TIME SERIES MODELING OF ALL ITEMS CONSUMER PRICE INDEX IN NIGERIA

BY

UMANAH, EYAEKOP EFIONG
PG/M.Sc/07/42873

DEPARTMENT OF STATISTICS
UNIVERSITY OF NIGERIA, NSUKKA

BEING A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS OF THE UNIVERSITY OF NIGERIA

March, 2010
CERTIFICATION

The work in this project is original and has not been in substance for any other degree of this university or any other university.

________________________________________
SUPERVISOR
DR. F.I. UGWUOWO
DEPARTMENT OF STATISTICS
UNIVERSITY OF NIGERIA, NSUKKA

________________________________________
HEAD OF DEPARTMENT
PROF. F.C. OKAFOR
DEPARTMENT OF STATISTICS
UNIVERSITY OF NIGERIA, NSUKKA

________________________________________
EXTERNAL EXAMINER
DEDICATION

My precious wife Ekaete and lovely children, Udeme, Iberedem and Idoremyin for their patience and understanding. And to the memory of my late parents Deacon and Mrs. Efiong Sampson Umanah.
ACKNOWLEDGEMENTS

Firstly, I would like to thank my Lord and saviour, Jesus Christ, for teaching me to number my days so that I may apply my heart to wisdom. I would especially like to thank Dr. F.I. Ugwuowo for his willingness to guide this study. I greatly appreciate the quality time and efforts he has given to me and my work. I express gratitude to my lecturers Prof. J.N. Adiche, Prof. P.I. Uche, Mr. W.I.E. Chukwu, Mr. E. P. Okubike for their invaluable insight. Thanks also to Mr. U.C. Nduka for his part in molding my research interests.

Worthy of mention is my lovely wife, Ekaete, my siblings Udeme, Iberedem and Idorenyin for their continued encouragement and ensuring that I was never in need.

Special thanks to my friends, roommates and course mates; especially victor Omonona, Charles Orhemba, Emmanuel Okon, Ngozi Orisakwe, Laurel Adima, Peter Umukoro, Chibuike, Elizabeth Okoro, James Adema, Patient Uka, Nse Uto, Anthony Ekpo, Emmanuel Umoh, Ifreke Udoidem, Amekan Ikpe, Nse Udoh, Nseobong Clement and several other colleagues in Nkrumah Hall and Odili Hall for making my stay in Nsukka a huge success.
ABSTRACT

Many time series are measured annually, either as average, total or index and such data follow the famous Autoregressive Integrated Moving Average (ARIMA) processes of Box and Jenkins (1976) and consequently time series data have been modeled by these processes. However, it has been observed that time series data exhibit some local fluctuations and variations that can be modeled by transcendental functions like the Fourier Series Model. The ARIMA process which involves the analysis of the series at distinct time is known as analysis in the time domain. The Fourier Series Model describes the value of the time series as a weighted sum of periodic functions of the form \( \cos(\omega t) \) and \( \sin(\omega t) \). This is known as the analysis in the frequency domain. In this work, these two models are fitted to the Nigerian All Items Consumer Price Index and comparisons were made based on their out-of-sample forecasts performance as measured by Average Percentage of Predicted Error (APE) and the Root Mean Square Error (RSME). The two models fitted were found to be adequate using the normal probability plot and quantile-quantile plot. Based on the APE and RMSE values, the ARIMA(1,1,0) or the differenced AR(1) model outperformed the Fourier Series Model.
CHAPTER ONE
INTRODUCTION

Consumer Price Index (CPI) is a measure of the average change overtime in the prices of consumer items, that is, goods and services that people buy for day-to-day living. The CPI uses data from survey of consumption pattern of households to produce a timely and precise average price change for the consumption sector of any economy like the Nigerian economy. The most important task in the production of CPI is the determination of the market basket of goods and services whose current price is usually compared with its base year price to measure changes in price.

Annual CPI has pattern in the forms of small local fluctuations and cyclic movements that have been recognised over the years. Upward or downward movements in CPI sometimes persist for a very long period and also tend to retrace back and forth as the case may be. As a result, world economies witness periods of boom and recession or downturn. In order to understand what causes this, local fluctuations and cyclic movements can be modelled in a variety of ways in accordance with Mooney, Jolliffe and Helms (2006). In this study we describe and compare two of these possibilities. Notwithstanding that this work was on Nigeria All Items Consumer Price Index (NAICPI) data, the comparisons are also relevant to other situations where data contain a seasonal, or other cyclic, pattern, for example in geology (Upton et al., 2003), biology (Batschelet, 1981) and atmospheric science (Wilks, 1995, Section 8.4).

The approach to modelling and forecasting time series such as the NAICPI data has been developed over several decades (see Grenander and Rosenblatt, 1957; Box and Jenkins, 1976; Makridakis and Wheelwright, 1993). The modelling and forecasting methods for both stationary and non-stationary time series have been applied to many different fields and many
successful results have been obtained in different areas. The method developed by Box, Jenkins and Reinsel (2008, p.78); that based on the modelling of time series by autoregressive integrated moving average (ARIMA) processes, is one of the two approaches in this work. The second approach is the frequency domain approach, which assumes that the time series is best regarded as a sum or linear superposition of periodic sine and cosine waves of different periods or frequencies. See, for example, Pollock (1999, p. 555) and Khuri (2003, p. 500).

1.1 STATEMENT OF THE PROBLEM

In this work, we shall fit two suitable models to the NAICPI data. We shall then compare ARIMA model with the Fourier series model. The error levels of the models will be critically assessed to select the best which will guide prediction and policy formulation and implementation in Nigeria.

1.2 OBJECTIVE OF THE STUDY

We shall build time series model for time and frequency domain, that is Box-Jenkins ARIMA model and Fourier series model. Specifically, the objectives of the study are:

i. To fit a Box-Jenkins ARIMA model to the NAICPI data;

ii. To fit a Fourier series model to the NAICPI data;

iii. To determine the best modeling approach given the two models;

iv. To forecast for some lead time using the two models.

1.3 DATA COLLECTION AND SOURCE

The NAICPI data were obtained and put together, for the purpose of this research work, from various publications of Federal Office of Statistics of Nigeria now National Bureau of Statistics of Nigeria. The data range from 1960 to 2008 for a total of 49 observations. The
annual data of NAICPI used for this study are purely time series data based on the fact that they were computed sequentially in time.

The rest of the work is organized in the following manner. In Chapter two, we will describe the definitions and of general time series models like AR, MA, ARMA and Fourier series models. Stationarity of time series data, acf, pacf, and spectrum, identification of models and estimation of model parameters will also be discussed in this Chapter. In Chapter three we will discuss fitting ARIMA model to the NAICPI data. Fitting Fourier series model to the NAICPI data is discussed in Chapter four. The final conclusion is made in Chapter five.
CHAPTER TWO

PRELIMINARIES

This chapter shall first review some relevant literature and then introduce the concept of time series analysis by giving brief overview of the models used in this study. The models are the Autoregressive Integrated Moving Average (ARIMA) and the Fourier Series (FS) models.

2.1 LITERATURE REVIEW

Numerous works have investigated the relative accuracy of alternative inflation forecasting models. One type of practice has been to compare the accuracy of survey respondents’ inflation forecasts relative to univariate time series models. Another approach is the methodology found in Fama (1975) and their extensions Fama and Gibbons (1984). This approach uses extracts from observed nominal interest rates, which is the market’s inherent expectation of inflation. Fama and Gibbons (1984) found that the interest-rate model yields inflation forecasts with a lower error variance than a univariate model. This conclusion was based on a univariate time series modelling of the real interest rates.

Engle (1982) introduced a new regression model with disturbances following an ARCH process. This newly developed model was used to estimate the means and variances of inflation in the UK. A natural generalization of the model introduced in Engle (1982) was proposed by Bollerslev (1986) and used as an empirical example relating the uncertainty of the inflation rate in the US.

Al-Eideh, Al-Refai and Sbeiti (2004) used Maximum Likelihood estimator within a lognormal diffusion process with closed form analytical solutions to obtain the monthly CPI forecasts for the period between 1970 and 2002. The quarterly estimates of inflation rates are obtained from these monthly forecasts rather than from quarterly data. This procedure significantly improved the estimates of inflation rates. The model also produced a superior fit
as compared to random walk and GARCH($p,q$)-M models. The adopted approach is found to be simple, economical and generally suitable for modelling stochastic processes that reflect aggregation over time stemming from many factors, and in which the transition path between consecutive states is relatively smooth.

Malliaris and Malliaris (1995) presented a decomposition of inflation and its volatility. According to the traditional quantity theory of money, the rate of inflation is decomposed into three components: the rate of change in the money supply, plus the rate of change in the velocity of circulation, minus the rate of change in real output. They derived a generalization of this decomposition by postulating that the rate of change of money supply, velocity, and output follow diffusion equations. Using stochastic calculus techniques, two expressions are obtained decomposing inflation and its volatility as a sum of several economically important terms. Two sets of U.S. data are used to illustrate these decompositions with actual numbers.

ARIMA models for forecasting the Irish inflation were outlined by Kenny, Meyler and Quinn (1998). The study considered two alternative approaches to the issue of identifying ARIMA models, namely, the Box-Jenkins approach and the objective penalty function methods. The emphasis was on forecast performance, which suggests that ARIMA forecast outperformed the other approach.

Nadal-De Simone (2000) estimated two time-varying parameter models for Chilean inflation rates. The study discovered that ARIMA models outperformed the two other models considered in that paper for short-term out-of-sample forecasts. However, this superiority diminishes in longer forecasts. Stockton and Glassman (1987) upon discovering similar results confirmed that simple ARIMA models do well in predicting inflation rates.

An autoregressive model with a deterministically shifting intercept was introduced by Gonzalez and Terasvirta (2008). This implies that the model has a shifting mean and is thus
nonstationary but stationary around a nonlinear deterministic component. The shifting intercept is defined as a linear combination of logistic transition functions with time as the transition variables. The number of transition functions was determined by selecting the appropriate functions from a possibly large set of alternatives using a sequence of specification tests. This selection procedure is a modification of a similar technique developed for neural network modelling by (White, 2006). A Monte Carlo experiment was conducted to show how the proposed modelling procedure and some of its variants work in practice. The paper contains two applications in which the results are compared with what is obtained by assuming that the time series used as examples may contain structural breaks instead of smooth transitions and selecting the number of breaks following the technique of Bai and Perron (1998).

Kang, Kim and Morley (2009) investigated the existence and timing of changes in U.S. inflation persistence. To do so, they developed an unobserved components model of inflation with Markov-switching parameters and measured persistence using impulse response functions based on the model. An important feature of their model is its allowance for multiple regime shifts in parameters related to the size and propagation of shocks. Inflation persistence depends on the configuration of these parameters, although it need not change even if the parameters change. Using the GDP deflator for the sample period of 1959-2006, Kang, Kim and Morley (2009) found that U.S. inflation underwent two sudden permanent regime shifts, both of which corresponded to changes in persistence. The first regime shift occurred around the collapse of the Bretton Woods system at the beginning of the 1970's and produced an increase in inflation persistence, while the second regime shift occurred immediately after the Volcker disinflation in the early 1980's and produced a decrease in inflation persistence. Meanwhile, consistent with the New Keynesian Phillips
Curve, the gap between inflation and its long-run trend displayed little or no persistence throughout the entire sample period.

Much of these works done in modelling and forecasting of Inflation, which is measured by Consumer Price Index, have used ARIMA models and perhaps made comparisons with some other models. In this study we shall follow the same approach. To this end, this study shall compare the performance of ARIMA models in forecasting the Nigerian All Items Consumer Price Index (NAICPI) with the Fourier series model.

2.2 OVERVIEW OF ARIMA MODELS

A general class of univariate models is the ARIMA model. An ARIMA model represents current values of a time series in terms of past values of itself, the autoregressive component, and past values of the error term, the moving average component. The integrated component refers to the number of times a series must be differentiated to induce stationarity.

2.2.1 Autoregressive AR(p) Models

A pure AR(p) process may be represented as follows, where \( X_t \) is modelled as lagged values of itself plus a ‘white noise’ error term.

\[
X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t.,
\]

This may be alternatively written as

\[
\phi(B)X_t = a_t,
\]

where \( \phi(B) \) is a \( p \)-order polynomial in the backshift operator, which is equal to

\[
(1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p), \text{ and } B \text{ is the backshift operator, such that } B^0 X_t = X_t,
\]

\[
B^1 X_t = X_{t-1}, \quad B^2 X_t = X_{t-2}, \ldots.
\]
A useful way of gaining insight into univariate processes is to consider their autocorrelation and partial autocorrelation functions (ACF and PACF). The ACF measures the ratio of the covariance between observations \(k\) lags apart and the geometric average of the variance of observations (i.e., the variance of the process if the process is stationary, as \(V(X_t) = V(X_{t-k})\)).

\[
\rho_k = \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t) \cdot \text{Var}(X_{t-k})}}.
\]  

(2.3)

The sample autocorrelation function (SACF) may be calculated as follows:

\[
\rho_k = \frac{\sum_{t=k+1}^{n} (X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n} (X_t - \bar{X})^2}.
\]  

(2.4)

However, some of the observed autocorrelation between \(X_t\) and \(X_{t-k}\) could be due to both being correlated with intervening lags. The PACF seeks to measure the autocorrelation between \(X_t\) and \(X_{t-k}\) correcting for the correlation with intervening lags. For example, consider an AR(1) process of the form \(X_t = 0.8 X_{t-1} + \epsilon_t\). The first order autocorrelation coefficient is 0.8. The autocorrelation coefficient for the second lag is 0.64, although the partial autocorrelation coefficient for the second lag is zero, as the process is an AR(1) process. In other words the autocorrelation between observations two lags apart is due only to the correlation between observations one lag apart which feeds through into the second lag. As the lag length increases the autocorrelation coefficient declines (at lag length \(k\) the autocorrelation coefficient is \((0.8)^k\)).

The PACF is calculated as the partial regression coefficient, \(\phi_{kk}\) in the \(k\)th order autoregression

\[
X_t = \phi_{k1}X_{t-1} + \phi_{k2}X_{t-2} + \ldots + \phi_{kk}X_{t-k} + \epsilon_t.
\]  

(2.5)

Thus, for an AR(p) process, \(\phi_{kk} = 0, \forall k > p\).
Some general properties of the ACF and PACF for AR processes can be observed by considering a simple AR(1) process given as

\[ X_t = \phi X_{t-1} + a_t. \]  

(2.6)

Note that the AR(1) model can be written as an infinite length MA process, providing \( \phi < 1 \).

Denote the AR(1) series as,

\[ (1-\phi B)X_t = a_t \]  

(2.7)

which gives

\[ X_t = (1-\phi B)^{-1}a_t \]  

(2.8)

which upon expansion and providing \( \phi < 1 \) yields

\[ X_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \ldots \]  

(2.9)

This result holds more generally so that any finite order stationary AR process may be expressed as an infinite order MA process. This duality between AR and MA processes is an important property which can often be exploited when attempting to identify ARMA models.

For the AR(1) process the value of the ACF at lag \( k \) is given by \( \phi^k \). The value of the autoregressive coefficient can yield some insight into the underlying data generating process. For example, higher values of \( \phi \) indicate a higher degree of persistence in the series. A negative autoregressive component indicates a process which oscillates around its mean value. For more general AR(p) models, the behaviour of the process is determined by the solution to the \( p \)-order polynomial \((1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)\) given by

\[ \phi(B) = (1-g_1 B)(1-g_2 B)\ldots(1-g_p B) = 0. \]  

(2.10)

For the process to be stationary it is a necessary and sufficient condition for the roots of the \( p \)-order polynomial to lie outside the unit circle.
2.2.2 Moving Average MA($q$) Models

An MA($q$) process may be represented as follows, where $X_t$ is modelled as the weighted average of a ‘white noise’ series,

$$X_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$  \hspace{1cm} (2.11)

or alternatively

$$X_t = \theta(B)a_t$$  \hspace{1cm} (2.12)

where $a_t$ is a white noise process, $\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)$ is a $q$-order polynomial in the backshift operator.

Note that the expected value of $X_t$ equals zero. Furthermore, the autocorrelation between $X_t$ and $X_{t+k}$ equals zero for $k$ greater than $q$. Thus the order of the MA process, $q$, indicates the ‘memory’ of the process. All MA processes are stationary, regardless of the coefficients of the model. However, to ensure invertibility of the model (i.e., that the finite order MA process can be written in terms of a stationary infinite order AR process) the roots of the MA polynomial must lie outside the unit circle. MA models can be particularly useful for representing some economic time series as they can handle random shocks such as strikes, weather patterns, etc..

2.2.3 Autoregressive Moving Average ARMA($p,q$) Models

An ARMA($p,q$) series may be represented as

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}$$  \hspace{1cm} (2.13)

or alternatively

$$\sum_{i=0}^{p} \phi_i X_{t-i} = \sum_{j=0}^{q} \theta_j a_{t-j},$$  \hspace{1cm} (2.14)
where $\phi_0$ and $\theta_0$ are equal to 1, or more compactly
\[ \phi(B)X_t = \theta(B)a_t. \] (2.15)

Using mixed ARMA models can be useful as it should usually be possible to represent a time series satisfactorily using fewer parameters than might be required with a pure AR or pure MA models.

### 2.2.4 Seasonal ARMA Models

Seasonal data may be also modelled. The numbers of seasonal AR and MA terms are usually denoted by $P$ and $Q$ respectively. Thus, a general seasonal ARMA model may be represented as,
\[ \phi(B)\Phi(B)X_t = \theta(B)\Theta(B)a_t, \] (2.16)

where $\Phi(B) = 1 - \Phi_{1S}B^{1S} - \Phi_{2S}B^{2S} - \ldots - \Phi_{PS}B^{PS}$,
\[ \Theta(B) = 1 + \Theta_{1S}B^{1S} + \Theta_{2S}B^{2S} + \ldots + \Theta_{QS}B^{QS} \]
and $S$ is the seasonal span, hence quarterly data $S = 4$ and for monthly data $S = 12$.

### 2.2.5 Autoregressive Integrated Moving Average (ARIMA) Models

The integrated component of an ARIMA model represents the number of times a time series must be differenced to induce stationarity. A general notation for ARIMA models is ARIMA $(p,d,q)(P,D,Q)$, where $p$ denotes the number of autoregressive terms, $q$ denotes the number of moving average terms and $d$ denotes the number of times a series must be differenced to induce stationarity. $P$ denotes the number of seasonal autoregressive components, $Q$ denotes the number of seasonal moving average terms and $D$ denotes the number of seasonal differences required to induce stationarity. This may be written as
\[ \phi(B)\Phi(B)\nabla^d\nabla_S^D Y_t = \theta(B)\Theta(B)a_t, \] (2.17)
where \( X_t = \nabla^d \nabla_S^d Y_t \), \( \nabla^d = (1 - B)^d \) represents the number of regular differences and \( \nabla_S^d = (1 - B^S)^d \) represents the number of seasonal differences required to induce stationarity in \( Y_t \).

Two important properties of the parameters of ARMA models are worth repeating. First, for an ARMA process to be stationary it is required that the modulus of the roots of the \( p \)-order AR polynomial be greater than unity. Second, for an ARMA model to be invertible (i.e., representable as a stationary infinite lag AR model) the roots of the \( q \)-order MA polynomial should also be greater than unity.

### 2.3 Fourier Series Models

Often one is interested in determining the frequency content of signals. Signals are typically represented as time dependent functions. Real signals are continuous, or analog signals. However, through sampling the signal by gathering data, the signal does not contain high frequencies and is finite in length. The data is then discrete and the corresponding frequencies are discrete and bounded. Thus, in the process of gathering data, one seriously affects the frequency content of the signal. This is true for simple a superposition of signals with fixed frequencies. The situation becomes more complicated if the data has an overall non-constant trend or even exists in the presence of noise.

As described in the last section (hopefully), we have seen that by restricting our data to a time interval \([0, T]\) for period \( T \), and extending the data to \((-\infty, \infty)\), one generates a periodic function of infinite duration at the cost of losing data outside the fundamental range. This is not unphysical, as the data is typically taken over a finite period of time. Thus, any physical results in the analysis can be obtained be restricting the outcome to the given period. In typical problems one seeks a representation of the signal, valid for \( t \in [0, T] \) as
\[ X_i = \frac{\alpha_0}{2} + \sum_{k=1}^{m} \{ \alpha_k \cos(w_k t) + \beta_k \sin(w_k t) \} \] (2.18)

where \( w_k = \frac{2\pi k}{T} \) is a multiple of the fundamental frequency \( w_1 = \frac{2\pi}{T} \).

\[ \alpha_k = \frac{2 \sum_{i=1}^{T} X_i \cos w_k t}{T}, \quad k = 0, 1, \ldots, m, \]

\[ \beta_k = \frac{2 \sum_{i=1}^{T} X_i \sin w_k t}{T}, \quad k = 0, 1, \ldots, m. \]

The values \( w_1, w_2, \ldots, w_m \) are called harmonic frequencies.

This two class of models will be fitted to the NAICPI data and compared based on the out-of-sample forecast performance in the following chapters.
CHAPTER THREE
FITTING ARIMA MODELS TO NAICPI

3.1 Exploratory Data Analysis

This research is motivated by the need to study and estimate the Fourier series model and model fitting of non-stationary Nigerian All Items Consumer Price Index (NAICPI) time series. The Fourier series transform and model fitting of this data can provide a better understanding of the dynamics of Nigerian economy. The dataset being considered in this study is taken from the bulletin of the National Bureau of Statistics, Nigeria from 1960 to 2008.

Let $X_t$ be the value of the NAICPI at time $t$ for $t = 1, 2, ..., T$. The original time series plot is provided in Figure 3.1. We also provided in Figure 3.2 and Figure 3.3 the sample autocorrelation function (acf) and partial autocorrelation function (pacf) plots respectively, which definitions can be found in Brockwell and Davies (1996, p.52). The summary Statistics of these data are given in Table 3.1.

The NAICPI time series plot in Figure 3.1 indicates local trends from 1960 to 1985 and a general most likely linear upward trend from 1987 to 2008. The slow decay of the NAICPI autocorrelation function plot in Figure 3.2 indicates short range or short memory. This implies that there is only fairly weak dependence between the NAICPI series, $X_t$, since most of the autocorrelation coefficients at different time lags after lag 8 are within the approximate 95% limits. However, the autocorrelation coefficients at larger lags are not within the approximate 95% limits, this shows that there is a probability of long memory in the series. The NAICPI partial autocorrelation function plot in Figure 3.3 indicates non-stationarity caused, perhaps, by the stochastic trend of integration of order 1, see Box and
Jenkins (1970, p.174). Such time series could be transformed to a stationary process by differencing it once [see Brockwell and Davies (1996), chapter 9].

Table 3.1: Summary Statistics of NAICPI time series data

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Values (Estimates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1653.6976</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2571.6170</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.7640</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.0720</td>
</tr>
</tbody>
</table>

The skewness coefficient is 1.7640 which suggests that NAICPI is right skewed, the density estimate plot of NAICPI in Figure 3.4 obtained by plotting the x- and y-coordinates from the output of Matlab 7.4 function ‘ksdensity’ agrees with this skewness coefficient. The value of the kurtosis is less than 3, meaning that it has a light tail than the standard normal distribution. Fig. 3.5 and Fig. 3.6 show the normal probability plot and normal quantile plot of the NAICPI series. The figures show that the series is not from a normal population.

\[ \text{Figure 3.1: NAICPI time series} \]
Figure 3.2: NAICPI Autocorrelation function plot

Figure 3.3: NAICPI Partial autocorrelation function plot
Figure 3.4: Density estimate plot for NAICPI

Fig. 3.5: Normal probability plot of NAICPI
Fig. 3.6: Normal quantile plot of NAICPI

3.2: Identification of the ARIMA(p,d,q) Model

From Figure 3.2 we discovered that the series is non-stationary so in order to make it stationary we take first-order differencing. Figure 3.7 and Figure 3.8 show the acf and pacf of the differenced series of NAICPI respectively. The acf plot of the differenced NAICPI shows no cut off while there is cut off in pacf plot of the differenced NAICPI, hence ARIMA(1,1,0) model is suspected and will be entertained. This suspicion is confirmed by a procedure in SPSS version 15.0 called the expert modeller. The expert modeller automatically identifies the best-fitting ARIMA for a time series, thus eliminating the need to identify an appropriate model through trial and error. For the NAICPI series the expert modeller identified ARIMA(1,1,0).
Figure 3.7: Differenced NAICPI autocorrelation function plot

Figure 3.8: Differenced NAICPI partial autocorrelation plot
3.3 Estimation of the ARIMA(1,1,0) model

The ARIMA(p,d,q) model is defined as

\[ \phi(B)(1-B)^d X_t = \phi_0 + \Theta(B)\alpha_t, \]  

(3.1)

where \( \phi(B) \) represents the autoregressive operator, \( \Theta(B) \) is the moving average operator, \( (1-B)^d \) is the differencing operator and \( \phi_0 \) is the mean of the series, \( p \) is order of autoregressive operator, \( q \) is the order of moving average and \( d \) is the order of differencing.

Following (3.1) the ARIMA(1,1,0) is written as

\[ (1-\phi B)(1-B)X_t = \phi_0 + \alpha_t, \]  

(3.2)

Expanding (3.2) we obtain

\[ \phi(B)X_t = \phi_0 + \alpha_t \]  

(3.3)

where \( \phi(B) = (1-\phi B) \). Equation (3.2) is an ARIMA(1,1,0) process but the parameters in this operator can be obtained by fitting an AR(1) to the differenced \((d = 1)\) series. Recall that a Gaussian AR(1) process takes the form

\[ X_t = \phi_0 + \phi X_{t-1} + \alpha_t \]  

(3.4)

with \( \alpha_t \approx i.i.d. N(0, \sigma^2) \). For this case the vector of population parameters to be estimated consists of \( \theta = (\phi_0, \phi, \sigma^2) \). For the AR(1) model defined in (3.4), the joint density of the observations \( X_1, \ldots, X_T \) can be written as the product of the conditional densities, conditioned on the first observation,

\[ f_{X_T, X_{T-1}, \ldots, X_2, X_1} (x_T, x_{T-1}, \ldots, x_2 / x_1; \theta) = \prod_{i=1}^{T} f_{X_i / X_{i-1}} (x_i / x_{i-1}; \theta), \]  

(3.5)

the objective then being to maximize with respect to the parameters \( \theta = (\phi_0, \phi, \sigma^2) \)

\[ \log f_{X_T, X_{T-1}, \ldots, X_2, X_1} (x_T, x_{T-1}, \ldots, x_2 / x_1; \theta) \]
\[
- \frac{(T-1)/2 \log(2\pi) - (T-1)/2 \log(\sigma^2)}{2\sigma^2}.
\]

The Time Series Modeller procedure in SPSS version 15 has function that implements the maximization of this log-likelihood function in (3.6). We shall then obtain the parameter estimates using the SPSS version 15 procedure.

It is often useful to divide the time series into an estimation, or historical, period and a validation period. Then a model is developed on the basis of the observations in the estimation (historical) period and a test is conducted to see how well it works in the validation period. By forcing the model to make predictions for points we already know (the points in the validation period), we get an idea of how well the model does at forecasting. There are 49 data points in the NAICPI series, the first 45 data points of this series will be used to develop a model and the remaining 4 data points will be used to test how well the developed model does at forecasting. Table 3.2 presents the estimates of the model parameters, together with their \(t\)-statistics and \(p\)-values. Since the sample is small, we have not tried to eliminate the insignificant parameters.

### Table 3.2: Estimated AR(1) model for NAICPI

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>(t)</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>193.569</td>
<td>1.361</td>
<td>0.181</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.839</td>
<td>8.101</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The model fitted is

\[
(1 - 0.839B)(1 - B)X_t = 193.569 + a_t.
\]

### 3.4: Diagnostics for the ARIMA(1,1,0) model

A fitted model must be examined carefully to check for possible model inadequacy. If the model is adequate, then the residual series should behave as a white noise. The autocorrelation function and the Ljung-Box \(Q\)-statistic of the residuals can be used to check
the closeness of $\hat{a}$, to a white noise. The Ljung-Box Q-statistic (Ljung and Box, 1978) is defined by:

$$Q = n(n+2)\sum_{j=1}^{K} \frac{r_j^2}{(n-j)}$$

(3.8)

where $n$ is the number of observations, $K$ is the largest lag used and $r_j$ is the sample autocorrelation function at lag $j$ of the residual series. When fitting ARMA($p,q$) models to data and testing the residuals to see if they are approximately white noise, under the null hypothesis that the series is ARMA($p,q$), $Q$ has approximately the $\chi^2$ distribution with $(K - p - q)$ degrees of freedom. For AR($p$) model, the Ljung-Box Q-statistic follows asymptotically a $\chi^2$ distribution with $(K - p)$ degrees of freedom. Here the number of degrees of freedom is modified to signify that $p$ AR coefficients are estimated. If a fitted model is found to be inadequate, it must be refined.

Consider the residual series of the fitted ARIMA(1,1,0) model for the NAICPI series. We have Ljung-Box $Q(18) = 9.075$ with $p$-value 0.938 based on its asymptotic $\chi^2$ distribution with 17 degrees of freedom. Thus we have no sufficient evidence to reject the null hypothesis of no serial correlation in the first 18 lags. Hence we conclude that the fitted model is adequate and $\hat{a}$ is white noise. This conclusion is buttressed further by the autocorrelation function of the residuals of the NAICPI series shown in Figure 3.9.
Figure 3.9: NAICPI residuals autocorrelation function plot

Fig. 3.10: Normal probability plot of the residuals
Fig. 3.11: Normal quantile plot of the residuals

Fig. 3.10 and Fig. 3.11 show the normal probability plot and normal quantile plot of the residuals from the ARIMA(1,1,0) fit. The plots show that fitting ARIMA(1,1,0) to NAICPI has actually reduced the series to white noise process, hence the model is adequate.

3.5 Forecasting

Forecasting is an important application of time series analysis. For AR(p) model, suppose that we are at the time index \( t \) and are interested in forecasting \( X_{t+l} \), where \( l \geq 1 \). The time index \( t \) is called the forecast origin and the positive integer \( l \) is the forecast horizon. Let \( \hat{X}_t(l) \) be the forecast of \( X_{t+l} \) using the minimum squared error loss function. In other words, the forecast \( \hat{X}_t(l) \) is chosen such that

\[
E[X_{t+l} - \hat{X}_t(l)] \leq \min_g E(X_{t+l} - g)^2, \tag{3.9}
\]

where \( g \) is a function of the information available at time \( t \) (inclusive). We referred to \( \hat{X}_t(l) \) as the \( l \)-step ahead forecast of \( X_t \) at the forecast origin \( t \).
**One-Step Ahead Forecast**

From the AR(\(p\)) model, we have

\[
X_{t+1} = \phi_0 + \phi_1 X_t + \cdots + \phi_p X_{t-p} + a_{t+1} .
\]

(3.10)

Under the minimum squared error loss function, the point forecast of \(X_{t+1}\) given the model and observations up to time \(t\) is the conditional expectation

\[
\hat{X}_t(1) = E(X_{t+1} \mid X_t, X_{t-1}, \ldots) = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i}
\]

(3.11)

and the associated forecast error is \(e_t(1) = X_{t+1} - \hat{X}_t(1) = a_{t+1}\). Consequently, the variance of the one-step ahead forecast error is \(Var[e_t(1)] = Var(a_{t+1}) = \sigma^2_a\). If \(a_t\) is normally distributed, then a 95\% one-step ahead interval forecast of \(X_{t+1}\) is \(\hat{X}_t(1) \pm 1.96 \times \sigma_a\).

**Two-Step Ahead Forecast**

From the AR(\(p\)) model, we have

\[
X_{t+2} = \phi_0 + \phi_1 X_{t+1} + \cdots + \phi_p X_{t-2p} + a_{t+2} .
\]

(3.11)

Taking conditional expectation, we have

\[
\hat{X}_t(2) = E(X_{t+2} \mid X_t, X_{t-1}, \ldots) = \phi_0 + \phi_1 \hat{X}_t(1) + \phi_2 X_t + \cdots + \phi_p X_{t-2p}
\]

(3.12)

and the associated forecast error is

\[
e_t(2) = X_{t+2} - \hat{X}_t(2) = \phi_1 [X_{t+1} - \hat{X}_t(1)] + a_{t+2} = a_{t+2} + \phi_1 a_{t+1}.
\]

(3.13)

The variance of the forecast error is \(Var[e_t(2)] = (1 + \phi_1^2) \sigma^2_a\). Interval forecasts of \(X_{t+2}\) can be computed in the same way as those for \(X_{t+1}\).

It is interesting to see that \(Var[e_h(2)] \geq Var[e_h(1)]\), meaning that as the forecast horizon increases the uncertainty in forecast also increases. This is in agreement with common sense that we are more uncertain about \(X_{t+2}\) than \(X_{t+1}\) at the time index \(h\) for a linear series.
**Multi-step Ahead Forecast**

In general, we have

\[ X_{i+l} = \phi_0 + \phi_1 X_{i+l-1} + \ldots + \phi_p X_{i+l-p} + a_{i+l}, \quad (3.14) \]

The \( l \)-step ahead forecast based on the minimum squared error loss function is the conditional expectation of \( X_{i+l} \) given \( \{X_{i-i}\}_{i=0}^{\infty} \), which can be obtained as

\[ \hat{X}_i(l) = \phi_0 + \sum_{i=1}^{p} \phi_i X_i(l-i), \quad (3.15) \]

where it is understood that \( \hat{X}_i(i) = X_{i+l} \) if \( i \leq 0 \). This forecast can be computed recursively using forecasts \( \hat{X}_i(i) \) for \( i = 1, \ldots, l-1 \). The \( l \)-step ahead forecast error is \( e_i(l) = X_{i+l} - \hat{X}_i(l) \).

The variance of the forecast error approaches the unconditional variance of \( X_i \).

Table 3.3 contains the one-step to four-step ahead forecasts and the standard errors of the associated forecast errors at the forecast origin year 2004 for the NAICPI series using an AR(1) model that was estimated using the first 45 observations. The actual NAICPI series are also given.

**Table 3.3: Multi Ahead Forecasts of an AR(1) model For NAICPI. The origin is 2004**

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td>7042.06</td>
<td>7680.51</td>
<td>8247.48</td>
<td>8754.43</td>
</tr>
<tr>
<td>Std. Error</td>
<td>5.13</td>
<td>6.70</td>
<td>6.70</td>
<td>6.70</td>
</tr>
<tr>
<td>Actual</td>
<td>7446.43</td>
<td>8059.59</td>
<td>8493.64</td>
<td>9477.23</td>
</tr>
</tbody>
</table>

Figure 3.7 shows the plot of the actual NAICPI series and their corresponding fitted values using AR(1) model. From the figure we noticed that the forecasts are very close to the actual values. Also the R-Squared for the fitted model is 0.991, which implies that AR(1) fitted to the NAICPI series is highly adequate. Other fit statistic is the RMSE which is equal to 161.24.
Figure 3.7: Plot of NAICPI series and corresponding forecasts
CHAPTER FOUR
FITTING FOURIER SERIES MODEL TO NAICPI

4.0: Introduction

Fourier analysis is particularly useful in areas such as signal and image processing, filtering, convolution, frequency analysis, and power spectrum estimation. Fourier analysis provides insight into the periodicities in data by representing the data using a linear combination of sinusoidal components with different frequencies. The amplitude and phase of each sinusoidal component in the sum determines the relative contribution of that frequency component to the entire signal. Fourier transforms can be applied to any time series, including non-stationary time series. We show in this chapter how Fourier series model can be used to estimate and forecast NAICPI data.

4.1 Fourier Representation of a Sequence

According to the basic result of Fourier analysis, it is always possible to approximate an arbitrary analytic function defined over a finite interval of the real line, to any desired degree of accuracy, by a weighted sum of sine and cosine functions of harmonically increasing frequencies. See, for example, Pollock (1999, p. 555) and Khuri (2003, p. 500).

Consequently, similar results apply in the case of sequences, which may be regarded as functions mapping from the set of integers onto the real line. For a sample of $T$ observations $x_1, x_2, \ldots, x_T$, it is possible to device an expression in the form

$$X_t = \frac{\alpha_0}{2} + \sum_{k=1}^{m} \{ \alpha_k \cos(w_k t) + \beta_k \sin(w_k t) \} \tag{4.1}$$

where $w_k = \frac{2\pi k}{T}$ is a multiple of the fundamental frequency $w_1 = \frac{2\pi}{T}$.
\[ \alpha_k = \frac{2 \sum_{t=1}^{T} X_t \cos w_k t}{T}, \quad k = 0, 1, \ldots, m, \]
\[ \beta_k = \frac{2 \sum_{t=1}^{T} X_t \sin w_k t}{T}, \quad k = 0, 1, \ldots, m. \]

The values \( w_1, w_2, \ldots, w_m \) are called harmonic frequencies. This model provides a decomposition of the time series into a set of cycles based on the harmonic frequencies. This expression is called the Fourier decomposition of \( X_t \), and the set of coefficients \( \{ \alpha_k, \beta_k \}; j=1, 2, \ldots, k \) are called the Fourier coefficients.

When \( T \) is even, we have \( m = T/2 \); it follows that \( \sin(w_0 t) = \sin(0) = 0 \), \( \cos(w_0 t) = \cos(0) = 1 \), \( \sin(w_m t) = \sin(\pi) = 0 \) and \( \cos(w_m t) = \cos(\pi) = (-1)^{j}. \)

Therefore, equation (4.1) becomes
\[ X_t = \alpha_0 + \sum_{k=1}^{m-1} \{ \alpha_k \cos(w_k t) + \beta_k \sin(w_k t) \} + \alpha_m (-1)^{j}. \] (4.2)

When \( T \) is odd, we have \( m = (T-1)/2 \); and equation (4.1) then becomes
\[ X_t = \alpha_0 + \sum_{k=1}^{m} \{ \alpha_k \cos(w_k t) + \beta_k \sin(w_k t) \}. \] (4.3)

4.2: Estimation of the Fourier Series Model

The expressions for \( \alpha_k \) \( (k = 1, 2, \ldots, m) \) and \( \beta_k \) \( (k = 1, 2, \ldots, m) \) were obtained by treating the model as a linear regression model with \( (2m+1) \) parameters and then fitting it to the \( (2m+1) \) observations by the method of least squares. See for example, Fuller (1976, chapter 7).
A standard way to use model for periodic behaviour, such as the Fourier Series Model, is to first of all start with a single harmonic in the model and then to add harmonics of basic cycle gradually until the R square improves. If sufficient terms are added, the data will be over-fitted, see Mooney et al (2006). The Curve Fitting Toolbox in MATLAB version 7 will be used to obtain the estimates of the parameters of the various models that will be fitted to the NAICPI data in this study.

4.3: Modelling and forecasting for the NAICPI data

In order to obtain the Fourier series model that will adequately describe the underlying behaviour of the NAICPI data, we tried five Fourier series models and picked the one with the highest R-square and minimum sum of squares for error. We fitted Fourier series models with one, two, three, four and five harmonic(s) to the data. We discovered that R-square and sum of squares for error increases and decreases respectively with increase in number of harmonics. Hence, the Fourier series model with five harmonics was selected. This model is given as

\[ X_t = \alpha_0 + \sum_{k=1}^{5} \{ \alpha_k \cos(w_k t) + \beta_k \sin(w_k t) \} . \]  (4.3)

Results from fitting the model in equation (4.3) to the NAICPI data are given in Table 4.1. The estimated Fourier series model fitted to the NAICPI data is adequate with only one insignificant parameter estimate. The adequacy of the fitted model is further buttressed by the value of the R square and the sum of squares for the error. The R square is 0.996 and the root mean square error is 113.8176.

To check the entertained the Fourier series model we consider the standardized residual series. The standardized residuals plot is shown in Fig. 4.1. The plot shows no particular pattern and hence the fitted Fourier series model is adequate. Also Fig. 4.2 and Fig. 4.3 show the normal probability plot and normal quantile plot of residuals obtained from
fitting the Fourier series model to NAICPI. The plots show that residuals are pure white noise.

Table 4.1: Estimated Fourier series model fitted to the NAICPI data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2963.984</td>
<td>271.449</td>
<td>10.919</td>
<td>.000</td>
</tr>
<tr>
<td>Cos(w)</td>
<td>1326.635</td>
<td>344.222</td>
<td>3.854</td>
<td>.000</td>
</tr>
<tr>
<td>Sin(w)</td>
<td>-4448.493</td>
<td>375.873</td>
<td>-11.835</td>
<td>.000</td>
</tr>
<tr>
<td>Cos(2w)</td>
<td>-2076.360</td>
<td>82.501</td>
<td>-25.168</td>
<td>.000</td>
</tr>
<tr>
<td>Sin(2w)</td>
<td>-1894.923</td>
<td>411.622</td>
<td>-4.604</td>
<td>.000</td>
</tr>
<tr>
<td>Cos(3w)</td>
<td>-1415.873</td>
<td>233.997</td>
<td>-6.051</td>
<td>.000</td>
</tr>
<tr>
<td>Sin(3w)</td>
<td>101.297</td>
<td>185.869</td>
<td>.545</td>
<td>.589</td>
</tr>
<tr>
<td>Cos(4w)</td>
<td>-592.731</td>
<td>162.751</td>
<td>-3.642</td>
<td>.001</td>
</tr>
<tr>
<td>Sin(4w)</td>
<td>655.094</td>
<td>72.990</td>
<td>8.975</td>
<td>.000</td>
</tr>
<tr>
<td>Cos(5w)</td>
<td>209.417</td>
<td>54.604</td>
<td>3.835</td>
<td>.001</td>
</tr>
<tr>
<td>Sin(5w)</td>
<td>346.526</td>
<td>66.121</td>
<td>5.241</td>
<td>.000</td>
</tr>
</tbody>
</table>

Fig. 4.1: Residuals plot of the fitted Fourier series model
Since the fitted Fourier series model is adequate in describing the behaviour of the NAICPI, we shall use it to make multi-step ahead forecast of the NAICPI data. Table 4.2
contains the several four steps (four years) ahead prediction of the NAICPI data. The actual NAICPI series are also given.

**Table 4.2: Multi Ahead Forecasts of the Fourier series model for NAICPI. The origin is 2004**

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td>7518.95</td>
<td>8799.36</td>
<td>10109.60</td>
<td>11311.30</td>
</tr>
<tr>
<td>Std. error</td>
<td>5.03</td>
<td>6.89</td>
<td>6.99</td>
<td>7.01</td>
</tr>
<tr>
<td>Actual</td>
<td>7446.43</td>
<td>8059.59</td>
<td>8493.64</td>
<td>9477.23</td>
</tr>
</tbody>
</table>

Figure 4.2 shows the plot of the actual NAICPI series and their corresponding fitted values using Fourier series model. From the figure we noticed that the forecasts are very close to the actual values but the tend deviate from the actual series as the forecast steps become long.

![Figure 4.2: Plot of NAICPI series and forecasts](image)

**Fig. 4.2: Plot of NAICPI series and forecasts**

### 4.4: Forecasting comparison

To evaluate the performance of the estimated models in describing NAICPI we compare their out-of-sample forecasts. The post-sample forecast comparisons are carried out as follows. First, we reserve the last four observations (4 years from 2005 to 2008) for forecast comparison. Secondly, both models used in forecasting are estimated using the first 45 observations as have mentioned earlier. Thirdly, for a given model the time origin of forecast
is \( t = 45 \). Fourthly, for a given model and the time origin \( t_0 \), we forecast the NAICPI series for the time points \( t_0+1, t_0+2, \ldots, 49 \). Therefore, we compute one to 4-step ahead forecasts. Such a scheme provides 4 one-step ahead forecasts. Finally, we summarize the forecast performance by considering the average percentage of predicted error (APE). Following Wong et al (2003) we define APE by

\[
APE = \frac{1}{4} \sum_{t=1}^{4} \left| \frac{\text{Forecast}(t) - \text{Actual}(t)}{\text{Actual}(t)} \right|. \tag{4.4}
\]

Table 4.3 below gives the results of the forecasting comparison.

<table>
<thead>
<tr>
<th></th>
<th>ARIMA(1,1,0)</th>
<th>Fourier series model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APE</strong></td>
<td>0.0207</td>
<td>0.0485</td>
</tr>
</tbody>
</table>

The four steps ahead predictions for the NAICPI data show that the APE of the AR(1) model is the minimum compared with the Fourier series model.
CHAPTER FIVE
SUMMARY AND CONCLUSION

In this project, we have fitted to the NAICPI data both the ARIMA and Fourier series models successfully. We discovered that both models can be used to describe and predict the NAICPI data adequately. The normal probability plots and normal quantile-quantile plots of the residuals show that residuals obtained from the fitted models are pure white noise. Based on the R square of the two models, for projection purposes, the Fourier series model is superior to the ARIMA model. The R square values for the two models are 0.996 for the Fourier series model and 0.991 for the ARIMA model. This superiority was achieved with a high cost of having to estimate eleven parameters. The four steps ahead predictions for the NAICPI data show that the APE of the ARIMA approach is the minimum compared with the Fourier series model. The values are 0.0485 for the Fourier series model and 0.0207 for the ARIMA model. Interest here is on minimizing the APE value. Also the Root Mean Square Error (RMSE) of the forecasts was used to compare the out-sample forecast performance of the fitted models. It is observed that the RMSE value for the Fourier series model is 1277.46, which is three times greater than that of ARIMA model (RMSE value is 471.7657). Consequently, based on the APE and RMSE values, the ARIMA(1,1,0) or the differenced AR(1) model is recommended for predicting the NAICPI data.
References


