PREDICTION OF HOUSING PRICES: AN APPLICATION OF THE ARIMA MODEL

BY

MICHAEL SICHANGI MUKOPI

School of Mathematics

College of Biological and Physical Sciences

University of Nairobi

A project submitted in Partial Fulfillment of the Requirement for the degree of Master of Science in Social Statistics

November 2012

DECLARATION

Candidate:

This project is my original work and has not been presented for a degree in any other university.

Signature ______________ Date 30/11/2012

Michael Sichangi Mukopi - I56/65128/2010

Supervisor:

This project has been submitted for examination with my approval as university supervisor

Signature ______________ Date 30|11|2012

Dr. Kipchirchir Isaac Chumba

School of Mathematics

University of Nairobi

P.O. BOX 30197 – 00100

NAIROBI

ACKNOWLEDGEMENT

I am very grateful to my supervisor (Dr. Kipchirchir Isaac Chumba) for his steadfast guidance right from the start of my project formulation. I would also like to thank the teaching and non-teaching staff in The School of Mathematics, University of Nairobi for their support particularly my M.Sc. Lecturers (Prof. M. Manene, Prof. J.A.M Otieno, Prof. P. Weke, Dr. Oeba Vincent Onguso and Mr. E. Obudho)

May I single out my my Employer Mobile Decisioning Africa Ltd for giving me support and encouragement for me to pursue my Master of Science degree in Social Statistics.

I appreciate the support from my family and above all, I thank the almighty God for granting me good health throughout the study period.

DEDICATION

This degree is dedicated to my family and my daughter Laila Nafula Mukopi.

ABSTRACT

This study uses an ARIMA model to provide out-of-sample forecasts for United States housing prices as represented by the Case Schiller Index during the financial crises timeline (between 2005 and 2009). The major findings are that the model fails to predict the peak/turning point of the financial crisis but successfully predicted declining prices since 2006:6; therefore the magnitude of the loss realized during that period could have been reduced had the models prediction been considered. The model predicted extremely negative one year ahead prices in 2008:2, 2008:3 and 2008:9 which explains the timeline of the collapse of the Bear Stearns and Lehman Brothers as well as the subsequent global financial meltdown.

ABBREVIATIONS AND ACRONYMS

ACF - Autocorrelation Function

ADL - Autoregressive Distributed Lag

AIC - Akaike Information Criterion

ANN - Artificial Neural Networks

AR - Autoregressive Model

ARCH - Autoregressive Conditional Heteroskedasticity

ARIMA - Autoregressive Integrated Moving Average

ARMA - Autoregressive Moving Average

FA-VAR - Factor Augmented Vector Autoregression

PACF - Partial Autocorrelation Function

S&P - Standard and Poors

VAR - Vector Autoregressive

VECM - Bayesian Vector Error Correction

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND INFORMATION

Housing can refer to the physical and environmental attributes of houses. It encompasses housing as an economic good, a social good, a social infrastructure, a political object and a body of knowledge. Therefore housing statistics can be described as a collection of numerical facts or data on the state of housing.

The statistics on housing prices are produced by exploiting a number of data sources and registers. The data covers the actually published statistical variables but in some statistics the processing necessitates the use of data on background variables that are retrieved from diverse registers that are quite exhaustive and of good quality. In this respect, you can have various sources of data on statistics for housing prices which includes; prices on dwelling, rents on dwelling and real estate prices. Housing price statistics (dwelling prices, real estate prices, dwelling rents) have been used to come up with housing price indices that help to distinguish better than before the real price developed from price changes caused by dwelling characteristics in different time periods.

The housing market differs in many ways from the text book model of a liquid asset market with exogenous fundamentals. This implies that housing prices can be influenced not only by general supply and demand conditions, but also by idiosyncratic factors such as the urgency of sale and the effects of the ownership transfer on the physical quality of the house. Other factors that may influence housing prices include; the incomes of potential buyers and the ability to make payments, borrow money, and the cost of borrowing money since a large percentage of homes purchased globally are purchased with a mortgage. Given the illiquid nature of the housing market, housing prices tend to decline in the case of increased foreclosures. Foreclosures transfer houses to financial institutions that must maintain and protect them until they are sold.

These foreclosed houses are likely to sell at low prices, both because they may be physically damaged during the foreclosure process, and because financial institutions have an intention to sell them quickly. In a liquid market, an asset can be sold rapidly with minimal impact on its price, but the characteristics of housing discussed above are a classic example of how urgent sales lower prices. The supply and demand effect on housing prices and the subsequent impact of foreclosures on mortgage related instruments can be observed by the perceived “credit crunch” that was observed between the years 2007 and 2008 in the United States.

The concept of a “credit crunch” has a long standing history that can be traced back to the Great Depression of the 1930’s and more recently to the recession of 1990-91 that was characterized by a decline in the supply of credit controlling for the stage of the business cycle. The term credit crunch can thus be used to explain the curtailment of credit supply in reaction to a decline in the value of bank capital and conditions imposed by regulators that require banks to hold more capital than they previously would have held. A milder version of a credit crunch can be referred to as a “credit squeeze” and arguably this was what was observed between 2007 and early 2008.

The recent global financial crisis experienced between the years 2007 and 2008 can be attributed to the effects of financial innovation in mortgage-backed securities. The crisis occurred when housing prices fell and loan defaults increased as a result of the rapid growth of the sub-prime mortgage market in the United States which was characterized by having non-standard mortgage facilities offered to individuals with nonstandard income and credit profiles – but it is really a crisis that occurred as a result of mispricing of the risk associated with these products. New assets were developed based on subprime and other mortgages and sold to investors in the form of repackaged securities for increased sophistication. These securities provided higher returns than conventional securities and were regarded as safe due to their high ratings.

These repackaged securities were however not as safe as they were deemed to be mainly because their value was closely tied to the movements in house prices. As earlier mentioned, the securities offered high returns compared to other investment vehicles when housing values rose but when housing values began to fall, foreclosures on mortgages increased which sent shockwaves through the financial markets and resulted to the financial crisis.

The earliest signs of the impending crisis were seen in early 2007 when in April of that year New Century Financial, a subprime specialist had filed for chapter 11 bankruptcy and laid off half its employees; and in early May 2007 the Swiss owned investment bank UBS had closed the Dillon Reed hedge fund after incurring huge mortgage-related losses. Although these seemed like isolated incidents at the time, that month Moody's (a rating agency) announced that it was reviewing the ratings of 6 asset groups based on 21 U.S subprime mortgage securitizations. In June 2007 Bear stearns supported two failing hedge funds, and in June and July three rating agencies all downgraded subprime related mortgage products from their "safe" AAA status. These events were later to develop into the full-scale credit-crunch of 2007-08.

Many pundits believe that the financial crisis could have been prevented if predictions from both simple and complex econometric models had been used.

1.2 PROBLEM STATEMENT

Before 2007, the subprime mortgage market saw an enormous growth trend which concurrently led to the introduction of repackaged debt securities of increasing sophistication. These securities offered high returns than other conventional securities but were tied to movements in housing prices. When housing prices rose so did the return on the securities but when prices fell, foreclosures on mortgages rose drastically which led to the credit crunch. Despite the availability of both complex and simple econometric tools that could have been use to predict the crisis, many analysts ignored such input which if used could have prevented the financial crisis all together.

1.3 OBJECTIVES OF THE STUDY

The objective of the study is to determine if an ARIMA model can be used to effectively predict both the movements of housing prices and specifically describe the timeline of the financial crisis that occurred in the U.S between the years 2007 and 2008.

1.4 JUSTIFICATION FOR THE STUDY

Many benefits can be derived from the standardization of social statistics data in the Kenyan context. Health statistics can be used to determine and improve the state of the healthcare system, crime statistics may be used to influence policies on management of crime in different regions and housing statistics can be used by governments to address issues on distribution of resources and can also give an indication of different dynamics on the economic state of a nation.

Given the benefits that can be derived from having a standardized approach with respect to the collection and maintenance of social statistics data, it is of glaring note that there exists a huge gap in the Kenyan system as far as data collection of these elements of social statistics is concerned. The data collected by a majority of the responsible agencies is inaccurate, irrelevant or too little to derive any significant inferences.

The recent global financial crisis as experienced in the United States can be used as a case study of the importance of having a standardized approach of collecting housing data. The crisis was mainly caused by an unexpected fall in housing prices and resulted to a global glut. Given that US housing prices are collected in the form of the Case-Shiller index, it can be shown that the crisis could have been predicted and avoided beforehand thus affirming the importance of having such data.

CHAPTER 2

LITERATURE REVIEW

Housing price forecasts are important in predicting mortgage defaults; property taxes, and other consumption investment and policy decisions (Kochin and Parks, 1982). Furthermore, studies by Case and Schiller (1989) indicate that housing prices are forecastable to a certain degree. It is therefore of great importance to determine which forecasting models can best describe future movements of housing prices.

Works by Larson (2011) compares the performance of different forecasting models on California housing prices and finds that multivariate theory models outperforms other time series models across a range of forecast comparison procedures. It is shown that incorporation of theoretic economic relationships into empirical forecasting models greatly improves forecast results. Although a myriad of market analysts tend to predominantly rely on multivariate theory models for their predictions, one of the critiques to the models is that they rely on functional assumptions to ascribe a form to fit the relationships of the variables (Lam et. al., 2008)

However, Artificial Neural Network models (ANN) are designed to capture functional forms automatically allowing the uncovering of hidden nonlinear relationships between the modeling variables. Results of the study conducted by Lam et. al. (2008) rightfully indicates that ANN models produced by recurrent back-propagation neural networks to produce housing pricing models for Hong Kong outperform multivariate models. These models also have the advantage of being relatively inexpensive to produce and can predict trend movements without the need to quantify some data. Despite these benefits, ANN models are notoriously "black-box" in nature and lack capability of explanation, they are built through learning and adjusted weights during training and hence have no definite formula and they require a huge amount of data for learning to achieve optimum results.

Simple Vector Autoregressive models (VAR) with error correction have been used to determine the causality between housing sales and prices (Zhou, 1996). Results indicated that there exists a bidirectional causality relationship between the two metrics. Price affects sales significantly but sales affects price weakly. The VAR model was then used to forecast sales and price for existing single-family housing during the period 1991 to 1994 by using a recursive method. The findings showed that the predictions for sales and price fit the data well.

Vitner and Iqbal (2009) used six different models to project changes in US home prices, including AR, ARIMA, Bayesian vector autoregression-level (VAR-level), Bayesian vector autoregression-difference (VAR-difference), Bayesian vector error correction model (VECM) and Bayesian factor augmented vector autoregression model (FA-VAR). The models were applied to the three most important measures for house prices in the US which are: FHFA purchase only index, S&P Case-Shiller index of house prices and the FHFA index of house values. Results of the study indicated that the Bayesian FA-VAR outperforms other models in terms of the out-of-sample root mean square error criteria.

Decision makers normally have multiple forecasts of a particular variable available to them, it would therefore be difficult to know how best to exploit the information available in individual forecasts. A solution to the problem as suggested by Drought and McDonald (2011) is to combine individual forecasts to produce a single summary forecast a concept known as model combination. Model combination approaches have the advantage of being more robust to the misspecification biases of individual forecasts. Since individual models are subject to different biases, a combination will average out the biases and improve forecast accuracy. The study conducted by Drought and McDonald generated a combined forecast using three approaches: equal weights, mean square error weights and model selection. Results showed that model selection techniques had lower root mean squared forecast errors than the combination methods. This result is at odds with the earlier assumption and works by Timmerman (2006) that simple averages often produce more accurate forecasts than the best performing model at the time.

Autoregressive conditional heteroskedasticity models (ARCH) can also be used to forecast housing prices. Studies on the housing price index of Shanghai from Jan 1999 to Oct 2003 by Gongliang and Fenjie (2003) show that ARCH models based on the autoregressive distributed lag model (ADL) are better at analyzing and forecasting the volatility and trend of an index. Although these models better forecast short-term changes than other commonly used methods during relatively high volatile periods they also have a tendency to overestimate the volatility impacts for forecasting mild price movements.

Various time-series methods for forecasting housing prices have been employed in a growing body of empirical studies (e.g., Brown et al., 1997; Pace et al., 2000; Gu, 2002). Due to the boom-bust cycles and the substantial transaction costs in the housing market, the literature predominantly assumes a nonlinear relationship linking housing prices to a set of publicly known factors or to unobservable states (Muellbauer and Murphy, 1997; Capozza et al., 2002; Miles, 2008). However, Crawford and Fratantoni (2003) indicate that a linear ARIMA model displays better out-of-sample forecasting of home prices than Markov-switching and GARCH models, although the Markov-switching model is superior for the in-sample fit. Case and Schiller (1989, 1990) were pioneers in their use of linear techniques to forecast housing prices. The view holding that simpler linear specifications produce better forecasting performances than complex and sophisticated methods is shared by Timmermann and Perez-Quiros (2001) in the context of stock returns. Moreover, due to its adaptive process, the forecasts from the ARIMA models are less affected by structural breaks (Clements and Hendry, 1996).

CHAPTER 3

METHODOLOGY

3.1 DATA

The data consists of the S&P (Standard and Poors) Case Shiller seasonal-adjusted 10-city composite index as the housing price from 1987:1 to 2009:1. The index is the leading measure for the US residential market, tracking movements in the value of residential real estate in 10 metropolitan regions. The index is calculated monthly using the repeated sales methodology which is widely considered as the most accurate way of measuring price changes for real estate. The methodology measures the movement in prices of single family homes by collecting data on actual sale prices of single family homes in their respective regions. When a house is resold at some period later, the new sale price is matched and compared to its first sale price. The difference in the "sales pair" is measured and recorded and finally aggregated into one index.

3.2 TIME SERIES MODELS

3.2.1 Autoregressive Models

An autoregressive process $\{Y_t\}$ of order p takes the form:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t \qquad \textbf{(3.1)}$$

Where Y_t is the value of the series at time t and Y_{t-1} , Y_{t-2} , …, Y_{t-p} are dependent on the previous values of the variable at specified time periods; α_1, α_2, … ,α_p are the regression coefficients and ε_t is the residual term that represents random events not explained by the model. It is abbreviated AR(p).

The Autoregressive model is capable in a wide variety of time series forecasting by adjusting the regression coefficients α. The difference between the Autoregressive models and other conventional regression models is with respect to the assumption of the independence of the error term. Since the independent variables are time-lagged values for the dependent variable, the assumption of uncorrelated error is easily violated.

3.2.2 Moving Average Models

The basic idea of Moving-Average model is firstly finding the mean for a specified set of values and then using it to forecast the next period and correcting for any mistakes made in the last few forecasts. A moving average process $\{Y_t\}$ of order q takes the form

$$Y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \qquad \textbf{(3.2)}$$

where Y_t is the value of the series at time t, β_1 , β_2 ,....., β_q are the weights applied to

ε_{t-1} , ε_{t-2},......, ε_{t-q} previous forecast errors and ε_t is the residual error. It is abbreviated by MA(q).

To specify a Moving-Average, the number and the value of the q moving average parameter β_1 through β_q have to be decided subject to the certain restrictions in value in order for the process to be stationary. The Moving-Average model works well with stationary data, a type of time series without trend or seasonality.

3.2.3 Autoregressive Moving Average Models

While the AR and MA models can be used for many data sets, there are some data for which they are not adequate, and a more general set of models are needed.

The time series $\{Y_t\}$ is said to be an autoregressive moving average process of orders p and q abbreviated ARMA (p,q) if Y_t satisfies the equation

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} \qquad \textbf{(3.3)}$$

This can be re-written as

$$h(\mathrm{L})\, Y_t = \mathrm{g}(\mathrm{L})\, \varepsilon_t \qquad \textbf{(3.4)}$$

where

$$h(\mathrm{L}) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$$

and

$$\mathrm{g}(\mathrm{L}) = 1 + \beta_1 L + \ldots + \beta_q L^q$$

are polynomials of order p and q respectively.

and L is a lag operator defined as

$$L^i Y_t = Y_{t-i}$$

The ARMA (p,q) model is stationary provided the roots of $h(\mathrm{L}) = 0$ lie outside the unit circle and invertible provided the roots of $g(\mathrm{L}) = 0$ lie outside the unit circle.

If $p \neq 0$ and $q \neq 0$ we have a mixed ARMA. However, if $p \neq 0$ and $q = 0$ we have a pure AR(p). whereas if $p = 0$. $q \neq 0$ we have a pure MA(q).

3.2.4 Autoregressive Integrated Moving Average Models

ARIMA modeling takes into consideration historical data and breaks it down to the Autoregressive process, where there is a memory of past events (e.g. the housing price this month is related to the housing price of last month, and so forth with a decreasing lag); an integrated process which accounts for stabilizing thus making the data easier to forecast; and a Moving Average (MA) of the forecast errors such that the longer the historical data the more accurate the forecasts will be over time. The three components are all combined and interact with each other and recomposed to form the ARIMA (p,d,q) model.

In general we say that $\{Y_t\}$ is an ARIMA process of order p, d, q if the d^{th} difference of Y_t is a stationary, invertible ARMA process of order p, q. Thus using the lag operator L

$$h(\mathrm{L})\,(1 - \mathrm{L}\,)^d Y_t = g(\mathrm{L})\varepsilon_t \qquad \textbf{(3.5)}$$

Where ε_t is white noise, $h(\mathrm{L})$ and $g(\mathrm{L})$ are polynomials of degree p and q respectively with all roots of the polynomial equations $h(\mathrm{L}) = 0$ and $g(\mathrm{L}) = 0$ lie outside the unit circle.

3.3 The autocovariance and autocorrelation functions

The autocovariance function of a stationary process $\{Y_t\}$ is

$$\Upsilon(\mathrm{k}) = \mathrm{cov}\{Y_t, Y_{t-k}\} \tag{3.6}$$

One consequence of the definition of stationarity is that Υ(k) does not depend on t. Since Υ(0) is the variance of each Y(t), we define the autocorrelation function as;

$$\rho(k) = \Upsilon(\mathrm{k}) / \Upsilon(0) \tag{3.7}$$

For a stationary process $\{Y_t\}$, k is necessarily an integer.

The autocorrelation function is an important, albeit incomplete, summary of the serial dependence within a stationary random function. General properties of $\rho(k)$ include;

(a) $\rho(0) = 1$

(b) $\rho(k) = \rho(-k)$,

(c) $-1 \leq \rho(k) \leq 1$,

(d) If Y_t and $Y_{t-k}$are independent, then $\rho(k) = 0$.

3.3.1 Estimating the autocorrelation function

ARMA and differenced ARIMA models are identified through patterns in their autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) . Both ACFs and PACFs are computed for the respective lags in the series and plotted. The ACF plot is defined as a graph of the sample autocorrelation co-efficients r_k

against the corresponding lags k, each r_k being defined as

$$r_k = g_k/g_0 \tag{3.8}$$

where

$$g_k = \sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})/\text{n} \tag{3.9}$$

is the kth sample autocovariance coefficient and $\bar{y} = \sum_{t=1}^{n} y_t$ is the sample mean.

The simplest pattern to detect from visual inspection is a cut-off with respect to an MA(q), that is, all r_k for k greater than some integer q are approximately zero. Smoothly decaying autocorrelations are more difficult to detect by visual inspection, except when the decay takes a particularly simple form such as the exponential decay associated with the AR(1) process. It is for this reason that we introduce a variant of the correlogram, the partial correlogram, which has a cut-off at lag p for an underlying AR(p) process. The correlogram and partial correlogram of a mixed ARMA (p,q) fails to cut-off.

3.4 THE PARTIAL AUTOCORRELATION FUNCTION

The PACF takes the form

$$\pi(k) = \text{Cov}\big(Y_t - \text{E}(Y_t \;/\; Y_{t+1}, Y_{t+2},\ldots, Y_{t+k-1}),\ Y_{t+k} - \text{E}(Y_{t+k}/\ Y_{t+1}, Y_{t+2},\ldots, Y_{t+k-1})\big) \tag{4.0}$$

where

$E(Y_t \,/\, Y_{t+1}, Y_{t+2},\ldots, Y_{t+k-1})$ and $E(Y_{t+k}/\, Y_{t+1}, Y_{t+2},\ldots, Y_{t+k-1})$ are the predictions of Y_t and Y_{t+k} given $Y_{t+1},\ldots, Y_{t+k-1}$ respectively.

in particular,

$$\pi(0) = 1$$

$$\pi(1) = \rho(1)$$

$$\pi(2) = \frac{\rho(2) - \rho((1))^2}{1 - \rho((1))^2}$$

3.5 FORECASTING WITH ARIMA MODELS

The Box-Jenkins methodology is used to identify and select a suitable model for the time series data as shown in the flow chart below

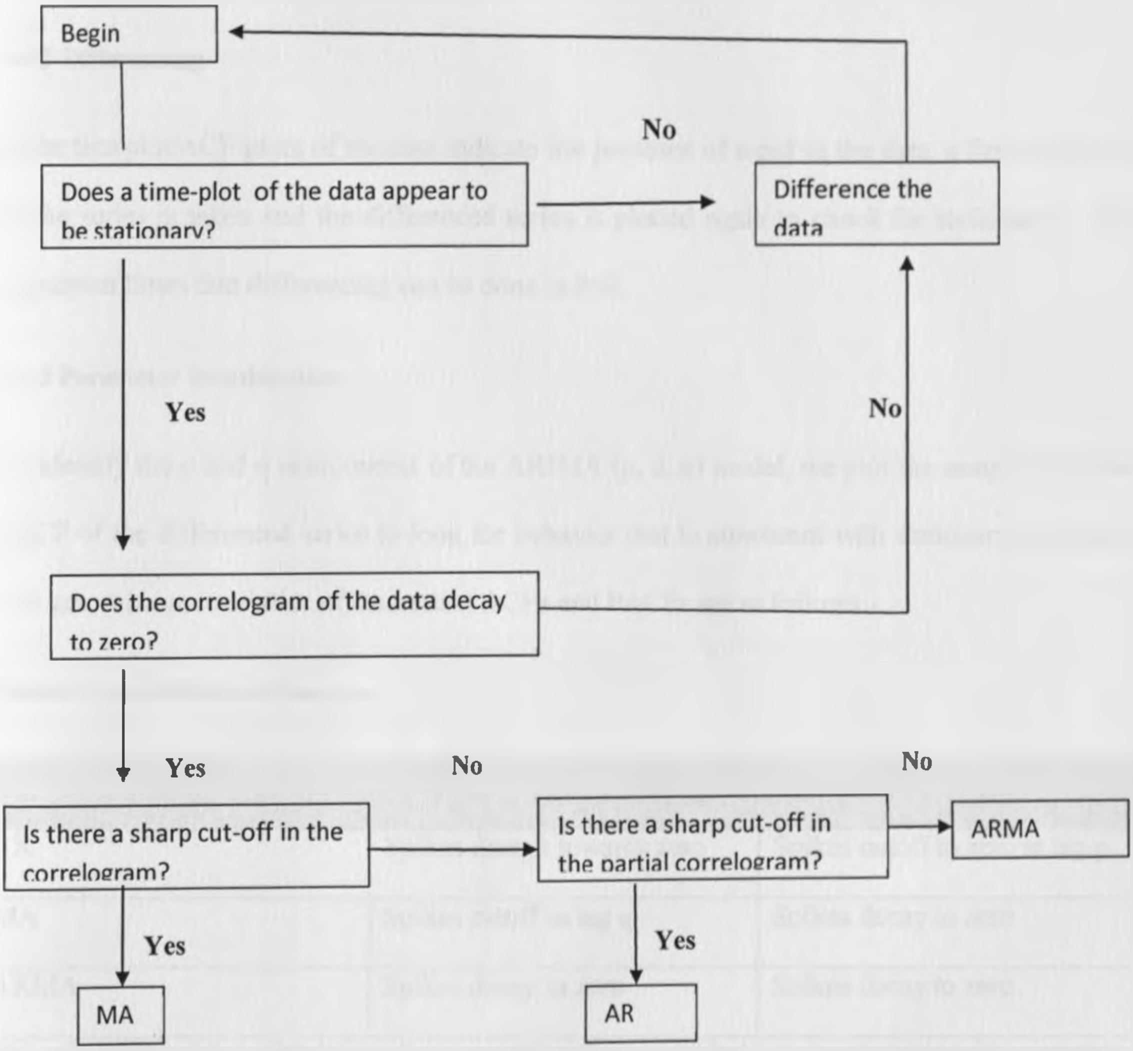

Begin
No
Does a time-plot of the data appear to be stationary?
Difference the data
Yes
No
Does the correlogram of the data decay to zero?
Yes
No
No
Is there a sharp cut-off in the correlogram?
Is there a sharp cut-off in the partial correlogram?
ARMA
Yes
Yes
MA
AR

Figure 3.1: Flowchart illustrating the identification process

3.4.1. Stationarity

The first stage of identification of an ARIMA process is to establish the stationarity of the series by examining the timeplot of the data, $\{Y_t\}$. If the data is not stationary, the timeplot appears to have a trend.

ACF plots can also be used to demonstrate non-stationarity if they have a linear decay.

3.4.2 Differencing

If the timeplot/ACF plots of the data indicate the presence of trend in the data, a first difference of the series is taken and the differenced series is plotted again to check for stationarity. The maximum times that differencing can be done is d=2.

3.4.3 Parameter Identification

To identify the p and q components of the ARIMA (p, d, q) model, we plot the sample ACF and PACF of the differenced series to look for behavior that is consistent with stationary processes. The general characteristics of theoretical ACFs and PACFs are as follows:-

Table 3.1: Identification of the order

Model	ACF	PACF
AR	Spikes decays towards zero	Spikes cutoff to zero at lag p.
MA	Spikes cutoff at lag q	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

Once the order of the ARIMA (p, d, q) model has been specified, we can then generate parameter estimates for the series. The list generated contains the coefficients, residuals and the Akaike Information Criterion (AIC).

3.4.4 Model Diagnostics

Diagnostic checking of the fitted models involves analysis of the residuals from the fit for any signs of non-randomness. Box Jenkins methodology requires examining the residuals of the actual values minus those estimated by the model, if such residuals are random it is assumed that the model is appropriate. If not another model is entertained, parameters estimated, and residuals checked for randomness.

Signs of non-randomness can be checked by plotting a diagnostic plot that contains a plot of the residuals, autocorrelation of the residuals and the p-values of the Ljung-Box statistic for the first 10 lags.

When faced with a group of candidate models, the Akaike Information Criterion can be used as a model selection condition. For the fitted ARIMA time series of length n, the AIC is defined as

$$\mathrm{AIC}=\ln(\sigma^2_{p,q})+2(p+q)/n \tag{3.10}$$

where $\sigma^2_{p,q}$ is the residual error variance from the fitted model. When comparing the fitted models the basic idea is that the smaller the AIC the better the fit. The AIC penalizes for additional model complexity with the addition of 2(p+q)/n.

3.5 SOFTWARE

The housing index data was analyzed using **R** software which is a widely used environment for statistical analysis. It is an open source software which is maintained by scientists for scientists. R has gained many users and contributors which increases the capabilities of the software by releasing add-ons (packages) that offer functionalities that suit users' needs. Various packages of the software were downloaded and used to perform the end to end process from model identification to the eventual prediction of the 10 city composite index for this analysis.

R is distributed by the "Comprehensive R Archive Network" (CRAN) and is available from the Url: http://cran.r-project.org

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 Stationarity Check

The first step was to check for the stationarity of the series which was done by generating the time-plot for the data as shown below

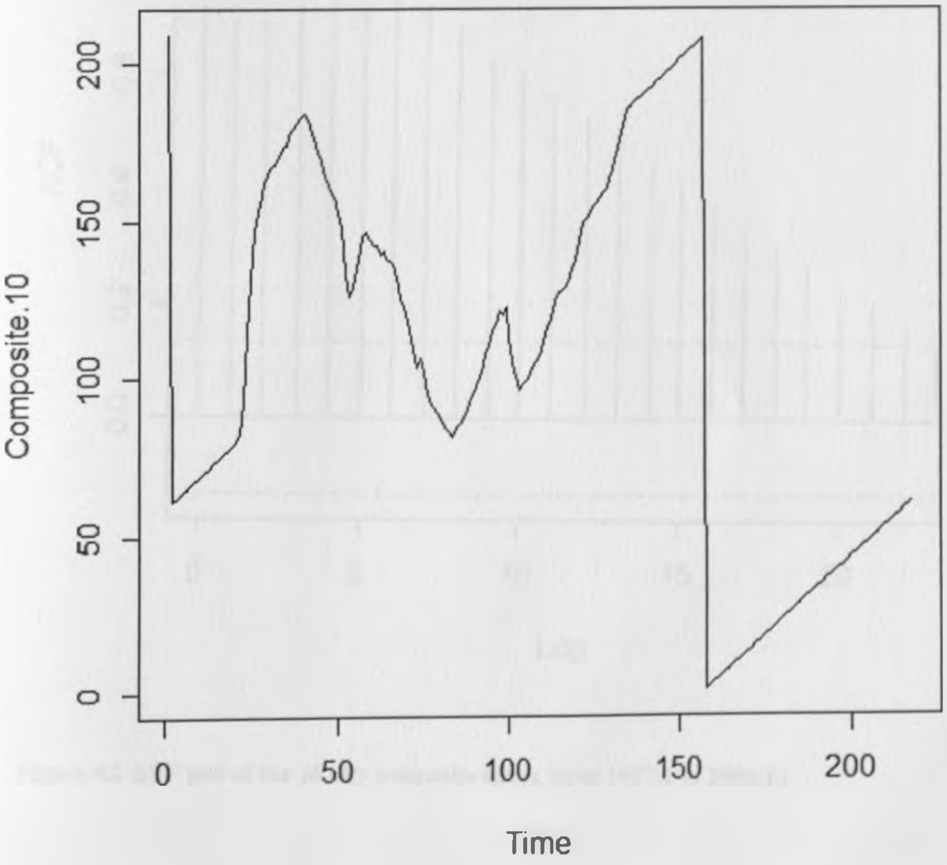


Figure 4.1 Time series plot of the 10 city composite index from 1987:1 to 2004:12

From the time series plot, there exists some evidence of trend which is indicative of non-stationarity in the 10 city composite index. This was also compounded by the ACF plot as shown below;

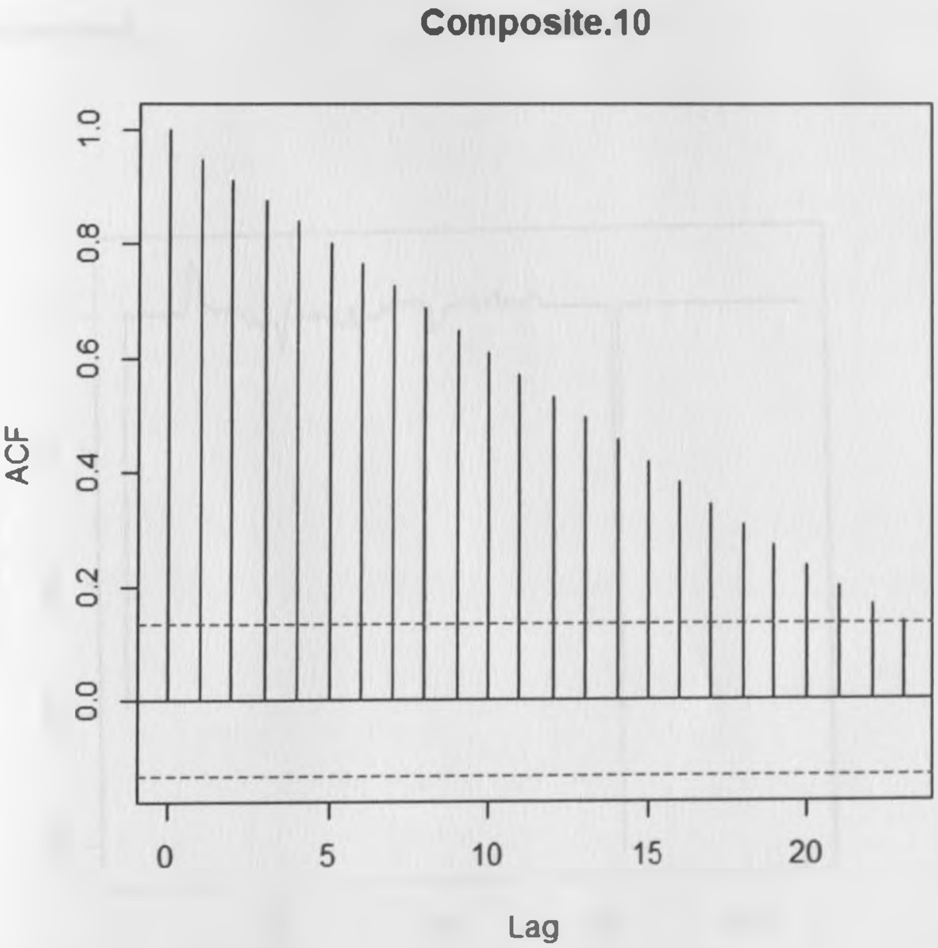


Figure 4.2 ACF plot of the 10 city composite index from 1987:1 to 2004:12

The ACF plot has a significantly linearly decaying pattern which indicates a non-stationary process.

4.2 Differencing the series and Parameter Identification

Given that the series was not stationary, the 10-city composite index was differenced once to eliminate trend.

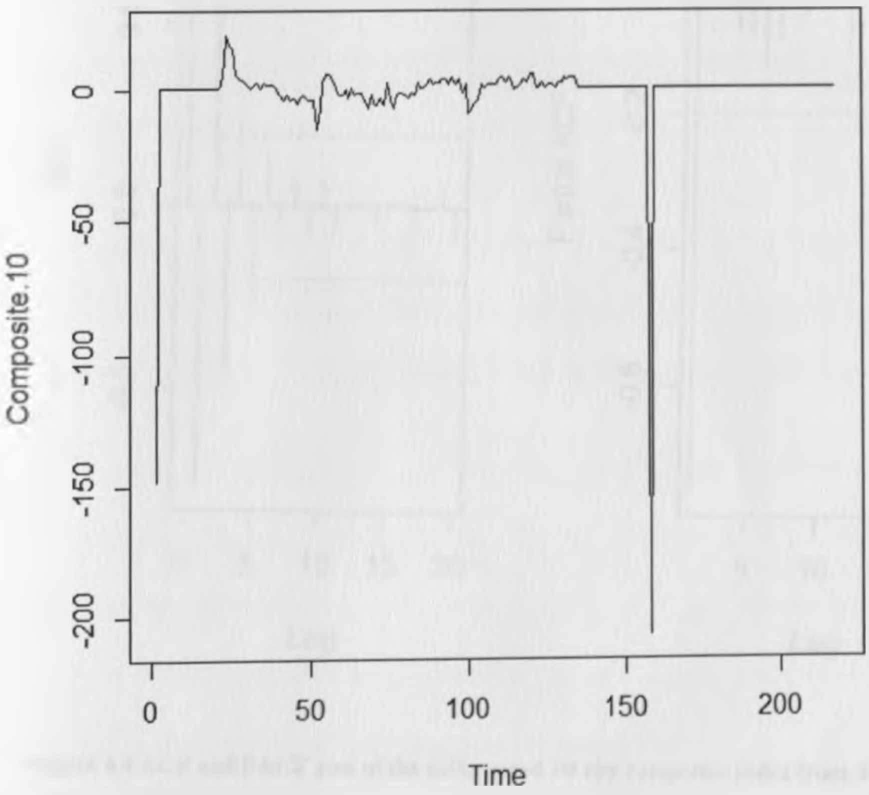


Figure 4.3 First differenced Time series plot of the 10 city composite index from 1987:1 to 2004:12

From figure 4.3 above it was found that the trend in the 10 – city composite index was eliminated by differencing the data once hence d = 1.

Given that the data was stationary, the ACF and PACF plots for the data were plotted to aid in identification of the p and q components in the ARIMA model as shown below;

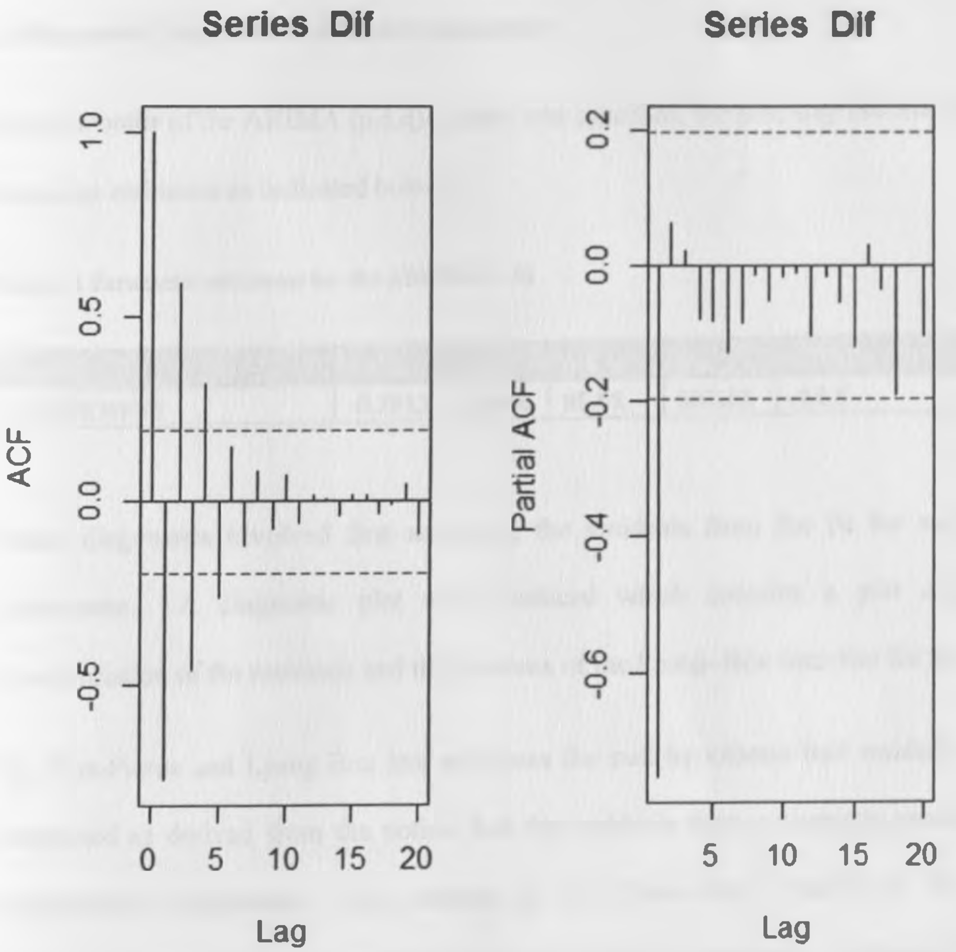


Figure 4.4 ACF and PACF plot of the differenced 10 city composite index from 1987:1 to 2004:12

From figure 4.4 we can see that the ACF plot of the differenced series decays more quickly. The sample PACF cuts off after lag 1. This behavior is consistent with a first-degree autoregressive-AR(1)-model. We thus specified and fit an ARIMA(1,1,0) model for the S&P Case Shiller 10-city composite index.

4.3 Parameter Estimation and Model Diagnostics

Once the order of the ARIMA (p,d,q)- model was specified, the next step involved generating the parameter estimates as indicated below;

Table 4.1 Parameter estimates for the ARIMA(1,1,0)

	AR(1)	S.E	AIC	BIC	LOG-LIKELIHOOD	Sigma ^2
CO-EFFICIENTS	-0.7912	0.0693	95.01	102.82	-44.5	0.1412

Model diagnostics involved first analyzing the residuals from the fit for any sign of non-randomness. A diagnostic plot was produced which contains a plot of residuals, the autocorrelation of the residuals and the p-values of the Ljung- Box statistics for the first 10 lags.

The Box-Pierce and Ljung Box test examines the null hypothesis that residuals are randomly distributed as derived from the notion that the residuals from a correctly specified model are independently distributed. After running the Box-Pierce test to the fitted 10-city composite index, we accepted the null hypothesis that the residuals are randomly distributed given a p-value of 0.4452. The diagnostic plot of residuals is shown below;

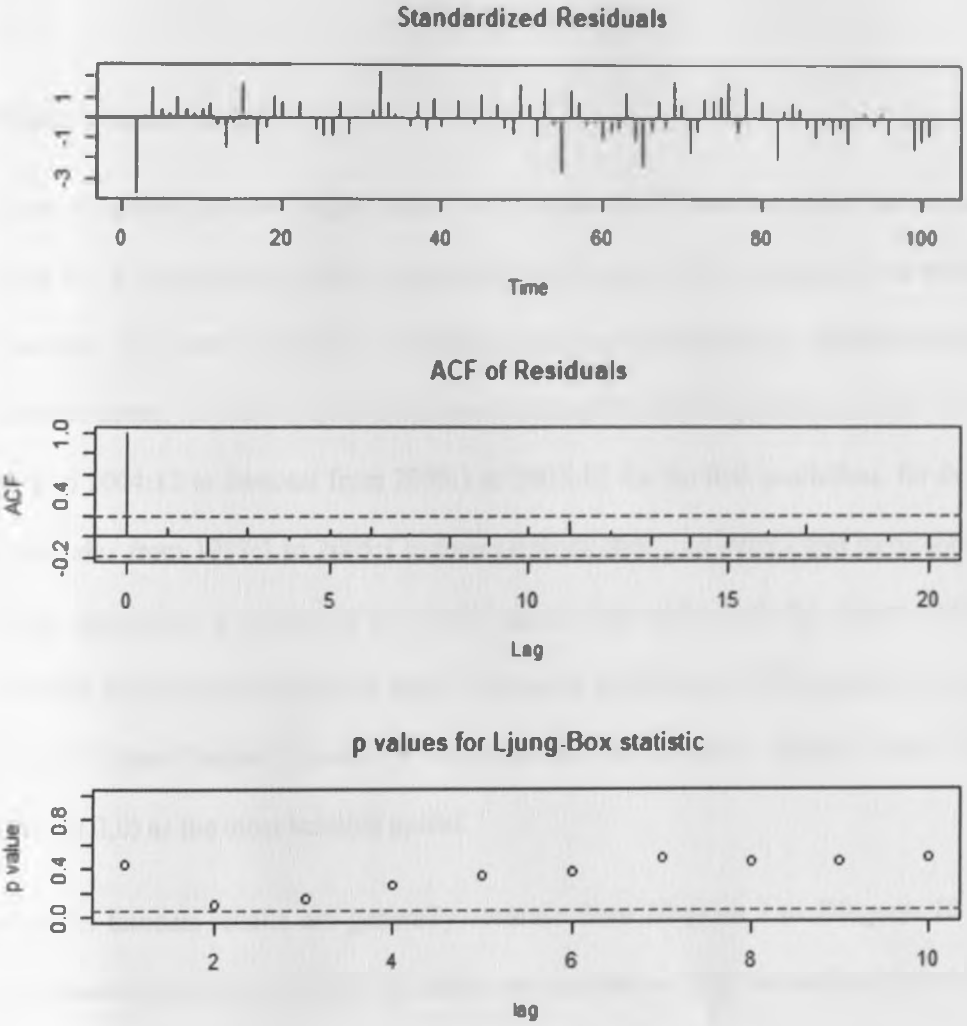


Figure 4.5 Diagnostic plot of residuals

4.4 Model Forecast Results

The out of sample forecast begins from 2005:1 where the in sample period lies between 1987:1 to 2004:12. The in sample is then expanded by one month and the parameters re-estimated based on the most up-to-date information available at the time of the forecast. Each forecast horizon is 12 months ahead. As shown from Diagrams 1 through 50 in the appendix, we use the index from 1987:1 to 2004:12 to forecast from 2005:1 to 2005:12 for the first prediction, for the second we use the index from 1987:1 to 2005:1 to forecast from 2005:2 to 2006:1 and so on thereafter. The ARIMA prediction is presented as a solid green line while both the upper and lower 95% confidence bounds are represented with a solid grey line. Each ARIMA prediction goes through the model identification process as outlined in the previous chapter where we use the ARIMA(1,1,0) as the most suitable model.

The model forecast results are generally accurate from Diagram 1 to Diagram 10 (forecasting made between 2004:12 and 2005:9), where we can observe that the realized housing prices are mostly within the 95% confidence interval of the 12 months out-of-sample forecasts. Beginning from Diagram 11(forecasting made in 2005:10) the out-of-sample results begin to deteriorate where only the first four months predictions are accurate. From Diagram 12 to 17, the model inaccurately predicts a rising trend in housing prices where it fails to detect the peak/turning point of the housing prices in 2006:4. In Diagram 18 (2006:5), one month after the turning point, it is surprising to see that not all realized prices lie within the models forecasting interval. In short the model does not predict the housing bubble until 2006:6.

In Diagram 19 (2006:6), two months past the turning point, the model accurately predicts the declining housing price with a modest declining rate. During the extraordinary turnaround of the housing market as seen in Diagram 20 to 29 (2006:7 and 2007:4), the model exhibits a downward trend, however for the most part the realized housing prices are seemingly below the models forecasting lower bound and the one year ahead forecasting price is still above 200. In Diagram 30 (2007:5), for the first time the model was able to predict a one year ahead price that is below 200. This means that at this time, the model begins to recognize the length and breadth of the housing slump. It is during this period in real time that the financial market began to show concern about mortgage-back security. From Diagrams 31 to 38 (2007:6 to 2007:12), the model predicts a declining trend but the realized decline is deeper than what is predicted.

Predictions made in Diagrams 39 and 40 (2008:2 and 2008:3), produce a lower price forecast than what is realized where the one-year-ahead forecasts lie between 140 and 150 (a significant decline from the peak). At this time the Bear Stearns collapsed. From Diagram 46 (2008:9), we can observe that for the first time the model predicts a much lower price of 130 in one year. It is in this month that Lehman Brothers' went bankrupt and the financial markets exploded. From this, we can learn that the timeline of the financial crisis was not driven by shocks due to a specific company's financial problem or due to government decision making but is rationally driven by a forecasting model. The forecasted values from the model have a direct effect on the values of the mortgage-related securities and therefore should determine the investor's investment decisions.

Diagrams 46 to 50 (2008:9 to 2009:1) show the worst outlook of the housing prices which coincides with the period where there was wide-spread panic of a global financial meltdown.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

There are five key insights that were derived from the study. First, the ARIMA model was not able to predict the turning point of the housing prices that occurred in 2006:4; it recognizes the peak after a two month lag i.e 2006:6. Second, the model predicts mild declining prices from 2006:6 to 2007:5; while the model fails to capture the gravity of the housing slump during this period, it explains the illogicality of longing mortgage-related assets during this period. Third, the model predicts deeper declining prices from 2007:6 to 2008:1 which reflects the beginning of the housing crisis but fails to capture the scale of the housing burst. Fourth, in 2008:2 and 2008:3, the model predicts significantly worse prices, which coincides with the collapse of the Bear Stearns in 2008:3. Finally, from 2008:9 to 2009:1, the model forecasts worsening housing prices which coincides with the meltdown of the global financial market.

5.2 RECOMMENDATIONS

The standardized approach of capturing housing price data in the U.S in the form of the Case-Shiller index has proven to be of significant worth given that such information can be used to make certain inferences/predictions that touch on the economic condition of a nation. In the Kenyan context, it is therefore of great importance that the government adopt a standardized approach to the capturing of such information as it bears a direct correlation to the performance of the economy. Predictions can then be made based on such information that consequently advice fiscal policy.

APPENDIX

Diagram 1

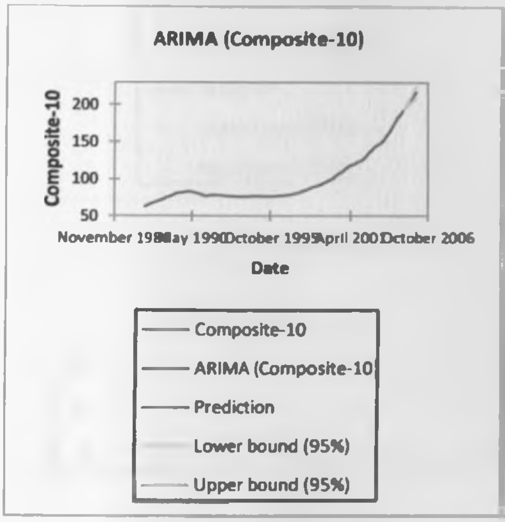
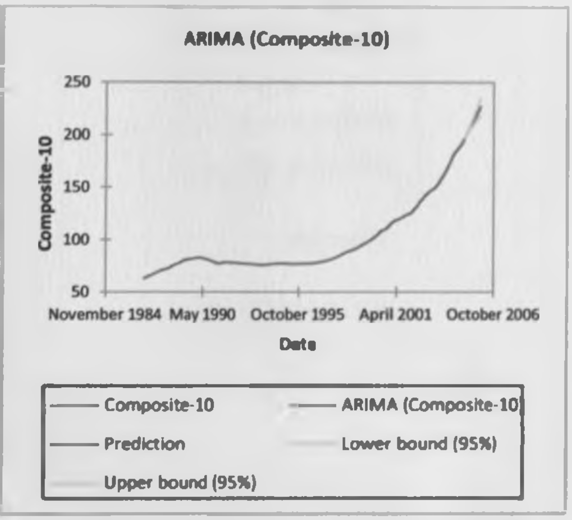


Diagram 2



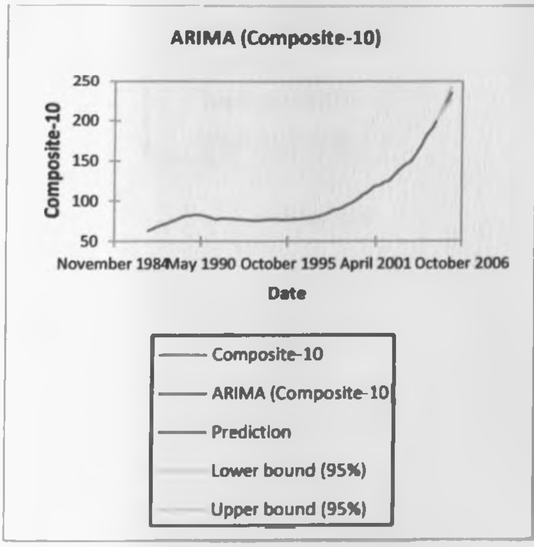


Diagram 3

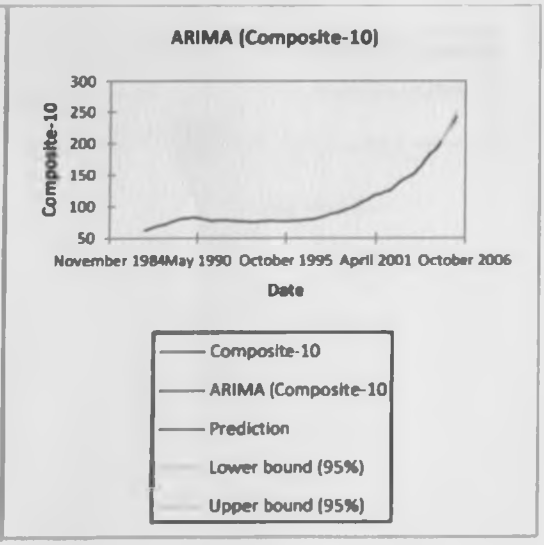


Diagram 4

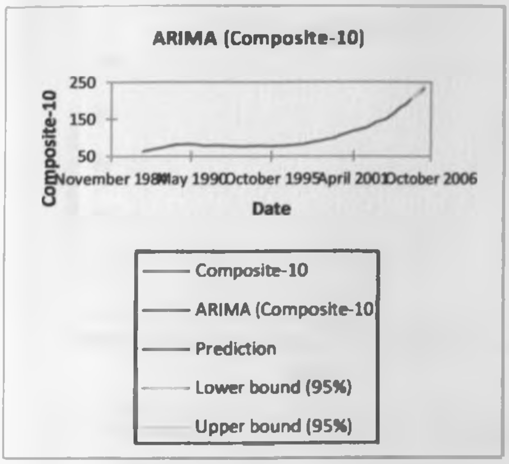


Diagram 5

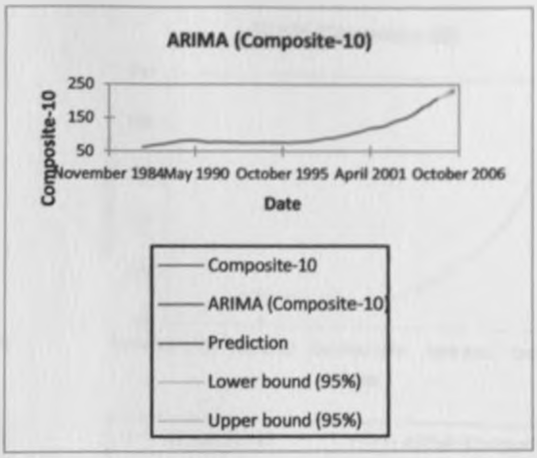


Diagram 6

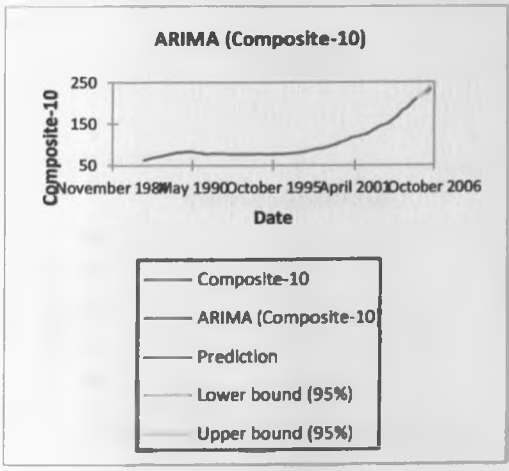


Diagram 7

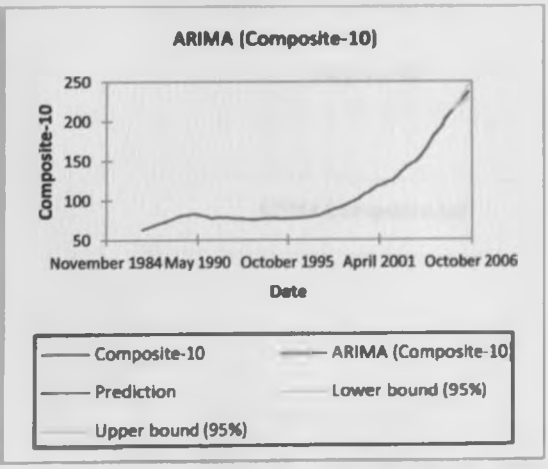


Diagram 8

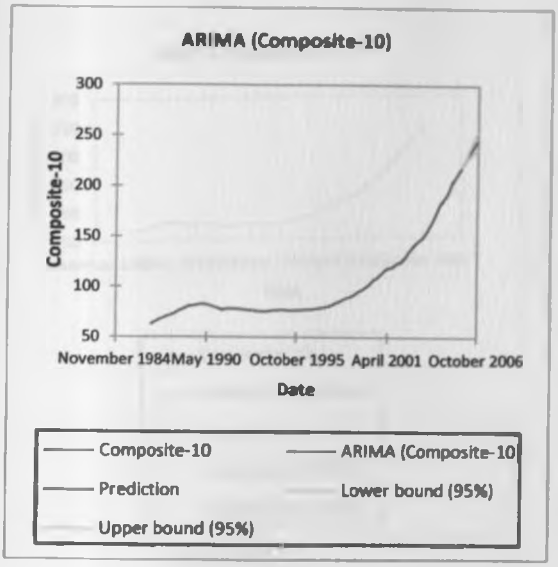


Diagram 9

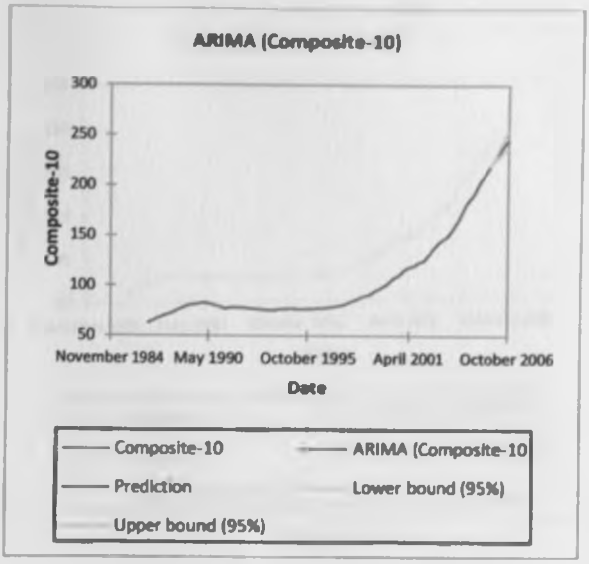


Diagram 10

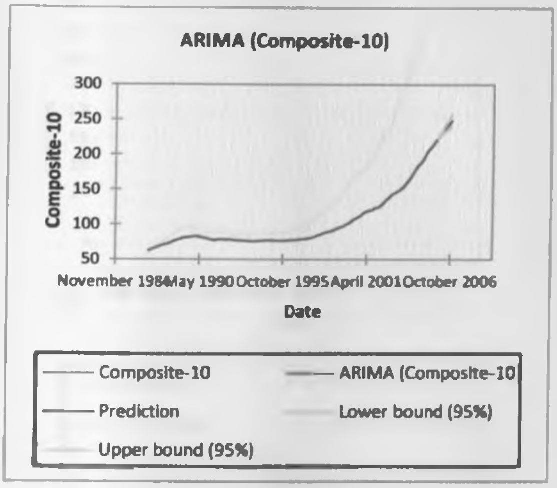


Diagram 11

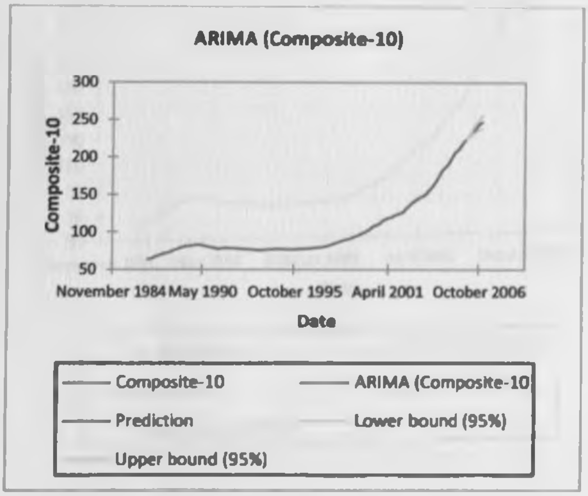


Diagram 12

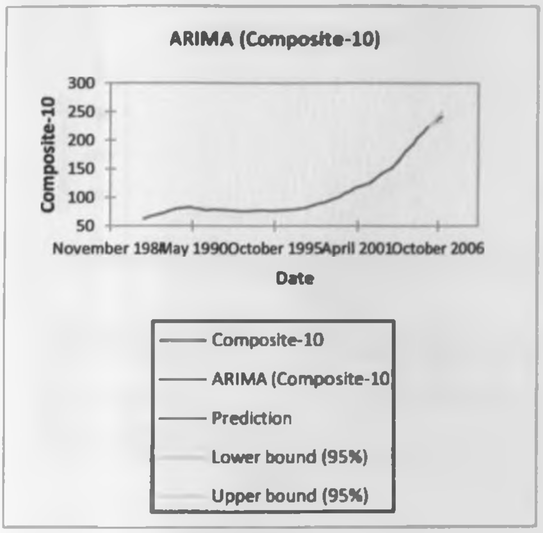


Diagram 13

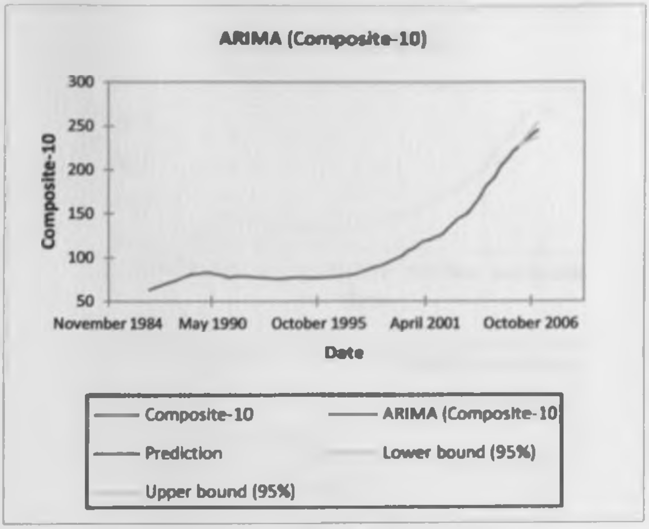


Diagram 14

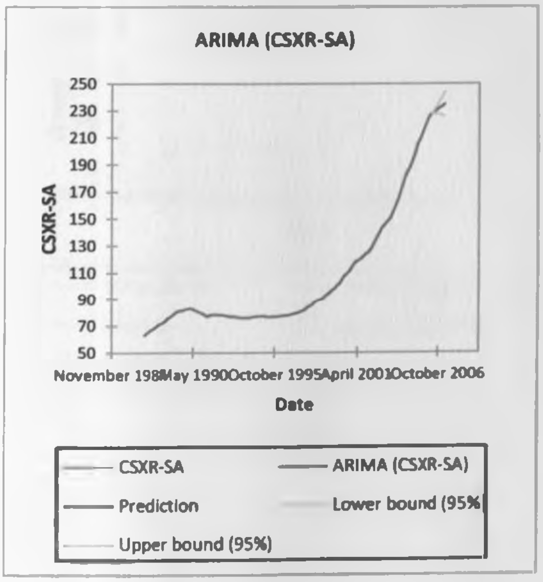


Diagram 15

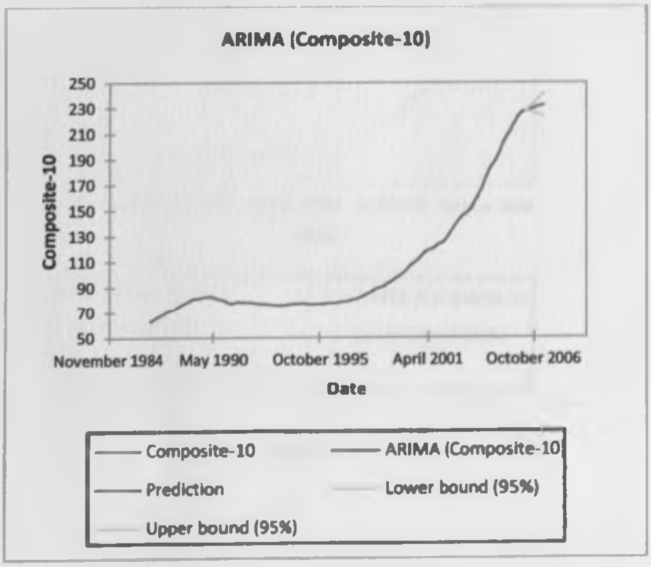


Diagram 16

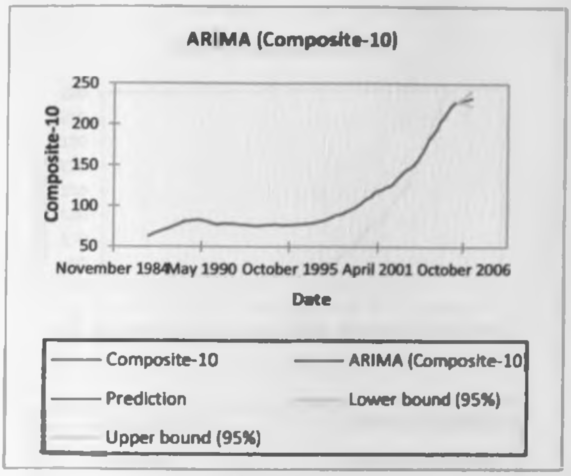


Diagram 17

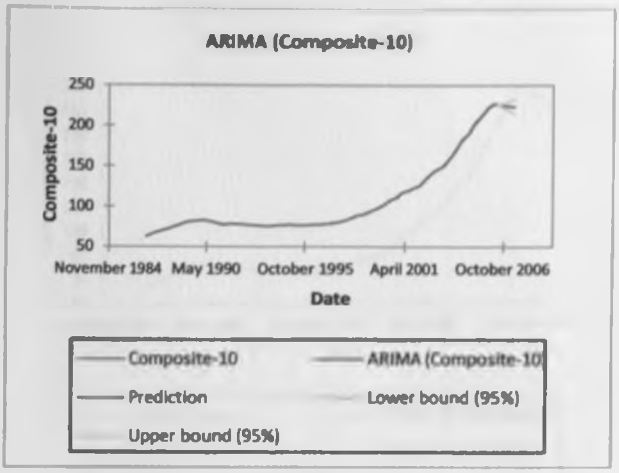


Diagram 18

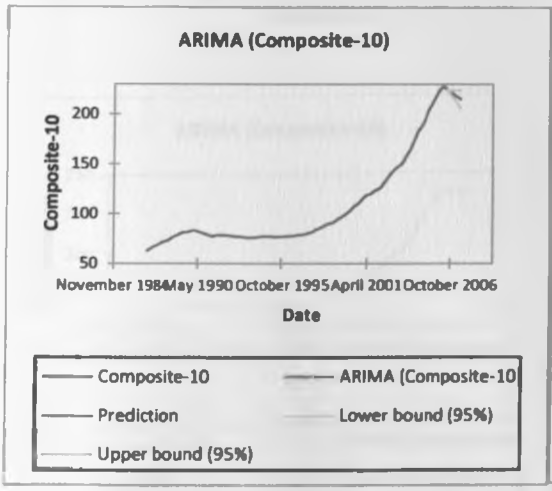


Diagram 19

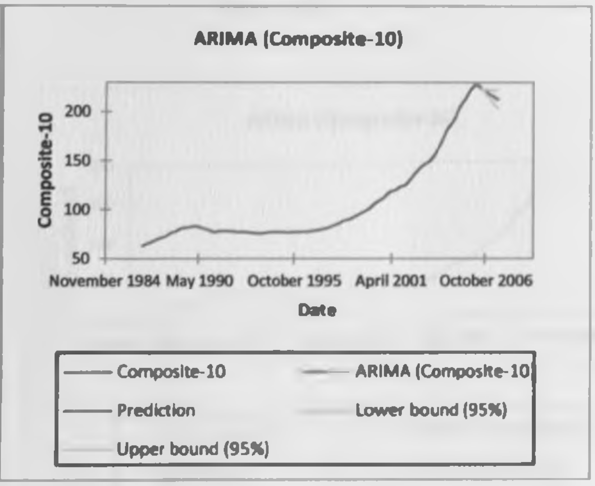


Diagram 20

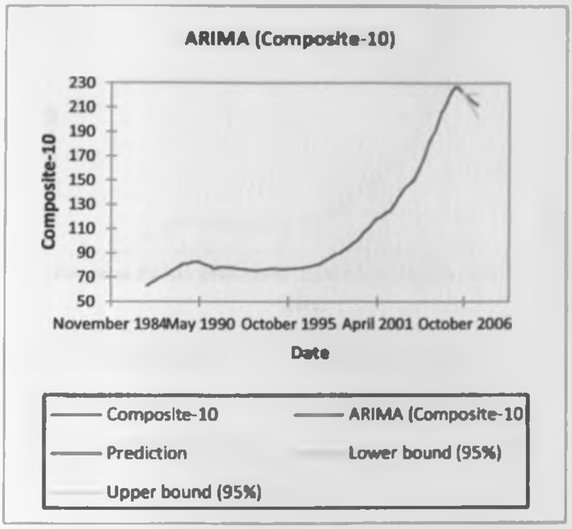


Diagram 21

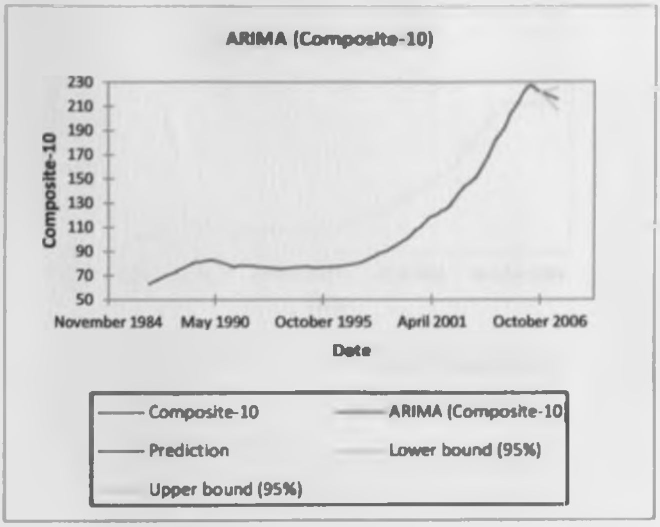


Diagram 22

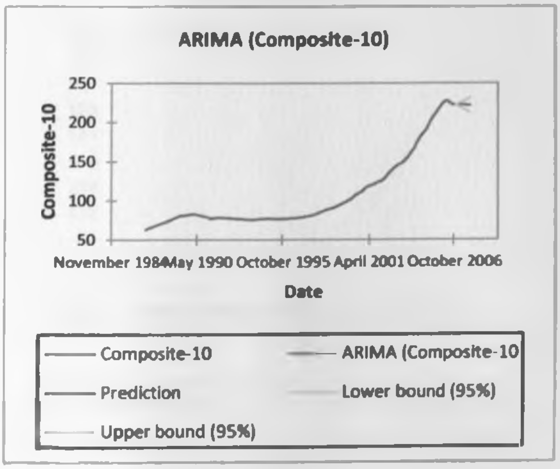


Diagram 23

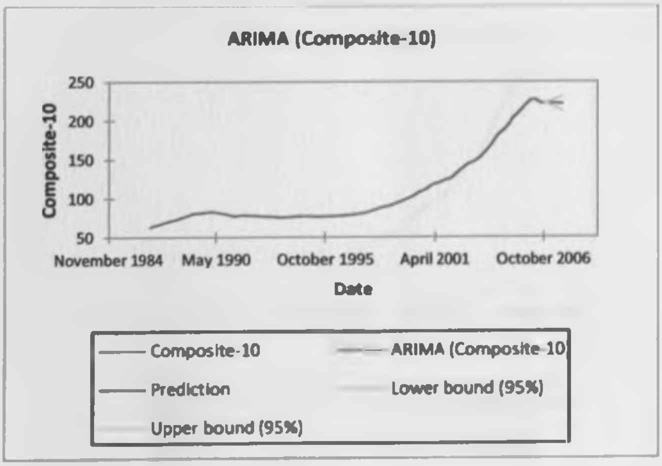


Diagram 24

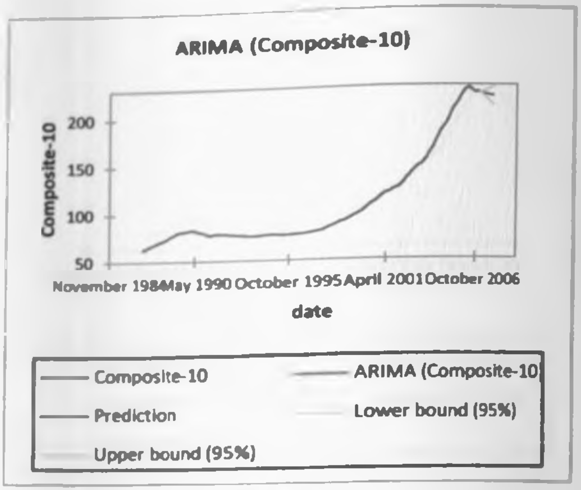


Diagram 25

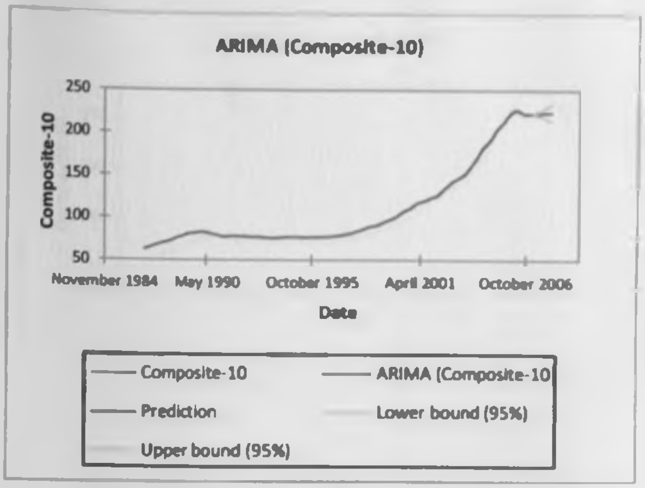


Diagram 26

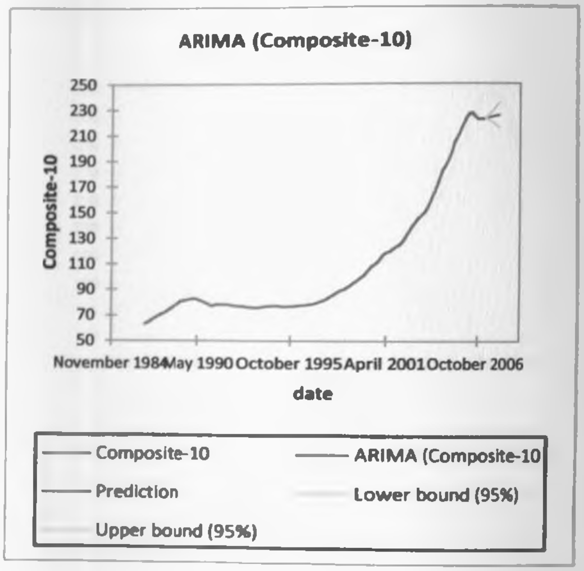


Diagram 27

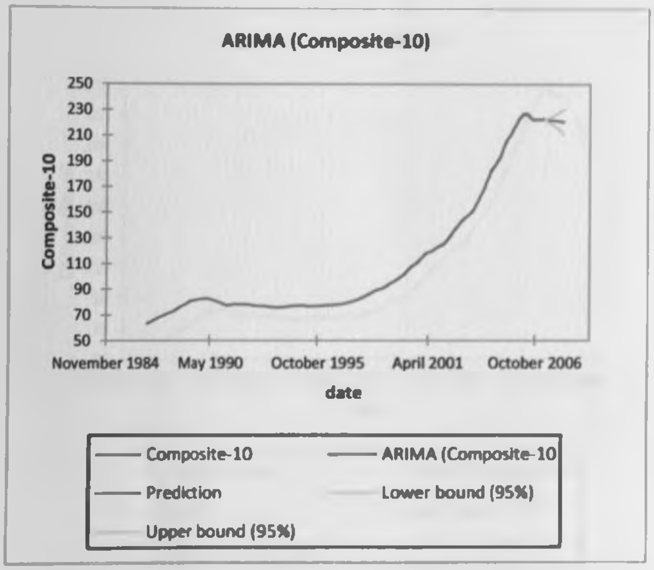


Diagram 28

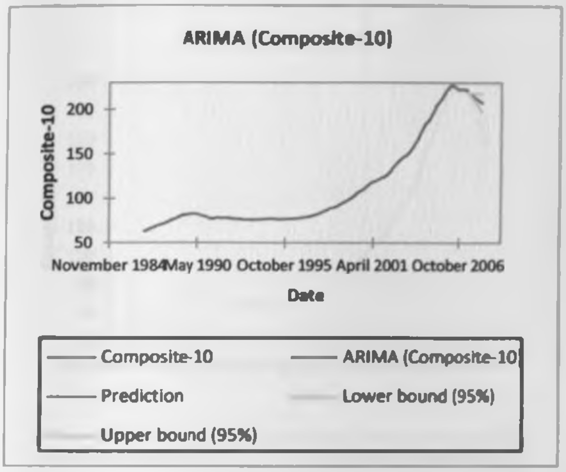


Diagram29

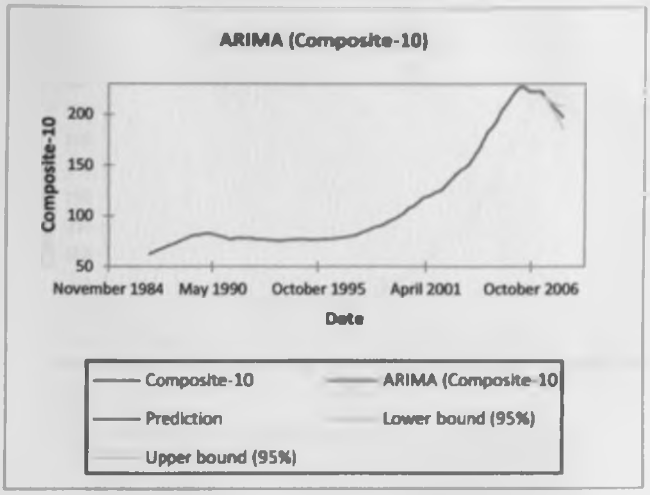


Diagram 30

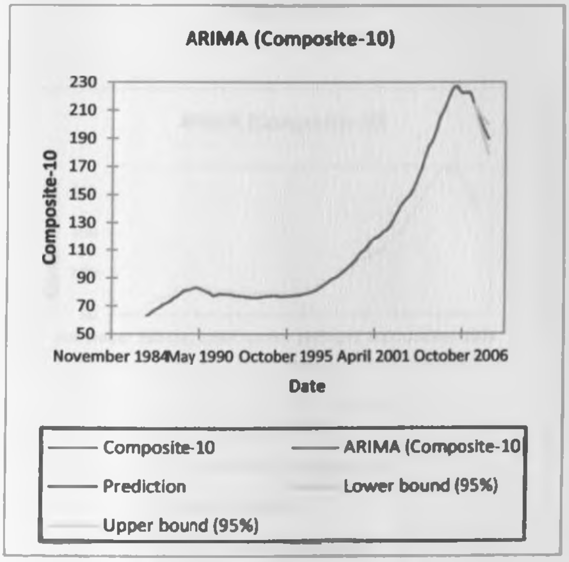


Diagram 31

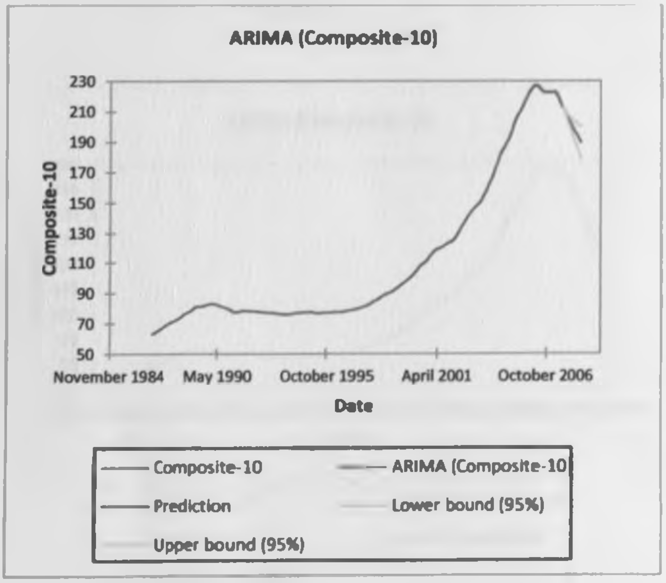


Diagram 32

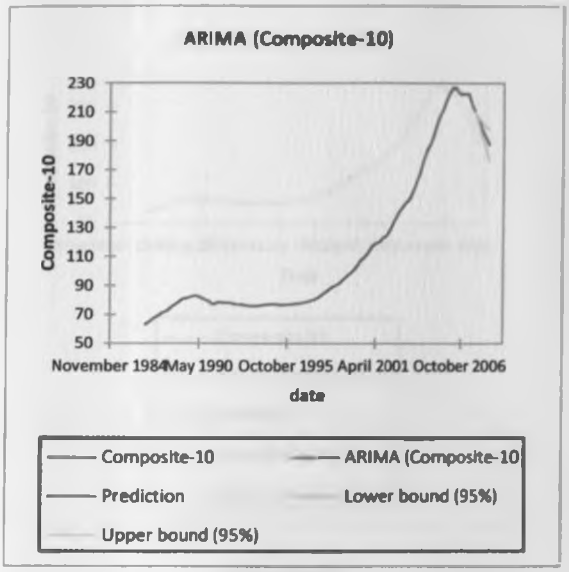


Diagram 33

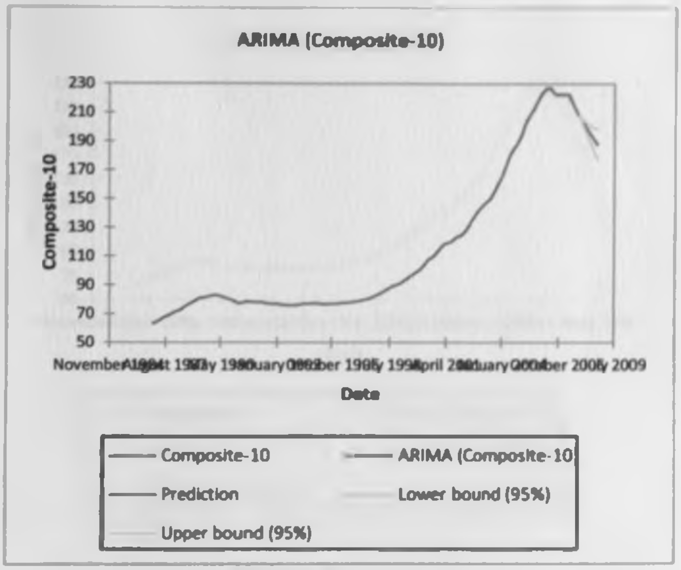


Diagram 34

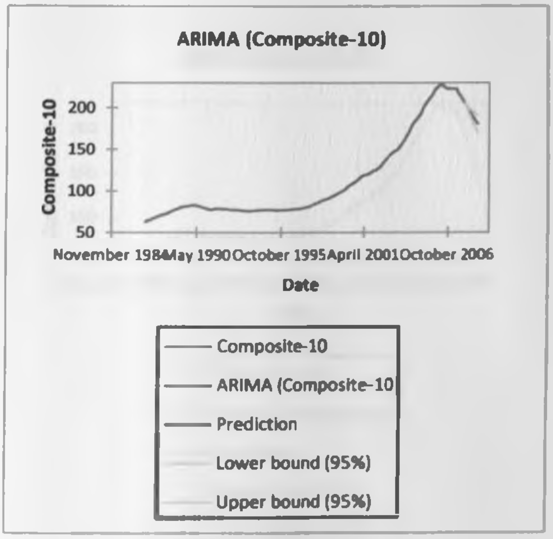


Diagram 35

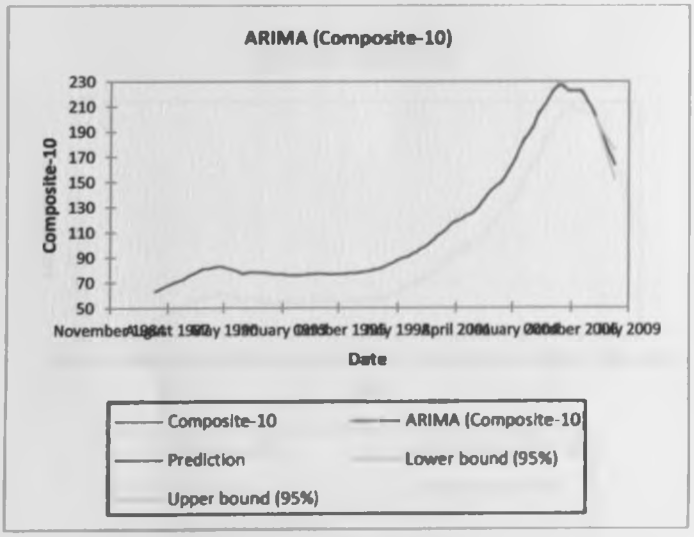


Diagram 36

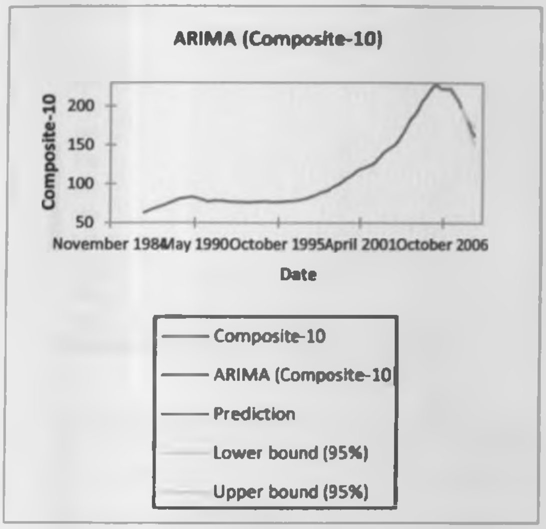


Diagram 37

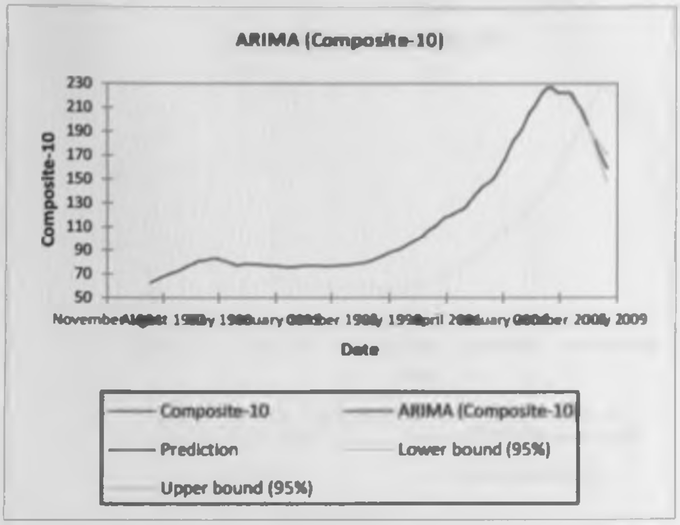


Diagram 38

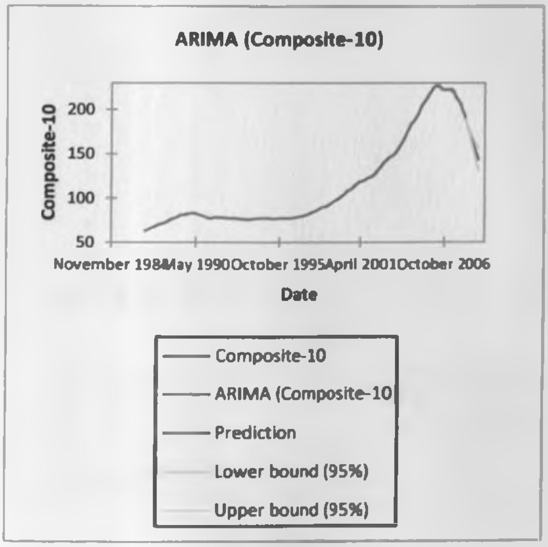


Diagram 39

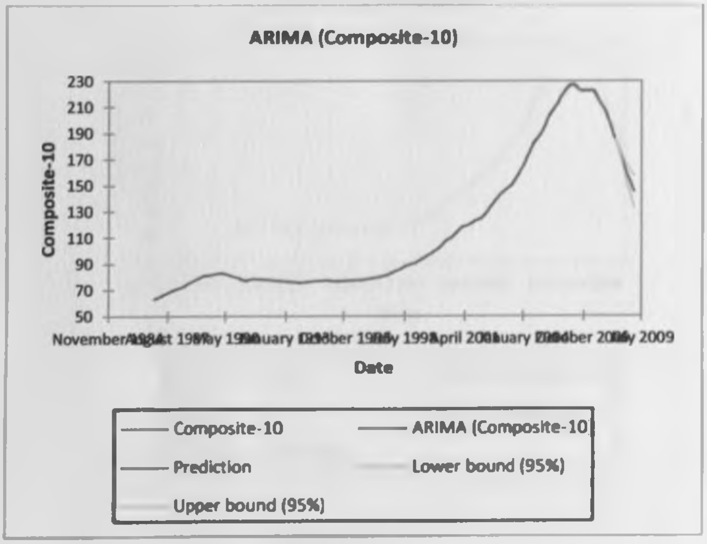


Diagram 40

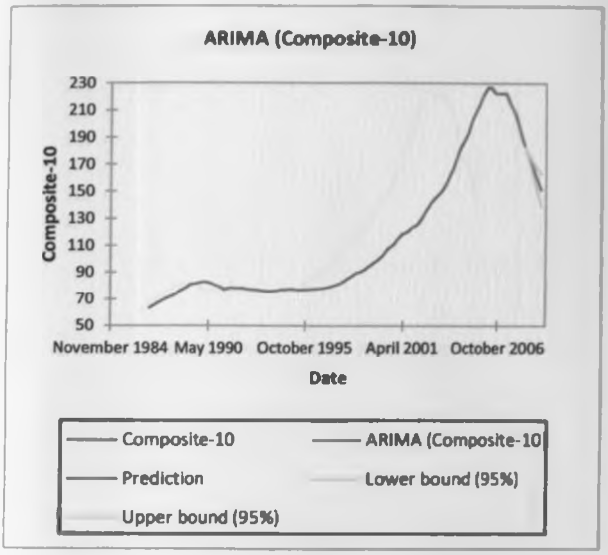


Diagram 41

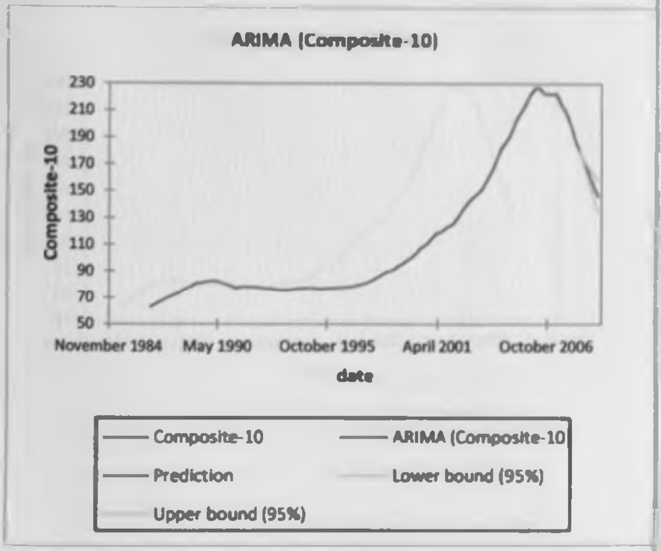


Diagram 42

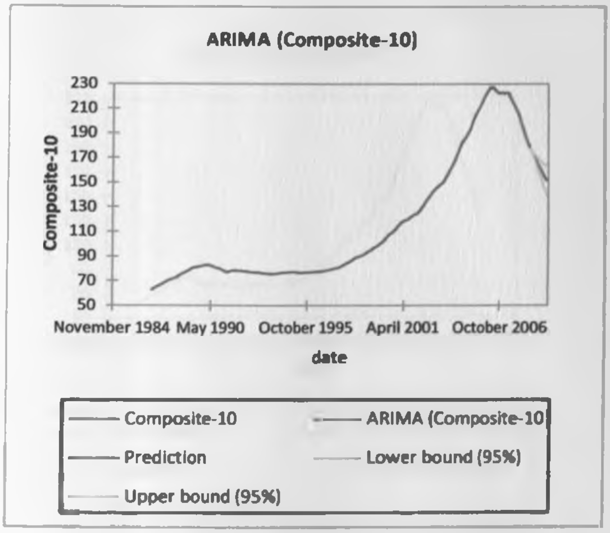


Diagram 43

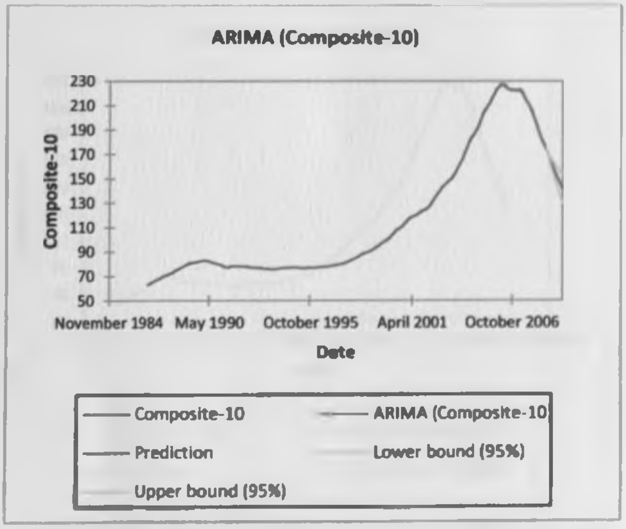


Diagram 44

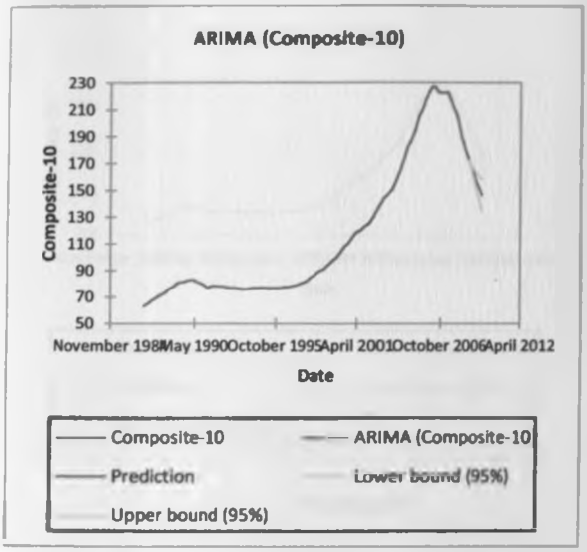


Diagram 45

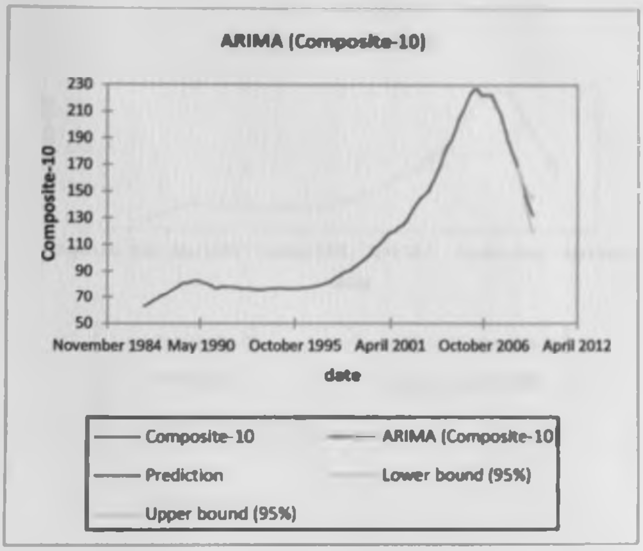


Diagram 46

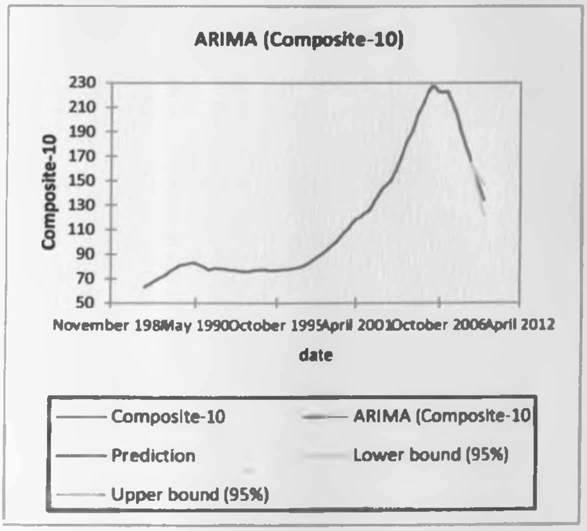


Diagram 47

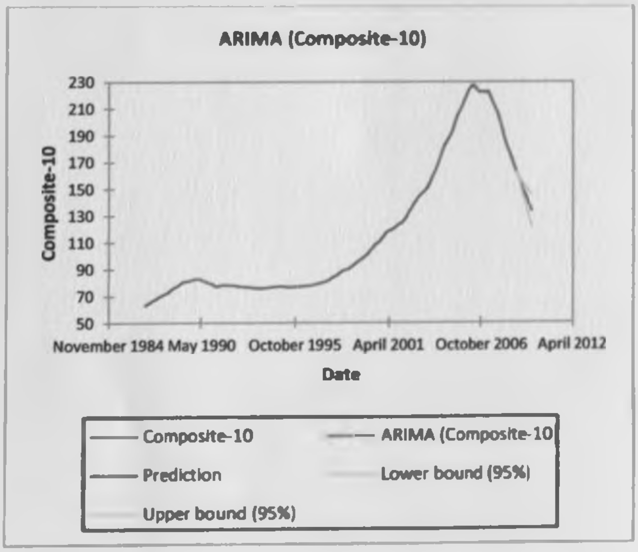


Diagram 48

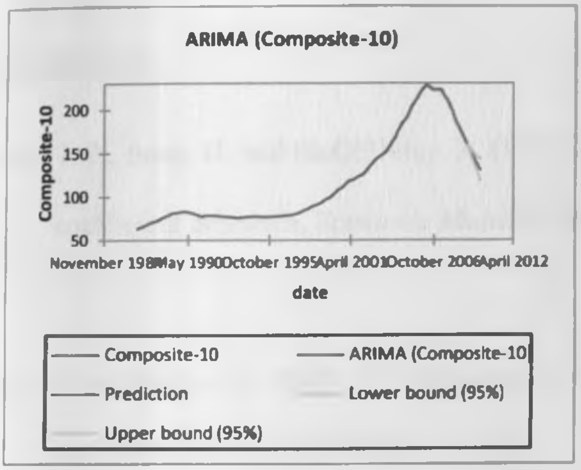


Diagram 49

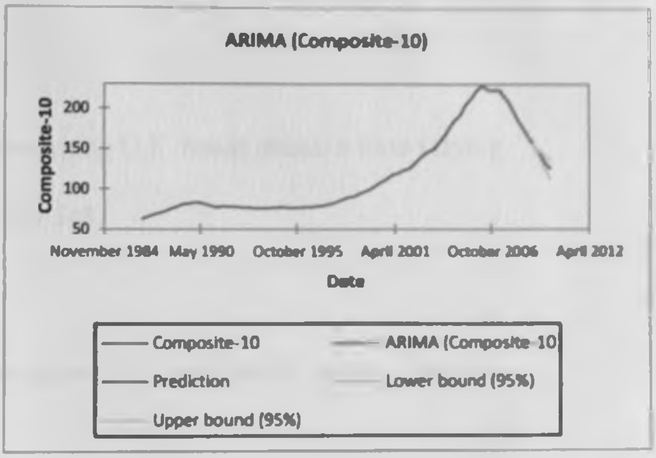


Diagram 50

REFERENCES;

Brown, J. P., Song, H. and McGillivray, A. (1997) Forecasting U.K. house prices: a time varying coefficient approach, *Economic Modeling*, **14**, 529-548.

Case, K. and Shiller, R. (1989) The efficiency of the market for single family homes, *American Economic Review*, **79**, 125-137.

Case, K. and Shiller, R. (1990) Forecasting prices and excess returns in the housing market, *Journal of the American Real Estate and Urban Economics Association*, **18**, 253-273.

Capozza, D, P Hendershott, C Mack and C Mayer (2002): "Determinants of Real House Price Dynamics", *NBER Working Paper no* 9262, October.

Crawford, G. and Fratantoni, M. (2003) Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices, *Real Estate Economics*, **31**, 223–243.

Drought, S. and McDonald, C. (2011). Forecasting housing price inflation: a model combination approach. *RBNZ working paper no* 7, October.

Ge X.J., Runeson, G. and Lam, K.C. (2003). Forecasting Hong Kong housing prices: an artificial neural network approach, paper delivered to the International Conference of Methodologies in Housing Research, Stockholm, Sweden 22-24 September.

Gu, A. Y. (2002) The predictability of house prices, *Journal of Real Estate Research*, **24**, 214-33.

Kochin, L. A. and Parks, R. W. (1982). Vertical equity in real estate assessment: A fair appraisal. *Economic Inquiry*, **20(4)**, 511–532.

Muellbauer, John and Anthony Murphy (1997), "Booms and Busts in the U.K. Housing Market," *The Economic Journal*, **107**, 1701–1727.

Zhou, G. Forecasting sales and price for existing single-family homes: "A VAR model with error correction", *Journal of real estate research*, **14** , 157- 165