YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS
MASTER OF APPLIED STATISTICS PROGRAMME

ARIMA MODELLING FOR MALARIA INFECTION IN KACHIN STATE

THEINGI AYE
MAS – 25

DECEMBER, 2019

YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS
MASTER OF APPLIED STATISTICS PROGRAMME

ARIMA MODELLING FOR MALARIA INFECTION IN KACHIN STATE

This thesis is submitted as a partial fulfillment towards
The Degree of Master of Applied Statistics (MAS)

BY

THEINGI AYE
MAS – 25

DECEMBER, 2019

YANGON UNIVERSITY OF ECONOMICS
DEPARTMENT OF STATISTICS

ARIMA MODELLING FOR MALARIA INFECTION IN KACHIN STATE

(JANUARY, 2011 TO DECEMBER, 2016)

This Thesis is submitted to Board of Examination as partial fulfillment of the requirement for degree of MAS (Master of Applied Statistics)

Approved by the Board of Examiners

Supervised by

Prof. Dr. Maw Maw Khin
Professor (Head of Department)
Department of Statistics
Yangon University of Economics

Submitted by

Theingi Aye
Roll No. 25
MAS. (Batch -1)
Yangon University of Economics

DECEMBER, 2019

ACCEPTANCE

Accepted by the Board of Examiners of the Department of Statistics, Yangon University of Economics in partial fulfillment for the requirement of the Master Degree, Master of Applied Statistics.

……..
(Chairperson)
Prof. Dr Tin Win
Rector
Yangon University of Economics

…..……..
(Chief Examiner/Supervisor)
Prof. Dr. Maw Maw Khin
Professor (Head of Department)
Department of Statistics
Yangon University of Economics

..……..
(Examiner)
Daw Win Win Nu
Associate Professor (Retd.)
Department of Economics
Yangon University of Distance Education

…..……..
(Examiner)
Prof. Dr. Mya Thandar
Professor
Department of Statistics
Yangon University of Economics

DECEMBER, 2019

ABSTRACT

This study attempts to model and to forecast malaria infection of Kachin State which had been affected malaria highest risk areas at 2004 in Myanmar. This study utilized monthly time series data from January, 2011 to December, 2016 and employed the well-known Box-Jenkins Seasonal ARIMA Modeling procedures. The objectives of this study are to study Malaria incidence in Kachin State, to examine the best fitted ARIMA model and to forecast the incidence of Malaria infection in Kachin State based on the best fitted model. Following the Box and Jenkins methodology, the time series modeling involves transformation of the data to achieve stationary followed by identification of appropriate models, estimation of model parameters, diagnostic checking of the assumption model and finally forecasting of future data values. SARIMA (1,0,0) x $(1,1, 0)_{12}$ was defined the best model to predict the future Malaria infection in Kachin state and forecasted the future values using the fitted model. The results of this paper indicate that over 50% of malaria incidence in Kachin State is decreased in 2017. That is why malaria incidence in Kachin State may be eliminated in 2020 although the Kachin State is not included in 2020 targeted areas for malaria elimination in Myanmar. There is also observed that the SARIMA model is capable of representing with relative precision the number of malaria infection in the next year.

ACKNOWLEDGEMENT

Firstly, I would like to express my deepest gratitude to Professor Dr. Tin Win, Rector of Yangon University of Economics and Professor Dr. Nilar Myint Htoo, Pro-rector of Yangon University of Economics, for supporting and offering the MAS programme which significantly contributed to my academic improvement and enhancement of professional knowledge.

I would like to convey my sincere thanks to my supervisor, Prof. Dr. Maw Maw Khin, Professor, Head of Department of Statistics and Director of Master of Applied Statistics, Yangon University of Economics, for her treasured and kind guidance, suggestions, encouragement and permission to carry out this thesis.

I would like to acknowledge my indebtedness to Professor Dr. Mya Thandar, Department of Statistics, Yangon University of Economics, for her valuable suggestions, helpful advices and recommendations to improve my thesis.

I would like to acknowledge my gratitude to Associate Professor Daw Win Win Nu (Retd.), Department of Economics, Yangon University of Distance Education for her valuable comments and suggestions in preparing this thesis.

I would like to express many thanks to all of my Professors and teachers at Yangon University of Economics for their kindness and teaching throughout my study in Yangon University of Economics.

I am also highly grateful to the responsible people at National Malaria Control Program (NMCP) in Nay Pyi Taw for providing the required data about Malaria for this study.

Finally, I would like to convey my deepest gratitude to my family members, my friends and my colleagues for their support, patience and encouragement throughout the course of this study.

CONTENTS

			Page
ABSTRACT			i
ACKNOWLEDGEMENT			ii
CONTENTS			iii
LIST OF TABLES			vi
LIST OF FIGURES			vii
ABBREVIATIONS			viii
CHAPTER			
CHAPTER	I	INTRODUCTION	1
	1.1	Rationale of the Study	1
	1.2	Objectives of the Study	2
	1.3	Method of Study	2
	1.4	Scope and Limitations of the Study	3
	1.5	Organization of the Study	3
CHAPTER	II	LITERATURE REVIEW	4
	2.1	Background of Malaria	4
	2.2	Reviews on Related Studies	4
CHAPTER	III	RESEARCH METHODOLOGY	7
	3.1	Components of time series	7
	3.2	Time Series Models	8
	3.3	Test of Seasonality	9
	3.4	Method of Finding Seasonal Variation	10
		3.4.1 Ratio to Moving Average Method	11
	3.5	Stationary Stochastic Process	11
		3.5.1 Mean and Variance of Stochastic Process	11
		3.5.2 Autocovariance and Autocorrelation Coefficients	12
		3.5.3 Partial Autocorrelation Function	12
	3.6	White Noise Processes	16

3.7 Autoregressive Processes 17
3.7.1 General p^{th} Order Autoregressive AR (p) Process 18
3.8 Moving Average Processes 19
3.8.1 General q^{th} Order Moving Average MA (q) Process 19
3.8.2 Autoregressive Moving Average ARMA (p, q) Processes 20
3.8.3 Autoregressive Integrated Moving Average ARIMA (p,d,q) Process 22
3.8.4 Seasonal Autoregressive Integrated Moving Average, SARIMA (p,d,q) x (P,D,Q)s Model 22
3.9 Model Building for a Time Series 23
3.9.1 Identification 23
3.9.2 Estimation of parameters 24
3.9.2.1 Maximum Likelihood Method 24
3.9.3 Diagnostic Checking 26
3.9.3.1 Model Selection Criteria 28
3.9.4 Forecasting 29
3.9.4.1 Minimum Mean Square Error Forecasts for ARIMA Models 29
3.9.4.2 Model Building and Forecasting for Seasonal Model 31

CHAPTER IV DATA ANALYSIS AND FORECASTING OF MALARIA INFECTION IN KACHIN STATE 32
4.1 Descriptive Statistics of the volumes of Malaria infection in Kachin State (January 2011 to December 2016) 33
4.2 Test of Seasonality for Malaria Infection in Kachin State 34
4.3 Model Identification for Malaria Infection in Kachin State 36
4.4 Parameter Estimation for SARIMA (1,0,0) x $(0,1,0)_{12}$ Model 45
4.5 Diagnostic Checking SARIMA (1,0,0) x $(0,1,0)_{12}$ Model 45
4.6 Parameter Estimation for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model 49
4.7 Diagnostic Checking SARIMA (1,0,0) x $(1,1,0)_{12}$ Model 49

4.8 Forecasting with SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Kachin Malaria Infection Series 53

CHAPTER V CONCLUSION 55

REFERENCE

APPENDIX

LIST OF TABLES

Table	Descriptions	Page No.
4.1	Volumes of Malaria Infection in Kachin State	33
4.2	ANOVA Table for Malaria Incidence in Kachin State	35
4.3	Seasonal Index for Malaria Infection in Kachin State	35
4.4	Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) of Malaria Incidence in Kachin state	38
4.5	Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) for Natural Logarithms of Malaria Incidence in Kachin state Ln (Z_t)	40
4.6	Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) of Seasonal Difference Transformed Series for Malaria Incidence in Kachin State	43
4.7	Model Parameters of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Incidence in Kachin State	45
4.8	Estimated Residuals ACFs and PACFs of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Incidence in Kachin State	46
4.9	Model Statistics of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Incidence in Kachin State	48
4.10	Model Parameters of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State	49
4.11	Estimated Residual ACFs and PACFs of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State	50
4.12	Model Statistics of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State	52
4.13	Forecast Values with 95% Limits for Malaria Infection in Kachin State	53

LIST OF FIGURES

Figure	Descriptions	Page No.
4.1	Volumes of Malaria Infection in Kachin State	34
4.2	The Correlograms for Sample ACF and PACF of Malaria Incidence in Kachin State	36
4.3	Transformed Volumes of Malaria Incidence in Kachin State	39
4.4	The Correlograms of Sample ACF and PACF for natural Logarithms Series of Malaria Incidence in Kachin State	41
4.5	Seasonal Differenced Log Transformed Volumes of Malaria Infection in Kachin State	42
4.6	The Autocorrelation and Partial Autocorrelation Function of Seasonal Transformed Series for Malaria Incidence in Kachin State	44
4.7	The Autocorrelation and Partial Autocorrelation Function of Residuals for SARIMA $(1,0,0) \times (0,1,0)_{12}$ Model	47
4.8	The Autocorrelation and Partial Autocorrelation Function of Residuals for SARIMA $(1,0,0) \times (1,1,0)_{12}$ Model	51
4.9	Forecast Values with 95% Limits for Malaria infection for SARIMA $(1,0,0) \times (1,1,0)_{12}$ Model	54

ABBREVIATIONS

ACF	Autocorrelation Function
ARIMA	Autoregressive integrated Moving Average
BIC	Bayessian Information Criterion
LCL	Lower Confidence Limit
MA	Moving Average
MAPE	Mean Absolute Percentage Error
NMCP	National Malaria Control Program
PACF	Partial Autocorrelation Function
SARIMA	Seasonal Autoregressive Integrated Moving Average
SPSS	Statistical Package for Social Science Software
UCL	Upper Confidence Limit

CHAPTER I
INTRODUCTION

1.1 Rationale of the Study

Malaria is an infectious disease caused by a parasite; it is spread by the bite of an infected mosquito. It is caused by the Plasmodium genus that is transmitted between humans by Anopheles mosquitoes (Thomas, 2014). Falciparum and P.vivax are the most common species that cause malaria in humans. Every year, 300 to 700 million people get infected. Malaria kills 1 million to 2 million people every year.

Globally, more countries are moving towards elimination: in 2016, 44 countries reported fewer than 10,000 malaria cases, up from 37 countries in 2010. In 2016, an estimated 216 million cases of malaria occurred worldwide (95% confidence interval [CI]: 196–263 million), compared with 237 million cases in 2010 (95% CI: 218–278 million) and 211 million cases in 2015 (95% CI: 192–257 million). Malaria continues to claim a significant number of lives: in 2016, 445000 people died from malaria globally, compared to 446000 estimated deaths in 2015. In 2016, WHO identified 21 countries with the potential to eliminate malaria by the year 2020.

Asia ranks second to Africa in terms of malaria burden. In 19 countries of Asia, malaria is endemic and 2.31 billion people or 62% of the total population in those countries are at risk of malaria. In 2010, WHO estimated around 34.8 million cases and 45,600 deaths due to malaria in Asia. Sri Lanka was declared malaria-free in 2016, becoming only the second country in Southeast Asia, after the Maldives, to successfully eliminate malaria. Apart from India, Indonesia, Myanmar, and Thailand, malaria-endemic countries have reported reductions in malaria incidence of more than 75% since 2000.

Malaria is one of the major public health problems in Myanmar about 1976. In year 1978, Peoples Health Plan was initiated and malaria control programme was integrated with Basic Health Services. Compared to 2002 data, cases and deaths were lower in 2003 and 2004. Year 2004 is the lowest recorded number of malaria cases in outpatient and inpatient department, malaria deaths as well as malaria morbidity rate and mortality rate in Myanmar during the last two and half decade period.

In term of year 2004 statistics, the areas of high malaria morbidity rate (per 1000 population) are Rakhine State (62.43), Chin State (46.41), Kayah State (28.92),

Kachin State (24.34) and Tanintharyi Division (21.46). High malaria mortality rate (per 100,000 population) were seen in Kayah State (14.00), Kachin State (8.80), Chin State (8.58), Shan State (7.55) and Tanintharyi Division (7.44). Myanmar has made impressive progress in malaria control during the past 5 years: 80 % reduction in the number of confirmed malaria cases has been registered from 2011 to 2016 (from 567,452 to 110,146 respectively) and 96 % reduction in the number of deaths attributable to malaria has been reported in the same period (from 581 to 21 respectively).

In term of world malaria report (2017), effective surveillance of malaria cases and deaths is essential for identifying the areas or population groups that are most affected by malaria. Malaria related cases had dropped significantly and there was zero malaria death since 2015. Although significant progress has been made in recent years, the malaria burden in Myanmar remains the highest among the six countries of the GMS (Greater Mekong sub region).

The ministry is trying to stop transmission of malaria in five regions; Yangon, Ayeyarwaddy, Bago, Mandalay and Magwe by 2020. These five regions are excluded the regions which has been the lowest recorded number of malaria cases, death, morbidity rate and mortality rate in Myanmar, 2004. Those lowest recorded areas of 2004 were Rakhine State, Chin State, Kayah State, Kachin State and Tanintharyi Division. Kachin state was the highest infectious area among the lowest recorded areas of 2004. This study is mainly focused on Kachin State in order to figure out why this area is not excluded in targeted areas of malaria elimination in 2020.

1.2 Objectives of the Study

The objectives of the study are

i) to study Malaria incidence in Kachin State

ii) to examine the best fitted ARIMA model of Malaria infection in Kachin State

iii) to forecast the incidence of Malaria infection in Kachin State based on the best fitted model.

1.3 Method of Study

Monthly time series secondary malaria data of Kachin State from January 2011 to December 2016 are used for this study. Box and Jenkins Method (SARIMA) model is applied for forecasting the incidence of Malaria in Kachin State, Myanmar.

Model identification is made based on autocorrelation (ACF) and partial autocorrelation function (PACF). The parameters are estimated by using the Least Square Method depending on the model. The adequacy of the models was verified by plots of the correlograms and ACF and PACF of the residuals and Ljung-Box test, which is a test for hypotheses of no correlation across a specified number of time lags. ACF of the residuals and Ljung-Box statistics are useful for testing the randomness of the residuals.

1.4 Scope and Limitations of the Study

This study is focus on malaria infected people in Kachin State of Myanmar. The study period is from January 2011 to December 2016. The required data are obtained from National Malaria Control Program (NMCP), Nay Pyi Taw, Myanmar.

1.5 Organization of the Study

This paper consists of five chapters; Chapter (1) illustrates introduction that include rational of the study, objective of the study, scope and limitation of the study, method of study and organization of the study. Chapter (2) mentions the description of literature review. Chapter (3) describes the research methodology which is used in this paper. Chapter (4) presents the findings of the study which are the regional trends of malaria infection in Kachin State, Myanmar and estimation of malaria infection in next years of Myanmar. Chapter (5) illustrates Conclusion of the study.

CHAPTER II

LITERATURE REVIEW

In this chapter, critically review the background of Malaria and literature of the previous research papers and articles which are relevant to the current study was presented.

2.1 Background of Malaria

Malaria is an infectious disease caused by a parasite, called Plasmodium that invades red blood cells and liver cells. The parasites are transferred to humans by the bite of an infected *Anopheles* mosquito. There are four different species of Plasmodium parasites which cause most of the malaria in humans: *Plasmodium vivax*, *Plasmodium falciparum*, *Plasmodium ovale*, and *Plasmodium malariae*, with some species causing more severe symptoms than others.

The severity of malarial illness depends largely on the immunological status of the person who is infected. Partial immunity develops over time through repeated infection, and without recurrent inflection, immunity is relatively short-lived.

Although mosquitoes can be found on every continent, malaria is only found in specific parts of the world like Sub-Saharan Africa, the Indian subcontinent, South Pacific Islands (Solomon Islands, Papua New Guinea) and Haiti (in the Caribbean). Mosquitoes used to transmit malaria like warm, wet tropical and subtropical regions. Malaria is indeed a great global health problem affecting approximately 106 countries, with half of the world's population at risk (3.3 billion people) (WHO, 2014).

2.2 Reviews on Related Studies

The available literature on the subject has been reviewed and presented under the following:

Anokye et al., (2018), studied namely, "Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence". The monthly malaria data were used from the Regional Health Directorate from January 2010 to December 2016. Trend of malaria prevalence was analyzed and compared by years and months. The Quadratic model was used for the forecasting of the half year incidence of Malaria while Auto regressive integrated moving average (ARIMA) (1, 1, 2) was

used for forecasting monthly malaria incidence for the years 2018 and 2019 in Kumasi Metropolis.

Anwar et al. (2016), who conducted a study namely “Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence”. This study employs data from Ministry of Public Health monthly reports from January 2005 to September 2015. Malaria incidence in Afghanistan was forecasted using autoregressive integrated moving average (ARIMA) models in order to build a predictive tool for malaria surveillance. Results indicate ARIMA models can be applied to forecast malaria patterns in Afghanistan, complementing current surveillance systems. The models provide a means to better understand malaria dynamics in a resource limited context with minimal data input, yielding forecasts that can be used for public health planning at the national level.

Babajide Sadiq (2015) conducted a research; “A time series analysis of malaria cases in Ogun State, Nigeria”to find out the changes in the trend of malaria cases and to know whether meeting the Millennium Development Goal for malaria. 10 years malaria data from 2004 to 2013 were used and a trend analysis was performed on the malaria cases. Then, rating to know if there is a monthly or yearly increase or decrease in malaria incidence and a time-series analysis (ARIMA model) was conducted to forecast malaria cases for 2014 and 2015.

Jserbr (2018), stated that “Time Series Analysis and Forecasting Model for Monthly alaria Infection by Box-Jenkins Techniques in Kass Zone, South Darfur State, Sudan”. Time series analysis has been extensively utilized in health fields, and epidemic diseases. The major goal of this study become to offer a malaria prediction model by means of the usage of Box-Jenkins statistics and historic malaria morbidity records for malaria-endemic areas in Kass zone. Sudan over a period of 4 years from January 2005 to December 2008, were analyzed by seasonal ARIMA model. The ARIMA forecast period is January 2008 to December 2008, there is deviation from month 1 and month 2. Prediction from month 9 to month 11 almost exact. Slight deviation in predicting in month 6 to month 9, over prediction is good.

Kumar V (2014), mentioned that “Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis”. This study was designed to forecast malaria cases using climatic factors as predictors in Delhi, India. Monthly malaria cases of the malaria clinic at Rural Health Training Centre (RHTC), Najafgarth, Delhi from January 2006 to December 2013. Autoregressive integrated moving average,

ARIMA (0,1,1) (0,1,0) (12), was the best fit model and it could explain 72.5% variability in the time series data. Seasonal adjusted factor (SAF) for malaria cases shows peak during the months of August and September.

Martinez, E.Z., et al. (2011), "A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil," Revista da Sociedade Brasileire de Medicina Tropical. The results of this article indicate that SARIMA models are useful tools for monitoring dengue incidence. They also observe that the SARIMA model is capable of representing with relative precision the number of cases in a next year.

O Ebhuoma (2018) had made a study namely "A Seasonal Autoregressive Integrated Moving Average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa". This study was using a clinically confirmed monthly malaria case dataset that was split into two. The first dataset (January 2005-December 2013) was used to construct a SARIMA model by adopting the Box-Jenkins approach, while the second dataset (January- December 2014) was used to validate the forecast generated from the best-fit model. Among three plausible models, the SARIMA (0,1,1) x $(0,1,1)_{12}$ was selected as the best-fit model. It could serve as a useful tool for modeling and forecasting monthly malaria cases in KZN. It could therefore play a key role in shaping malaria control and elimination efforts in the province.

Takyi Appiah, S., Otoo, H. and Nabubie, I.B. (2015), studied namely "Times Series Analysis of Malaria Cases in EjisuJuaben Municipality". The number of malaria cases in the Ejisu-Juaben Municipality was modeled statistically to find the best model for forecasting the disease for a two-year period. The Box-Jenkins approach was applied to Secondary data from the municipality to determine the best model fit. From the model obtained, the forecast was found to have an oscillatory trend for some period and then remain constant for the period of two years from 2014 and 2016.

CHAPTER III
RESEARCH METHODOLOGY

A time series is a continuous set of observations that are ordered in equally spaced in travel (e.g., one per month). Time series is anything which is observed sequentially over the time at regular interval like hourly, daily, weekly, monthly, quarterly etc. The data of time series can be categorized as stationary and non-stationary data in term of its trend presence or absence. No trend in data leads to non-stationary of time series.

3.1 Components of Time Series

There are four common time series patterns. They are Trend patterns, Seasonal patterns, Cyclical patterns and Random patterns.

Trend Pattern

A trend is a general increase or decrease in a time series that lasts for approximately seven or more periods (e.g., seven months, where seven is a crude rule of thumb). Trends are caused by long-term population changes, growth during product and technology introductions, changes in economic conditions, and so on.

Seasonal Pattern

Seasonal series result from events that are periodic and recurrent (e.g., monthly changes recurring each year). Common seasonal influences are climate, human habits, holidays, repeating promotions, new-product announcements, and so on. Seasonality can occur many different ways, for example, by week of the year, month of the year, day of the month, day of the week (e.g., telephone usage by hour).

Cyclical Pattern

Economic and business expansions (increasing demand) and contractions (recessions and depressions) are the most frequent cause of cyclical influences on time series. These influences most often last for two to five years and recur, but with no known period.

Random Pattern

Random time series are the result of many influences that act independently to yield nonsystematic and nonrepeating patterns about some average value. Purely random series have a constant mean and no systematic patterns.

3.2 Time Series Models

A time series model can be expressed as some combination of these components. The model is simply a mathematical statement of the relationship among the four components.

Two types of models are commonly associated with time series;

(i) the additive model and

(ii) the multiplicative model

The additive model is constructed by adding four components. That is

$Y_t = T_t + S_t + C_t + R_t$

The multiplicative model is constructed by multiplying four components. That is

$Y_t = T_t \times S_t \times C_t \times R_t$

where, Y_t is the value of the time series for time period

T_t is the trend value

S_t is the seasonal variation

C_t is the cyclical variation

R_t is the random variation for the same time period.

The additive model is usually used when the seasonal swing of time series does not change with time interval 't'. This model suffers from the somewhat unrealistic assumption that the component of each other.

The multiplicative model is usually used when the seasonal swing of time series is charging with time variable 't'. In this model, only trend is expressed in the original units, and seasonal, cyclical, and random variations are stated in terms of percentages.

Monthly or quarterly time series may show seasonal effects within years. Seasonality means a tendency to repeat a pattern of behavior over a seasonal period, generally one year. Seasonal series are characterized by a display of strong serial correlation at a seasonal lag, that is, the lag corresponding to the number of observations per seasonal lag. Seasonal time series usually display time to time changes over the years, showing also within year variations. It is useful to understand the actual situation and is used for short term planning.

3.3 Test of Seasonality

There are two main reasons of isolating the seasonal component or element. They are

(i) to study seasonal variations

(ii) to eliminated them.

In the study of seasonality, seasonal variation for each month of the year is usually considered. As such, the following statistical model is employed

$$y_{ij} = \mu + a_i + b_j + e_{ij} \quad ; 1 \leq i \leq n \, , \, 1 \leq j \leq k$$

Where , μ = general mean/ unknown constant

a_i = effect of i^{th} year (i = 1, 2, 3, …, n)

b_j = effect of j^{th} month (j = 1, 2, 3, …, n)

e_{ij} = random error

y_{ij} = observed value of y at j^{th} month of i^{th} year.

Hence, it is assumed that e_{ij} are independently and normally distributed with mean zero and constant variance σ^2.

i.e. $e_{ij} \sim IN(0, \sigma^2)$

In general,

$H_0 : b_1 = b_2 = \ldots = b_{12} = 0$

There exist no monthly effects. i.e. b_j is zero or there is no seasonality.

H_1 : At least one b_j is not equal to zero.

There exists monthly effect and there is seasonality.

Test is obtained from the following computation procedure and analysis of variance (ANOVA) table shown below.

SST = Total Sum of Square = $\sum_i \sum_{ij} y_{ij}^2 - (C.T)$

SSM = Sum of Square due to months = $\frac{1}{n} \sum_{j=1}^{12} R_j^2 - (C.T)$

SSY = Sum of Square due to years = $\frac{1}{12} \sum_{i=1}^{n} C_j^2 - (C.T)$

SSE = Error Sum of Squares = SST – SSM – SSY

Where C.T = G^2/nk, G = $\sum_{i=o}^{n} \sum_{j=1}^{k} y_{ij}$ = Grand Total

After computation of SST, SSM, SSU and SST – SSM – SSY, the following ANOVA table is constructed.

Source	Sum of Square	Degree of Freedom	Mean Square	F-Ratio
Due to Months Due to Years Error	SSM SSY SSE	k-1 n-1 (n-1) (k-1)	MSM = SSM / k-1 MSY = SSY / n-1 MSE = SSE / (n-1) (k-1)	F_1= MSM/MSE F_2= MSY/MSE
Total	SST	kn-1		

At 100 (1- α) % level of significant, the critical value from the F table for the degree of freedom (k-1) and (k-1) (n-1) is given by, $K= F_{\alpha, (k-1), (k-1)(n-1)}$.

If $F \geq K$, reject H_0 and it is decided that there exists seasonality.

If $F < K$, accept H_0 and it is decided that there exists no seasonality.

3.4 Method of Finding Seasonal Variation

Seasonal variation is measured in terms of an index, called a seasonal index. It is an average that can be used to compare an actual observation relative to what it would be if there were no seasonal variation. An index value is attached to each period of the time series within a year. This implies that if monthly data are considered there are 12 separate seasonal indices, one for each month. There exist different methods for measuring the seasonal variation of a time series. The methods have been developed to meet different objectives of estimating seasonal and the assumed models of the time series. The seasonal pattern itself is important in the application of these methods since most of the methods assume that the seasonal pattern is constant or stable.

In finding the index of seasonal variation as seasonal measures, it should be noted that the index must

(a) Measure all the variation is the series that is seasonal in character, and

(b) Measure nothing but the seasonal variation

A seasonal index thus consists of a series of percentage figures, averaging 100, which shows the relative level of the series for the various months, quarters or weeks of the year. An index of seasonal variation can be constructed by expressing each item in the time series as a percent of the average monthly or quarterly value for the year.

3.4.1 Ratio to Moving Average Method

A seasonal index is a measure of how a particular season compares with the average season. Seasonal indices are calculated so that their average is 1. This means that the sum of the seasonal indices equals the number of seasons.

The steps for the computation of the seasonal index by the Ratio to Moving Average method are shown as below. (Steiner, 1956)

1. Find the twelve months centered moving averages. This is equivalent to a moving average of thirteen months with weights $\frac{1}{24}$ (1,2,2, … , 2,2,1)

By finding twelve months centered moving averages, we eliminate the seasonality, since the seasonal pattern is periodic with a period of twelve months. Also, it will eliminate the random component or irregular movements. Therefore, the centered twelve month moving averages are the approximates of trend and cyclical components.

2. Compute the ratio to moving average values, that is, the original data is divided by its appropriate moving average value. There, the first and last six months may not be obtained.

By this step, the trend and cyclical components are removed from the original data and the ratios are the values due to seasonal and random components. They are called specific seasonal. (Steiner, 1956)

3. Compute the averages of these ratios referring to the same months. These averages are the crude seasonal index values.

This step involves two different purposes: the elimination of the random components and averaging the seasonal relatives referring to the same months.

4. Adjust the crude seasonal index.

In multiplicative mode, the total seasonal index values have to be equal to twelve (or 1200 percent) for monthly series. Therefore, the crude seasonal index is adjusted to get a total of twelve (or 1200 percent).

3.5 Stationary Stochastic Process

3.5.1 Mean and Variance of Stochastic Process

The stochastic process Z_t has a constant mean,

$$\mu = \mu_t = E(Z_t) = \int_{-\infty}^{+\infty} z_t p(z_t) dz_t \qquad (3.5.1)$$

This defines the level about which it fluctuates, and a constant variance.

$$\sigma_z^2 = \sigma_{z_t}^2 = E(z_t - \mu_t)^2 = \int_{-\infty}^{+\infty}(z_t - \mu_t)^2 p(z_t)dz_t \quad (3.5.2)$$

which measures its spread about this time level.

The mean μ of the stochastic process can be estimated by the sample mean

$$\bar{Z}_t = \frac{1}{N}\sum_{t=1}^{N} Z_t \quad (3.5.3)$$

and the variance $\sigma_{z_t}^2$ can be estimated by the sample variance

$$\hat{\sigma}_{z_t}^2 = \frac{1}{N}\sum_{t=1}^{N}(Z_t - \bar{Z}_t)^2 \quad (3.5.4)$$

If the $\mu_t, \sigma_{z_t}^2$ do not depend on t (or) they are constant values, the stochastic process is called strictly stationary process.

3.5.2 Autocovariance and Autocorrelation Coefficients

The covariance between Z_t and Z_{t+k} is called the auto-covariance at lag k and is defined by

$$\gamma_k = Cov[Z_t, Z_{t+k}] = E[Z_t - \mu][Z_{t+k} - \mu] \quad (3.5.5)$$

and the correlation between Z_t and Z_{t+k} as

$$\rho_k = \frac{Cov[Z_t, Z_{t+k}]}{\sqrt{Var[Z_t]}\sqrt{Var[Z_{t+k}]}} = \frac{\gamma_k}{\gamma_o} \quad (3.5.6)$$

where $Var(Z_t) = Var(Z_{t+k}) = \gamma_o$. As function of k, γ_k is called the autocovariance function and ρ_k is called the autocorrelation function.

For a stationary process the autocovariance function γ_k and the autocorrelation function ρ_k have the following properties.

1. $\gamma_o = Var[Z_t], \rho_o = 1$
2. $|\gamma_k| \leq \gamma_o, |\rho_k| \leq 1$
3. $\gamma_k = \gamma_{-k}$ and $\rho_k = \rho_{-k}$ for all k.

3.5.3 Partial Autocorrelation Function

The following conditional correlation

Cor $(Z_t, Z_{t+k} | Z_{t+1}, \ldots, Z_{t+k-1})$ (3.5.7)

and is usually referred to as the partial autocorrelation in time series analysis.

Consider a stationary process $[Z_t]$and, assume that E (Z_t) = 0. Let the linear dependence of Z_{t+k} on Z_{t+1},Z_{t+2},.....,and Z_{t+k-1} be defined as the best linear estimate in the mean square sense of Z_{t+k} as linear function of Z_{t+1}, Z_{t+2}, ..., and Z_{t+k-1}. That is, if $\hat{Z}_{t+k}$ is the best linear estimate ofZ_{t+k}, then

$$\hat{Z}_{t+k}=\alpha_1 Z_{t+k-1}+\alpha_2 Z_{t+k-2}+\cdots+\alpha_{k-1}Z_{t+1}, \quad (3.5.8)$$

Where $\alpha_1(1\le i\le k-1)$ are the mean squared linear regression, coefficients obtained from minimizing

$$\mathrm{E}(Z_{t+k}-\hat{Z}_{t+k})^2=\mathrm{E}(Z_{t+k}-\alpha_1 Z_{t+k-1}-\cdots-\alpha_{k-1}Z_{t+1})^2. \quad (3.5.9)$$

The routine minimization method through differentiation gives the following linear system of equations

$$\gamma_i=\alpha_1\gamma_{i-1}+\alpha_2\gamma_{i-2}+\ldots.+\alpha_{k-1}\gamma_{i-k+1}(1\le i\le k-1). \quad (3.5.10)$$

Hence,

$$\rho_i=\alpha_1\rho_{i-1}+\alpha_2\rho_{i-2}+\cdots+\alpha_{k-1}\rho_{i-k+1} \qquad (1\le i\le k-1). \quad (3.5.11)$$

In terms of matrix notation, the above system of becomes

$$\begin{bmatrix}\rho_1\\ \rho_2\\ \cdots\\ \rho_{k-1}\end{bmatrix}=\begin{bmatrix}1 & \rho_1 & \rho_2 & \cdots\rho_{k-2}\\ \rho_1 & 1 & \rho_1 & \cdots\rho_{k-3}\\ \vdots & \vdots & \vdots & \vdots\\ \rho_{k-2}\rho_{k-3}\rho_{k-4}\cdots & & 1 & \end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \cdots\\ \alpha_{k-1}\end{bmatrix} \quad (3.5.12)$$

Similarly,

$$\hat{Z}_t=\beta_1 Z_{t+1}+\beta_2 Z_{t+2}+\cdots+\beta_{k-1}Z_{t+k-1} \quad (3.5.13)$$

Where $\beta_i(1\le i\le k-1)$ are the mean squared linear regression, coefficients obtained by minimizing

$$\mathrm{E}(Z_t-\hat{Z}_t)^2=\mathrm{E}(Z_t-\beta_1 Z_{t+1}-\cdots-\beta_{k-1}Z_{t+k-1})^2 \quad (3.5.14)$$

Hence,

$$\begin{bmatrix}\rho_1\\ \rho_2\\ \cdots\\ \rho_{k-1}\end{bmatrix}=\begin{bmatrix}1 & \rho_1 & \rho_2 & \cdots\rho_{k-2}\\ \rho_1 & 1 & \rho_1 & \cdots\rho_{k-3}\\ \vdots & \vdots & \vdots & \vdots\\ \rho_{k-2}\rho_{k-3}\rho_{k-4}\cdots & & 1 & \end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \cdots\\ \beta_{k-1}\end{bmatrix} \quad (3.5.15)$$

This implies that $\alpha_i=\beta_i(1\le i\le k-1)$.

It follows that the partial autocorrelation between $Z_t$and Z_{t+k} will equal the ordinary autocorrelation between $(Z_t-\hat{Z}_t)$ and$(Z_{t+k}-\hat{Z}_{t+k})$. Thus, letting $P_k$denote the partial autocorrelation between Z_t andZ_{t+k} , having

$$P_k=\frac{Cov[(Z_t-\hat{Z}_t),(Z_{t+k}-\hat{Z}_{t+k})]}{\sqrt{Var(Z_t-\hat{Z}_t)}\sqrt{Var(Z_{t+k}-\hat{Z}_{t+k})}} \quad (3.5.16)$$

Now,$Var\left(Z_{t+k}-\hat{Z}_{t+k}\right)=E[(Z_{t+k}-\alpha_1 Z_{t+k-1}-\cdots-\alpha_{k-1}Z_{t+1})^2]$

$$=E[Z_{t+k}(Z_{t+k}-\alpha_1 Z_{t+k-1}-\cdots-\alpha_{k-1}Z_{t+1})^2]$$

$$-\alpha_1 E[Z_{t+k-1}(Z_{t+k}-\alpha_1 Z_{t+k-1}-\ldots-\alpha_{k-1}Z_{t+1})]$$

$$\ldots - \alpha_{k-1}E[Z_{t+1}(Z_{t+k} - \alpha_1 Z_{t+k-1} - \cdots - \alpha_{k-1}Z_{t+1})]$$

$$= E[Z_{t+k}(Z_{t+k} - \alpha_1 Z_{t+k-1} - \ldots - \alpha_{k-1}Z_{t+1})]$$

Since all other remaining terms reduce to zero by virtue of equation (3.5.10).

Hence,

$$Var(Z_{t+k} - \hat{Z}_{t+k}) = Var(Z_t - \hat{Z}_t) = \gamma_0 - \alpha_1\gamma_1 - \ldots - \alpha_{k-1}\gamma_{k-1}. \quad (3.5.17)$$

Next, using the fact that $\alpha_i = \beta_i (1 \le i \le k-1)$, Having

$$Cov[(Z_t - \hat{Z}_t), (Z_{t+k} - \hat{Z}_{t+k})]$$

$$= E[(Z_t - \alpha_1 Z_{t+1} - \cdots - \alpha_{k-1}Z_{t+k-1})(Z_{t+k} - \alpha_1 Z_{t+k-1} - \cdots - \alpha_{k-1}Z_{t+1})]$$

$$= E[(Z_t - \alpha_1 Z_{t+1} - \cdots - \alpha_{k-1}Z_{t+k-1})Z_{t+k}]$$

$$= \gamma_k - \alpha_1\gamma_{k-1} - \cdots - \alpha_{k-1}\gamma_1 . \quad (3.5.18)$$

Therefore,

$$P_k = \frac{\gamma_k - \alpha_1\gamma_{k-1} - \cdots - \alpha_{k-1}\gamma_1}{\gamma_0 - \alpha_1\gamma_1 - \cdots - \alpha_{k-1}\gamma_{k-1}} = \frac{\rho_k - \alpha_1\rho_{k-1} - \cdots - \alpha_{k-1}\rho_1}{1 - \alpha_1\rho_1 - \cdots - \alpha_{k-1}\rho_{k-1}}. \quad (3.5.19)$$

Solving the system in (3.12) for α_i by Cramer's rule gives

$$\alpha_i = \frac{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{i-2} & \rho_1 & \rho_i & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \cdots & \rho_{i-3} & \rho_2 & \rho_{i-1} & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & \rho_{k-i} & \rho_{k-1} & \rho_{k-i-2} & \cdots & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{i-2} & \rho_1 & \rho_i & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \cdots & \rho_{i-3} & \rho_2 & \rho_{i-1} & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & \rho_{k-i} & \rho_{k-i-1} & \rho_{k-i-2} & \cdots & 1 \end{vmatrix}} \quad (3.5.20)$$

as the ratio of two determinants. The matrix in the numerator is the same as the symmetric matrix in the denominator except for its i^{th} column being replaced by $(\rho_1, \rho_2, \ldots, \rho_{k-1})$. Substituting α_i in (3.20) to equation (3.19) and multiplying both the numerator and denominator of (3.19) by the determinant

$$\begin{vmatrix} 1 & \rho_1 & \cdots & \rho_{k-2} \\ \rho_1 & 1 & \cdots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{vmatrix},$$

the resulting P_k in (3.19) can be easily seen to equal the ratio of the expansion of the following expression in terms of the last column,

$$P_k = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \qquad (3.5.21)$$

The partial autocorrelation can also be derived as follows. Consider the regression model, where the dependent variable Z_{t+k} from a zero mean stationary process is regressed on k lagged variables $Z_{t+k-1}, Z_{t+k-2}, \ldots,$ and Z_t, i.e.,

$$Z_{t+k} = \emptyset_{k1} Z_{t+k-1} + \emptyset_{k2} Z_{t+k-2} + \cdots + \emptyset_{kk} Z_t + e_{t+k}, \qquad (3.5.22)$$

Where $\emptyset_{ki}$ denotes the i^{th} regression parameter and e_{t+k} is a normal error term uncorrelated with $Z_{t+k-j}$for$j \geq 1$. Multiplying Z_{t+k-j} on both sides of the above regression equation and taking the expectation, getting

$$\gamma_j = \emptyset_{k1}\gamma_{j-1} + \emptyset_{k2}\gamma_{j-2} + \cdots + \emptyset_{kk}\gamma_{j-k} \qquad (3.5.23)$$

and hence, $$\rho_j = \emptyset_{k1}\rho_{j-1} + \emptyset_{k2}\rho_{j-2} + \cdots + \emptyset_{kk}\rho_{j-k} \qquad (3.5.24)$$

For j=1, 2, …, k, having the following system of equations: it can be written as follows

$$\rho_1 = \emptyset_{k1}\rho_0 + \emptyset_{k2}\rho_1 + \cdots + \emptyset_{kk}\rho_{k-1}$$

$$\rho_2 = \emptyset_{k1}\rho_1 + \emptyset_{k2}\rho_0 + \cdots + \emptyset_{kk}\rho_{k-2}$$

$$\vdots$$

$$\rho_k = \emptyset_{k1}\rho_{k-1} + \emptyset_{k2}\rho_{k-2} + \cdots + \emptyset_{kk}\rho_0$$

Using Creamer's rule successively for k = 1,2, …,

$$\emptyset_{11} = \rho_1$$

$$\emptyset_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}$$

$$\emptyset_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

$$\vdots$$

$$\emptyset_{kk} = \frac{\begin{vmatrix} 1 & P_1 & P_2 & \cdots P_{k-2} P_1 \\ P_1 & 1 & P_1 & \cdots P_{k-3} P_2 \\ \vdots & \vdots & \vdots & \vdots \\ P_{k-1} P_{k-2} P_{k-3} & \cdots & P_1 & P_k \end{vmatrix}}{\begin{vmatrix} 1 & P_1 & P_2 & \cdots P_{k-2} P_{k-1} \\ P_1 & 1 & P_1 & \cdots P_{k-3} P_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ P_{k-1} P_{k-2} P_{k-3} & \cdots & P_1 & 1 \end{vmatrix}} \qquad (3.5.25)$$

Comparing Equation (3.25) with (3.21), it can be seen that $\emptyset_{kk}$equals ρ_k. Thus, the partial autocorrelation between Z_t and Z_{t+k} can also be obtained as the regression coefficient and Z_t in (3.22). Because $\emptyset_{kk}$ has becomes a standard notation for the partial autocorrelation between $Z_t$and Z_{t+k}in time series. As a function of k, $\emptyset_{kk}$is usually referred to as the partial autocorrelation function (PACF).

3.6 White Noise Processes

In process $\{\alpha_t\}$is called a white noise process if it is a sequence of uncorrelated random variables from a fixed distribution with constant mean E(a_t) = μ_a , assumed to be 0, constant variance Var (a_t) = σ_a^2 and γ_k = Cov (a_t, a_{t+k}) = 0 for all $k \neq 0$. A white noise process $\{\alpha_t\}$ is stationary with the autocovariance function

$$\gamma_k = \begin{cases} \sigma_a^2, & k=0, \\ 0, & k\neq 0, \end{cases} \qquad (3.6.1)$$

The autocorrection function

$$\rho_k = \begin{cases} 1 & k=0, \\ 0, & k\neq 0, \end{cases} \qquad (3.6.2)$$

The partial autocorrection function

$$\emptyset_{kk} = \begin{cases} 1 & k=0, \\ 0, & k\neq 0, \end{cases} \qquad (3.6.3)$$

By definition ρ_0 = $\emptyset_{00}$ =1 for any process, when the autocorrection and partial auto correlations, refer only to ρ_k and $\emptyset_{kk}$ for $k \neq 0$. The basic phenomenon of the white noise process is that ACT and PACT are identically equal to zero.

3.7 Autoregressive Processes

Based on a finite number of available observations, a finite order parametric model was constructed to describe a time series process. In this chapter, the autoregressive modes were described as a special case. These models are useful in describing a wide variety of time series. The characteristics of each process in terms of autocorrelation and partial autocorrelation functions will also be discussed in this section.

In time series analysis, there are two useful representations to express a time series process. The first one is to write a process $\hat{Z}_t$ in an autoregressive (AR) representation, in which regress the value of $\hat{Z}$ at time t on its own past values plus a random shock, i.e.,

$$\hat{Z}_t = \pi_1 \hat{Z}_{t-1} + \pi_2 \hat{Z}_{t-2} + \cdots + a_t \qquad (3.7.1)$$

or equivalently,

$$\pi(B)\hat{Z}_t = a_t \qquad (3.7.2)$$

where $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$, and $1+\sum_{j=1}^{\infty} |\pi_j| < \infty$. The autoregressive representation is a very useful model for the mechanism of forecasting. In the autoregressive representation of a process, if only a finite number of π weights are nonzero, i.e.,$\pi_1 = \emptyset_1, \pi_2 = \emptyset_2, \ldots, \pi_p = \emptyset_p$ and $\pi_k = 0$ for k> p, then the resulting process is said to be an autoregressive process (model) of order p, which is denoted as AR (p). It is given by

$$\hat{Z}_t = \emptyset_1 \hat{Z}_{t-1} + \emptyset_2 \hat{Z}_{t-2} + \cdots + \emptyset_p \hat{Z}_{t-p} + a_t \qquad (3.7.3)$$

$$\emptyset_p(B)\hat{Z}_t = a_t, \qquad (3.7.4)$$

or

where $\emptyset_p(B) = (1 - \emptyset_1 B - \cdots - \emptyset_p B^p)$, and$a_t$ is random error or disturbance term. Since $\sum_{j=1}^{\infty} |\pi_j| = \sum_{j=1}^{\infty} |\emptyset_j| < \infty$, the process is always invertible. To be stationary, the roots of $\emptyset_p$ (B) = 0 must lie outside of the unit circle.

The AR processes are useful in describing situations in which the present value of a time series depends on its preceding values plus a random shock. First, consider the following simple autoregressive models.

3.7.1 General p^{th} Order Autoregressive AR (p) Process

The p^{th}order autoregressive process AR (p) is

$$\left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\right)\hat{Z}_t = a_t \quad (3.7.5)$$

or

$$\hat{Z}_t = \emptyset_1 \hat{Z}_{t-1} + \emptyset_2 \hat{Z}_{t-2} + \cdots + \emptyset_p \hat{Z}_{t-p} a_t \quad (3.7.6)$$

(a) Autocovariance Function of General AR (p) Process

To find the autocovariance function, both sides of equation (3.7.6) is multiply by $\hat{Z}_{t-k}$

$$\hat{Z}_{t-k}\hat{Z}_t = \emptyset_t \hat{Z}_{t-k}\hat{Z}_{t-1} + \cdots + \emptyset_p \hat{Z}_{t-k}\hat{Z}_{t-p} + \hat{Z}_{t-k} a_t$$

and take the expected value

$\gamma_k = \emptyset_1 \gamma_{k-1} + \cdots + \emptyset_p \gamma_{k-p}$k>0, where E ($a_t \hat{Z}_{t-k}$) = 0 for k>0.

(b) Autocorrelation Function of General AR(p) Process

The following recursive relationship for the autocorrelation function;

$$\rho_k = \emptyset_1 \rho_{k-1} + \cdots + \emptyset_p \rho_{k-p} \text{k>0.} \quad (3.7.7)$$

From (3.32), the ACF ρ_k is determined by the difference equation

$\emptyset_p(B)\rho_k = \left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\right)\rho_k = 0$fork>0. Hence, it can be written as

$$\emptyset_p(B) = \prod_{i=1}^{m}(1 - G_i B)^{d_i},$$

where $\sum_i^m d_i = p$, and ${G_i}^{-1}$ (i=1,2,...,m) are the roots of multiplicity d_i of $\emptyset_p(B) = 0$. Using the difference equations results, as follows,

$$\rho_k = \sum_{i=1}^{m} G_i^k \sum_{j=0}^{di-1} A_{ij} k^j \quad (3.7.8)$$

If $d_i = 1$ for all ${G^{-1}}_i$, are all distinct and the above reduces to

$$\rho_k = \sum_{i=1}^{p} A_i G_i^k \text{ k} > 0 \quad (3.7.9)$$

For a stationary process,$\left|G_i^{-1}\right| > 1$and $|G_i| < 1$. Hence, the ACF ρ_k tails off as a mixture of exponential decays and/or damped sine waves depending on the roots of $\emptyset_p(B) = 0$. Damped sine waves appear if some roots are complex.

(c) Partial Autocorrelation Function of General AR(p) Process

By using the fact that $\rho_k = \emptyset_1\rho_{k-1} + \emptyset_2\rho_{k-2} + \cdots + \emptyset_p\rho_{k-p}$ for k>0, it can obviously be seen that when k>p the last column of the matrix in the numerator of $\emptyset_{kk}$ in (3.25) can be written as linear combination of previous column of the same matrix. Hence, the *PACF* $\emptyset_{kk}$ will vanish after lag p.

3.8 Moving Average Processes

The characteristics of moving average processes in terms of the autocorrelation and partial functions will be discussed as follows. A process $\hat{Z}_t$ is a linear combination of a sequence of uncorrelated random variables, that is,

$$\hat{Z}_t = \mu + a_1 + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \qquad (3.8.1)$$

$$= \mu + \sum_{j=0}^{\infty} \psi_j \, a_{t-j},$$

where $\{a_t\}$ is which white noise process with mean zero and variance σ_a^2.

In the moving average representation of a process, if only a finite number of ψ weights are nonzero, that $\psi_1 = -\theta_1, \psi_2 = -\theta_2, \ldots, \psi_q = -\theta_q$ and $\psi_k = 0$ for k > q, then the resulting process is said to be a moving average process or model of order q and is denoted as MA (q). It is given by

$$\hat{Z}_t = a - \theta_1 a_{t-1} - \theta_1 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad \text{or}$$

$$\hat{Z}_t = \theta(B) a_1$$

where $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \emptyset_q B^q)$.

Because $1 + {\theta_1}^2 + {\theta_2}^2 + \cdots + {\theta_q}^2 < \infty$, a finite moving average process is always stationary. This moving average process is invertible if the roots of $\theta(B) = 0$ lie outside of the unit circle.

Moving average processes are useful to describe a phenomenon in which events produce an immediate effect that only lasts for only short periods of time. To discuss other properties the MA(q) process, let us first consider the following simpler cases.

3.8.1 General q^{th} Order Moving Average MA (q) Process

The general q^{th} order moving average process is

$$\hat{Z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)\, a_t, \qquad (3.8.2)$$

For this general MA (q) process, the variance is

$$\gamma_0 = \sigma_a{}^2 \sum_{j=0}^{q} \theta_j{}^2, \qquad \text{where } \theta_0 = 1.$$

(a) Autocovariance Function of the MA(q) Process

$$\gamma_k = \begin{cases} \sigma_a{}^2(-\theta_k + \theta_1\theta_{k-1} + \cdots + \theta_{q-k}\theta_q) & k = 1,2,\ldots,q, \\ 0, & k > q \end{cases}$$

(b) Autocorrelation Function of the MA(q) Process

$$\rho_k = \begin{cases} \dfrac{-\theta_k + (\theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q)}{1 + \theta_1{}^2 + \cdots + \theta_q{}^2} & k = 1,2,\ldots,q, \\ 0, & k > q \end{cases}$$

The autocorrelation function of an MA (q) process cuts off after lag q. This important property enables us to identify whether a given time series is generated by a moving average process.

3.8.2 Autoregressive Moving Average ARMA (p, q) Processes

The following useful mixed autoregressive moving average ARMA (p, q)processes

$$\emptyset_p(B)\hat{Z}_t = \theta_q(B)a_t, \qquad (3.8.3)$$

where,

$$\emptyset_p(B) = 1 - \emptyset_1 B - \cdots - \emptyset_p B^p, \quad \text{and} \quad \theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q.$$

For the process to be invertible, it requires that the roots of $\theta_q(B) = 0$ lie outside the unit circle. To be stationary, it requires that the roots of $\emptyset_p(B) = 0$ lie outside the unit circle. Assuming that $\emptyset_p(B) = 0$ and $\theta_q(B) = 0$ share no common roots this process refers to an ARMA (p, q) process or model, in which p and q are used to indicate the orders of the associated autoregressive and moving average polynomials, respectively.

The stationary and invertible ARMA process can be written in a pure autoregressive representation discussed in Section (3.4), i.e.,

$\hat{Z}_t = \psi(B)a_t$

where,

$$\psi(B) = \frac{\theta_q(B)}{\emptyset_p(B)} = (1 + \psi_1 B + \psi_2 B^2 + \cdots).$$

(a) Autocorrelation Function of the ARMA (p, q) Process

To derive the autocorrelation function, the equation (3.37) is rewritten as

$$\hat{Z}_t = \emptyset_1 \hat{Z}_{t-1} + \cdots + \emptyset_p \hat{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

and multiplied by $\hat{Z}_{t-k}$on both sides

$$\hat{Z}_{t-k}\hat{Z}_t = \emptyset_1 \hat{Z}_{t-k}\hat{Z}_{t-1} + \cdots + \emptyset_p \hat{Z}_{t-k}\hat{Z}_{t-p} + \hat{Z}_{t-k}a_t - \theta_1 \hat{Z}_{t-k}a_{t-1} - \cdots - \theta_q \hat{Z}_{t-k}a_{t-q}$$

Take the expected value to obtain

$$\gamma_k = \emptyset_1 \gamma_{k-1} + \cdots + \emptyset_p \gamma_{k-p} + E\left(\hat{Z}_{t-k}a_t\right) - \theta_1 E\left(\hat{Z}_{t-k}a_{t-1}\right) - \cdots - \theta_q E\left(\hat{Z}_{t-k}a_{t-q}\right).$$

Because $E\left(\hat{Z}_{t-k}a_{t-i}\right) = 0 \qquad for\ k > i,$

$$\gamma_k = \emptyset_1 \gamma_{k-1} + \cdots + \emptyset_p \gamma_{k-p} k \geq (q+1) \qquad (3.8.4)$$

and hence,

$$\rho_k = \emptyset_1 \rho_{k-1} + \cdots + \emptyset_p \rho_{k-p} k \geq (q+1) \qquad (3.8.5)$$

Equation (3.39) satisfies the p^{th} order homogeneous difference equation given by (3.16). Therefore, the autocorrelation function of an ARMA (*p*,*q*) model tails off after lag q just like an AR(*p*) process, which depends only on the autoregressive parameters in the model. However, the first q autocorrelations $\rho_q, \rho_{q-1}, \cdots, \rho_1$depend on both autoregressive and moving average parameters in the model and serve as initial value for the pattern. This distinction is useful for model identification.

(b) Partial Autocorrelation Function of the ARMA (p, q) Process

Because the ARMA process contains the MA process as a special case, its PACF will also be a mixture of exponential decays and/or damped sine waves depending on the roots of $\emptyset$ (B) = 0 and θ (B) = 0.

Non-Stationary Processes

In previous section the stationary processes have been discussed. However, many applied time series, particularly those arising from economic and business areas, are non-stationary. With respect to the class of covariance stationary processes, non-stationary time series can occur in many different ways. They could have non-constant means μ_1, time varying second moments such as nonconstant variance σ2t, or have both of these properties.

3.8.3 Autoregressive Integrated Moving Average ARIMA (p,d,q) Process

The autoregressive integrated moving average process can be defined as

$$\emptyset_p(1-B)^d Z_t = \emptyset(B)\nabla^d Z_t = \theta_0 + \theta(B)a_t \qquad (3.40)$$

Where, $\emptyset_p(B) = 1 - \emptyset_1 - \emptyset_2 B^2 - \cdots - \emptyset_p B^p \theta_q(B) = 1 - \theta_1 - \theta_2 B^2 - \cdots - \theta_q B^q$

In what follows:

1. $\emptyset_p$ (B) will be called the autoregressive operator: it is assumed to be stationary, that is, the roots of $\emptyset_p$ (B) =0 lie outside the unit circle.
2. $(1-B)^d = \emptyset$ (B) ∇^dwill be called the generalized autoregressive operator: it is a nonstationary operator with d of roots of $(1-B)^d = 0$equal to unity.
3. $\theta_q(B)$will be called the moving average operator; it is assumed to be invertible, that is, the roots of $\theta_q(B)$=0 lie outside the unit circle.

   Where d = 0, the model (3.40) represents a stationary process. The requirements of stationary and invertibility apply independently, and in general, the operators $\emptyset_p(B)$and$\theta_q(B)$ will not be of the same order.

3.8.4 Seasonal Autoregressive Integrated Moving Average, SARIMA$(\mathbf{p,d,q}) \times (\mathbf{P,D,Q})_s$ Model

The ARIMA model is for non-seasonal non-stationary data. Box and Jenkins have generalized this model to deal with seasonality. The theoretical justification for modeling univariate time series of traffic flow data as seasonal ARIMA processes is founded in the time series theorem known as the world decomposition, which applies to discrete time data series that are stationary about their mean and variance. Therefore, it is also necessary to support an assertion that an appropriate seasonal difference will induce stationarity.

The generalized form of SARIMA $(\mathrm{p,d,q}) \times (\mathrm{P,D,Q})_s$model can be written as: $\emptyset_p(B)\emptyset_p(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\theta_Q(B^s)a_t$ (3.8.6)

Where:

$\emptyset_p(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$

$\emptyset_p(B^s) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^P$

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

$\theta_Q(B^s) = 1 - \theta_1 B - \theta_2 B^{2s} - \cdots - \theta_Q B^{Qs}$

Where:

$p, d, and\ q$are the order of non-seasonal AR, differencing and MA respectively.

$P, D, and\ Q$is the order of seasonal AR, differencing and MA respectively.

$Z_t$represents time series data at period t.

B represents backshift operator defined by $BZ_t = Z_{t-1}$.

$(1 - B)^d$represents non-seasonal difference.

$(1 - B^s)^d$represents seasonal difference.

s represents seasonal order (s=12 for monthly data)

$a_t$represents white noise process at period t. It is identically and normally distributed with mean zero, variance σ^2; and $cov\,(e_t, e_{t-k}) = 0\ \forall\ k \neq 0$, that is, $\{e_t\} \sim WN\ (0, \sigma^2)$.

From a practical perspective, fitted seasonal ARIMA models provide linear state transition equations that can be applied recursively to produce single and multiple interval forecasts. Furthermore, seasonal ARIMA models can be readily expressed in state space form, thereby allowing adaptive Kalman filtering techniques to be employed to provide a self-tuning forecast model.

3.9 Model Building for a Time Series

3.9.1 Identification

Model identification refers to the methodology in identifying the required transformation such as variance stabilizing transformation and differencing transformations, the decision to include the deterministic parameter θ_0 when d >= 1 and the proper order of p and q for the model. The purpose of identification is to determine the differencing required to produce stationary and the order of seasonality and non-seasonality of Autoregressive (AR) and Moving Average (MA) operators for the series. Generally, model identification is an explanatory process and analysis done is based upon previous result. Identification consists of specifying the suitable structure Autoregressive Integrated Moving Average (ARIMA) and the order of the model. Identification is sometimes done by looking at the autocorrelation function (ACF) and partial autocorrelation function (PACF) to determine whether the observations are stationary or not. Once stationary is achieved, the second ARIMA parameter d, is simply the number of time series is differenced to achieve stationary. Afterward is the identification of the order of AR and MA, pure AR and MA

processes have characteristics signature in the ACF and PACF. The steps use to identify AR and MA and their orders are simplified in the table.

Characteristic Behavior of ACF, PACF of AR, MA and ARMA Processes

Process	ACF	PACF
AR(p)	Infinite (damped exponentials and /or damped sine waves). Tail off according to $\rho_j = \emptyset_1\rho_{j-1}+\emptyset_2\rho_{j-2}+\ldots+\emptyset_p\rho_{j-p}$	Finite Spike at lag 1 through p, then cut off.
MA(q)	Finite Spike at lag 1 through q, then cuts off	Infinite (dominated by damped exponentials and/or damped sine waves) Tail off.
ARMA(p,q)	Infinite (damped exponentials and/ or damped sine waves after first q-p lags). Irregular pattern at lag 1 through q, then tails off according to $\rho_j = \emptyset_1\rho_{j-1}+\emptyset_2\rho_{j-2}+\ldots+\emptyset_p\rho_{j-p}$	Infinite (dominated by damped exponentials and/ or damped sine waves after first p-q lags). Tail off

Source: Box, G.E.P and Jekins, G.M (1976) "Time Series Analysis Forecasting and Control"

3.9.2 Estimation of Parameters

The second step is to estimate the co-efficient of the model. After the model is identified for a given time series it is important to obtain efficient estimates of the parameters. To obtain the estimate of the parameters, the least squares method and maximum likelihood estimates are used. Maximum Likelihood method is used in this study.

3.9.2.1 Maximum Likelihood Method

The maximum likelihood method has been widely used in estimation.

(a) Conditional Maximum Likelihood Estimation

For the general stationary ARMA *(p,q)* model

$$\hat{Z}_t = \emptyset_1\hat{Z}_{t-1} + \cdots + \emptyset_p\hat{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \qquad (3.9.1)$$

where, $\hat{Z}_t = Z_t - \mu \; and \; (a_t) are$ independently and independently normally distributed random variable with mean zero and variance σ_a^2, the joint probability density of a = $(a_1, a_2, \dots, a_n)$ is given by

$$P(a|\emptyset, \mu, \theta, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} \exp\left[\frac{1}{2\sigma_a^2} \sum_{i=0}^{n} \sigma_a^2\right]. \quad (3.9.2)$$

and then, a_t can be described as

$$a_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + \hat{Z}_t - \emptyset_1 \hat{Z}_{t-1} - \cdots - \emptyset_p \hat{Z}_{t-p}, \quad (3.9.3)$$

These can write the likelihood function of the parameter $(\emptyset, \mu, \theta, \sigma_a^2)$.

Let Z = $(Z_1, Z_2, \dots, Z_n)'$ and assume that the initial conditions

$Z_* = (Z_{1-p}, \dots, Z_{-1}, Z_0)'$ and $a_* = (a_{1-q}, \dots, a_{-1}, a_0)'$ are known. The conditional log-likelihood function is

$$\ln L_*(\emptyset, \mu, \theta, \sigma_a^2) = -\frac{n}{2} \ln 2\pi\sigma_a^2 - \frac{S_*(\emptyset, \mu, \theta)}{2\sigma_a^2}$$

where, $S_*(\emptyset, \mu, \theta) = \sum_{t=1}^{n} a_t^2$ $((\emptyset, \mu, \theta | Z_*, a_*, Z)$ (3.9.4)

is the conditional sum of squares function. The quantities of $\hat{\emptyset}$, $\hat{\mu}$, and $\hat{\theta}$, which maximum equation **(3.68)** are called the conditional maximum likelihood estimators. Because of ln L_* $(\emptyset, \mu, \theta, \sigma_a^2)$ contains the data only through $S_*(\emptyset, \mu, \theta)$, these estimators are the same as the conditional least squares estimators got from minimizing the conditional sum of squares function $S_*(\emptyset, \mu, \theta)$, which don't contain the parameter σ_a^2.

By assuming $a_p = a_{p-1} = \cdots = a_{p+1-q} = 0$ and replacing Z_t by the sample mean $\bar{Z}$ the conditional sum of squares function $S_*(\emptyset, \mu, \theta)$ can be written as becomes

$$S_*(\emptyset, \mu, \theta) = \sum_{t=p+1}^{n} a_t^2(\emptyset, \mu, \theta | Z)$$

After obtaining the parameter estimates $\hat{\emptyset}$, $\hat{\mu}$, and $\hat{\theta}$, the estimate $\hat{\sigma}_a^2$ is calculated from

$$\hat{\sigma}_a^2 = \frac{S_*(\emptyset, \mu, \theta)}{\text{d.f.}}$$

Where the number of degrees of freedom d.f equals the number of terms used in the sum of $S_*(\hat{\emptyset}, \hat{\mu}, \hat{\theta})$ minus the number of parameters estimated.

(b) Unconditional Maximum Likelihood Estimation and Back casting Method

A further improvement in estimation, Box, Jenkins, and Reinsel (1994) suggest the following unconditional log-likelihood function:

$$\ln L(\emptyset, \mu, \theta, \sigma_a^2) = -\frac{n}{2} \ln 2\pi\sigma_a^2 - \frac{S(\emptyset, \mu, \theta)}{2\sigma_a^2} \quad (3.9.5)$$

where S($\emptyset, \mu, \theta$) is the unconditional sum of squares function given by

$$S(\emptyset, \mu, \theta) = \sum_{t=-\infty}^{n} [E(a_t|\emptyset, \mu, \theta, Z)]^2 \quad (3.9.6)$$

And $E(a_t|\emptyset, \mu, \theta, Z)]^2$ is the conditional expectation of a_t given $\emptyset, \mu, \theta,$ and Z.

The quantities $\hat{\emptyset}, \hat{\mu}$ and $\hat{\theta}$ that maximize function **(**3.79**)** are called unconditional maximum Likelihood estimators. Again, Since lnL($\emptyset, \mu, \theta, \sigma_a^2$) involves the data only through S($\emptyset, \mu, \theta$), these unconditional maximum likelihood estimators are equivalent to the unconditional least squares estimators obtained by minimizing S($\emptyset, \mu, \theta$).

3.9.3 Diagnostic Checking

Time series model building is an interactive procedure. It starts with model identification and parameter estimation. After parameter estimation, it has to assess model adequacy by checking whether the model assumptions are satisfied. The basic assumption is that the $\{a_t\}$ are white noise. That is, a_t's are uncorrelated random shocks with zero mean and constant variance. For any estimated model, the residuals a_t's are estimates of this unobserved white noise a_t's. Hence, model diagnostic checking is accomplished through a careful analysis of the residual series $\hat{a}_t$. Because this residual series is the product of parameter estimation, the model diagnostic checking is usually contained in the estimation phase of a time series package.

(1) To check whether the errors are normally distributed, one can construct a histogram of the standardized residuals $\hat{a}_t$ / $\hat{\sigma}_t$ and compare it with the χ.

(2) To check whether the variance is constant, one can examine the plot of residuals or evaluate the effect of different λ value in Box-Cox method.

(3) To check whether the residuals are approximately white noise, one can compute the sample ACF and sample PACF of the residuals to see whether they do not form any pattern and are all statistically insignificant.

The Ljung-Box (Q) test is considered as a diagnostic tool that is used to test the lack of fit of a time series. This uses the entire residual sample ACF's as a unit to check the null hypothesis.

Hypothesis H_0 : $\rho_1 = \rho_2 = \ldots = \rho_k = 0$

H_1 : At least one autocorrelation are not equal.

Test statistic : $Q = n(n+2)\sum_{k=1}^{K} (n-k)^{-1} \hat{\rho}^2_k$

Critical value : $K = \chi^2_{(\alpha, k-m)}$

where, m = the number of parameter estimated in the model.

Based on the residual results, if the model is inadequate, a new model can be easily derived.

Ljung-Box portmanteau Q statistics, Q is the test of null hypothesis which specifies that the ACF does not differ from zero up to lag k. It is evaluated as chi-square with k-m degree of freedom, where k is the number of lags examined and m is the number of parameters estimated. The second test is to examine the ACF and the PACF plot of the first difference of the residual.

The third step is to check the adequacy of the model. This step is also called diagnostic checking or verification. Diagnostic checking consists of evaluating the adequacy of the estimated model. It is important to ensure that the estimated parameters are statistically significant. Usually the model fitting process is guided by the principle of parsimony by which the best mode is the simplest possible model- the model with a fewer parameters- that adequately describe the data. An adequate model satisfies these four conditions:

(a) The estimates of all the parameters must differ significantly from zero

(b) All AR parameter estimates must be within the "bounds of stationary". This guarantees that the model is stationary about its mean.

(i) Bound of stationary

The requirement of the bound of stationary is:

the absolute value of $ø < 1$, $(-1 < ø < 1)$. If $ø = 1$, it becomes ARIMA (0,1,0) which is non-stationary. If $ø > 1$, the past values of Y_{t-k} and e_{t-k} have greater and greater influence on Y_t, it implies the series is non-stationary with an ever increasing mean. To sum up, if Bound of Stationary does not hold, the series is no autoregressive; it is either drifting or trending, and first-difference should be used to model the series with stationary.

Auto-regressive Process: ARIMA (p,0,0)

$Y_t = \theta + ø_1 Y_{t-1} + ø_2 Y_{t-2} + \ldots + ø_p Y_{t-p} + e_t$ (or)

$Y_t = ø_1 Y_{t-1} + ø_2 Y_{t-2} + \ldots + ø_p Y_{t-p} + e_t$

(c) All MA parameters estimate must lie within the bounds of "invertibility". This is the MA along to stationary to AR model. Where the model is re-express as infinite series as AR terms, inevitability guarantees that this series converges.

(ii) Bound of Invertibility

The requirement of the bound of invertibility is:

the absolute value of θ is less than 1 ($-1 < \theta < 1$). If not hold, the model is non-stationary. Moving Average Process: ARIMA (0,0, q)

$$Y_t = \theta + ø_1 e_{t-1} + ø_2 e_{t-2} + \ldots + ø_q e_{t-q}$$

The important feature of ARIMA (0,0, q) is that the variables of e_{t-1} to e_{t-q} are unobserved and have to estimated using the available sample data. In practice, it is usual to keep q at a small value, and it is often set at 1 or 2.

For models of the same orders, that is AR (i) and MA (j), the bounds of invertibility place limit on that are identical to those on by the bounds of invertibility.

(d) Residual must not differ significantly from a series of purely random error (White noise) with mean zero. For White noise the theoretical ACF and PACF are both zero at all lags. For residual, the calculated standard error tends to over-estimate the true standard error (Monserud, 1986).The simplest way of checking the best model is to use goodness of fit statistics such as the real adjusted R-square mean absolute error, sum of square of error normalize BIC (Bayesian Information Criteria) and the residual plot of ACF and PACF. In summary, the best model is the one with relatively small of BIC, relatively small of mean absolute error, relatively small of sum of squares error, relatively high adjust R-square and Random pattern of the plot of the ACF and PACF.

3.9.3.1 Model Selection Criteria

Numerous criteria for model comparison are introduced in this selection. Model identification tools such as ACF and PACF used only for identifying adequate models. For a given data set, when there are multiple adequate models, the selection criterion is normally based on summary statistics from residuals computed from a fitted model. Some model selection criteria are based on residuals, BIC (Bayesian's Information Criteria), AIC (Akaike's Information Criteria).

In this study, coefficient of determination (R^2 or r^2) and Mean Absolute Percent Error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics, the Bayesian information criterion (BIC) and the estimated parameters are used for criteria of tentative ARIMA model. The higher value of R^2 and the lower value of MAPE and normalized BIC and significance of the parameters are used as criteria for selecting the most adequate model among all feasible models.

3.9.4 Forecasting

Forecasting is the process of estimating future based on the analysis of past and present data or behaviour. Time series data is important in predicting something which is changing over the time using past data. The goal of time series analysis is to estimate the future value using the behaviours in the past data.

3.9.4.1 Minimum Mean Square Error Forecasts for ARIMA Models

The general nonstationary ARIMA (p, d, q) model with $d \neq 0$, i.e.,

$\emptyset(B)(1-B)^d Z_t = \theta(B)\, a_t$,

Where $\emptyset(B) = (1 - \emptyset B - \ldots - \emptyset_p (B^{\emptyset})$ is a stationary AR operator and $\theta(B) = (1 - \theta_1 B - \ldots - \theta_q (B^q)$ is an invertible MA operator, respectively. Although for this process the mean and the second order moment such as the variance and the autocovariance functions vary over time, the complete evolution of the process is completely determined by a finite number of fixed parameters. Hence, the forecast of the process can be viewed as the estimation of a function of these parameters and obtain the minimum mean square error forecast using a Bayesian argument. It is well known that using this approach with respect to the mean square error criterion, which corresponds to a squared loss function, when the series is known up to time n, the optimal forecast of Z_{n+l} is given by its conditional expectation $E(Z_{n+l} \; Z_n, Z_{n-1}, \ldots)$. The minimum mean square error forecast for the stationary ARMA model discussed earlier is, of course, a special case of the forecast for the ARIMA (p, d, q) model with $d = 0$.

To derive the variance of the forecast for the general ARIMA model, we rewrite the model at time t + l in an AR representation that exist because the model is invertible, Thus,

$$\pi(B)\, Z_{t+l} = a_{t+l}, \qquad (3.9.7)$$

Where

$$\pi(B) = 1 - \sum_{i=0}^{n} \pi_j B^j = \frac{\emptyset(B)(1-B)^d}{\theta(B)} \qquad (3.9.8)$$

or, equivalently

$$Z_{t+l} = \sum_{i=0}^{n} \pi_j Z_{t+l\sim j} + a_{t+l} \qquad (3.9.9)$$

Following Wegman (1986), we apply the operator

$$1+ \psi B + \ldots + \psi_{l-1} B^{t-1} \quad (3.9.10)$$

To (3.9.10) and obtain

$$\sum_{j=0}^{\infty} \sum_{k=0}^{l-1} \pi_j Z_{t+l-j-k} + \sum_{k=0}^{l-1} \psi_k \ a_{t+l-1} = 0, \quad (3.9.11)$$

Where π_0 = -1 and ψ_0 = 1. It can be easily shown that

$$\sum_{j=0}^{\infty} \sum_{k=0}^{l-1} \pi_j Z_{t+l-j-k} = \pi_0 Z_{t+l} + \sum_{m=1}^{l-1} \sum_{i=0}^{m} \pi_{m-i} \psi_i Z_{t+l-m} + \sum_{j=1}^{\infty} \sum_{i=0}^{l-1} \pi_{l-1+j-i} \psi Z_{t-j+1}. \quad (3.9.12)$$

Choosing ψ weights so that

$$\sum_{i=0}^{m} \pi_{m-i} \psi_i = 0, \text{ for } = 1,2, \ldots, l\text{-}1 \quad (3.9.13)$$

We have

$$Z_{t+l} = \sum_{j=1}^{\alpha} \pi_j^{(l)} Z_{t-j+1} + \sum_{i=0}^{l-1} \psi_i \ a_{t+l-i}, \quad (3.9.14)$$

Where

$$\pi_j^{(l)} = \sum_{i=0}^{l-1} \pi_{l-1+j-1} \psi_i. \quad (3.9.15)$$

Thus, given Z_t, for $t \leq n$, we have

$$\hat{Z}_n(l) = E (Z_{n+l} \mid Z_t, t \leq n)$$

$$= \sum_{j=1}^{\alpha} \pi_j^{(l)} Z_{n-j+1}, \quad (3.9.16)$$

Because E ($a_{n+j} \mid Z_t, t \leq n$) = 0, for j > 0. The forecast error is

$$e_n (l) = Z_{n+1} - \hat{Z}_n(l)$$

$$= \sum_{i=0}^{l-1} \psi_j \ a_{n+l-j}, \quad (3.9.17)$$

Where the ψ_j weights, by (3.9.13), can be calculated recursively from the π_j weights as follow:

$$\psi_j = \sum_{i=0}^{j-1} \pi_{j-i} \psi_i, \ j = 1, \ldots, l-1. \quad (3.9.18)$$

Because E($e_n(l) \mid Z_t, t \leq n$) = 0, the forecast is unbiased with the error variance

$$\text{Var } (e_n(l)) = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2. \quad (3.9.19)$$

For a normal process, the (1- α) 100% forecast limits are

$$\hat{Z}_n(l) \pm N_{\alpha/2} [1 + \sum_{j=0}^{l-1} \psi_j^2]^{1/2} \sigma_a \quad (3.9.20)$$

Where $N_{\alpha/2}$ is the standard normal deviate such that P(N > $N_{\alpha/2}$) = $\alpha/2$.

3.9.4.2 Model Building and Forecasting for Seasonal Model

Seasonal models are special forms of ARIMA models, the model identification, parameter estimation, diagnostic checking, and forecasting for these models follow the same general methods of ARIMA models which are already introduced in this chapter.

CHAPTER IV
DATA ANALYSIS AND FORECASTING OF MALARIA INFECTION IN KACHIN STATE

This paper is to study the incidence of malaria in Kachin State from January 2011 to December 2016. The monthly data are kindly supported from National Malaria Control program (NMCP) Myanmar. These data are analyzed by using Box-Jenkins Method. There are four steps in this method which are model identification, parameters estimation, diagnostic checking and forecasting.

The data of malaria infection cases was used as a platform for creating the ARIMA models. ARIMA models are used in stationary time series analysis. Theoretically, if the mean of the series and the covariance among its observations do not change over time and do not follow any trend, a time series is said to be stationary. Practically, most time series are non-stationary.

In order to fit stationary models, it is indispensable to get rid of the non-stationary source of variation. ARIMA models are one of the solutions to overcome the limitation of being non-stationary. Moreover, ARIMA models are the best models for forecasting a time series in theory. Box and Jenkins have generalized this model to deal with seasonality namely Seasonal ARIMA (SARIMA) model.

In term of econometric methodology, whether time series are stationary or not is needed to determine. The correlogram is used to test stationary in this study. By observation, the data are not stationary because the series display a long-term pattern and the mean is not zero. On the other hand, the data have the seasonality as well. That is why, the seasonality test is conducted to the monthly figures using ANOVA.

Firstly, malaria infection data are described and analyzed. And then, the model building processes are made for the study. The ACF and PACF function are used to identify the order p and q of the model in the identification stage. In term of the nature of p and q of the model, six feasible ARIMA models are occurred. The coefficient of determination (R^2 or r^2), mean absolute percentage error (MAPE), the Bayesian information criterion (BIC) and the parameters of the model are used as selection criteria for the most fitted model of malaria incidence data in Kachin State.

4.1 Descriptive Statistics of the Volumes of Malaria Infection in Kachin State (January 2011 to December 2016)

The volume of incidence, minimum, maximum and average of monthly malaria infection in Kachin State from January 2011 to December 2016 are shown in Table (4.1). Over the six-year period, the table indicates that the average number of malaria infected people in 2011 and 2012 are the peak and its of 2016 is the least. It shows that the malaria infection in Kachin State are declined. The maximum and minimum infection of 2011 are 5482 and 1166 and its of 2012 are 6932 and 1262.

Comparing every month of a year, the average incidence of malaria is peak on June and July. Over the six-year period of June, the maximum and minimum infection are 5327 and 802. The maximum malaria infected month is every July of the study years. The maximum infection of those month is 6932 and its minimum infection is 539.

Table (4.1) Volumes of Malaria Infection in Kachin State

Year / Month	2011	2012	2013	2014	2015	2016	Min	Max	Average
Jan	1890	1797	1478	724	489	221	221	1890	1100
Feb	1441	1545	1167	733	465	179	179	1545	922
Mar	1166	1262	925	621	407	132	132	1262	752
Apr	2109	1579	1057	617	454	293	293	2109	1018
May	3458	2515	1867	972	705	434	434	3458	1659
Jun	5327	5181	4228	1949	1400	802	802	5327	3148
Jul	5482	6932	4024	2301	1537	539	539	6932	3469
Aug	4872	5514	3855	2063	1302	334	334	5514	2990
Sep	4328	4141	2977	1349	970	208	208	4328	2329
Oct	3517	3022	2143	1013	719	131	131	3517	1758
Nov	2971	2292	1862	877	604	138	138	2971	1457
Dec	2261	1805	1187	769	409	106	106	2261	1090
Min	1166	1262	925	617	407	106	106	1262	747
Max	5482	6932	4228	2301	1537	802	802	6932	3547
Average	3235	3132	2231	1166	788	293	293	3235	1808

Sources: National Malaria Control Program (NMCP), Nay Pyi Taw

Figure (4.1) indicates that the series has a seasonal pattern with peaks and valley in the same month of the year. There are regular patterns with different magnitudes. Therefore, the test for seasonality of the Malaria Infection in Kachin State data series is conducted as follows.

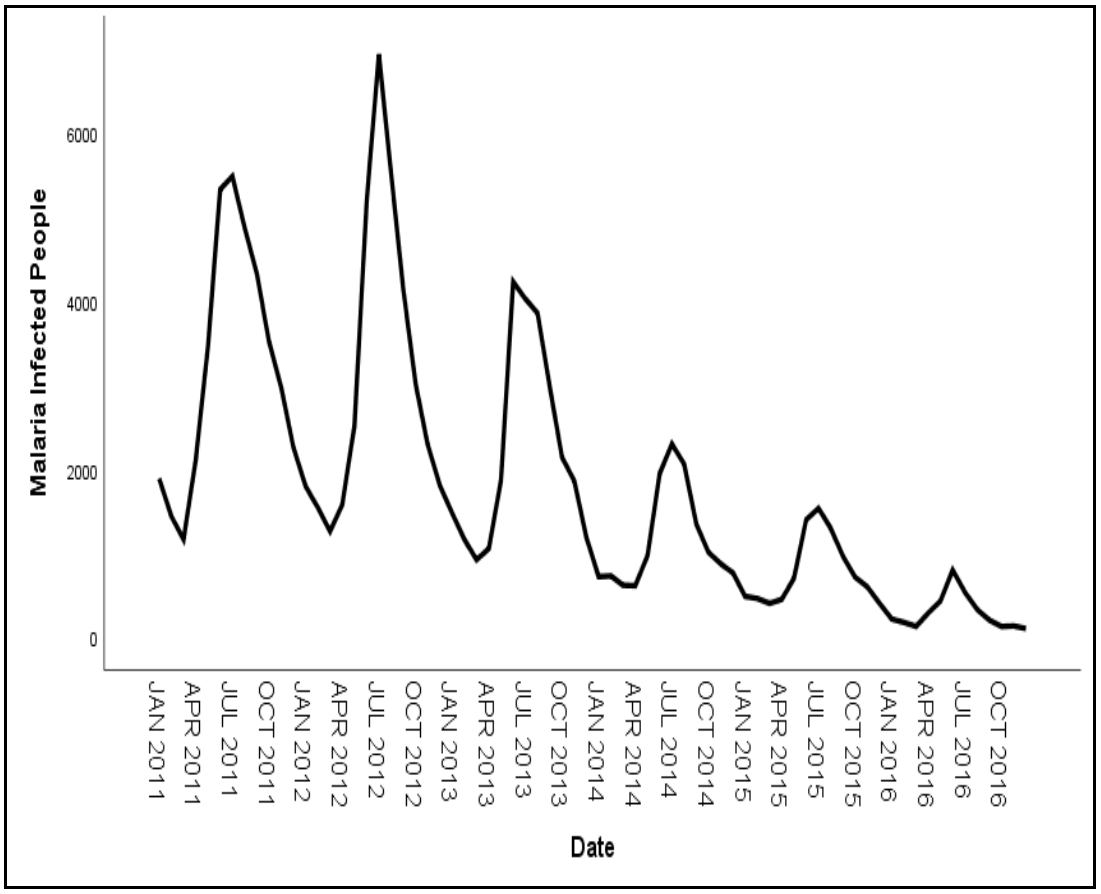


Figure (4.1) Volumes of Malaria Infection in Kachin State

4.2 Test of Seasonality for Malaria Infection in Kachin State

Seasonality are testing by using ANOVA table. The results for testing the seasonality in number of malaria incidence in Kachin State from 2011 to 2016 are illustrated in Table (4.2).

Hypothesis

Null Hypothesis : H_0 = There is no seasonality (no monthly effect).

Alternative Hypothesis : H_1 = There is seasonality (monthly effect).

Table (4.2) ANOVA Table for Malaria Incidence in Kachin State

	Sum Square	Degree of Freedom	Mean Square	F-ratio
Due to Month	59477005.71	11	5407000.519	12.27721
Due to Year	92590835.13	5	18518167.03	42.0476
Error	26864983.04	61	440409.5581	
Total	178932823.9	77		

In term of the above ANOVA table, at 5 % level of significance, the critical value of $K=F_{(0.05,11,61)}$ is 1.9174. The computed value of F is 12.27721 which is greater than K-value (1.9174). According to the decision rule, the result is lead to reject H_0 which is no monthly affect in data. Therefore, there is seasonality in Malaria Infection of Kachin State data series. As seasonality is existed in data series, seasonal index is calculated as follow.

Table (4.3) Seasonal Index for Malaria Infection in Kachin State

Month	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**mean**	**Seasonal Index**
Jan		0.55	0.58	0.46	0.53	0.43	0.51	0.4703
Feb		0.47	0.48	0.51	0.53	0.41	0.48	0.4422
Mar		0.39	0.39	0.47	0.48	0.35	0.41	0.3827
Apr		0.49	0.46	0.50	0.55	0.87	0.57	0.5293
May		0.80	0.83	0.82	0.88	1.42	0.95	0.8765
Jun	1.65	1.66	1.92	1.69	1.80	5.47	2.37	2.1851
Jul	1.70	2.24	1.87	2.03	2.04		1.98	1.8249
Aug	1.50	1.80	1.82	1.85	1.78		1.75	1.6179
Sep	1.34	1.37	1.43	1.23	1.36		1.35	1.2428
Oct	1.11	1.02	1.05	0.94	1.04		1.03	0.9524
Nov	0.95	0.79	0.98	0.84	0.92		0.90	0.8270
Dec	0.71	0.66	0.68	0.77	0.69		0.70	0.6505
Total							**12.99**	**12**

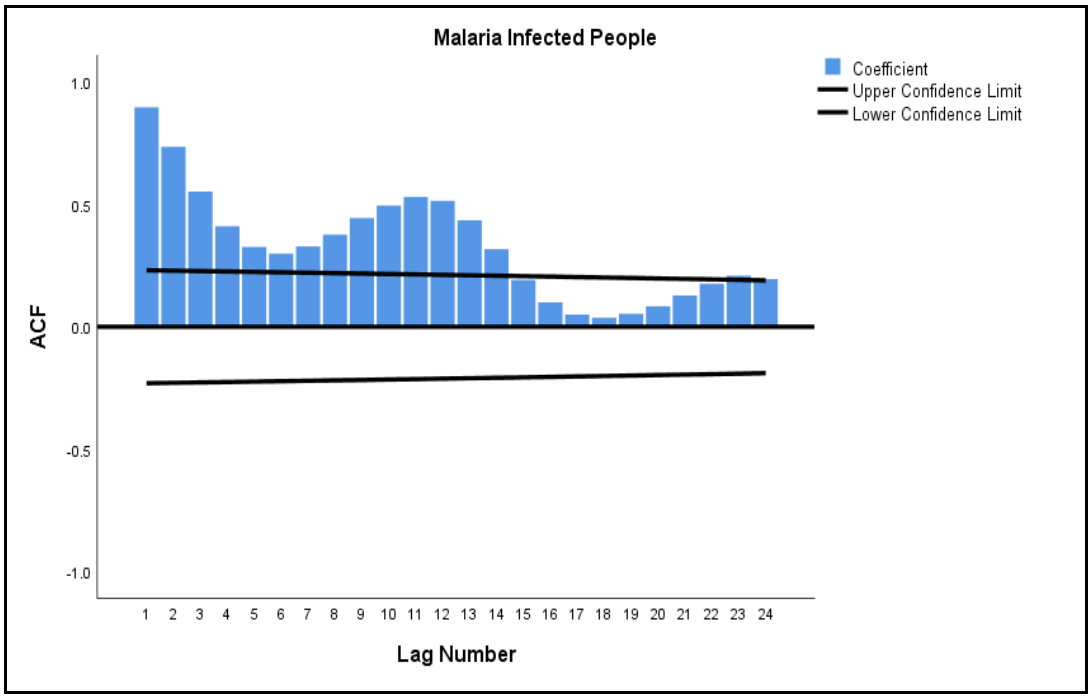
The seasonal index of the data is shown in Table (4.3). The seasonal index is a ratio and it has an average value of 1. In this study, the seasonal index of January to

May and October to December are lower than 1. The seasonal index for June is 2.1851, it means that June is 218 percent of monthly average. The seasonal index for malaria infection was highest on June in Kachin State. Seasonal indices of July, August and September are 1.825, 1.618 and 1.243 respectively which mean July is 183 percent, August is 162 percent and September is 124 percent of monthly average. It means that Malaria infection in Kachin State from June to September of every year is greater than monthly average which indicates that the data has seasonality.

After proving the original data series has seasonality and finding the seasonal index, the ARIMA model building is performed as follow.

4.3 Model Identification for Malaria Infection in Kachin State

Model identification refers to the methodology in identifying the required transformations. The sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) are shown in Table 4.4 (a) and (b). The correlograms for sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF) were found for the original series Z_t and displayed in Figure (4.2) along with the confidence limits.



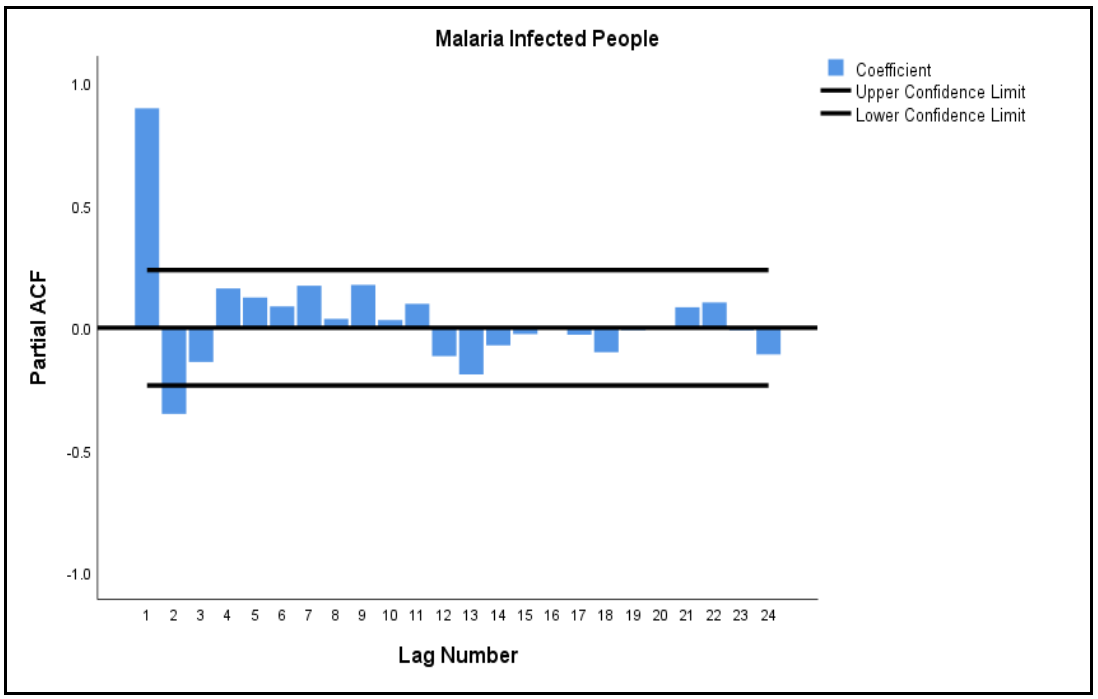


Figure (4.2) The Correlograms for Sample ACF and PACF of Malaria Incidence in Kachin State

According to the above Figure (4.2), the ACF of original series Z_t shows damped sine waves and partial autocorrelation function (PACF) cuts off after lag 2 of malaria incidence in Kachin State. Therefore, Malaria Infection in Kachin State data series seems to fit an AR (2) model.

However, the series may not be stationary both in the mean and the variance. The plot of monthly malaria infection in Kachin State between January 2011 to December 2016 indicates both that the mean level depends on time and that variance decreases as the mean level decreases. The Log transformation is needed to make stationarity in variance.

Table (4.4) Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) of Malaria Incidence in Kachin state (Z_t)

(a) $\hat{\rho}_k$ For $\{Z_t\}$

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
Z_t	1-12	0.886	0.656	0.415	0.227	0.113	0.060	0.076	0.150	0.276	0.440	0.595	0.658
	St-E	0.115	0.115	0.114	0.113	0.112	0.111	0.110	0.110	0.109	0.108	0.107	0.106
	13-24	0.579	0.399	0.194	0.037	-0.054	-0.090	-0.081	-0.041	0.027	0.116	0.202	0.234
	St-E	0.105	0.104	0.103	0.103	0.102	0.101	0.100	0.099	0.098	0.097	0.096	0.095

(b) $\hat{\emptyset}_{kk}$ For $\{Z_t\}$

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
Z_t	1-12	0.886	-0.594	0.104	0.073	0.016	0.007	0.225	0.111	0.251	0.273	0.133	-0.243
	St-E	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
	13-24	-0.237	-0.022	-0.033	0.066	-0.030	-0.064	-0.081	-0.104	-0.081	-0.085	-0.005	-0.031
	St-E	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118

To remove non stationarity, logarithms transformation is performed to stabilize the variance of a time series. The transformed volumes of Malaria Infection in Kachin State are described in Figure (4.3)

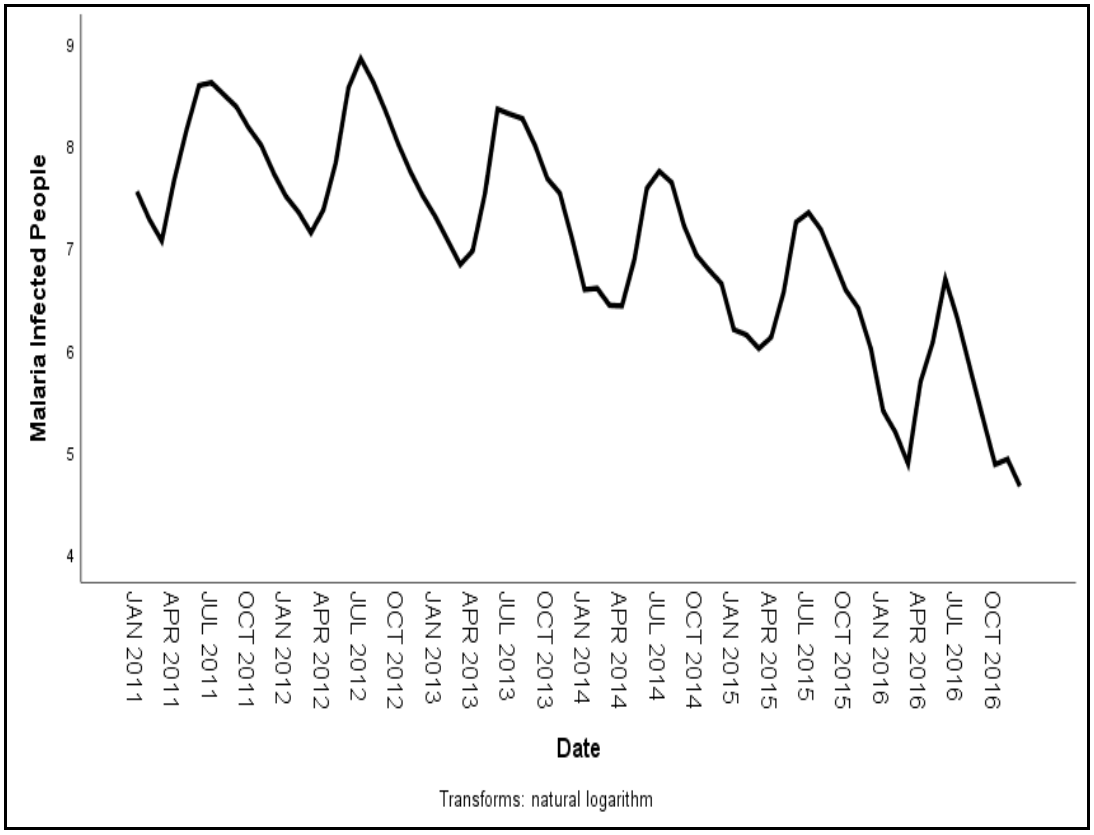


Figure (4.3) Transformed Volumes of Malaria Infection in Kachin State

The sample ACF and sample PACF for natural logarithms series Ln (Z_t) are shown in Table 4.5 (a) and (b) and they were displayed in Figure (4.4).

Table (4.5) Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (PACF) for Natural Logarithms of Malaria Incidence in Kachin state Ln (Z_t)

(a) $\hat{\rho}_k$ For Ln (Z_t)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
Ln (Z_t)	1-12	0.897	0.736	0.553	0.411	0.326	0.299	0.328	0.377	0.445	0.496	0.531	0.515
	St-E	0.115	0.115	0.114	0.113	0.112	0.111	0.110	0.110	0.109	0.108	0.107	0.106
	13-24	0.435	0.317	0.192	0.100	0.050	0.037	0.053	0.084	0.128	0.176	0.209	0.195
	St-E	0.105	0.104	0.103	0.103	0.102	0.101	0.100	0.099	0.098	0.097	0.096	0.095

(b) $\hat{\emptyset}_{kk}$ For Ln (Z_t)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
Ln (Z_t)	1-12	0.897	-0.352	-0.140	0.161	0.124	0.087	0.172	0.037	0.175	0.031	0.098	-0.116
	St-E	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
	13-24	-0.191	-0.072	-0.026	-0.004	-0.029	-0.100	-0.010	-0.006	0.084	0.103	-0.010	-0.109
	St-E	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118

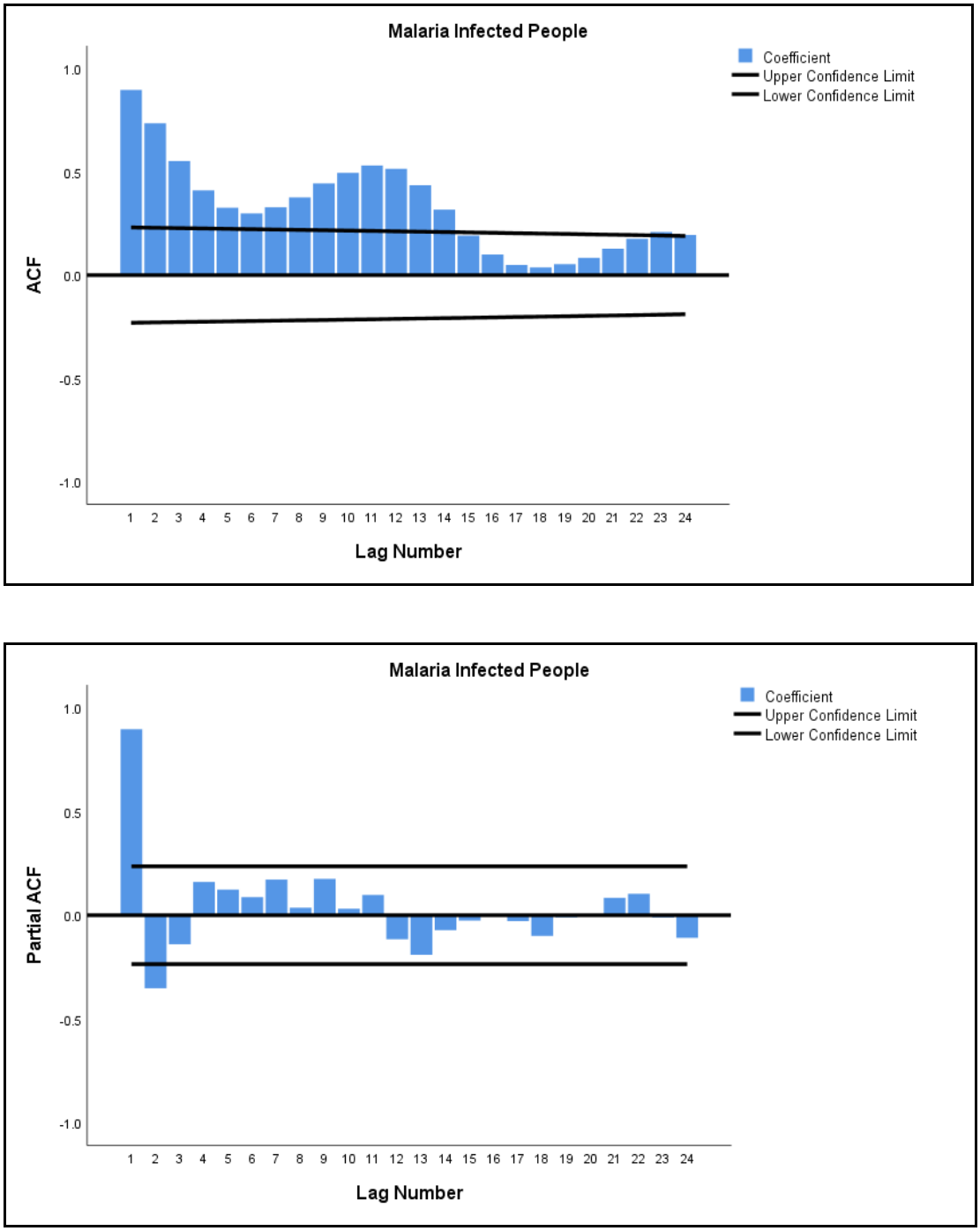


Figure (4.4) The Correlograms of Sample ACF and PACF for Natural Logarithms Series of Malaria Incidence in Kachin State

The ACF of Ln (Z_t) series show damped sine wave and there is significant spike of PACF at Lag 1 and 2. Therefore, the log transformed data series seems to fit an AR (2) Model. But the log transformed series is not stationary in the mean. To remove non-stationary, the log transformed data series is needed to compute as seasonal differencing because of existing seasonality. It means that it leads to remove

changes in the level of the time series. The seasonal differenced log transformed volumes of Malaria Infection in Kachin State are shown in Figure (4.5).

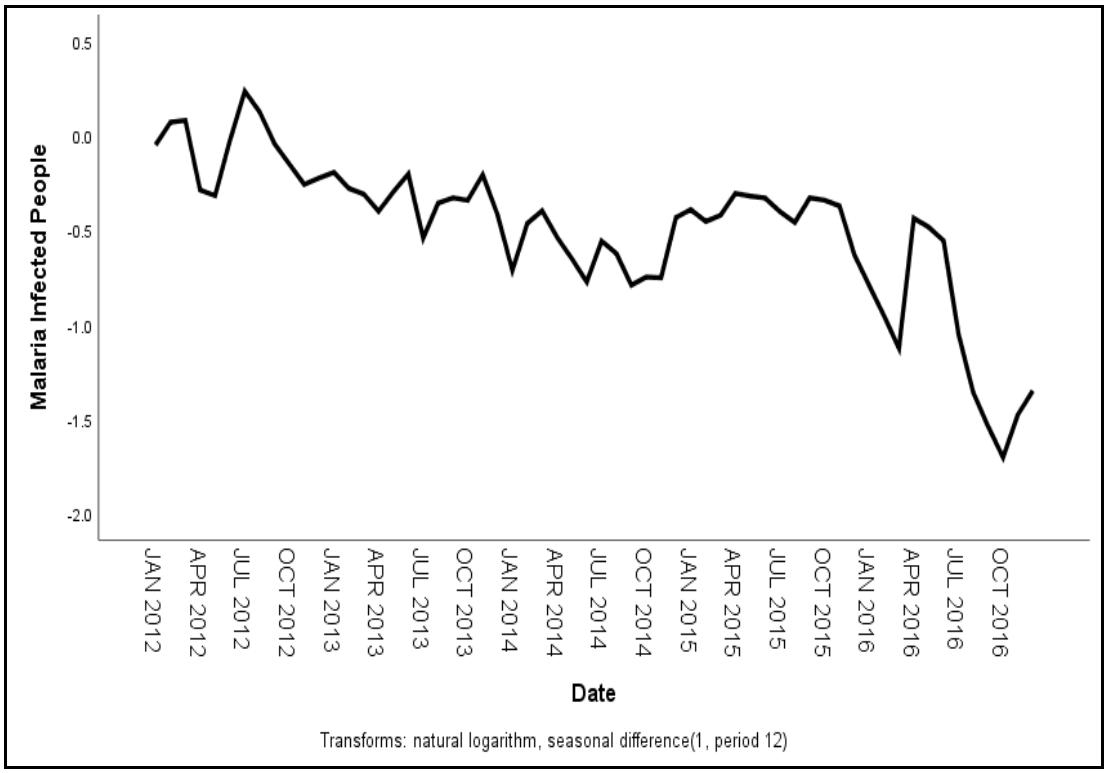


Figure (4.5) Seasonal Differenced Log Transformed Volumes of Malaria Infection in Kachin State

The sample ACF and PACF of seasonal differenced log transformed series are shown in Table 4.6 (a) and (b) and they are displayed in Figure 4.6.

Table (4.6) Sample Autocorrelation Function (ACF) and Sample Partial Autocorrelation Function (ACF) of Seasonal Difference Transformed Series for Malaria Incidence in Kachin State (W_t)

(a) $\hat{\rho}_k$ For W_t from $(1-B^{12})$ Ln (Z_t)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
W_t	1-12	0.840	0.642	0.468	0.377	0.348	0.301	0.274	0.241	0.208	0.113	0.033	-0.029
	St-E	0.126	0.125	0.124	0.123	0.122	0.120	0.119	0.118	0.117	0.116	0.115	0.114
	13-24	-0.039	-0.017	-0.012	-0.007	-0.041	-0.048	-0.025	0.000	0.020	0.017	0.033	0.064
	St-E	0.112	0.111	0.110	0.109	0.108	0.106	0.105	0.104	0.102	0.101	0.100	0.098

(b) $\hat{\emptyset}_{kk}$ For W_t from $(1-B^{12})$ Ln (Z_t)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
W_t	1-12	0.840	-0.212	-0.029	0.161	0.089	-0.122	0.113	0.000	-0.035	-0.240	0.086	-0.042
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129
	13-24	0.045	0.011	0.004	0.000	-0.090	0.089	0.133	-0.072	-0.007	-0.009	0.077	0.033
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129

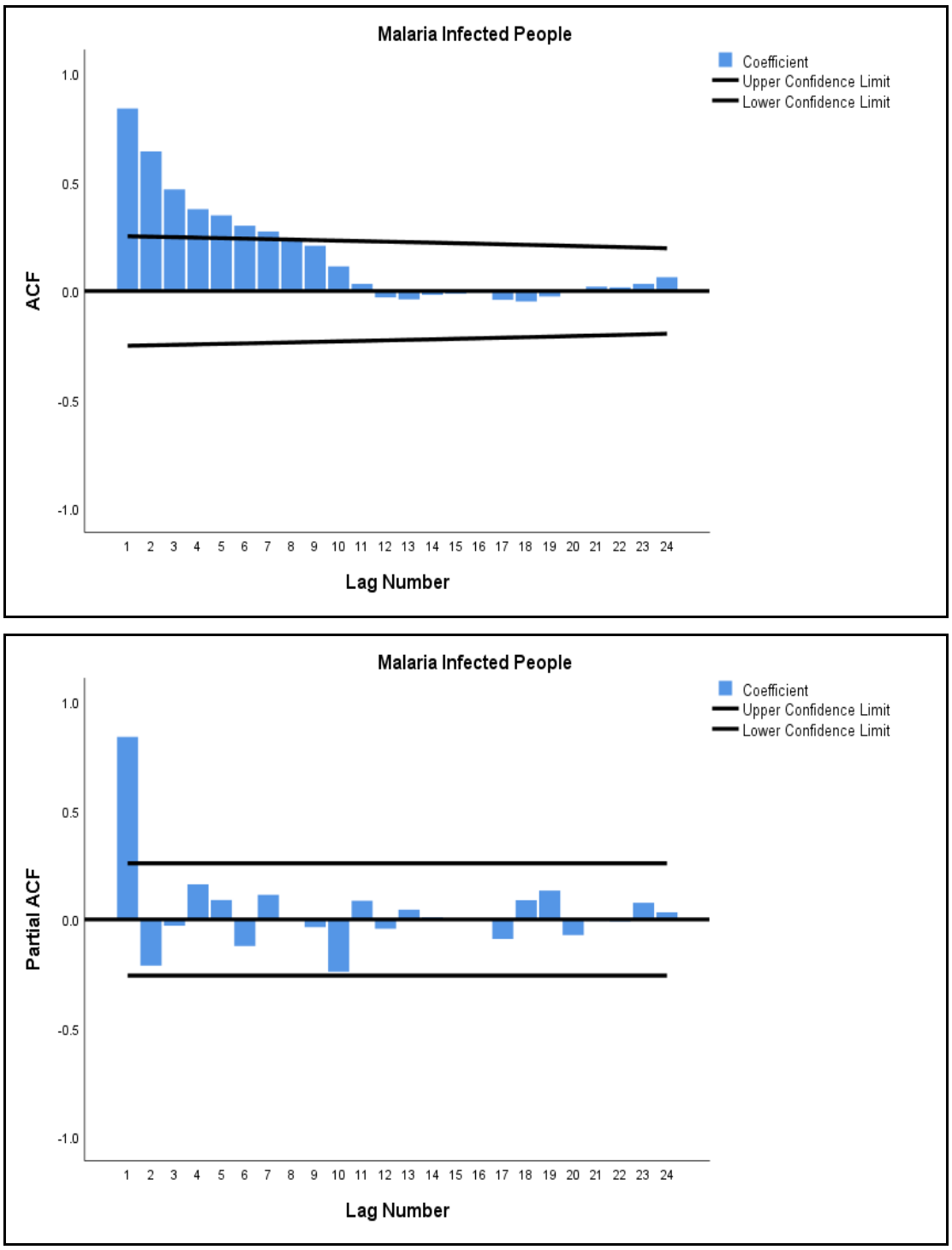


Figure (4.6) The Autocorrelation and Partial Autocorrelation Function of Seasonal Transformed Series for Malaria Incidence in Kachin State.

As the ACF decays at multiples of seasonal period 12 implies that a seasonal differencing $(1-B^{12})$ is made to achieve stationary. As a result of this, the sample ACF decays exponantially and sample PACF is significant spike at only Lag 1 which suggest SARIMA (1,0,0) x $(0,1,0)_{12}$ may be for this data series. Therefore, estimated parameters are needed to figure out for the tentative model.

4.4 Parameter Estimation for SARIMA (1,0,0) x $(0,1,0)_{12}$ Model

The parameters of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Infection in Kachin State are described in Table (4.7).

Table (4.7) Model Parameters of SARIMA (1,0,0) x $(0,1,0)_{12}$ model for Malaria Incidence in Kachin state

ARIMA Model Parameters							
Malaria Infected People-Model_1				Estimate	SE	T	Sig.
Malaria Infected People	Natural Logarithm	Constant		-0.547	0.244	-2.238	0.029
		AR	Lag 1	0.908	0.063	14.528	0
		Seasonal Difference		1			

It can be seen that the estimated parameter of $\emptyset_1$ is 0.908 which is less than one, supporting the required stationary and invertibility condition. Since its p-value is zero, there is no evidence to reject the null hypothesis. It is statistically significant at 0.01 significance level.

SARIMA (1,0,0) x $(0,1,0)_{12}$ model is

$(1 - \emptyset_1 B)\ (1-B^{12})\ \text{Ln}\ (Z_t) = \theta_0 + a_t$

the estimated model is

$(1-0.908)\ (1-B^{12})\ \text{Ln}\ (Z_t) = -0.547 + a_t$

With respect to the estimated results of malaria infected people in Kachin State, the feasible model is SARIMA (1,0,0) x $(0,1,0)_{12}$ model for the data series of Malaria Incidence in Kachin State.

4.5 Diagnostic Checking for SARIMA (1,0,0) x $(0,1,0)_{12}$ Model

The estimated residual ACFs and PACF for the above model are illustrated in Table 4.8 (a) and (b).

Table (4.8) Estimated Residual ACFs and PACFs of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Incidence in Kachin State

(a)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
ρ_k	1-12	0.153	-0.108	-0.228	-0.138	0.206	0.107	0.051	0.018	0.025	0.016	-0.009	-0.140
	St-E	0.129	0.132	0.134	0.140	0.142	0.147	0.148	0.149	0.149	0.149	0.149	0.149
	13-24	-0.038	0.067	-0.005	0.102	-0.082	-0.107	-0.008	0.019	0.045	-0.082	-0.044	-0.019
	St-E	0.151	0.151	0.152	0.152	0.153	0.153	0.155	0.155	0.155	0.155	0.156	0.156

(b)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{\emptyset}_{kk}$	1-12	0.153	-0.134	-0.197	-0.092	0.211	-0.018	0.029	0.090	0.096	-0.016	0.008	-0.137
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129
	13-24	-0.009	0.024	-0.097	0.086	-0.056	-0.071	0.024	0.043	-0.039	-0.088	0.031	-0.036
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129

The residuals of ACF and PACF for the tentative SARIMA (1,0,0) x $(0,1,0)_{12}$ Model are described in Figure (4.7).

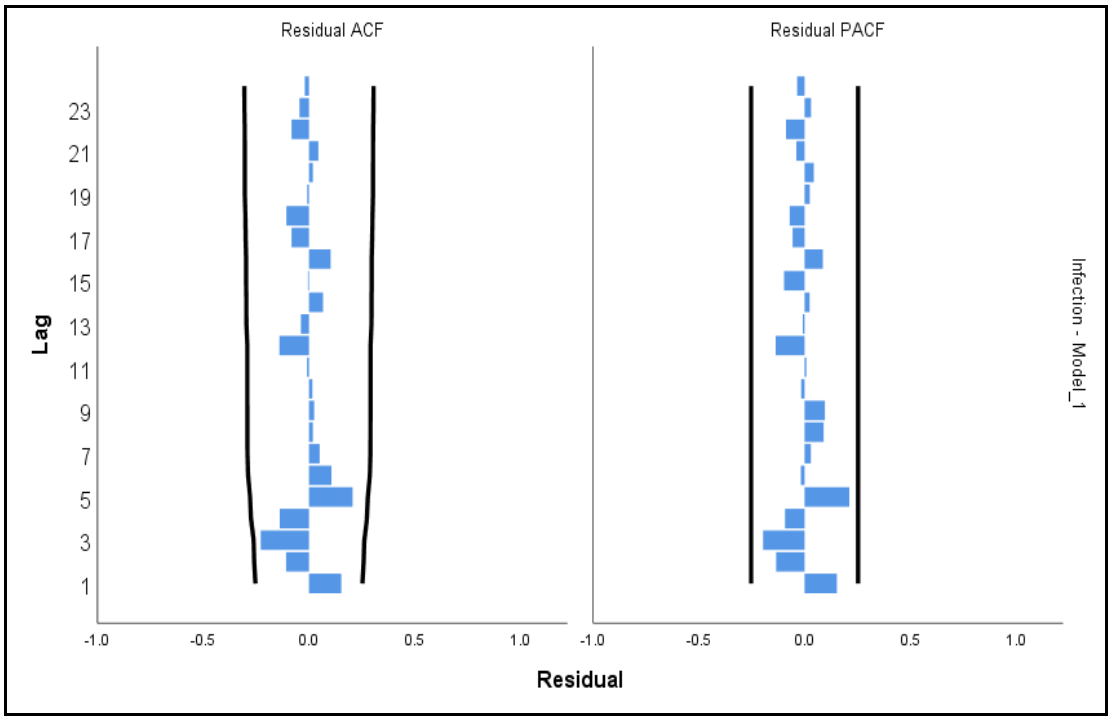


Figure (4.7) The Autocorrelation and Partial Autocorrelation Function of Residuals for SARIMA (1,0,0) x $(0,1,0)_{12}$ Model

According to Figure (4.7), the residuals values of the ACF and PACF for the Malaria Incidence are all small and fall within the two limits, lower confidence limit (LCL) and upper confidence limit (UCL) as well as exhibit no pattern, it can be said that the residual series are white noise process. It means SARIMA (1,0,0) x $(0,1,0)_{12}$ model is adequate to represent the seasonally log transformed data series of Malaria Incidence in Kachin State.

On the other hand, the autocorrelation among residuals are checked by using the test statistic (Q).

$H_0 : \rho_1 = \rho_2 = \ldots = \rho_k = 0$ (There is no autocorrelation among residuals.)

The detail summary residual values for SARIMA (1,0,0) x $(0,1,0)_{12}$ model is shown in Table (4.9).

Table (4.9) Model Statistics of SARIMA (1,0,0) x $(0,1,0)_{12}$ Model for Malaria Incidence in Kachin State

Model Statistics									
Model	Number of Predictors	Model Fit statistics				Ljung-Box Q (18)			Number of Outliers
		Stationary R-squared	R-squared	MAPE	Normalized BIC	Statistics	DF	Sig.	
Malaria Infected People-Model_1	0	.759	.920	14.977	12.192	15.268	17	.576	0

As the result of above table, the observed value of Q is 15.268 and it is not significant since p-value is 0.576 which is greater than 0.05. It means there is no autocorrelation among residuals. Thus, the model SARIMA (1,0,0) x $(0,1,0)_{12}$ is adequate.

In addition, another possible model, SARIMA (1,0,0) x $(1,1,0)_{12}$ Model, to represent seasonal log transformed Malaria Incidence in Kachin State data series is also fitted as follow.

4.6 Parameter Estimation for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model

The parameters of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State are shown in Table (4.10).

Table (4.10) Model Parameters of ARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State

ARIMA Model Parameters							
Malaria Infected People-Model_1				**Estimate**	**SE**	**T**	**Sig.**
Malaria Infected People	Natural Logarithm	Constant		-0.536	0.238	-2.252	0.028
		AR	Lag 1	0.927	0.061	15.250	0.000
		AR, Seasonal	Lag 1	-0.300	0.179	-1.676	0.099
		Seasonal Difference		1			

The results are providing that the seasonal autoregressive order one, $Ø_1$ was estimated to be 0.927 (SE= 0.061) and seasonal $Φ_1$ was estimated to be -0.300 (SE= 0.179) and statistically significant at 0.1 significance level. The estimated parameters of $Ø_1$ and $Φ_1$ are less than one, supporting the required stationary and invertibility conditions.

SARIMA (1,0,0) x $(1,1,0)_{12}$ model is

$(1 - Φ_1 B^{12})\ (1 - Ø_1 B)\ (1-B^{12})\ \text{Ln}\ (Z_t) = \theta_0 + a_t$

the estimated model is

$(1 + 0.3B^{12})\ (1-0.927)\ (1-B^{12})\ \text{Ln}\ (Z_t) = -0.536 + a_t$

With respect to the estimated results of malaria infected people in Kachin State, SARIMA (1,0,0) x $(1,1,0)_{12}$ model is a tentative model for data series of Malaria Incidence in Kachin State.

4.7 Diagnostic Checking for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model

The estimated residuals of ACF and PACF for the tentative SARIMA (1,0,0) x $(1,1,0)_{12}$ Model are described in Table 4.11 (a) and (b).

Table (4.11) Estimated Residual ACFs and PACFs of SARIMA (1,0,0) x (1,1,0)$_{12}$ Model for Malaria Incidence in Kachin State

(a)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
ρ_k	1-12	0.138	-0.106	-0.238	-0.123	0.189	0.114	0.076	0.001	0.034	-0.059	-0.002	-0.009
	St-E	0.129	0.132	0.133	0.140	0.142	0.146	0.147	0.148	0.148	0.148	0.149	0.149
	13-24	-0.005	0.044	-0.077	0.098	-0.003	-0.073	-0.048	0.016	0.040	-0.068	-0.050	-0.071
	St-E	0.149	0.149	0.149	0.149	0.150	0.150	0.151	0.151	0.151	0.152	0.152	0.152

(b)

Lag	k	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{\emptyset}_{kk}$	1-12	0.138	-0.127	-0.211	-0.079	0.184	0.003	0.053	0.067	0.111	-0.081	0.032	-0.012
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129
	13-24	-0.031	-0.003	-0.084	0.123	-0.035	-0.088	-0.004	0.073	-0.049	-0.114	0.006	-0.032
	St-E	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129	0.129

The residuals ACF and PACF of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model are shown in Figure (4.8).

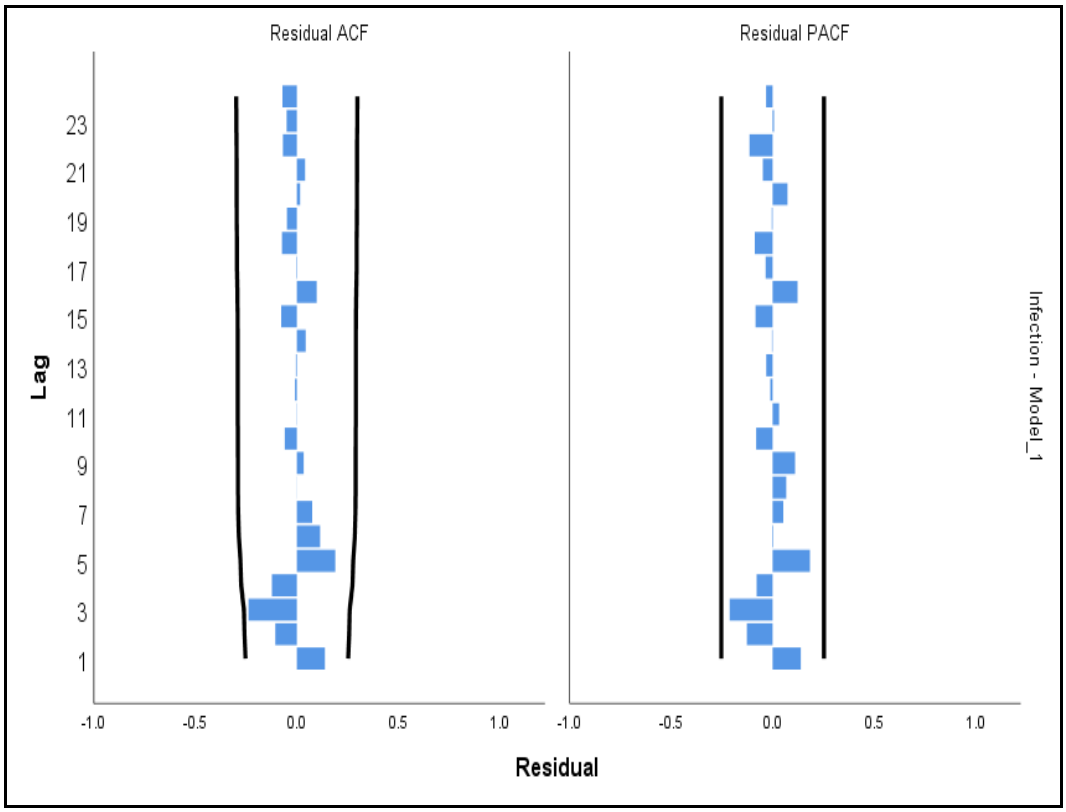


Figure (4.8) The Autocorrelation and Partial Autocorrelation Function of Residuals for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model

According to Figure (4.8), the residuals values of ACF and PACF for the Malaria incidence are all small and lie inside the confidence limits, lower confidence limit (LCL) and upper confidence limit (UCL), as well as exhibit no pattern. This suggested that the residuals are white noise. SARIMA (1,0,0) x $(1,1,0)_{12}$ model is adequate to perform the seasonally log transformed data series of Malaria Incidence in Kachin State.

On the other hand, the autocorrelation among residuals are checked by using the test statistic (Q).

$H_0 : \rho_1 = \rho_2 = \ldots = \rho_k = 0$ (There is no autocorrelation among residuals.)

The detail summary residual values for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model is shown in Table (4.12).

Table (4.12) Model Statistics of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Malaria Incidence in Kachin State

Model Statistics									
Model	Number of Predictors	Model Fit statistics				Ljung-Box Q (18)			Number of Outliers
		Stationary R-squared	R-squared	MAPE	Normalized BIC	Statistics	DF	Sig.	
Malaria Infected People-Model_1	0	0.771	0.928	14.083	12.168	12.641	16	0.699	0

As the result of above table, the observed value of Q is 12.641 and it is not significant since p-value is 0.699 which is greater than 0.05. It means that there is no autocorrelation among residuals.

By comparing these two feasible Models, SARIMA (1,0,0) x $(0,1,0)_{12}$ Model and SARIMA (1,0,0) x $(1,1,0)_{12}$ Model, the value of R-squared for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model is slightly larger than the other one. On the other hand, the values of MAPE and Normalized BIC are smaller than other one as well. That is why, SARIMA (1,0,0) x $(1,1,0)_{12}$ Model is adequate to fit and used to forecast the malaria infection of Kachin State next year.

4.8 Forecasting with SARIMA (1,0,0) x $(1,1,0)_{12}$ Model for Kachin Malaria Infection Series

The forecast values for 12 months period from January to December, 2017 are shown in Table (4.13) and displayed in Figure (4.9).

Table (4.13) Forecast Values with 95% Limits for Malaria Infection in Kachin State

Year	Forecast values	95% Limits		Actual Malaria data for 2017
		LCL	UCL	
Jan 2017	65	44	93	64
Feb 2017	59	35	96	71
Mar 2017	49	26	86	62
Apr 2017	94	46	174	73
May 2017	150	68	289	213
Jun 2017	298	129	595	386
Jul 2017	243	101	499	297
Aug 2017	172	69	363	186
Sep 2017	118	46	253	106
Oct 2017	81	31	176	99
Nov 2017	82	30	181	54
Dec 2017	63	23	140	20

The results indicate that the predicted values of SARIMA (1,0,0) x $(1,1,0)_{12}$ Model is not very close to the true value but all of them fall within lower confidence level (LCL) and upper confidence level (UCL) which proved the reliability of data series. According to these forecasting results, Malaria infection of Kachin state lead to eliminate in near future.

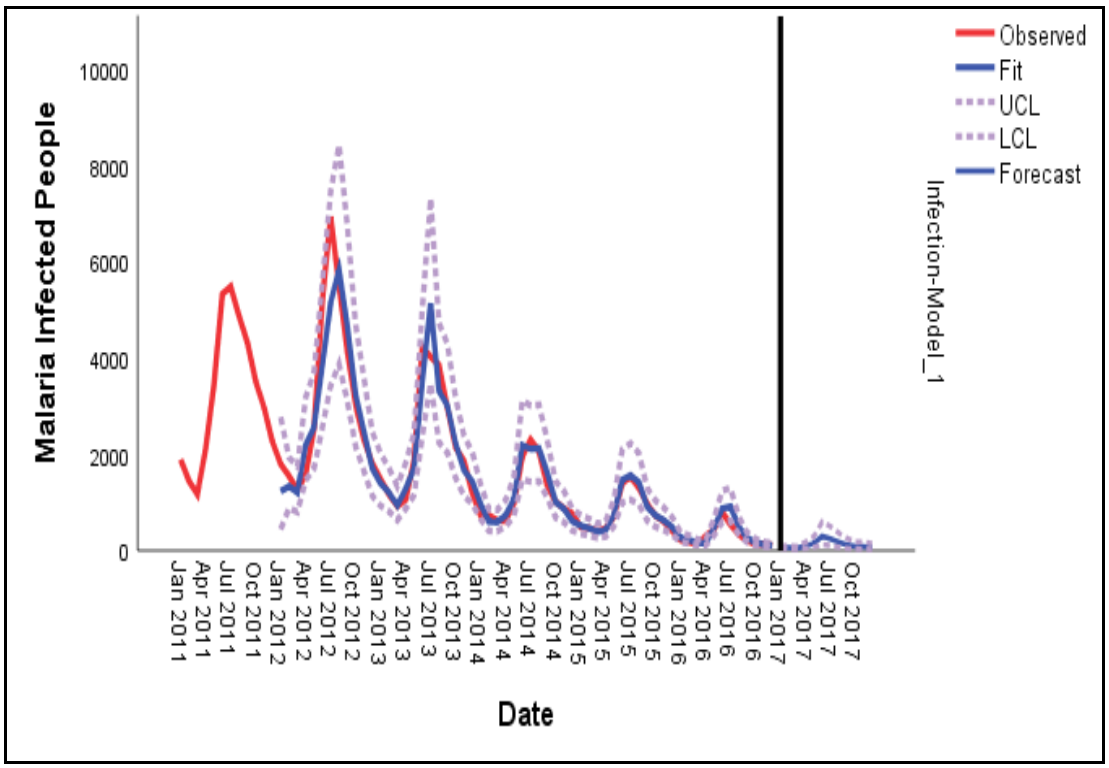


Figure (4.9) Forecast Values with 95% Limits for Malaria infection for SARIMA (1,0,0) x $(1,1,0)_{12}$ Model

CHAPTER V
CONCLUSION

This study focuses on the modeling and forecasting of Malaria Infection in Kachin State by modeling data from January, 2011 to December, 2016 using time series SAREMA model. SARIMA model can be obtained by using four iteratively Box-Jenkins steps and provide the prediction of the malaria infected number of people in Kachin State. Following Box and Jenkins methodology, the time series modeling involves for transformation of the data to achieve stationary, followed by the identification of appropriate models, estimation of model parameters, diagnostic checking of the assumption model and finally forecasting of the future data values.

In model identification process, test of seasonality is conducted by using ANOVA. Theoretical and estimated autocorrelation function (ACF) and partial autocorrelation function (PACF) play important role in the construction of SARIMA model. The estimated residuals are analyzed using the ACF and PACF to diagnose if the residuals are consistent with the hypothesis that the residuals are white noise.

Non-constant variance is removed by performing a natural log transformation. Then, trend in the series is removed by taking seasonal differencing. The results indicate that SARIMA (1,0,0) x $(1,1,0)_{12}$ Model is the fitted models. The model was also be able to represent the past data with MAPE, R^2 and normalize BIC are 14.083, 0.928 and 12.168 for malaria infection in Kachin state. As forecasting is essential for planning and operational control in a variety of areas, forecasting is made based on the best fitted SARIMA model.

In term of “Myanmar Times” publication on July 2017, The Public Health Department is aiming to make five regions; Yangon, Ayeyarwaddy, Bago, Mandalay and Magwe to free from malaria by 2020. Anti-malaria campaigns are especially being implemented in those regions according to the department.

Evaluation and forecasting the volume of malaria infection in Kachin State are significantly declined. The results of malaria infection in Kachin state is decreased over 50 % of infection (3517 to 1631) from 2016 to 2017. Although the Kachin State is not included in 2020 targeted areas for eliminating Malaria Infection in Myanmar, Malaria Incidence in Kachin State might be eliminated in 2020. It is because the finding of this study is shown that malaria incidence in Kachin State was declining.

That is why it can be clearly realized how the implementation of anti-malaria activities were continuously conducted by the responsible people.

The SARIMA model was used the data of malaria infection of Kachin State from January, 2011 to December, 2016, which contained 72 observations. Next 12 months was forecasted by the SARIMA $(1,0,0) \times (1,1,0)_{12}$ model well reflected the trend in the malaria infection of Kachin State. Results are indicated that SARIMA model was capable of representing the number of malaria infection in the following month with relative precision. That is why forecasting using time series models are useful of policy-making, supports for the planning and future analysis in Myanmar's health industry.

REFERENCES

Anokye, R., Enoch, A., Owusu, I., & Edmund Isaac, O. (2018). *Time series analysis of malaria in Kumasi: Using ARIMA models to forecast future incidence*.

Anwar, M. Y., Lewnard, J. A., Parikh, S., & Pitzer, V. E. (2016). *Time series analysis of malaria in Afghanistan: Using ARIMA models to predict future trends in incidence*. https://doi.org/10.1186/s12936-016-1602-1

Appiah, T., Otoo, S., & Nabubie, H. (2015). *Times Series Analysis Of Malaria Cases In Ejisu-Juaben Municipality*. *4*(06).

Babajide Sadiq MPH, D. P., & Perry Brown MSPH, D. P. (2015). A time series analysis of malaria cases in Ogun State, Nigeria. *International Scholars Journals, 3 (9)*, 245–259.

Box, G.E.P. and Jenkins, G.M. (1976) Time Series Analysis for Forecasting and Control. Holden-Day, San Francisco.

Hlaing, M. M. S. (2015). *Seasonal ARIMA Modeling of Airways Transport and Rail Ways Transport Series in Myanmar (2007-2013)*. Yangon University of Economics.

Htike, A. Z. (2015). *An Application of Seasonal ARIMA Models to Selected Time Series of Myanmar*. Yangon University of Economics, Yangon.

JSERBR. (n.d.). Time Series Analysis and Forecasting Model for Monthly alaria Infection by Box-Jenkins Techniques in Kass Zone, South Darfur State, Sudan. *Nyala Technizal College, Nyala, South Darfur, Sudan*. Retrieved from www.jsaer.com

Kumar, V. (2014). *Forecasting malaria cases using climatic factors in Delhi, India: A time series analysis*. Rural Health Training Center, Najafgarth, Delhi, India.

Martinez, E. Z. (2011). A SARIMA Forecasting Model to Predict the Number of Cases of Dengue in Campinas, State of Sao Paulo, Brazil. *Revista Da Sociedade Brazileira de Medicina Tropical*, *44*(4).

Mon, P. P. (2011). *Application of Seasonal ARIMA Models*. Yangon Institute of Economics, Yangon.

O Ebhuoma, Gebreslasie, M., & Magubane, L. (2018). *A Seasonal Autoregressive Integrated Moving Average (SARIMA) forecasting model to predict monthly malaria cases in KwaZulu-Natal, South Africa*. *Vol. 108, No. 7*.

Terence, C. M. (2019). (n.d.). *Applied Time Series Analysis*. Loughborough University, United Kingdom: Candice Janco.

Thanda, M. (1997). *Seasonal Models for Monthly Transort Time Series of Myanmar*. Yangon Institute of Economics.

Tint, S. T. Z. (2010). *Time Series Forecasting Using Holi-Winters Exponential Smooothing*. Yangon Institute of Economics, Yangon.

Tomer, D., & Majumder, A. (2018). *Forecasting malaria incidences in India using the ARIMA model*.

WHO. (2017). *WORLD MALARIA REPORT 2017*. World Health Organization.

WHO. (2010). *WORLD MALARIA REPORT 2010*. World Health Organization.

Wei, William W.S., W. (2006). *Time Series Analysis: Univariate and Multivariate Methods* (Second Edition), Pearson Education, Inc., U.S.A.

APPENDIX

Calculation of Seasonal Index

Time	Malaria Infection	Centered Moving Average (CMA)	Malaria infection / CMA	Seasonal Index
Jan_2011	1890			**0.4703**
Feb_2011	1441			**0.4422**
Mar_2011	1166			**0.3827**
Apr_2011	2109			**0.5293**
May_2011	3458			**0.8765**
Jun_2011	5327	3231.29	1.65	**2.1851**
Jul_2011	5482	3231.75	1.70	**1.8249**
Aug_2011	4872	3240.08	1.50	**1.6179**
Sep_2011	4328	3222.00	1.34	**1.2428**
Oct_2011	3517	3160.63	1.11	**0.9524**
Nov_2011	2971	3115.25	0.95	**0.8270**
Dec_2011	2261	3169.58	0.71	**0.6505**
Jan_2012	1797	3256.75	0.55	0.4703
Feb_2012	1545	3275.71	0.47	0.4422
Mar_2012	1262	3247.29	0.39	0.3827
Apr_2012	1579	3198.38	0.49	0.5293
May_2012	2515	3151.08	0.80	0.8765
Jun_2012	5181	3118.79	1.66	2.1851
Jul_2012	6932	3089.75	2.24	1.8249
Aug_2012	5514	3059.96	1.80	1.6179
Sep_2012	4141	3024.17	1.37	1.2428
Oct_2012	3022	2975.42	1.02	0.9524
Nov_2012	2292	2908.71	0.79	0.8270
Dec_2012	1805	2747.83	0.66	0.6505
Jan_2013	1478	2557.54	0.58	0.4703
Feb_2013	1167	2439.92	0.48	0.4422
Mar_2013	925	2354.79	0.39	0.3827
Apr_2013	1057	2300.25	0.46	0.5293
May_2013	1867	2256.58	0.83	0.8765
Jun_2013	4228	2199.42	1.92	2.1851
Jul_2013	4024	2149.92	1.87	1.8249
Aug_2013	3855	2119.17	1.82	1.6179
Sep_2013	2977	2088.17	1.43	1.2428
Oct_2013	2143	2032.54	1.05	0.9524
Nov_2013	1862	1900.29	0.98	0.8270
Dec_2013	1187	1733.54	0.68	0.6505
Jan_2014	724	1587.08	0.46	0.4703
Feb_2014	733	1444.58	0.51	0.4422
Mar_2014	621	1329.67	0.47	0.3827
Apr_2014	617	1241.54	0.50	0.5293

Time	Malaria Infection	Centered Moving Average (CMA)	Malaria infection / CMA	Seasonal Index
May_2014	972	1183.08	0.82	0.8765
Jun_2014	1949	1155.88	1.69	2.1851
Jul_2014	2301	1134.92	2.03	1.8249
Aug_2014	2063	1114.83	1.85	1.6179
Sep_2014	1349	1099.13	1.23	1.2428
Oct_2014	1013	1081.21	0.94	0.9524
Nov_2014	877	1047.21	0.84	0.8270
Dec_2014	769	992.50	0.77	0.6505
Jan_2015	489	928.96	0.53	0.4703
Feb_2015	465	881.46	0.53	0.4422
Mar_2015	407	853.42	0.48	0.3827
Apr_2015	454	829.79	0.55	0.5293
May_2015	705	803.42	0.88	0.8765
Jun_2015	1400	777.25	1.80	2.1851
Jul_2015	1537	754.17	2.04	1.8249
Aug_2015	1302	730.79	1.78	1.6179
Sep_2015	970	712.63	1.36	1.2428
Oct_2015	719	694.63	1.04	0.9524
Nov_2015	604	658.42	0.92	0.8270
Dec_2015	409	591.92	0.69	0.6505
Jan_2016	221	510.00	0.43	0.4703
Feb_2016	179	437.92	0.41	0.4422
Mar_2016	132	381.67	0.35	0.3827
Apr_2016	293	337.75	0.87	0.5293
May_2016	434	305.71	1.42	0.8765
Jun_2016	802	146.54	5.47	2.1851
Jul_2016	539			1.8249
Aug_2016	334			1.6179
Sep_2016	208			1.2428
Oct_2016	131			0.9524
Nov_2016	138			0.8270
Dec_2016	106			0.6505