.935	43.981	41.088	42.934

6.1.1.1.2 Forecasting using the proposed model

Table 6.2 the result of the forecasting SES model for the years (2017-2020) for Khartoum yearly cases

Model		**2017**	**2018**	**2019**	**2020**
Khartoum-Model_1	**Forecast**	4162	4162	4162	4162
	UCL	16625	20564	23725	26442
	LCL	-8301	-12240	-15402	-18119

6.1.1.2 Al-Gadarif

6.1.1.2.1 Comparison of the predictability of the two models

Table 6.3 the result of the comparison between the ARIMA (0,2,0) and Brown Exponential smoothing for Al-Gadaref yearly cases

	ARIMA (0,2,0)	**Exponential Smoothing *α=0.875***
MAPE	18.019	16.400

6.1.1.2.2 Forecasting using the proposed model

Table 6.4 the result of the forecasting using BES for the years (2017-2020) for Al-Gadaref yearly cases

Model		**2017**	**2018**	**2019**	**2020**
Al_Gadaref-Model_1	**Forecast**	8561.04	10203.60	11846.17	13488.73
	UCL	12076.72	17292.00	23184.27	29668.88
	LCL	5045.35	3115.20	508.06	-2691.42

6.1.1.3 Sennar

6.1.1.3.1 Comparison of the predictability of the four models

Table 6.5 the result of the comparison between the four models for Sennar yearly cases

Model	ARIMA (1,1,0)	ARIMA (0,1,1)	ARIMA (1,1,1)	Holt Exponential Smoothing α=.856
MAPE(%)	30.688	26.097	27.035	22.684

6.1.1.3.2 Forecasting using the proposed model

Table 6.6 the result of the forecasting using ARIMA (0,1,1) model for the years (2017-2020) for Sennar yearly cases

Model		2017	2018	2019	2020
Sennar-Model_1	**Forecast**	21921.51	22812.23	23739.14	24703.71
	UCL	38653.26	40612.47	42662.48	44807.56
	LCL	11429.79	11743.78	12068.80	12405.15

6.1.1.4 Display the yearly forecasting results using ArcMap

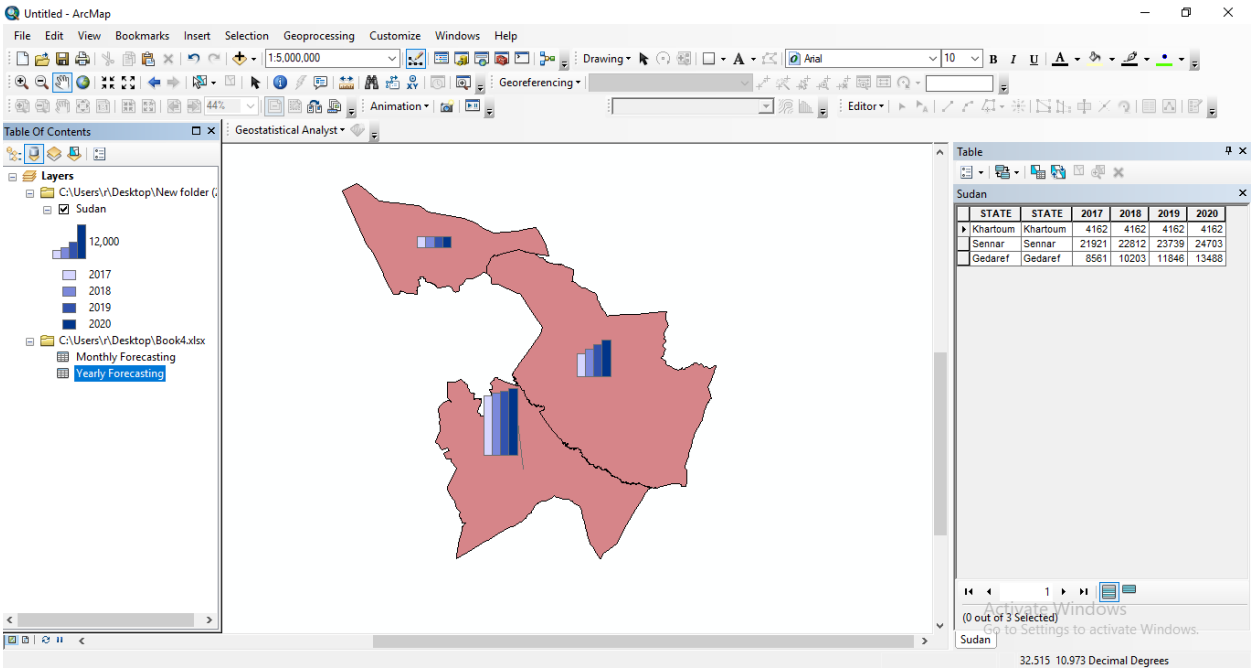

Figure 6.1 The yearly (1996-2016) forecasting results using ArcMap

6.1.2 Seasonal Model - Monthly

6.1.2.1 Khartoum

6.1.2.1.1 Comparison of the predictability of the four models

Table 6.7 the result of the comparison between the four models for Khartoum monthly cases

Model	ARIMA (1,1,0)(0,1,0)	ARIMA (0,1,1)(0,1,0)	ARIMA (0,1,1)(0,1,0)	Winters' Multiplicative
MAPE %	26.462	27.194	26.687	15.082

6.1.2.1.2 Forecasting using the proposed model

Table 6.8 the result of the forecasting winters' Multiplicative model for the years (2015-2018) models for Khartoum monthly cases

Model		Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018
Khartoum-Model_1	**Forecast**	960.67	1188.29	987.00	1090.09	1367.49
	UCL	1426.84	1667.93	1479.64	1624.79	2012.81
	LCL	494.50	708.64	494.37	555.39	722.17

6.1.2.2 Al-Gadaref

6.1.2.2.1 Comparison of the predictability of the four models

Table 6.9 the result of the comparison between the four models for Al-Gadaref monthly cases

Model	ARIMA (1,1,0)(0,1,0)	ARIMA (0,1,1)(0,1,0)	ARIMA (1,1,1)(0,1,0)	SSES
MPAE %	32.627	27.952	27.985	14.473

6.1.2.2.2 Forecasting using the proposed model

Table 6.10 the result of the forecasting Simple Exponential Smoothing model for the years (2015-2018) models for Al-Gadaref monthly cases

Model		Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018
Al_Gadaref-Model_1	**Forecast**	1563.21	2221.86	2366.20	2387.73	2393.88
	UCL	2826.15	4099.05	4450.85	4575.86	4670.81
	LCL	783.95	1083.83	1123.66	1104.66	1079.71

6.1.2.3 Sennar

6.1.2.3.1 Comparison of the predictability of the four models

Table 6.11 the result of the comparison between the four models for Sennar monthly cases

Model	ARIMA (1,1,0)(0,1,0)	ARIMA (0,1,1)(0,1,0)	ARIMA (1,1,1)(0,1,0)	SSES
MPAE %	36.637	33.090	33.139	20.233

6.1.2.3.2 Forecasting using the proposed model

Table 6.12 the result of the forecasting ARIMA (1,1,0)(0,1,0) model for the years (2015-2018) models for Sennar monthly cases

Model		Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018
Sennar-Model_1	Forecast	180.48	212.27	173.67	168.68	205.27
	UCL	377.70	469.85	445.91	464.02	615.83
	LCL	73.53	79.45	51.15	44.07	45.73

6.1.2.4 Display the monthly forecasting results using ArcMap

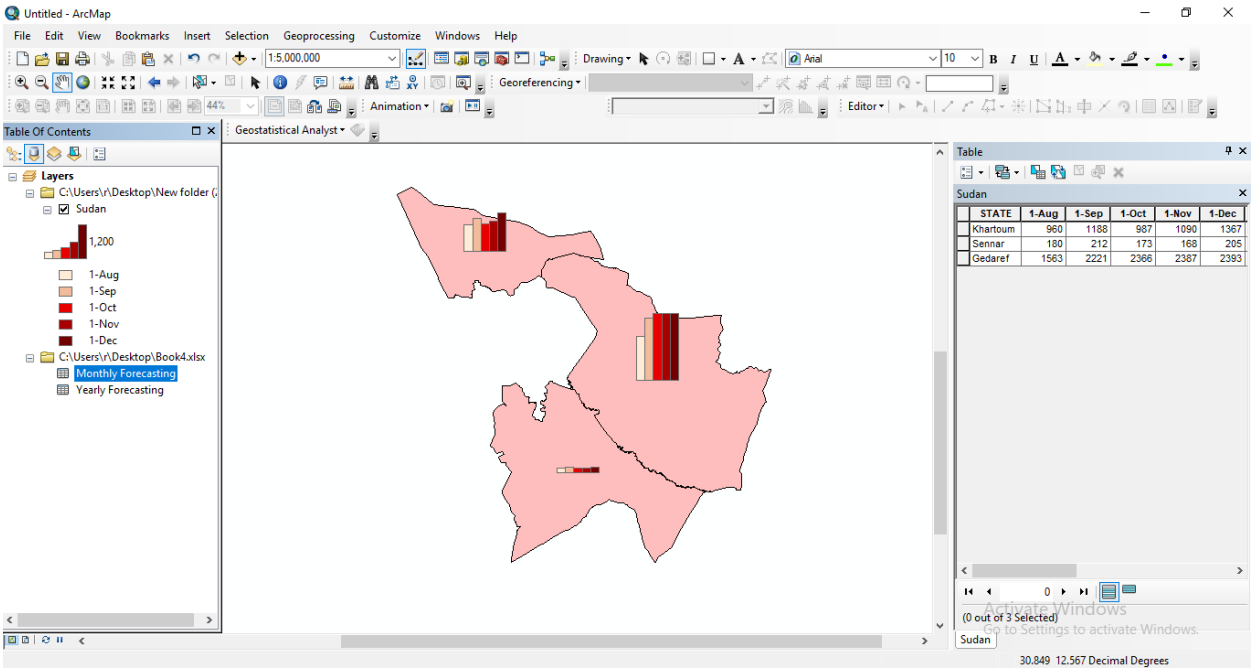

Figure 6.2 The Monthly (2015-2018) forecasting results using ArcMap

6.2 Discussion

In this study the ARIMA and Exponential models were applied for three states (Khartoum, Al-Gadaref and Sennar).

6.2.1 Non-Seasonal Model- Yearly

6.2.1.1 Khartoum

In Khartoum state in the estimating step the p value of the AR coefficient of the first model ARIMA (1,2,0) and the p value of the alpha of the fourth model SES is less than 5% which indicates that the correlations are significantly different from zero. But the p value of the MA coefficient of the second model ARIMA (0,2,1) and the p value of the ARIMA coefficients of the third model ARIMA (1,2,1) is more than 5% which indicates that the correlations aren't significantly different from zero. Therefore, the diagnosis step didn't apply to the last two models. And in the diagnosis step the p value of the AR coefficient of the first model ARIMA (1,2,0) and the p value of the alpha of the fourth model SES is more than 5% which indicates that there is no self-correlation between the errors.

The comparison of the predictability of the first and the fourth models showed that the SES is best model between the four models with very small different from ARIMA (1,2,0) to present and forecast Khartoum State yearly Malaria cases.

6.2.1.2 Al-Gadaref

In Al-Gadaref state since the values of AR and MA were zero that is mean the p values of the AR and MA coefficients had no effect in the estimating step. While the p value of the alpha of the fourth model Brown Exponential Smoothing (BES) is less than 5% which indicates that the correlation is significantly different from zero. Whilst in the diagnosis step the p value of the AR coefficient of the model ARIMA (0,2,0) and the p value of the alpha of the fourth model BES is more than 5% which indicates that there is no self-correlation between the errors.

The comparison of the predictability of the ARIMA (0,2,0) and the BES models showed that the BES is best model between the four models to present and forecast Khartoum State yearly Malaria cases.

6.2.1.3 Sennar

In Sennar state in the estimating step the p value of the AR coefficient of the first model ARIMA (1,1,0) and the p value of the MA coefficient of the second model ARIMA (0,1,1) is less than 5% which indicates that the correlations are significantly different from zero. But the p value of the ARIMA coefficients of the third model ARIMA (1,1,1) and the p value of the alpha of the fourth model Holt Exponential Smoothing (HES) is more than 5% which indicates that the correlations aren't significantly different from zero. Therefore, the diagnosis step didn't apply to the last two models. And in the diagnosis step the p value of the AR coefficient of the first model ARIMA (1,1,0) and the p value of the second model ARIMA (0,1,1) is more than 5% which indicates that there is no self-correlation between the errors.

The comparison of the predictability of the first and the second models showed that the ARIMA (0,1,1) is best model between the four models to present and forecast Khartoum State yearly Malaria cases.

6.2.2 Seasonal Model – Monthly

6.2.2.1 Khartoum

In Khartoum state in the estimating step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0), the p value of the second model ARIMA (0,1,1)(0,1,0) and the p value of the winters' multiplicative coefficients of the fourth model are less than 5% which indicates that the correlations are significantly different from zero. But the p value of the third model ARIMA (1,1,1)(0,1,0) is more than 5% which indicates that the correlations isn't significantly different from zero. Therefore, the diagnosis step didn't apply to the last model. And in the diagnosis step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0) is more than 5% which indicates that there is no self-correlation between the errors. While the p value of the second model ARIMA (0,1,1)(0,1,0) and the p values of the fourth model winters' multiplicative is less than 5% which indicates that there is self-correlation between the errors.

The comparison of the predictability of the four models showed that the ARIMA (1,1,0)(0,1,0) is best model between the models to present and forecast Khartoum State monthly Malaria cases.

6.2.2.2 Al-Gadaref

In Al-Gadaref state in the estimating step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0) and the p value of the MA coefficient of the second model ARIMA (0,1,1)(0,1,0) is less than 5% which indicates that the correlations are significantly different from zero. But and the p value of the ARIMA coefficients of the third model ARIMA (1,1,1)(0,1,0) and the p values of the fourth model SSES is more than 5% which indicates that the correlations aren't significantly different from zero. Therefore, the diagnosis step didn't apply to the last two models. And in the diagnosis step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0) and the p value of the of the second model ARIMA (0,1,1)(0,1,0) is more than 5% which indicates that there is no self-correlation between the errors.

The comparison of the predictability of the first and the second models showed that the ARIMA (0,1,1)(0,1,0) is best model between the four models to present and forecast Al-Gadaref State monthly Malaria cases.

6.2.2.3 Sennar

In Sennar state in the estimating step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0) is less than 5% which indicates that the correlations are significantly different from zero. But the p value of the MA coefficients of the second model ARIMA (0,1,1)(0,1,0), the p value of the ARIMA coefficient of the third model ARIMA (1,1,1)(0,1,0) and the p value of the fourth model SSES is more than 5% which indicates that the correlations aren't significantly different from zero. Therefore, the diagnosis step didn't apply to the last three models. And in the diagnosis step the p value of the AR coefficient of the first model ARIMA (1,1,0)(0,1,0) is more than 5% which indicates that there is no self-correlation between the errors.

The comparison of the predictability of the four models showed that the ARIMA (1,1,0)(0,1,0) is best model between the models to present and forecast Sennar State monthly Malaria cases.

CHAPTER SEVEN

CONCLUSION & RECOMMENDATION

7.1 Conclusion

Time series is anything which is observed sequentially over the time at regular interval like hourly, daily, weekly, monthly, quarterly etc. Time series data is important when you are predicting something which is changing over the time using past data. In time series analysis the goal is to estimate the future value using the behaviors in the past data.

Using time series found that the simple models (AR and MA) represented the data better in Khartoum, Al-Gadaref and Sennar states in the seasonal data and Sennar in the non-seasonal data. While the Exponential Smoothing and Mixed model (ARIMA) are better in representing Khartoum and Al-Gadaref non-seasonal data. Which prove that; not all data can be representing using the same forecasting model.

7.2 Recommendation

1. Expand the study by increasing the number of the states used in the study.
2. Increase the number of the yearly data.
3. Adding other variables like temperature, humidity or rainwater level, in order to produce more accurate results.
4. Establish a program to enter the data online from the all states of Sudan with the inclusion of predication equations to predict data directly, either monthly or annually.

Reference

1. Carlos Castillo-Chavez, S. B.-A. (2002). *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: Models, Methods, and Theory.* New York: Springer Science+Business Media.

2. Dongmei Chen, B. M. (2015). *Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases.* New Jersey: John Wiley & Sons Inc.

3. Christaki, E. (2015). New technologies in predicting, preventing and. *Virulence*.

4. Heesterbeek, K. D. (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 1-21.

5. Norstrøm, M. (2001). Geographical Information System (GIS) as a Tool in Surveillance and Monitoring of Animal Diseases. *Acta*.

6. Yves M. Tourre, D. F.-A.-P. (2007). *GIS and High-Resolution Remote Sensing Improve Early Warning Planning for Mosquito-Borne Epidemics.* Esri.

7. *mrcindia.* (n.d.). Retrieved April 12, 2018, from www.mrcindia.org/.../Epidemiology%20&%20Clinical%20Studies.pdf

8. Meentemeyer, R. H. (2006). GIS-Based Epidemiological Modeling of an Emerging Forest Disease: Spread of Sudden Oak Death Across California Landscapes. *The Sixth California Oak Symposium: Today's.* California.

9. Constantinos, S. a. (2014). Mathematical modeling of infectious disease dynamics. *Virulence*, 37-41.

10. Brooke Wikgren, H. K.-P.-k. (2014). Modeling the distribution of the North Atlantic right whale Eubalaena glacialis off coastal Maine . *Endangered Species Research*, 21-31.

11. Chiyaka, J. M. (2010). *Infectious Disease Modelling Research Progress.* New York: Nova Science Publishers, Inc.

12. Altobelli, J. (2011). *Lippincott's guide to infectious diseases.* Lippincott William & Wilkins.

13. Grace, C. (2003). *Medical Management of Infectious Disease.* NY: Marcel Dekker Inc.

14. Durvasula, V. S. (2013). *Dynamic Models of Infectious Diseases.* New York: Springer Science+Business Media .

15. Ray, K. J. (2004). *Sherris Medical Microbiology an Introduction to Infectious Diseases.* New York: The McGraw-Hill Companie Inc.

16. Fidell, B. G. (2007). Time Series Analysis. In *Using Multivariate Statistics.* Pearson Education. Inc.

17. Nugus, S. (2006). *Financial Planning using Excel Forcasting Planning and Budgeting Techniques.* Great Britain: Elsevier.

18. Sayad, S. (n.d.). Retrieved May 23, 2018, from Saed Sayad: www.saedsayad.com/docs/Time%20Series%20and%20Forecasting.pdf

19. *Bista.* (2016, May 31). Retrieved May 28, 2018, from www.bistasolutions.com: https://www.bistasolutions.com/resources/blogs/5-statistical-methods-for-forecasting-quantitative-time-series/

20. (n.d.). Time Series Analysis and Forecasting. McGraw-Hill Education.

21. *Esri.* (2009, August). Retrieved March 18, 2018, from esri.com: http://esri.com/library/bestpractices/early-detection.pdf