Short-Term Forecasting of Power Flows over Major Pacific Northwestern Interties: Using Box and Jenkins ARIMA Methodology

Piyush S. Paretkar

Thesis submitted to the Faculty of
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering

Dr. Virgilio Centeno, Chair
Dr. Lamine Mili
Dr. Yilu Liu
Dr. Kaiyan Jin (External)

10/01/08
Blacksburg, VA

Keywords: Pacific AC Intertie, Pacific DC Intertie, Northern Intertie, Box and Jenkins ARIMA methodology, Transfer Function Models, Hydropower, Deregulation

Short-Term Forecasting of Power Flows over Major Pacific Northwestern Interties: Using Box and Jenkins ARIMA Methodology

Piyush S. Paretkar

ABSTRACT

The deregulation of the Electricity Sector in US has led to a tremendous increase in the inter-regional wholesale electricity trade between neighboring utilities or regions. For instance, the generation deficit regions may choose to import power from surplus regions; thus the wholesale electricity market prices in the regions are also affected by the dynamics of its electricity trade with other regions. Valuable insights into such imports/exports ahead of time have become crucial market intelligence for the various academicians and the market players associated with the industry. In this thesis, the task of short-term forecasting of the power flows over three major transmission interties of the Pacific Northwest region, namely the Pacific AC Intertie, the Pacific DC Intertie and the Northern Intertie, is successfully accomplished. The Pacific AC and the Pacific DC interties connect the Pacific Northwest region of US with the state of California. The Northern Intertie is the only intertie connecting the British Columbia region in Canada with the Pacific Northwest US. Box-Jenkins ARIMA (Auto Regressive Integrated Moving Average) and Transfer function methodologies are used as the statistical tools to identify the forecasting models in this thesis. The data requirement for all of the models is restricted to publicly available data.

To My Family

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my Advisor, Dr. Virgilio Centeno, for giving me the opportunity to work with him and for being my constant source of guidance, support and encouragement for the past 2 years. Dr. Centeno has been a great source of inspiration during the entire period and also in the time to come. His knowledge of his field and his zeal for his work are exceptionally inspiring. Without his advices and persistent help, this work would not have been possible.

I would like to thank my Co-Advisor, Dr. Lamine Mili, for his valuable ideas and suggestions throughout this research work. Dr. Mili's brilliant knowledge of Statistics and his unique approach of solving problems have been instrumental in the accomplishment of this work. The time spent with him at the Northern Virginia campus and the various brainstorming sessions we had proved very helpful over the course of this work.

Classes I took with Dr. Yilu Liu during my first year at Virginia Tech were sources of valuable information and a good exposure to executing independent projects. I would like to thank her very much for her trust in me and for being on my graduate committee.

I am very grateful to EnvaPower Inc. for sponsoring this research project in entirety and for their trust in me. Special thanks to Dr. Kaiyan Jin of EnvaPower Inc. for consenting to be on my graduate committee on a very short notice and for reviewing my thesis. I would like to acknowledge my gratitude to Dr. Kaiyan Jin, Dr. Gustav Beerel and Dr. Crispin Miller of EnvaPower Inc. for their valuable support and suggestions and for providing me with some crucial sources of data during the course of this work

A large degree of emotional support and encouragement has been received from the many friendships I have made during my stay at Virginia Tech. Special thanks to my fellow graduate labmates at the Power Lab, all of my roommates and friends who have shared in many of the moments that made my years at Virginia Tech memorable.

I owe my deepest appreciation to my family members and relatives for their constant support and encouragement throughout my life and academic year.

Last but not the least, I would like to thank God for his divine help and guidance throughout my life.

Piyush S. Paretkar

Virginia Polytechnic and State University

September 2008

CONTENTS

List of Figures

List of Tables

Chapter 1 – Introduction

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

Deregulation has been one of the most radical changes undergone by the US power industry in the past 20 years. Deregulation began around 1996 after the FERC (Federal Energy Regulatory Commission) issued Orders 888 and 889, with the stated objective to "remove impediments to competition in wholesale trade and to bring more efficient, lower cost power to the Nation's electricity consumers" [1]. Prior to deregulation, the US electric utilities were vertically integrated. One utility was responsible for generation, transmission and distribution of the electric power to the customers in its territory of operation. Though most of these utilities were self-sufficient in themselves, better economics was dictated by interconnecting many utilities using tie-lines (or interties); thus the US power system became a mesh of many utilities interconnected with each other. The interconnected power system has many inherent advantages. The reserve requirements of two utilities connected to each other with tie-lines is reduced; the tie-lines can be used to provide assistance in events of power

shortage or blackouts and most importantly, the tie-line serves to increase the inertia of the two interconnected systems, thus making them more secure against system failures and oscillations. These tie-lines also provided one major advantage from a pure economic point of view: the interconnected utilities can trade power with each other and make profits.

Prior to deregulation, almost all of the trade contracts between the interconnected utilities were bilateral contracts; the price and the MW quantities were mutually set days in advance. Based on import-export schedules, the utilities could plan for dispatching its own generators with the objective to minimize the total cost of production, abiding with the other constraints of the system. Since only one utility was responsible for dispatching the total generation in its territory, there was little uncertainty involved in the process of calculating its optimum trade requirements.

Beginning in 1996, the US electric industry has underwent a radical change in its ownership and functioning. Deregulation required the vertically integrated utilities to either sell off its generators or transfer its ownership to private parties. As a result, many private generators, called Independent Power Producers (IPPs), were formed. Utilities can no longer operate their transmission networks, though they may still retain the ownership. The transmission networks of the neighboring utilities were bundled and a non-profit organization, called ISO (Independent System Operator) was formed by the FERC's recommendation to look after the combined operation, control and monitoring of the transmission network. This allowed fair access to the transmission systems to all the parties involved, fostering open market competition. Eight such ISO's and four RTOs (Regional Transmission Operators are same as ISOs but are bigger and look over wider

areas that cross state borders) are functioning in different parts of US today. Some ISO's and RTO's also act as a marketplace in wholesale power and various ancillary services. For instance, California ISO (CAISO) runs the real-time energy imbalance market (spot market), ancillary service market and the Transmission usage markets (also called FTRs, Firm Transmission Rights). On each trading day, the private generators and exporters bid their hourly price and MW quantities into the market. These bids are matched against the corresponding hourly demand requirements and the market is cleared at a price called the MCP (Market Clearing Price). Thus, bidding strategies from various private generators and exporters leads to uncertainty in predicting future electricity prices in the CAISO markets. Bonneville Power Administration (BPA) and British Columbia Hydro (BCH) are two major exporters to California from the Pacific Northwest region. Both BPA and BCH are predominantly hydropower dominated sources; almost all of the power they export to California is surplus hydro generation in these regions, which is relatively cheaper than almost all other sources of generation in California. Thus, the availability of exports from BPA and BCH also impacts the market prices in California. Consequently, the need to accurately forecast these exports has gained ever-increasing importance in the deregulated environment. The knowledge of the forecasts of exports not only helps in predicting the future prices; but it also forms a major constituent of the Market Intelligence, which is useful to all the market players. It also provides useful insights into the generation and demand patterns in the trading regions themselves. Thus, it's also useful to Academicians from a planning point of view.

Power Engineers have for decades successfully forecasted Electricity loads and prices by using various statistical techniques. Weron [2] provides an in-depth review of

various statistical tools used to forecast electricity loads and prices. Some of the advanced applications include Multivariate Adaptive Regression Splines, an adaptive non-parametric regression approach, which is used in [3] to identify the model to forecast hourly Ontario energy price. Multi-equation regression model with a diagonal first-order stationary vector auto-regressor (VAR) approach of forecasting electricity load are used in [4]. Various time series techniques are proposed in literature to forecast electricity loads and prices. Obradovic [5] provides a review of some of the time series methods used in forecasting electricity prices. GARCH (Generalized Auto-regressive Conditional Heteroskedastic) methodology has been applied to forecast the electricity prices in California markets in [6]. Many references exist in literature with application of ARIMA methodology in electricity load and price forecasting. Contreras [7] presents an application of ARIMA methodology to predict next-day electricity prices in California. An ARIMA approach to price forecasting with accuracy improvement by predicted errors is presented in [8]. In [9], ARIMA is used with ANN to identify a combined forecasting model for electricity loads. However, not much statistical work is done directly in some other application areas of the Power sector, specifically in the forecasting of power flows on transmission lines. Hitherto, most of the works done in forecasting power flows on transmission lines are based on results of complex power flow models [10]-[12]; these models suffer from various limitations. The knowledge of entire network system topology is required to build the power flow model for greater accuracy. In most situations, the knowledge of entire network is not available and may not even be public domain knowledge. Secondly, the computation time requirements are quite high. Also, to predict the future values of power flows, the knowledge of various issues related to the

network system, such as its operation procedures, maintenance schedules, protection equipment settings, line and equipment ratings, measure of losses in the system, reactive power dynamics etc. are necessary. The data requirements increase enormously with the size of the power system and level of accuracy sought. Highly meshed networks of present-day deregulated electric power systems aggravate these limitations even further.

1.1 Contribution of the work

The main objective of this research is to apply statistical forecasting techniques to solve one of the crucial problems faced by the deregulated and interconnected Electric Power Systems of modern period, namely, to predict with reasonable accuracy the power flows on major transmission interties from insufficiently available public-domain data. These statistical techniques don't suffer from the limitations of complex power flow models; they are less data intensive, computationally efficient and highly adaptive. Since most of the major interties are responsible for huge amounts of imports/exports trade between neighboring utilities or power markets, the main advantage of these models is to gain a better understanding of the imports/exports pattern of these regions. Knowledge of imports/exports pattern eventually provides more insight into the generation and demand patterns in the regions itself.

In this work, we demonstrate the application of statistical forecasting techniques to forecast the power flows on transmission lines using the specific problem of forecasting exports to California from the hydropower dominated Pacific Northwest region. Specifically, we developed and assessed short-term forecasting models, to forecast on a horizon of one day ahead, the power flow of three major interties of the

Western US, namely the Pacific AC intertie, the Pacific DC intertie and the Northern Intertie connecting Canada with US. The Pacific AC intertie is the major AC intertie connecting the Pacific Northwest with California and has a north-to-south transfer capacity of 4,800 MW. The Pacific DC intertie is the major DC link between the Pacific Northwest and California, with north-to-south transfer capacity of 3,100 MW. The Northern Intertie is the only one connecting the British Columbian province of Canada with Northwestern US. Consequently, it is wholly responsible for the power exports from British Columbia to US with north-to-south transfer capacity of 3,100 MW. These three interties are responsible for the major bulk of power exports to California from the Pacific Northwest region.

Box-Jenkins Auto-Regressive Integrated Moving Average (ARIMA) methodologies are used to solve the problem at hand. They are found to outperform other techniques such as decomposition methods, multiple regression methods, smooth filtering methods etc.

Some of the challenges encountered in the present work were:

1) Investigating the Pacific Northwest and California system and making assumptions about the dynamics behind their trade transactions using the limited available public domain data,
2) Working with the highly volatile and noisy nature of the power flow data on the Pacific Northwest-California interties,
3) Identification of the relevant inputs for the various forecasting models, keeping the constraint of using minimum data which is publicly-available, after several trials with many sets of data,

4) Formulating the appropriate form of the transfer function for the forecasting models to achieve better accuracy, using a blend of Explanatory and Time Series Analysis techniques.

Some of the quantitative contributions in this work are:

1) Collecting and examining the various sources of available data and selecting the relevant data for the various models built in this thesis,
2) Analyzing the trends and patterns present in the data and comprehending the association between the various variables,
3) Fitting ARIMA models and improving their performance by blending them with regression transfer function models,
4) Achieving a high R-square (in the range of 0.96) and low Mean Absolute Percentage Error for the forecasting models and demonstrating their usefulness.

It is helpful to present a brief outline of the chapters that follow:

Chapter 2 serves as the background of the present work. It provides an overview of the hydropower sector of Pacific Northwest including BPA and BCH, the generation mix in the region and some relevant figures. The generation mix in California and the trend of California imports from the Pacific Northwest are discussed next, with short notes on the three major tie-lines which are responsible largely for these imports.

Chapter 3 provides an elaborate overview of the Box-Jenkins ARIMA model. The basic terminology and the steps of identifying ARIMA models are covered. ARIMA models are used in identifying the forecasting models of later chapters.

Chapter 4 deals with the forecasting of power exports to California via the Pacific AC and Pacific DC interties. Suitable inputs for the ARIMA transfer function are selected and used to identify the short-term forecasting model. An alternate model, with some difference in fitting procedure, is built in Chapter 5 and the results are compared.

Chapter 6 deals with forecasting the 'CAISO portion of Pacific AC intertie' flows using the forecasts generated in chapter 4. CAISO has the right to schedule only a portion of the Pacific AC intertie through its market. This chapter focuses on forecasting only that portion of Pacific AC intertie.

Chapter 7 deals with forecasting the power exports from British Columbia to US through the Northern Intertie. These exports are used as one of the inputs in identifying the models of chapters 4 and 5.

Chapter 8 concludes the thesis with the inferences drawn and provides comments and recommendations concerning possible areas for future additional research work.

Chapter 2 - Background: Pacific Northwest Region and Exports to California

The Pacific Northwest region of US includes all of Washington, Oregon, Idaho, part of British Columbia and adjoining parts of Alaska, Montana, Yukon Territory and California. This chapter provides a brief overview of the hydropower sectors in the Pacific Northwest and California regions, and some facts and figures about the scenario of exports to California from the hydro-power dominated Pacific Northwest region.

2.1 Hydropower in the Pacific Northwest

Electricity production in the Pacific Northwest (PNW) is dominated by hydropower since the late 1880s, when the first hydropower turbine was installed on a Columbia River tributary. At present (June 2006), the total installed capacity of the region stands at approximately 53,000 MW [13]. Hydropower accounts for about 64% of the total supply capacity in the Pacific Northwest region. Most of the rest comes from coal-fired steam plants and natural gas plants in the region. The region has only one operating nuclear power plant, which accounts for about 2% of the generation capacity. Figs 2.1 and 2.2 provide the percentage breakdown of the total power supply capacity in PNW region by the type of fuel [13].

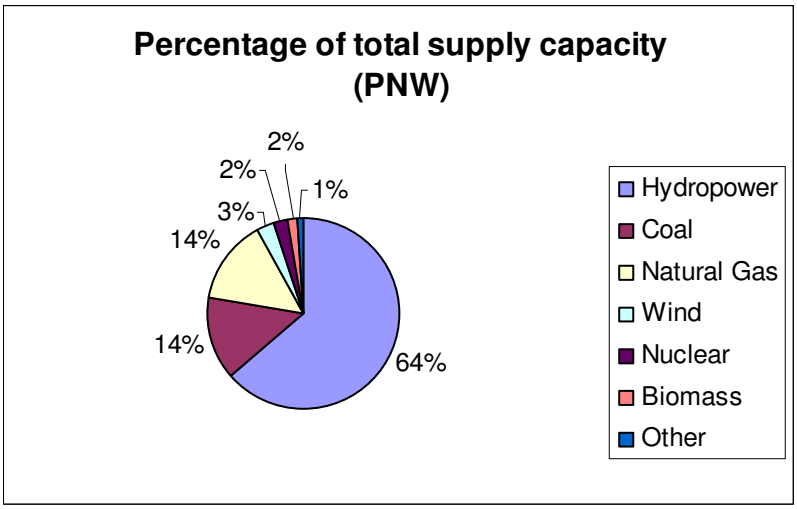

Figure 2-1: Percentage of Total Supply Capacity by Fuel Type in PNW (2006)

Northwest Power Supply Capacity by Fuel Type

Type of Fuel	Capacity in Megawatts	Percentage of Total Supply
Hydropower	33,562	63.6%
Coal	7,505	14.3%
Natural Gas	7,562	14.2%
Wind	895	3.0%
Nuclear	1,588	2.3%
Biomass	1200	1.7%
Other	486	0.9%
Totals	52,798	100%

Figure 2-2: Total Supply Capacity Breakdown by Fuel in PNW (2006) (Source: *Electricity generation for the PNW, NPCC, 2006*)

Most of the region's hydroelectric power is generated by the 40-odd major power plants out of the total of 58 dams on the Columbia and Snake rivers [14]. Out of these, 23 major hydropower plants are federal projects; 10 are owned and operated by the Bureau

of Reclamation, 13 are owned by the Corps of Engineers and the remaining 17 are owned and operated by local utilities and private companies. Grand Coulee Dam situated on the Columbia River, with installed capacity of approximately 6800 MW, is the largest electric generating facility in the country.

About one-third of the power used in PNW comes from Bonneville Power Administration (BPA), a Federal Agency responsible for marketing power from 31 Federal hydro projects in the Columbia River basin, one non-Federal nuclear plant and several other small non-federal power plants [15]. The combined installed capacity of the 31 hydropower plants stands at 20,460 MW. Hydropower accounts for over 90% of the

Figure 2-3: Hydroelectric Dams on Columbia River (Source: *Hydropower's importance to the Northwest, NRP, 2006*)

total generation capacity in BPA territory. 21 plants are owned and operated by the US Army Corps of Engineers (UACE) and the remaining 10 plants by US Bureau of Reclamation (USBR). Some of the major hydroelectric power plants under BPA include Grand Coulee (6,800 MW), Chief Joseph (2,500 MW) and John Day (2,160 MW) [15]-[16].

The second largest hydropower source in the PNW region is British Columbia Hydro in Canada. British Columbia Hydro (BCH), one of the major exporters of hydro-electric power to California, is the largest utility in the British Columbia province of Canada, serving approximately 95% of the population of British Columbia. Over 90% of power generation in BCH comes from hydro-electric plants [17]. Most of the electricity is produced at the 30 hydroelectric plants, with total capacity of 9,716 MW, on the Columbia and Peace rivers. Between 43,000 and 54,000 GWh of electricity is generated annually depending on the water conditions to meet the demands of the region and the surplus is exported to other regions, including the Southwest US. G.M.Shrum (2,730 MW) and Peace Canyon (700 MW) are the two largest generating stations and are both located on Peace River. They produce about 29% of the BC Hydro's electricity requirements [17].

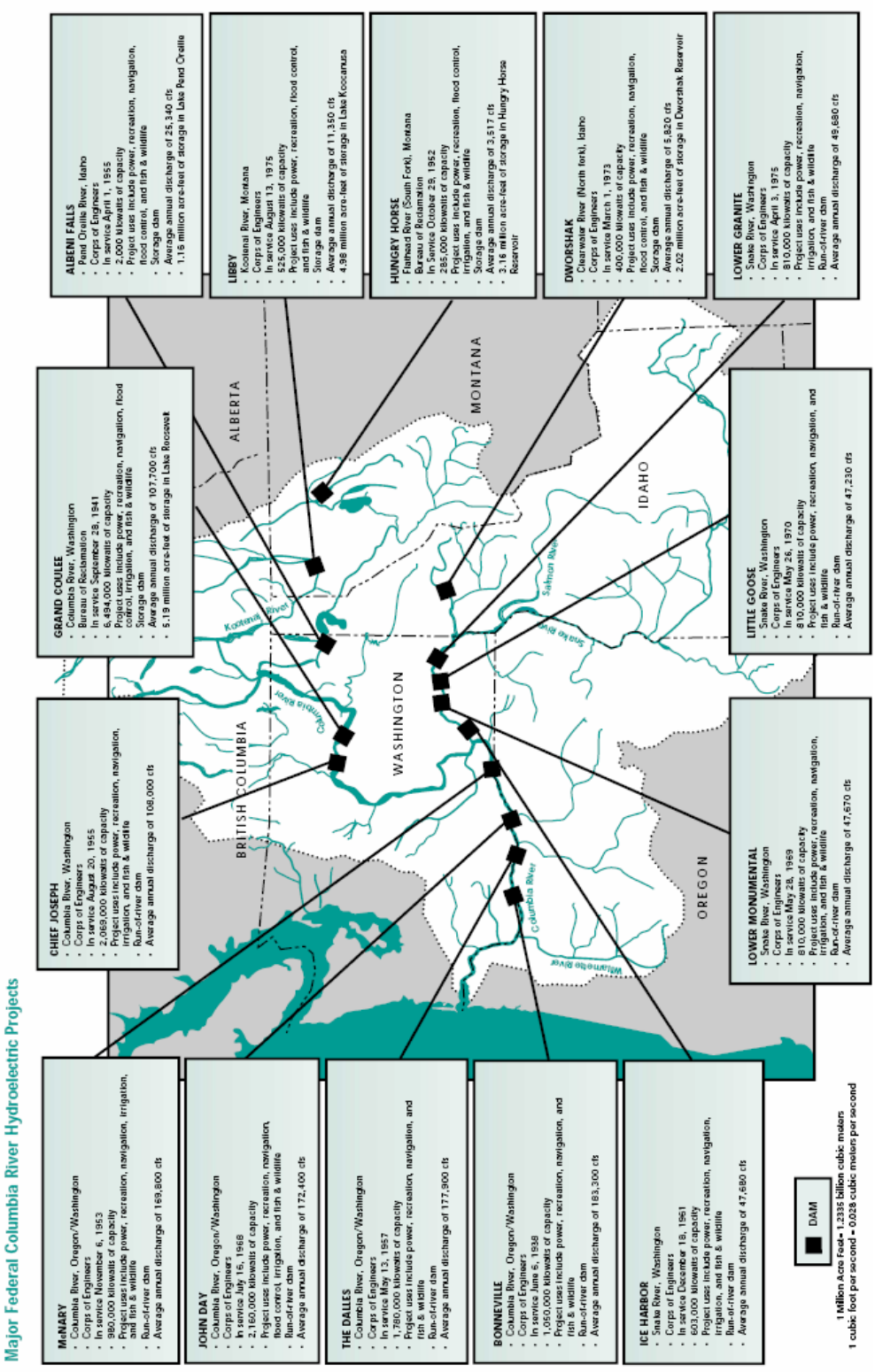

Figure 2-4: Major Federal Columbia River Hydroelectric Project (Source: *The Columbia River System Inside Story, USBR, 2001*)

2.2 California: Generation mix and Exports from Pacific Northwest

The total installed capacity of California stands at approximately 56,800 MW [18]. The generation mix of California for the year 2007 is shown in Fig 2-5.

The profile of monthly energy generation by fuel type for year 2007 is shown in Fig 2-6 [19]. Natural gas dominates the fuel mix in California power production. Hydro-electric production accounted for less than 12% of total generation in the state. In year 2007, California produced 69% of the electricity it used; 8% was imported from Pacific Northwest and 23% was imported from US southwest (Fig 2-7).

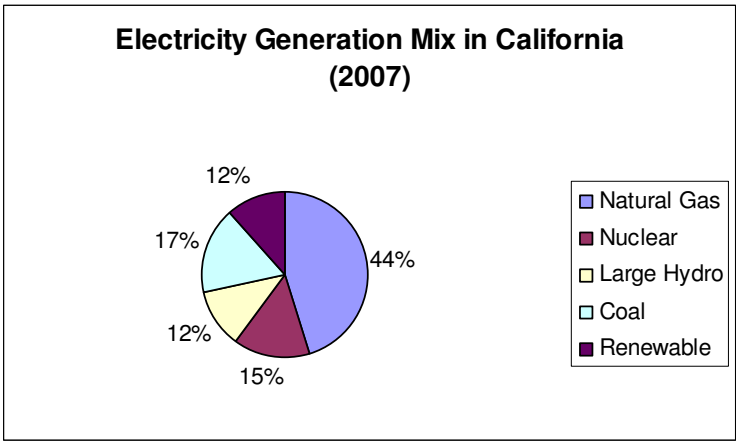

Figure 2-5: Electricity Generation Mix in California

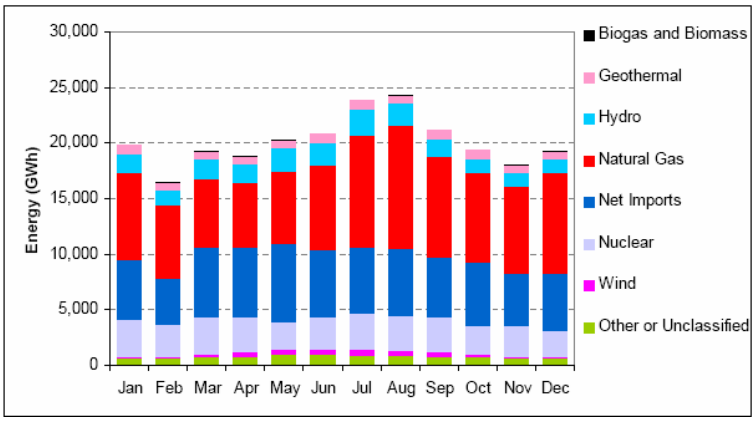

Figure 2-6: Monthly Energy Generation by fuel type in CA (2007)

The trend of average California annual imports, exports and the net imports for years 2003 through 2007 is shown in Fig 2.8. As depicted in Fig 2.9, the imports continue to play a key role in meeting the demand in California. Majority of the imports from Pacific Northwest region comes from surplus hydro-electric generation in the BPA and the BCH.

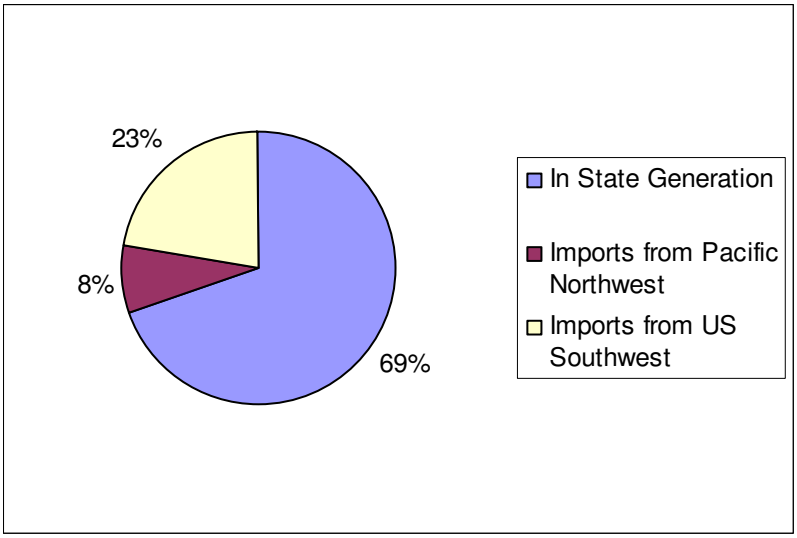

Figure 2-7: Exports to California (2007)

Figure 2-8: Trend of Imports and Exports of California (2003-2007)

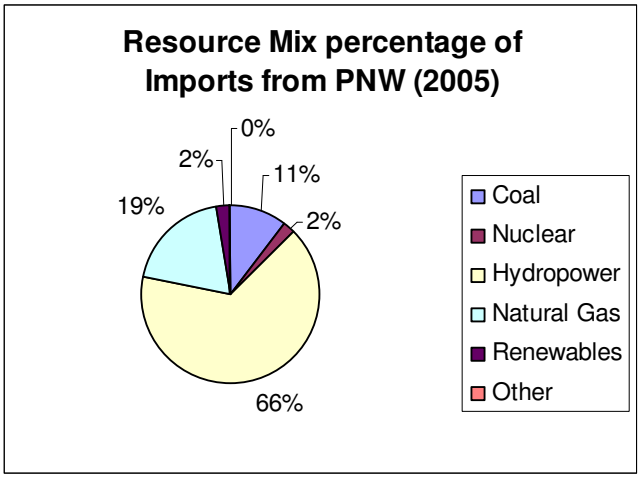

Figure 2-9: Resource Mix Percentage of Imports from Pacific Northwest

The PNW system is designed to meet all its requirements even in a critically dry hydro year; in all other years it has surplus generation to export to other regions in the WECC. The PNW can't keep all the surplus hydropower to itself; it's a winter-peaking region while the peak run-offs occur during the spring. This is in contrast to the summer peaking California region, which benefits from these exports. Shown in Fig 2.10 is the hourly power flows on PNW-CA combined interties for the year 2007 [20].

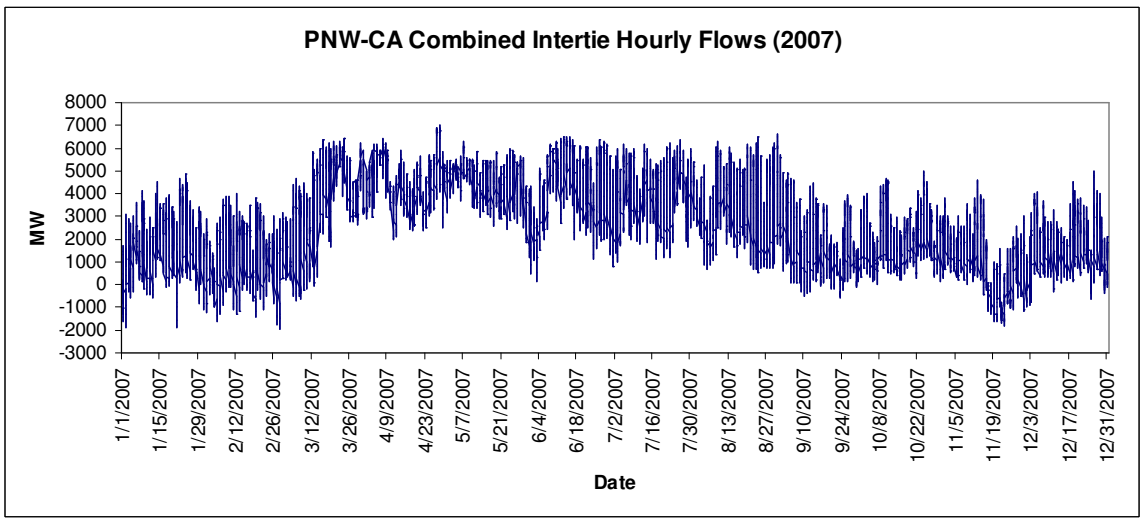

Figure 2-10: Hourly Power Flows on PNW-CA interties (COI and PDCI)

Analysis of the flows show that California imports daily peak power throughout the year from Pacific Northwest region. It imports more power during the spring and summer seasons, when the California system is peaking and the Pacific Northwest loads are low. The power imports are diurnal in nature; with high imports during the on-peak period (16 on-peak hours of the day) and relatively low imports during off-peak periods (8 off-peak hours). California may also export some power during off-peak hours to PNW during the winter seasons.

Almost all of these hydro-electric exports from Pacific Northwest reaches California through two major interties starting from BPA region, collectively called the Pacific Intertie: Pacific AC intertie (PACI) and the Pacific DC intertie (PDCI).

The Pacific AC intertie comprises of three 500 kV AC transmission lines. Its northern portion, which enters the California borders, is called the California Oregon Intertie (COI). The COI is also called Path 66 in California power markets. Two of the

Figure 2-11: PACI and PDCI Interties

500 kV transmission lines that form the Path 66 originate at Malin substation in BPA region and terminate at the Round Mountain substation in Northern California. The third 500 kV transmission line originates at the Captain Jack substation in BPA, to the west of Malin substation, and terminates at the Olinda substation in Northern California. The AC lines in Oregon are primarily owned by BPA. In California, the AC lines are owned and/or shared between various utilities: PGE, SCE, WAPA, SDGE, SMUD, LADWP and a consortium of other utilities known collectively as the California Oregon Transmission Project (COTP) [21]. The North-to-South transfer capacity of Path 66 is 4,800 MW, out of which 1,600 MW is in the COTP, also known as the Third AC line. The California ISO has contracts to share only 33 MW out of 1,600 MW of COTP transmission capacity [22]. Thus, the total import capacity available to CAISO from path 66 is approximately 3233 MW. The Pacific DC intertie (also called Path 65) consists of a parallel bipolar +/- 500 kV DC transmission line, with its northern end at the Celilo Converter station in BPA and southern end at the Sylmar Converter station near Los Angeles. The DC line is owned mostly by BPA in Oregon and equally owned by SCE and LADWP utilities in Southern California [21]. The North-to-South transfer capacity of the DC line is 3100 MW.

All of the trade transactions between British Columbia Hydro and California (or BPA) pass through the BPA region via a major intertie, called the Northern Intertie. The Northern Intertie connects BCH with BPA and it consists of two Ingledow-Cluster 500 kV transmission lines that cross the US-Canadian border at Blaine, Washington and two other transmission lines, Boundary-Nelway 230 kV and Boundary-Waneta 230 kV lines, that cross the border north of Spokane, Washington. The 500 kV lines form the “Western

Northern Intertie" while the 230 kV portion forms the "Eastern Northern Intertie" [23]. The Northern Intertie is reserved and scheduled as a single path. The total north-to-south capacity of Northern Intertie is 3100 MW.

Thus, depending on seasonal conditions, the Pacific Intertie is capable of collectively transferring up to 7,900 MW of surplus power generation in the PNW region to California markets, which is over 14% of the total installed MW capacity of California. As previously noted, most of these exports are during the spring and summer seasons, when the demand in California is high. Consequently, these exports play some role in setting the electricity prices in California ISO markets, which runs a short-term wholesale power market. Forecasting these short-term exports thus becomes crucial market intelligence to all the players involved with the California power market.

Chapter 3 - Overview of Box-Jenkins ARIMA methodology: Identifying Box-Jenkins models

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

This chapter provides an overview of the Box-Jenkins ARIMA methodology, a sophisticated time series analysis technique, and the ARIMA transfer function methodology, which is a blend of the explanatory and time series methodologies. Both, ARIMA and the ARIMA transfer function methodology, are the main tools used in identifying the forecasting models of later chapters. Software package SAS [33] is used as the platform to identify these models.

Section 3.1 introduces the basic terminology of Box-Jenkins methodology and the various classes of ARIMA models, followed by a discussion on the concept of 'Stationarity' of time series in section 3.2. In Section 3.3, the three-stage iterative method of fitting Box-Jenkins ARIMA models to a time series is explained, along with a discussion on the various statistical tests to measure the goodness of fit of such models. Section 3.4 looks at the procedure of forecasting using general ARIMA models. Seasonal ARIMA models and ARIMA transfer function models are covered in sections 3.5 and 3.6, respectively.

3. 1 Box-Jenkins ARIMA models: Definition and Terminology

Box-Jenkins ARIMA models [24]-[30] use historical values of a single variable to forecast its future values; hence they are classified as univariate methods. The variable of interest must be separated by equally spaced time intervals to apply Box-Jenkins methodology.

Let's consider a discrete time series of n equally spaced observations in time: $y_t = y_1, y_2, y_3, y_4, \ldots\ldots\ldots\ldots y_{n-1}, y_n$. The basic essence of Box-Jenkins methodology is that it considers the observed time series, y_t, to be the outputs of an unobservable 'black box' process. The inputs to this black box are a series of independent random shocks a_t, as illustrated in Fig 3.1.

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

For statistical purposes, these random shocks are assumed to be normally distributed with zero mean and a constant variance. This sequence is typically referred to as 'white noise' [24]. Thus, the Box-Jenkins approach views a time series as the result of transformation of a white noise process using a black box, which is nothing more than a linear filter.

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

1) The previous and current inputs (white noise or random shocks).
2) The previous output values of the time series under study, $y_{t-1}, y_{t-2}, \ldots$, in varying proportion.

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

Specifically, the Box-Jenkins approach proposes a simple linear form for the observed time series values [30]:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots + \varphi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \ldots - \theta_q a_{t-q} \qquad (3.1)$$

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

where $\Phi(B) = \left(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p\right)$,

$\Theta(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \ldots\ldots\ldots\ldots \theta_q B^q\right)$,

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

d = order of differencing

Equation 3.1 shows the current output as a linear weighted sum of previous outputs and inputs. Note that only "p" nonzero output terms and "q" nonzero input

terms are required to produce the current output. Thus only a finite number of recent inputs and outputs will have a statistically significant effect on the current output.

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

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

1) Autoregressive (AR) models:

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

   $$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots + \varphi_p y_{t-p} + a_t \qquad (3.3)$$

   or, $\Phi(B)\, y_t = a_t$ (3.4)

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

2) Moving Average (MA) models:

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

   $$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \ldots - \theta_q a_{t-q} \qquad (3.5)$$

   or, $y_t = \Theta(B)\, a_t$ (3.6)

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

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

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

3.2 Stationarity of time series

The Box-Jenkins methodology requires that the time series under analysis be "stationary" in both mean and variance. For a formal definition of stationarity, the reader is referred to [24]. In simplest non-statistical terms, the concept of stationarity can be explained as follows:

1) If the mean of the plotted series varies over time, the series is considered non-stationary in mean. If there is no evidence of a change in mean level over time, then the series is considered mean-stationary.
2) If the plotted series shows no obvious change in the variance over time, then the series is considered to be stationary in variance, otherwise it is considered to be non-stationary in variance.

One of the other advantages of Box-Jenkins model is that it can be applied to non-stationary series after making them stationary using some sort of transformation. In order to induce mean stationarity in the (mean non-stationary) data, typically, a concept called 'differencing' is used. A difference of order one (or second or higher orders) is all that is required to achieve mean stationarity in the majority of cases. A difference of order one means that each value of the time series is subtracted from the immediate previous value:

$$w_t = \bigtriangledown y_t = y_t - y_{t-1}$$

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

Therefore, $\nabla = (1 - B)$, the difference operator.

The result is a new time series w_t , having one less observation than the original y_t series. A difference of order two implies that the first order differenced series is differenced again resulting in:

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

The result again is a new time series k_t , having two less observations than the original series y_t. This can be generalized to d^{th} order differencing, where d is the order of differencing required to achieve mean stationarity.

After modeling the d^{th} order differenced series with an appropriate ARMA model, to reclaim the modeled values corresponding to the original undifferenced series, it is necessary to reverse the differencing transformation and "integrate" d times. This is represented by "I" in the acronym ARIMA and the order of integration is same as the order of differencing ('d' in this case).

Next, in order to make the series stationary in variance, if required, a different class of transformations can be carried out like the logarithmic transformation (taking log of the original data), square root, cubic root etc.

$w_t = \ln(y_t)$ (Natural logarithmic transformation)

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

If the series is not positive throughout, it can be made positive by adding a suitable constant c to each observation of the series before the transformation is carried out.

One of the best methods available to detect the proper transformation required to reduce heteroscadicity in data is using the Box-Cox transformation methodology, a general class of transformation which includes all other transformations mentioned earlier as special cases. The Box-Cox transformation also makes the data more normal distribution-like [33].

$$\left.\begin{aligned} w_t &= \frac{(y_t + c)^{\lambda} - 1}{\lambda} \quad if \quad \lambda \neq 0 \\ w_t &== \ln(y_t + c) \quad if \quad \lambda = 0 \end{aligned}\right\} \tag{3.7}$$

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

3.3 Steps in analyzing data and identifying ARIMA models

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

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

At the identification stage, the historical data of the time series of interest is statistically analyzed and an appropriate subclass of models from the general ARIMA (p,d,q) family is selected. The approach can be summarized as follows:

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

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

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

The basic tools for model identification (steps (b) and (c)) are the graphs of estimated Sample Autocorrelation Function (ACF) and the estimated Sample Partial Autocorrelation Function (PACF) obtained from the series. These graphs are used not only to help guess the form of the model, but also to obtain approximate estimates of the parameters (using Yule-Walker equations [24]), which are useful at the estimation stage to provide starting values for iterative procedures employed during the estimation of final parameters.

For a time series $y_t, t \geq 1$, the autocorrelation coefficient at lag k is [25]:

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

The sample k^{th} order autocorrelation is,

$$r_k = \frac{\sum_{t=1}^{n-k} \left(z_t - \bar{z}\right)\left(z_{t+k} - \bar{z}\right)}{\sum_{t=1}^{n} \left(z_t - \bar{z}\right)^2} \tag{3.8}$$

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

The sample partial autocorrelation coefficients can be computed as [25]:

$$r_{11} = r_1 \ ,$$

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} \left(r_{k-1,j}\, r_{k-j} \right)}{1 - \sum_{j=1}^{k-1} \left(r_{k-1,j}\, r_j \right)} \quad , \quad (k = 2, 3, \ldots\ldots) \tag{3.9}$$

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

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

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

These opposite characteristics are used to identify the type and order of AR, MA or ARMA processes in the data. In addition, other patterns may also be present in the ACF and PACF plots which can help to identify the true orders of AR and MA coefficients. The interested reader is referred to [24] for a detailed discussion on the identification of ARIMA models.

In practice the idealized procedure of significant spikes is confounded by sampling error in the estimated ACF and PACF and proper identification can become quite difficult depending on specific cases. Thus it requires some experience and judgment to identify a proper tentative form of the model.

Step 2- Estimation of the model parameters:

After a tentative form of the model is identified, the AR and MA parameters need to be estimated in the best possible manner. There are fundamentally two ways of getting final estimates:

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

b) Iterative improvement – choose a preliminary estimate obtained from the identification procedure (Yule-Walker equations [24]) and use an efficient nonlinear least-squares algorithm, called Marquardt algorithm, to refine the estimate iteratively.

Several methods exist to solve for the AR and MA coefficients using nonlinear square estimation: the Maximum Likelihood Method, Unconditional Least Squares Method and the Conditional Least Squares Method. In the present work, Conditional least squares method was employed since its computationally faster and under the assumption of normally distributed random shocks in the model, the Least squares parameter estimates are either exactly equal to or very nearly Maximum Likelihood estimates. Detailed information on these estimation methods can be found in [24]. In this work, the estimation of parameters was performed on SAS software package.

The Maximum likelihood method:

Let the ARIMA model be represented as [33]:

$\Phi(B)\left(W_t - \mu_t\right) = \Theta(B)\, a_t$,

where $W_t = \nabla^d y_t$ and μ_t is the mean parameter μ plus any exogenous inputs, if present in the model.

The log likelihood function can be written as:

$$-\frac{1}{2\sigma^2} x' \psi^{-1} x - \frac{1}{2} \ln|\psi| - \frac{n}{2} \ln\left(\sigma^2\right), \tag{3.10}$$

where vector $x = \begin{bmatrix} W_1 \\ W_2 \\ . \\ . \\ W_n \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_n \end{bmatrix}$, $\sigma^2 \psi$ is the variance of x as a function of the ϕ and θ parameters.

The maximum likelihood estimate (MLE) of σ^2 is:

$s^2 = \frac{1}{n-r} x^I \psi^{-1} x$, where r is the number of parameters in the model.

Thus, the log likelihood concentrated with respect to σ^2 can be taken up to additive constants as,

$$-\frac{n}{2}\ln\left(x^{I}\,\psi^{-1}\,x\right)-\frac{1}{2}\ln\left(|\psi|\right)$$

Let H be a lower triangular matrix with positive diagonal elements such that $HH^{I}=\psi$ and let the vector $e=H^{-1}x$.

The concentrated log likelihood with respect to σ^2 is now of the form,

$$-\frac{n}{2}\ln\left(e^{I}\,e\right)-\frac{1}{2}\ln\left(|H|\right), \tag{3.11}$$

or, $$-\frac{n}{2}\ln\left(|H|^{\frac{1}{n}}\,e^{I}\,e|H|^{\frac{1}{n}}\right)$$

Thus, the MLE is produced by using the Marquardt algorithm to minimize the sum of squares, $|H|^{\frac{1}{n}}\,e^{I}\,e|H|^{\frac{1}{n}}$, with the subsequent analysis of residuals done using e as the vector of residuals.

Unconditional Least Squares method:

The objective function to minimize is,

$$\sum_{t=1}^{n}\tilde{a}_t^2=\sum_{t=1}^{n}\left(x_t-C_tV_t^{-1}(x_1,\ldots\cdots,x_{t-1})^{I}\right)^2, \tag{3.12}$$

where C_t is the covariance matrix of x_t and $(x_1,x_2,\ldots\cdots,x_{t-1})$ and V_t is the variance matrix of $(x_1,x_2,\ldots\cdots,x_{t-1})$. It can be shown that,

$$\sum_{t=1}^{n}\tilde{a}^2=x^{I}\,\psi^{-1}\,x=e^{I}\,e$$

Thus, the unconditional least squares estimates are obtained by minimizing the sum of squared residuals.

Conditional Least Squares method:

The series x_t can be written in terms of its previous observations as:

$$x_t = a_t + \sum_{i=1}^{\infty} \pi_i x_{t-i} \ ,$$

where the π weights are obtained from the ratio of the Autoregressive and Moving Average polynomials,

$$\frac{\Phi(B)}{\Theta(B)} = 1 - \sum_{i=1}^{\infty} \pi_i B^i$$

The objective function to minimize is:

$$\sum_{t=1}^{n} \tilde{a}_t^2 = \sum_{t=1}^{n} \left(x_t - \sum_{t=1}^{\infty} \tilde{\pi}_i x_{t-i} \right)^2 \qquad (3.13)$$

where the unobserved past values of x_t are set to 0 and $\tilde{\pi}_i$ are computed from the estimates of ϕ and θ at each iteration.

In addition, at the estimation stage, there are certain conditions which the parameter estimates must meet in order for the model to be acceptable. All of the estimated parameters must be statistically significant. This can be verified using standard t-tests. Non-significant parameters must be dropped from the tentative model. Most importantly, the parameters must also satisfy certain inequality relations, called the stationarity and invertibility conditions [24]. The stationarity condition applies only to the AR parameters and the invertibility conditions apply only to the MA parameters. These conditions, in a combined fashion, check for the boundedness of the generated model. Finally, the correlation matrix of estimated parameters must be studied to ensure that the parameter estimates are not too highly correlated.

If satisfactory results are not obtained, an alternate model should be identified and re-estimated, using the possible insight gained during the estimation trials.

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

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

In particular, the following has to be performed:

1) A study of the residual series obtained after fitting the model to the data to see if any pattern remains accounted for. The ACF and PACF plots of the residual series help in detecting any unaccounted pattern.
2) A study of the sampling statistics of the current optimum solution to check if any further simplification of the model is possible.

The residuals left over after fitting an ARIMA model should ideally be just random noise (white noise) with zero mean and constant variance. The following statistical tests for lack of fit were used in the present work to check for the randomness of the residuals:

1) **ACF and PACF plots of the residuals**: The ACF of the residuals obtained after fitting a proper model to the data must show no significant autocorrelations at any lag order. Similarly, the PACF plot of the residuals must show no significant spikes at any lag order. Absence of any significant spikes in the residual ACF and PACF plots demonstrate proper fitting. However, in practice, there maybe a few spikes which are close to significance. One might expect approximately 1 lag in

every 20 lags to be statistically significant by chance alone for a 95% confidence limit test. Such spikes may not be a big concern; though their position of lag order also matters in deciding their importance and proper judgment should be used.

2) **Ljung-Box Chi-Square test**: Another measure of check for the randomness of residuals is using the Ljung-Box Chi-Square test. The null hypothesis is that the set of autocorrelations for residuals is white noise. This statistic measures the significance of residual autocorrelations as a set and points out if they are collectively significant [33]:

H_0 **:** The data is random

H_1 : The data is not random

It is computed as:

$$\chi_m^2 = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \tag{3.14}$$

where n is the size of sample, r_k is the sample autocorrelation at lag k, and the m is the number of lags being tested.

Each chi-square statistic is computed for all lags up to the indicated value and is not independent of the preceding chi-square values.

If α is the significance level, the null hypothesis is rejected if:

$\chi_m^2 > \chi_{1-\alpha,f}^2$, the α-quantile of the Chi-square distribution with f degrees of freedom.

Apart from these tests to check residual randomness, more tests need to be carried out on the model itself to check its adequacy and best fit. Two of the most important of such criteria are the Akaike's Information Criteria (AIC) and Schwarz's Bayesian

Criteria (SBC). The AIC and SBC are used to compare competing models fit to the same series. The model with smaller AIC and SBC values is a statistically better fit.

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

   It is computed as:

   $$AIC = -2\, loglikelihood + 2k \text{ ,} \tag{3.15}$$

   where k = number for parameters in the model ,

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

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

   $$AIC = n\left[\ln\frac{2\pi R}{n} + 1\right] + 2k \quad , \qquad R = \sum_{i=1}^{n} \tilde{a}_i^2$$

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

2) **Schwarz's Bayesian Criteria (SBC):** It is also called Bayesian Information Criteria (BIC). SBC is also a statistical tool for model selection, which penalizes

over fitting of estimation. The model with lower SBC is generally the best fit. It is computed as:

$$SBC = -2\, loglikelihood + k \ln(n) \tag{3.16}$$

For normally and independently distributed residuals,

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

The SBC criterion penalizes free parameters more heavily than the AIC criterion.

It must be noted that both AIC and SBC tests may not generally point to a common model as the best fit. In such cases, proper judgment is required in choosing the best fit for the model. In this work, a Monte-Carlo simulation of the validity of these tests was performed and the results are provided in section 3.7.

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

section and if they are comparable with respect to fitting adequacy and yielding white noise residuals, the model with minimum number of parameters must be preferred. More on the principle of parsimony can be found in [24].

3.4 Forecasting using ARIMA models

Once an adequate and satisfactory model is fitted to the series of interest, forecasts can be generated using the model. Consider the general ARIMA model of equation (3.1).

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots + \varphi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \ldots - \theta_q a_{t-q}$$

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

$$y_{t+1} = \varphi_1 y_t + \varphi_2 y_{t-1} + \ldots\ldots + \varphi_p y_{t-p+1} + a_{t+1} - \theta_1 a_t - \theta_2 a_{t-1} \ldots\cdot\cdot - \theta_q a_{t-q+1} \qquad (3.17)$$

Except a_{t+1}, the random shock at time $t+1$, all other parameters are known. Thus, setting $a_{t+1} = 0$, its true expected value, the one-step ahead forecasts can be generated. Similarly, future forecasts can be 'bootstrapped' using these obtained forecasts and setting the unrealized random shocks to 0 for each case. Also, the 95% confidence interval for the forecasts can be calculated [28].

3.5 Seasonal ARIMA models

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

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

In general, a time series often may contain both non-seasonal and seasonal components. Though the time series may be deseasonalized and a non-seasonal ARIMA model maybe fitted to the remainder, experience suggests that Box-Jenkins methodology provides good forecasts of periodic data series. Thus, it may be advisable to leave the seasonal component in the data and fit a general class of ARIMA model which accounts for both seasonality and non-seasonality [25]. Such a general ARIMA model can be represented by the form ARIMA $\left(p,d,q\right)\left(P,D,Q\right)^{s}$. This is commonly referred to as a seasonal ARIMA multiplicative model and it is represented by:

$$\Phi\left(B^{s}\right)\Phi\left(B\right)\left(1-B^{s}\right)^{D}\left(1-B\right)^{d}\, y_{t} = \Theta\left(B^{s}\right)\Theta\left(B\right)a_{t} \tag{3.18}$$

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

3.6 ARIMA Transfer Functions

In most cases, one or more exogenous inputs may affect the variable of interest; but may not completely determine it. In such cases, a blend of Explanatory and Time series methods, in the form of ARIMA transfer functions, is advised. These models combine the characteristics of the univariate ARIMA models and multiple regression models. Much of the theory and procedure behind identifying ARIMA transfer functions is common with univariate ARIMA methods and will not be repeated here.

A general ARIMA transfer function with multiple inputs can be represented by:

$$Z_t = v(B)U_t + w(B)V_t + \ldots\ldots + N_t \tag{3.19}$$

Where, Z_t = the output series

$U_t, V_t, \ldots$ = the input series

N_t = noise term

and, $v(B) = \left(v_0 + v_1 B + v_2 B^2 + \ldots\ldots + v_k B^k\right)$,

$w(B) = \left(w_0 + w_1 B + w_2 B^2 + \ldots\ldots + w_f B^f\right)$ etc., are the transfer functions of the corresponding inputs.

The order of the transfer functions, $k, f, \ldots$,etc., may sometimes be quite large and thus may not always be parsimonious. Thus, an alternate parsimonious form of the above transfer function model can be represented as [24]:

$$z_t = \frac{\omega_1(B)}{\delta_1(B)} u_{t-b_1} + \frac{\omega_2(B)}{\delta_2(B)} v_{t-b_2} + \ldots\ldots\ldots + \frac{\theta(B)}{\phi(B)} a_t \tag{3.20}$$

where, in general,

$\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \ldots\ldots\ldots - \omega_s B^s$,

$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 + \ldots\ldots + \delta_r B^r$,

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 + \ldots\ldots + \theta_q B^q$,

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 + \ldots\ldots + \phi_p B^p$,

z_t = the transformed and differenced value of Z_t ,

$u_t, v_t, \ldots$= the transformed and differenced values of $U_t, V_t, \ldots$, respectively,

a_t = residual terms, and ,

r, s, p, q and b are constants.

Of key interest here is to first identify the correct sets of inputs for the model. After the inputs are identified, their relation with the variable of interest is determined by the following procedure:

An appropriate ARIMA model is identified and estimated for the input series. The prewhitening of input series is carried out, which requires the removal of the entire pattern present in the data by filtering the input series with appropriate ARIMA filter. This ARIMA filter is then applied to the output series. The residual series of the prewhitened input series and the corresponding 'prewhitened' output series are cross-correlated with each other. The cross-correlation coefficients are studied to identify the relationship between the prewhitened input and the 'prewhitened' output series. (Note that the 'prewhitened' output series may not be random white noise as in the case of prewhitened input series; hence the quotation marks) The same relationship, obtained for the prewhitened input and 'prewhitened' output series, holds for the input and output series, since the same ARIMA filter was employed on both the series.

If there are multiple inputs, the above procedure is repeated. However, the procedure assumes that the different inputs should be independent of each other.

Once the transfer function between the input series and the output series is identified, equation (15) or (16) can be used to compute the noise series (loosely defined as that portion of the output which is not explained by the inputs):

$$n_t = z_t - \frac{\omega_1(B)}{\delta_1(B)} u_{t-b_1} - \frac{\omega_2(B)}{\delta_2(B)} v_{t-b_2} - \ldots\ldots\ldots \tag{3.21}$$

A suitable univariate ARIMA model can be fit to the noise series by using the three step iterative process of Box-Jenkins, resulting in,

$$n_t = \frac{\theta(B)}{\phi(B)} a_t \tag{3.22}$$

Thus, the final model becomes,

$$z_t = \frac{\omega_1(B)}{\delta_1(B)} u_{t-b_1} + \frac{\omega_2(B)}{\delta_2(B)} v_{t-b_2} + \ldots\ldots\ldots + \frac{\theta(B)}{\phi(B)} a_t, \tag{3.23}$$

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

3.7 AIC and SBC Simulations

In Section 3.3, the AIC and SBC tests of model selection were described. It was mentioned that SBC test is a superior test than AIC test under certain conditions (SBC test penalizes free parameters heavily than AIC test). In some of the cases in the later chapters, the SBC test was given preference over the AIC test in selecting the model. In this appendix, a sample ARMA(1,2) process was simulated 1000 times (by varying the random shocks) and the AIC and SBC values for 5 competing models were noted for each run. It was noted that the SBC test (minimum SBC criteria) pointed to the correct model more often than AIC test (minimum AIC criteria) for mixed ARMA processes.

3.7.1 Simulation results for a ARMA(1,2) process

A sample ARMA(1,2) (or P1Q2) process was considered from the source[25]:

$(1-0.801B)\ y_t = \left(1-0.423B+0.225B^2\right) a_t$

Five competing models were selected: P1Q2 (correct model), P3, Q3, P2Q2 and P2Q1.

The following simulation results were obtained:

Table 3.2 Simulation runs (1000 times) for competing models

Model	No. of times AIC was minimum	No. of times SBC was minimum
P1Q2 (Correct model)	238	249
P3	201	217
Q3	131	140
P2Q2	249	198
P2Q1	181	196

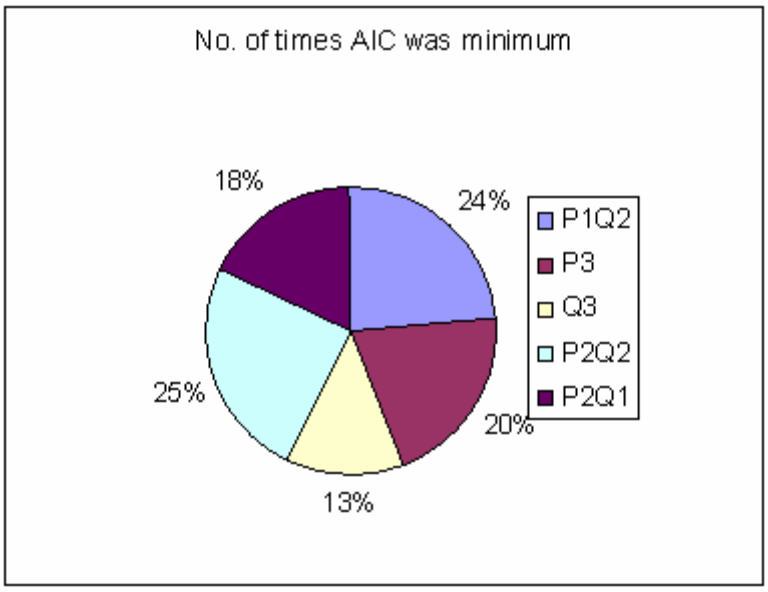

Figure 3-2: Percentage times AIC was minimum for the competing models

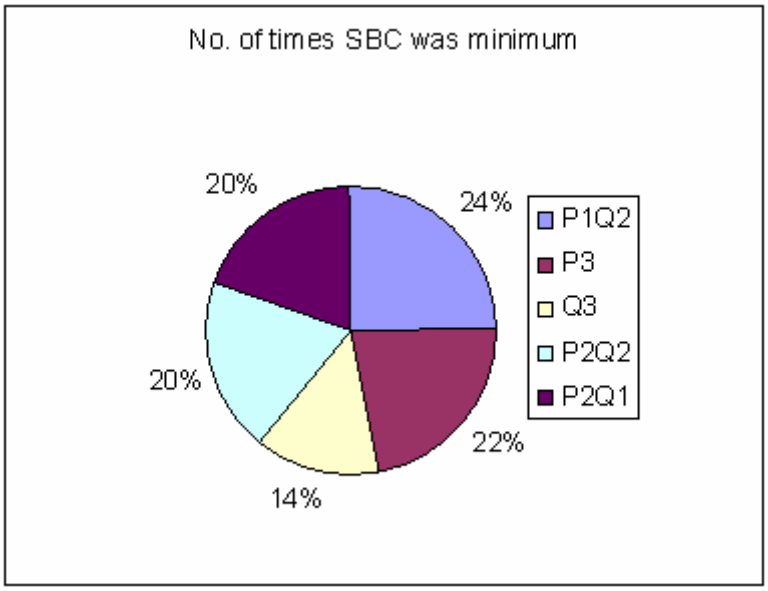

Figure 3-3: Percentage times SBC was minimum for the competing models

Chapter 4 – Forecasting the Pacific Northwest Exports to California

4.1 Definition of problem

The California-Oregon Intertie (COI) and the Pacific DC Intertie (PDCI) are collectively responsible for the flow of several thousand MWs of hydro-generation exports to California from the Pacific Northwest region (Section 2.2, Chapter 2). Since hydro resources are abundantly available in this region, the exports are cheaper than other sources of generation in California for most part. Thus, they form a valuable resource of generation for California.

In this chapter, a suitable model is identified to forecast the exports to California from Pacific Northwest via the COI and PDCI interties on a short-term basis. Since the dynamics of exports change from day-to-day, it is proposed that a different model be identified for each day of the week. The results of this work are illustrated by identifying a model which produces forecasts for the 16 on-peak hours of subsequent Thursday. Thursday was chosen on a pure arbitrary basis, since it forms the middle day of trading week in CAISO markets. The inputs used in the model are the historical data for each week's Thursday's 16 on-peak hours from Jan 1, 2006 through May 31, 2008.

In section 4.2, the various inputs to the model are identified. In section 4.3, the actual model is identified to forecast these exports. The model is statistical in nature and its functioning is validated in section 4.4 with various tests of significance. Sections 4.5 and 4.6 present the results of fitting and forecasting, respectively. Section 4.7 presents the conclusions of the chapter. The SAS programs are provided in Appendix A.

4.2 Selection of inputs

As outlined in chapter-2, the bulk of hydro-generation in the WECC area is situated in the BPA and BC Hydro (BCH) region. Different combinations of information available in public-domain were analyzed and specifically, the combined power exports to California via the COI and PDCI interties (henceforth referred by ACDC (t) series) were found to depend mainly on:

1) The surplus hydro generation (after the regional demand is met) in the BPA region (referred by BPA (t) series),
2) The exports from BCH to the BPA region (referred by BC (t) series), and
3) The changing trend of total demand in the California ISO (CAISO) system (referred by CA (t) series).

Table 4.1 presents the correlation coefficients between the various quantities.

For the input series BPA(t), the historical hourly generation values of 19 major BPA hydro-generators in Washington and Oregon states were considered for this study. The list of generators and their name plate capacities are provided in Appendix B.

Table 4-1: Correlation Coefficients Table

Series	ACDC(t)	BPA(t)	BC(t)	CA(t)
ACDC(t)	1			
BPA(t)	0.68	1		
BC(t)	0.275	-0.234	1	
CA(t)	0.483	0.145	0.43	1

For the BC(t) input series, the historical exports from BCH to BPA territory in US were used. The historical hourly flows on the Northern intertie were considered for this study.

For the CA(t) input series, the CAISO system demand was used in order to capture the trends in the demand of California system and to incorporate the temperature dependency of the power exports to California.

4.3 Identifying the Forecasting model

Box-Jenkins ARIMA methodology was employed to fit the final models. A detailed discussion on Box-Jenkins ARIMA methodology can be found in chapter 3. Chapter 3 also discussed the procedure followed to identify an ARIMA transfer function. The specific case of fitting ARIMA Transfer function to forecast the power exports to California (ACDC(t)) is presented in this section. The identification of the ARIMA model will be carried out according to the following steps:

4.3.1 Transformation of the inputs and output series

For illustration purpose, plots of the three input series to the model are shown in Figs 4-1 to 4-3 for the months of Jan- May 2008. Also shown in Fig 4-4 is the plot of actual power exports to California for the same period, which is the output of the proposed Transfer function model.

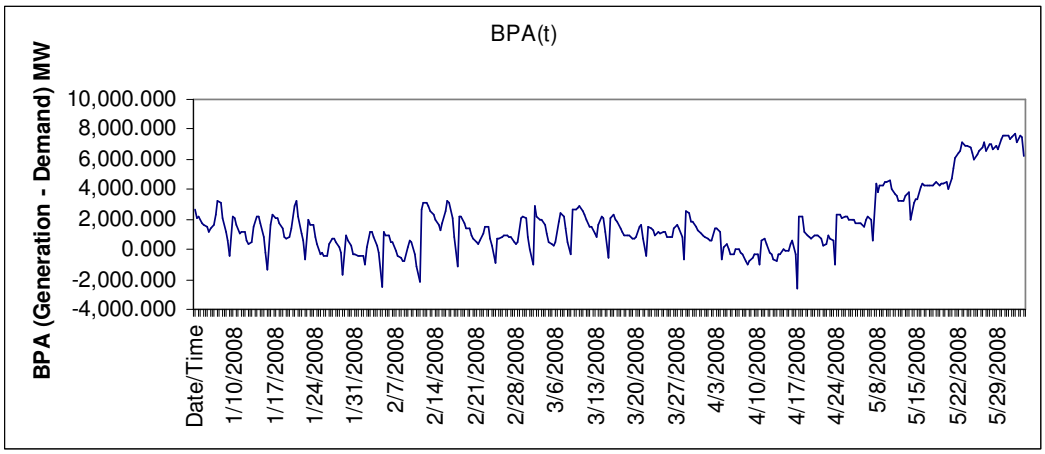

Figure 4-1: Plot of BPA(t), Jan-May 2008

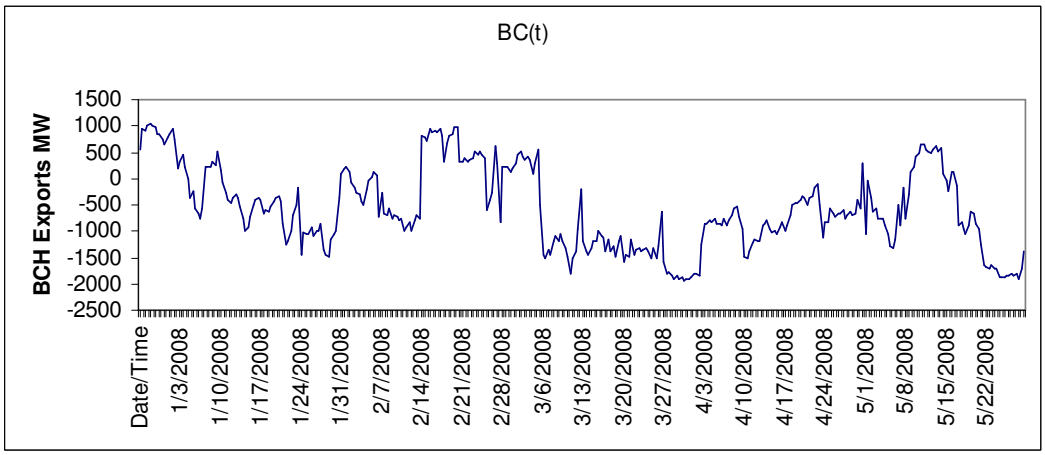

Figure 4-2: Plot of BC(t), Jan-May 2008

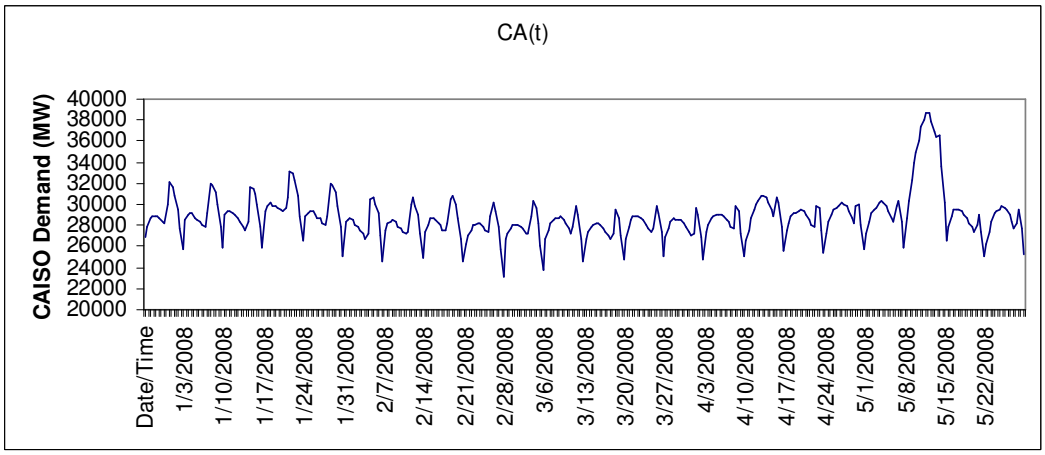

Figure 4-3: Plot of CA(t), Jan-May 2008

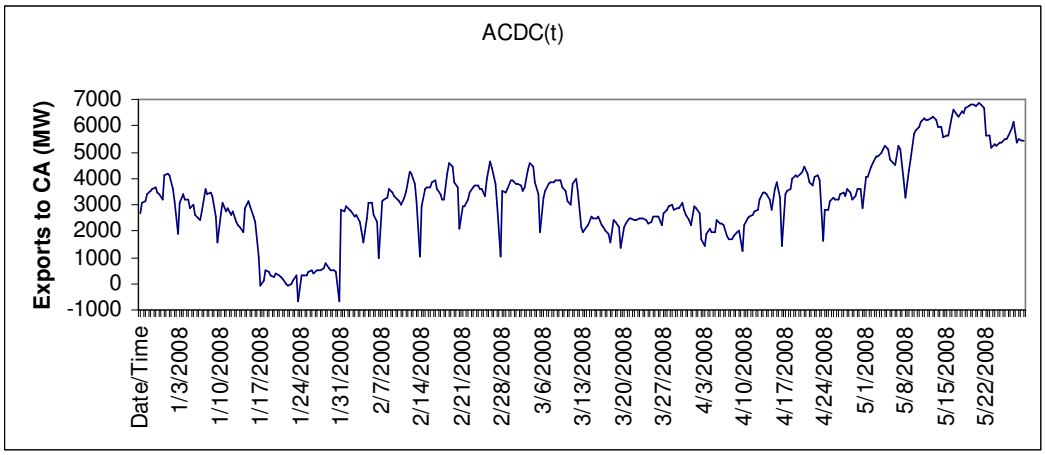

Figure 4-4: Plot of ACDC(t), Jan-May 2008

The preliminary requirement of ARIMA methodology is that the series should be stationary in both mean and variance (Section 3.2, chapter 3). Each of the BPA(t), BC(t), CA(t) and ACDC(t) series were standardized to a common metric using

$$\left(\frac{x_t - \mu}{\sigma}\right),$$

where x_t = observation at time t, μ = mean of the series, σ = standard deviation of the series. A positive constant, 10, was added to observations of the series and Box-Cox transformation was applied. A range of λ values were tried and the one which provided the optimal transformation was selected. Specifically, the series was transformed with each λ value and an autoregressive model was fit to the series. The likelihood of the data under each autoregressive model was computed and the λ producing the maximum likelihood was chosen as the optimal coefficient of transformation [33].

Table 4.6 presents the optimal coefficients that were obtained for the set of series. Tables 4-2- 4.5 provides the results of related computations over a range of values of λ.

Table 4-2: Optimal transformation for ACDC(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
2.50	-731.78	0.14	1475.56	1509.12
2.45	-730.63	0.14	1473.26	1506.82
2.39	-729.59	0.13	1471.17	1504.73
2.34	-728.65	0.13	1469.30	1502.85
2.29	-727.82	0.13	1467.64	1501.20
2.24	-727.10	0.13	1466.21	1499.76
2.18	-726.49	0.13	1464.99	1498.54
2.13	-726.00	0.13	1463.99	1497.55
2.08	-725.61	0.13	1463.22	1496.78
2.03	-725.34	0.13	1462.68	1496.23
1.97	-725.18	0.13	1462.36	1495.91
1.92	**-725.13**	**0.13**	**1462.27**	**1495.82**
1.87	-725.21	0.13	1462.41	1495.97
1.82	-725.39	0.13	1462.79	1496.34
1.76	-725.70	0.13	1463.39	1496.95
1.71	-726.12	0.13	1464.24	1497.79
1.66	-726.66	0.13	1465.32	1498.88
1.61	-727.32	0.13	1466.64	1500.20
1.55	-728.10	0.13	1468.21	1501.76
1.50	-729.01	0.13	1470.02	1503.57

Table 4-3: Optimal transformation for BPA(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
2.00	-94.96	0.06	201.93	235.48
1.95	-94.47	0.06	200.94	234.50
1.89	-94.09	0.06	200.18	233.73
1.84	-93.81	0.06	199.62	233.18
1.79	-93.65	0.06	199.29	232.85
1.74	**-93.59**	**0.06**	**199.17**	**232.73**
1.68	-93.64	0.06	199.28	232.83
1.63	-93.80	0.06	199.59	233.15
1.58	-94.07	0.06	200.13	233.69
1.53	-94.44	0.06	200.89	234.44
1.47	-94.93	0.06	201.86	235.41
1.42	-95.52	0.06	203.05	236.60
1.37	-96.23	0.06	204.45	238.01
1.32	-97.04	0.06	206.08	239.63
1.26	-97.96	0.06	207.92	241.48
1.21	-98.99	0.06	209.98	243.54
1.16	-100.13	0.06	212.26	245.82
1.11	-101.38	0.06	214.76	248.31
1.05	-102.74	0.06	217.47	251.02
1.00	-104.20	0.07	220.40	253.96

Table 4-4: Optimal transformation for BC(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
1.00	-1124.91	0.18	2261.82	2295.38
0.95	-1124.39	0.18	2260.79	2294.34
0.89	-1123.95	0.18	2259.91	2293.46
0.84	-1123.59	0.18	2259.18	2292.73
0.79	-1123.30	0.18	2258.60	2292.15
0.74	-1123.08	0.18	2258.17	2291.72
0.68	-1122.94	0.18	2257.89	2291.44
0.63	**-1122.88**	**0.18**	**2257.76**	**2291.31**
0.58	-1122.89	0.18	2257.78	2291.34
0.53	-1122.98	0.18	2257.95	2291.51
0.47	-1123.14	0.18	2258.27	2291.82
0.42	-1123.37	0.18	2258.74	2292.29
0.37	-1123.68	0.18	2259.35	2292.91
0.32	-1124.06	0.18	2260.12	2293.67
0.26	-1124.51	0.18	2261.03	2294.58
0.21	-1125.04	0.18	2262.08	2295.64
0.16	-1125.64	0.18	2263.29	2296.84
0.11	-1126.32	0.18	2264.63	2298.19
0.05	-1127.06	0.18	2266.13	2299.68
0.00	-1127.88	0.18	2267.77	2301.32

Table 4-5: Optimal transformation for CA(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
4.50	370.53	0.20	-729.07	-695.51
4.45	372.08	0.20	-732.16	-698.61
4.39	373.44	0.20	-734.88	-701.33
4.34	374.61	32747.03	-737.23	-703.68
4.29	375.60	6.68	-739.20	-705.65
4.24	376.41	0.76	-740.81	-707.26
4.18	377.03	77.44	-742.05	-708.50
4.13	377.46	0.49	-742.93	-709.37
4.08	377.72	0.22	-743.44	-709.89
4.03	**377.80**	**0.50**	**-743.60**	**-710.05**
3.97	377.70	0.17	-743.40	-709.85
3.92	377.43	0.13	-742.85	-709.30
3.87	376.97	0.11	-741.95	-708.39
3.82	376.35	0.10	-740.70	-707.14
3.76	375.55	0.10	-739.10	-705.54
3.71	374.58	0.10	-737.16	-703.60
3.66	373.44	12.86	-734.87	-701.32
3.61	372.13	0.13	-732.25	-698.70
3.55	370.65	0.08	-729.29	-695.74
3.50	369.00	0.07	-725.99	-692.44

Table 4-6: Optimal transformation coefficients

Input Series	Box-Cox Coefficient	Output Series	Box-Cox Coefficient
BPA(t)	1.73684	ACDC(t)	1.92105
BC(t)	0.63158		
CA(t)	4.02632		

The set of series were transformed using the optimal coefficients (eqn 3.7, Chapter 3) to correct the non-stationarity in variance. In order to check for mean-stationarity in the set of series, the autocorrelation and partial autocorrelation plots were analyzed. These are presented in Figs. 4.5-4.8.

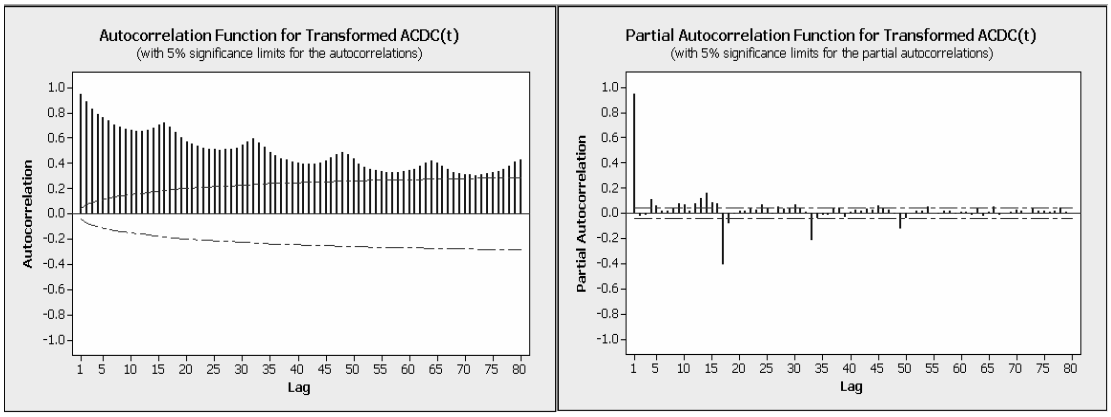

Figure 4-5: ACF and PACF plots of transformed ACDC(t)

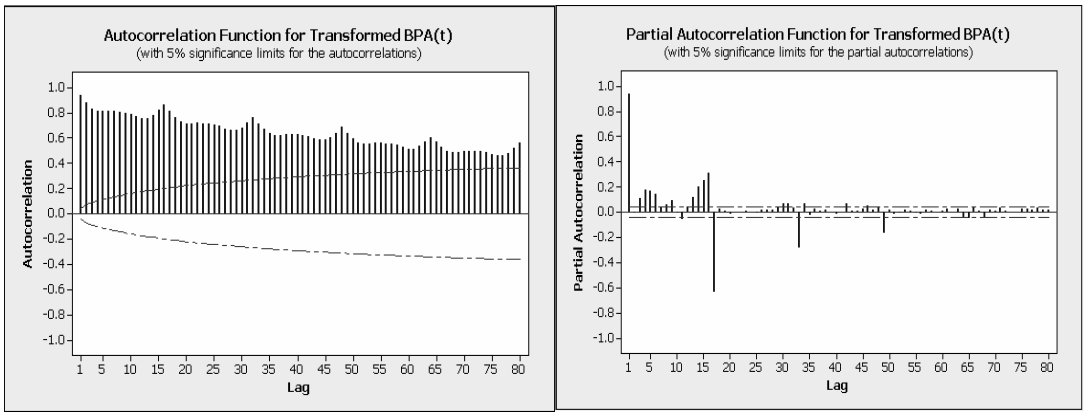

Figure 4-6: ACF and PACF plots of transformed BPA(t)

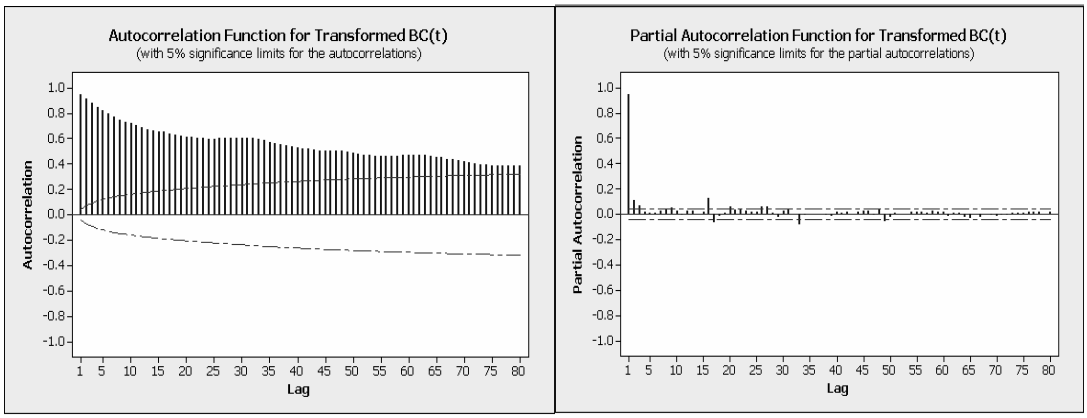

Figure 4-7: ACF and PACF plots of transformed BC(t)

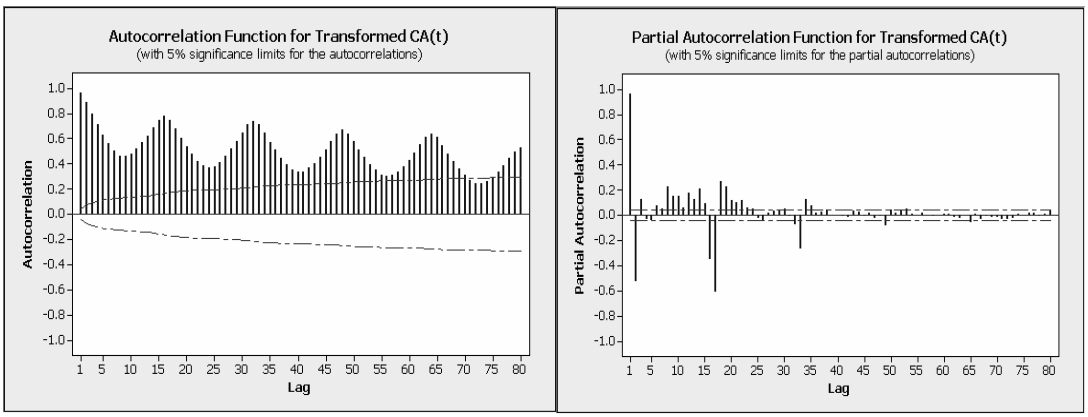

Figure 4-8: ACF and PACF plots of transformed CA(t)

Analysis of the autocorrelation plots revealed that each of the series is mean non-stationary in non-seasonal and seasonal orders (Section 3.2, Chapter 3). Thus, suitable differencing was required for each of the series to render mean-stationarity. First order non-seasonal and seasonal differencing was found to be sufficient for each series (Table 4.7).

Figs. 4.9-4.12 presents the autocorrelation and partial autocorrelation plots of the series obtained after differencing with the optimal non-seasonal and seasonal orders. From the figures, it can be deduced that each of the differenced series is mean-stationary.

Table 4-7: Optimal differencing orders

Input series	**Differencing order (d/D)** **d = non-seasonal** **D = seasonal**	**Output series**	**Differencing order (d/D)** **d = non-seasonal** **D = seasonal**
BPA(t)	d =1; D =1	ACDC(t)	d =1; D =1
BC(t)	d =1; D =1		
CA(t)	d =1; D =1		

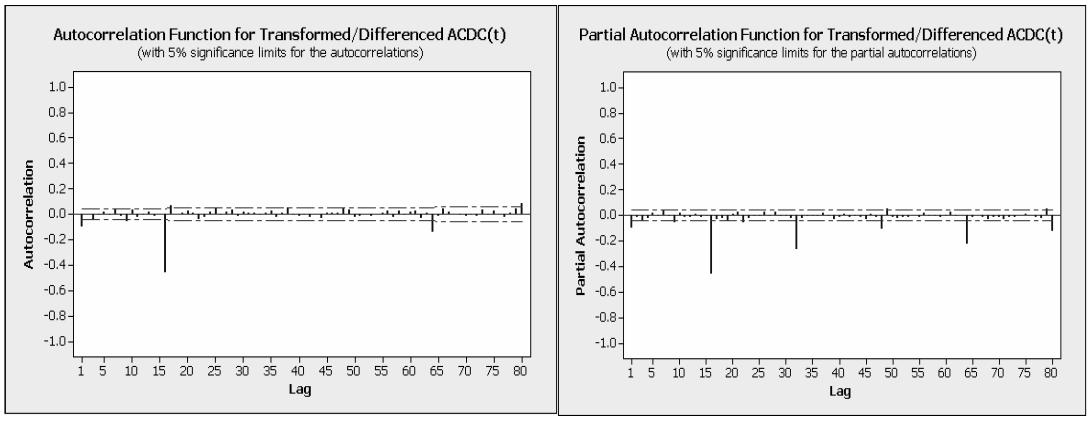

Figure 4-9: ACF and PACF plots of transformed/differenced ACDC(t)

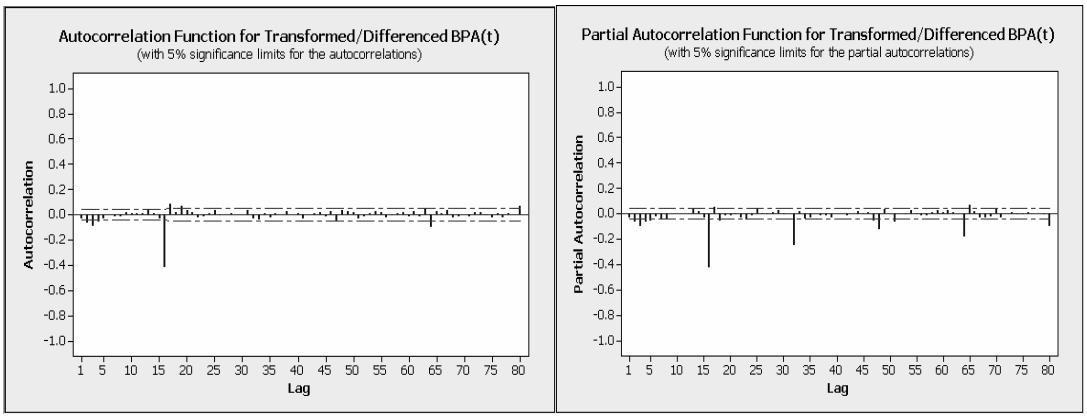

Figure 4-10: ACF and PACF plots of transformed/differenced BPA(t)

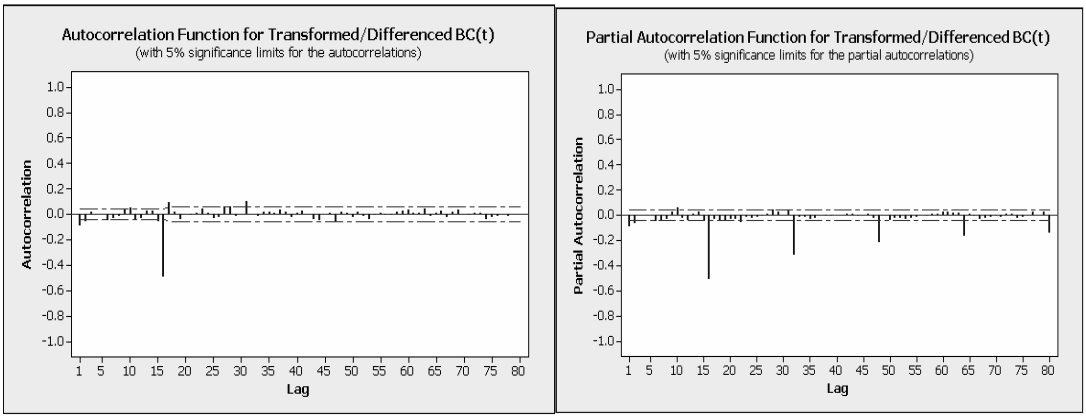

Figure 4-11: ACF and PACF plots of transformed/differenced BC(t)

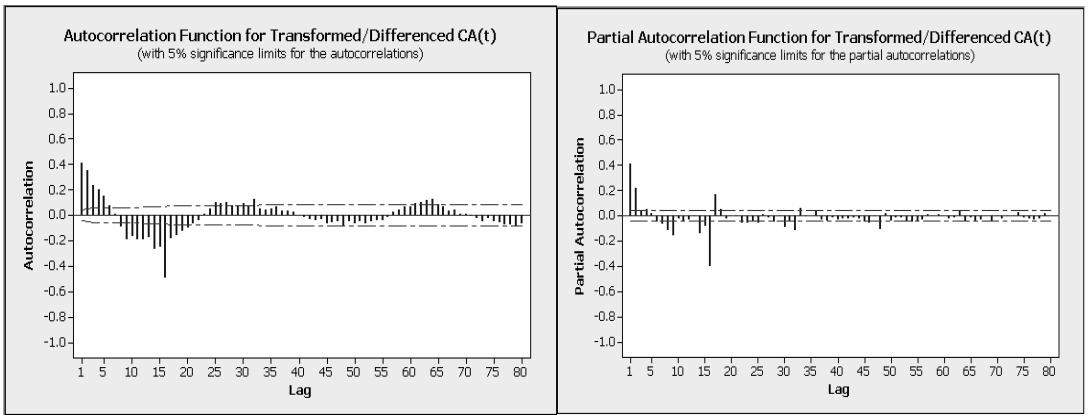

Figure 4-12: ACF and PACF plots of transformed/differenced CA(t)

4.3.2 Prewhitening the input series

The set of input series was individually prewhitened in this step. Prewhitening of an input series is a procedure to remove all identifiable patterns present in the series, leaving white noise as the residual series. The prewhitening filters are basically ARMA filters (note that no differencing is required at this stage since the set of input series is already differenced in step (1), hence ARMA and not ARIMA models are fitted for prewhitening). Thus, the three-stage analysis procedure of ARMA model fitting (Section 3.3, Chapter 3), viz., the identification, estimation and diagnostic checking, was carried out for each series in turn and suitable filters were fitted.

The final ARIMA prewhitening filters obtained for the set of input series are:

For the BPA(t) series:

$$\left(1-\theta_2 B^2-\theta_3 B^3-\theta_4 B^4-\theta_5 B^5-\theta_8 B^8-\theta_{16} B^{16}-\theta_{17} B^{17}\right)\left(1-\theta_{16} B^{16}-\theta_{64} B^{64}\right) \quad (4.1)$$

For the BC(t) series:

$$\left(1-\theta_1 B^1-\theta_2 B^2-\theta_4 B^4-\theta_6 B^6-\theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16}\right) \quad (4.2)$$

For the CA(t) series:

$$\left(1-\theta_1 B^1-\theta_2 B^2-\theta_3 B^3-\theta_4 B^4-\theta_7 B^7-\theta_8 B^8-\theta_9 B^9-\theta_{10} B^{10}-\theta_{11} B^{11}-\right.$$

$$\left.\theta_{12} B^{12}-\theta_{13} B^{13}-\theta_{14} B^{14}-\theta_{15} B^{15}-\theta_{16} B^{16}\right)\left(1-\theta_{16} B^{16}-\theta_{64} B^{64}\right) \quad (4.3)$$

which can be compactly represented in an alternate format (Table 4.8). The parameter estimates are given in Tables 4.9, 4.12 and 4.15. It can be seen from the t-test results that all the parameters are significant at 95% significance level.

Table 4-8: Prewhitening filter forms for the Input series

Series	Series Name	Prewhitening Filter Form
Input	**BPA(t)**	P=0 Q=(2 3 4 5 8 16 17) (16 64)
Input	**BC(t)**	P=0 Q=(1 2 4 6 15) (16)
Input	**CA(t)**	P=0 Q=(1 2 3 4 7 8 9 10 11 12 13 14 15 16) (16 32)

Verification of the validity of prewhitening filters for the set of input series was done by analyzing the autocorrelation and partial autocorrelation plots of the set of residual series (Figs. 4.13 – 4.16) and by conducting various statistical tests to measure the goodness of fit (Tables 4.9 – 4.17).

For the three input series, the autocorrelation and partial autocorrelation plots of the residual series are shown in Figs. 4.13-4.16. Also included is the table of parameter estimates, correlation between parameter estimates and Ljung-Box Chi-square test results for each of the series (Section 3.3, Chapter 3).

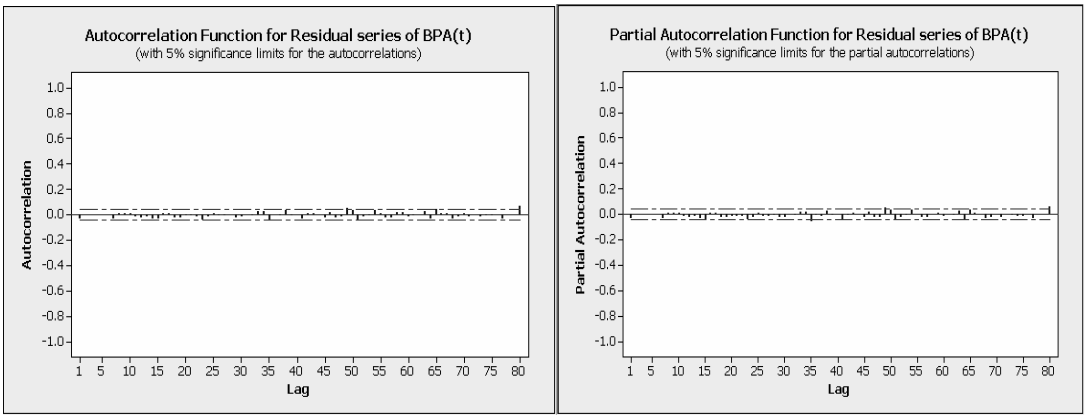

Figure 4-13: ACF and PACF plots of residual series of BPA(t)

Table 4-9: Parameter estimates of BPA(t) series

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.10537	0.02229	4.73	<.0001	2
MA1,2	0.12909	0.02210	5.84	<.0001	3
MA1,3	0.08516	0.02215	3.84	0.0001	4
MA1,4	0.05214	0.02246	2.32	0.0203	5
MA1,5	0.05611	0.02237	2.51	0.0122	8
MA1,6	-0.06882	0.03039	-2.26	0.0237	16
MA1,7	-0.07419	0.02200	-3.37	0.0008	17
MA2,1	0.71388	0.02339	30.52	<.0001	16
MA2,2	0.06430	0.01857	3.46	0.0005	64

Table 4-10: Correlation matrix of parameter estimates of BPA(t) series

Correlations of Parameter Estimates									
Parameter	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5	MA1,6	MA1,7	MA2,1	MA2,2
MA1,1	1.000	0.009	-0.132	-0.143	-0.023	0.047	0.054	-0.016	0.013
MA1,2	0.009	1.000	-0.007	-0.124	-0.082	0.016	0.053	0.012	0.018
MA1,3	-0.132	-0.007	1.000	0.017	-0.112	0.028	0.020	-0.007	0.015
MA1,4	-0.143	-0.124	0.017	1.000	-0.135	0.014	0.015	-0.007	-0.010
MA1,5	-0.023	-0.082	-0.112	-0.135	1.000	-0.079	-0.023	0.015	0.004
MA1,6	0.047	0.016	0.028	0.014	-0.079	1.000	-0.062	-0.690	0.391
MA1,7	0.054	0.053	0.020	0.015	-0.023	-0.062	1.000	0.075	-0.024
MA2,1	-0.016	0.012	-0.007	-0.007	0.015	-0.690	0.075	1.000	-0.559
MA2,2	0.013	0.018	0.015	-0.010	0.004	0.391	-0.024	-0.559	1.000

Table 4-11: Ljung-Box Chi-square test results for BPA(t) series

Autocorrelation Check of Residuals										
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations						
6	.	0	.	-0.034	-0.003	0.001	0.003	0.003	0.002
12	6.09	3	0.1072	-0.032	0.007	0.005	0.005	-0.017	-0.022
18	13.08	9	0.1591	-0.019	-0.030	-0.037	0.012	0.011	-0.024
24	19.22	15	0.2040	-0.026	-0.010	-0.012	-0.017	-0.039	-0.019
30	21.58	21	0.4242	0.012	-0.010	-0.009	-0.002	-0.025	-0.014
36	29.63	27	0.3310	0.000	-0.009	0.027	0.021	-0.052	0.004
42	33.90	33	0.4238	-0.010	0.030	-0.008	0.001	-0.031	0.009
48	37.68	39	0.5301	0.012	0.002	-0.022	0.019	-0.023	-0.018
54	52.71	45	0.2006	0.052	0.034	-0.043	-0.018	-0.005	0.036
60	57.06	51	0.2601	0.005	-0.026	-0.022	0.014	0.020	-0.018
66	65.39	57	0.2085	-0.006	-0.004	0.029	-0.036	0.041	0.012
72	69.10	63	0.2789	0.006	-0.032	-0.020	0.007	-0.018	-0.002
78	72.47	69	0.3644	-0.016	-0.011	-0.007	0.004	-0.034	0.003

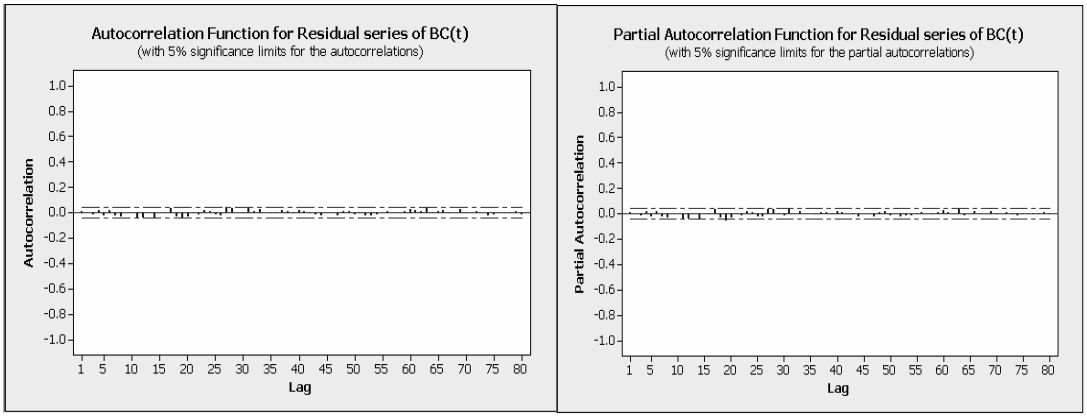

Figure 4-14: ACF and PACF plots of residual series of BC(t)

Table 4-12: Parameter estimates of BC(t) series

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.16165	0.02219	7.28	<.0001	1
MA1,2	0.08880	0.02238	3.97	<.0001	2
MA1,3	0.06137	0.02215	2.77	0.0057	4
MA1,4	0.06804	0.02202	3.09	0.0020	6
MA1,5	0.11591	0.02204	5.26	<.0001	15
MA2,1	0.83590	0.01252	66.77	<.0001	16

Table 4-13: Correlation matrix of parameter estimates of BC(t) series

Correlations of Parameter Estimates						
Parameter	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5	MA2,1
MA1,1	1.000	-0.178	0.001	0.013	0.042	-0.077
MA1,2	-0.178	1.000	-0.121	-0.067	-0.014	0.041
MA1,3	0.001	-0.121	1.000	-0.122	0.036	-0.002
MA1,4	0.013	-0.067	-0.122	1.000	-0.004	0.001
MA1,5	0.042	-0.014	0.036	-0.004	1.000	-0.138
MA2,1	-0.077	0.041	-0.002	0.001	-0.138	1.000

Table 4-14: Ljung-Box Chi-square test results for BC(t) series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.007	0.004	-0.014	0.016	-0.028	0.013
12	12.57	6	0.0503	-0.028	-0.030	-0.002	0.004	-0.040	-0.039
18	20.56	12	0.0572	-0.001	-0.045	0.001	-0.005	0.033	-0.029
24	28.63	18	0.0531	-0.048	-0.033	-0.006	-0.014	0.019	0.007
30	35.42	24	0.0624	-0.018	-0.021	0.039	0.032	-0.003	-0.007
36	40.28	30	0.0996	0.041	0.008	0.022	0.004	-0.006	-0.009
42	41.93	36	0.2292	0.015	0.006	-0.008	0.021	0.006	-0.005
48	45.00	42	0.3474	-0.017	-0.022	-0.003	-0.003	-0.026	0.005
54	48.99	48	0.4332	0.012	-0.017	-0.005	-0.024	-0.026	-0.016
60	50.75	54	0.6005	0.004	0.012	-0.002	0.001	0.011	0.024
66	54.55	60	0.6745	0.014	0.006	0.034	-0.007	0.009	0.018
72	56.29	66	0.7973	-0.010	-0.000	0.023	-0.012	-0.003	0.009
78	58.89	72	0.8665	-0.003	-0.027	-0.015	-0.010	-0.011	-0.010

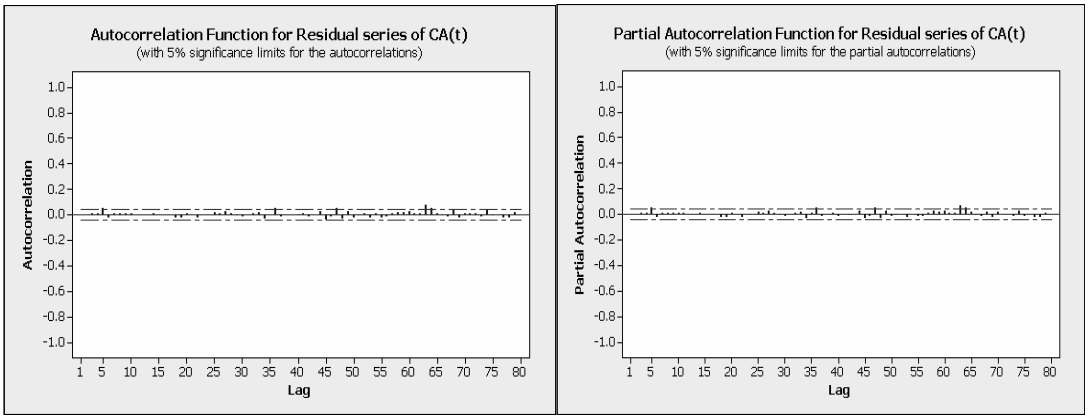

Figure 4-15: ACF and PACF plots of residual series of CA(t)

Table 4-15: Parameter estimates of CA(t) series

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	-0.24994	0.02233	-11.19	<.0001	1
MA1,2	-0.16853	0.02293	-7.35	<.0001	2
MA1,3	-0.07861	0.02277	-3.45	0.0006	3
MA1,4	-0.0737	0.02221	-3.32	0.0009	4
MA1,5	0.11261	0.02177	5.17	<.0001	7
MA1,6	0.17021	0.02149	7.92	<.0001	8
MA1,7	0.2771	0.0216	12.83	<.0001	9
MA1,8	0.18069	0.02207	8.19	<.0001	10
MA1,9	0.16596	0.02263	7.33	<.0001	11
MA1,10	0.12538	0.02251	5.57	<.0001	12
MA1,11	0.07646	0.02231	3.43	0.0006	13
MA1,12	0.17435	0.02269	7.68	<.0001	14
MA1,13	0.10357	0.02364	4.38	<.0001	15
MA1,14	0.12835	0.04846	2.65	0.0081	16
MA2,1	0.46658	0.05552	8.4	<.0001	16
MA2,2	-0.06071	0.02948	-2.06	0.0396	32

Table 4-16: Correlation matrix of Parameter estimates of CA(t) series

Correlations of Parameter Estimates																
Parameter	**MA1,1**	**MA1,2**	**MA1,3**	**MA1,4**	**MA1,5**	**MA1,6**	**MA1,7**	**MA1,8**	**MA1,9**	**MA1,10**	**MA1,11**	**MA1,12**	**MA1,13**	**MA1,1**	**MA2,1**	**MA2,2**
MA1,1		.229	.139	.06	.004	0.147	0.199	0.295	0.184	0.171	0.12	0.073	0.165	0.005	0.032	
MA1,2	.229		.243	.139	0.092	0.063	0.21	0.261	0.322	0.202	0.179	0.103	0.078	0.025	0.039	.032
MA1,3	.139	.243		.224	0.021	0.149	0.117	0.25	0.278	0.331	0.192	0.151	0.098	.021	0.042	.046
MA1,4	.06	.139	.224		0.02	0.035	0.162	0.1	0.238	0.237	0.289	0.126	0.121	0.006	0.021	.005
MA1,5	.004	0.092	0.021	0.02		.157	.061	0.036	0.022	0.157	0.077	0.202	0.195	0.109	0.01	0.023
MA1,6	0.147	0.063	0.149	0.035	.157		.142	.043	0.03	0.065	0.171	0.119	0.195	0.072	0.019	0.006
MA1,7	0.199	0.21	0.117	0.162	.061	.142		.165	.084	0.016	0.034	0.15	0.05	0.15	.081	0.095
MA1,8	0.295	0.261	0.25	0.1	0.036	.043	.165		.207	.092	0.002	0.025	0.066	0.151	.147	0.107
MA1,9	0.184	0.322	0.278	0.238	0.022	0.03	.084	.207		.229	.116	.029	.037	0.25	.221	0.167
MA1,10	0.171	0.202	0.331	0.237	0.157	0.065	0.016	.092	.229		.218	.115	.08	0.222	.244	0.195
MA1,11	0.12	0.179	0.192	0.289	0.077	0.171	0.034	0.002	.116	.218		.211	.145	0.261	.284	0.207
MA1,12	0.073	0.103	0.151	0.126	0.202	0.119	0.15	0.025	.029	.115	.211		.273	0.316	.382	0.245
MA1,13	0.165	0.078	0.098	0.121	0.195	0.195	0.05	0.066	.037	.08	.145	.273		0.298	.413	0.27
MA1,14	0.005	0.025	.021	0.006	0.109	0.072	0.15	0.151	0.25	0.222	0.261	0.316	0.298		0.892	.624
MA2,1	0.032	0.039	0.042	0.021	0.01	0.019	.081	.147	.221	.244	.284	.382	.413	0.892		0.714
MA2,2		.032	.046	.005	0.023	0.006	0.095	0.107	0.167	0.195	0.207	0.245	0.27	.624	0.714	

Table 4-17: Ljung-Box Chi-square test results for CA(t) series

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	.	0	.	0.002	0.002	0.008	0.011	0.047	-0.025
12	.	0	.	0.005	0.006	0.007	0.008	0.004	0.002
18	7.64	2	0.0219	0.004	0.006	0.005	0	0.003	-0.024
24	10.02	8	0.2635	-0.021	0.005	-0.001	-0.027	-0.003	0
30	13.22	14	0.5093	0.019	0.011	0.025	0.009	-0.008	-0.018
36	21.62	20	0.3614	0	0.006	0.016	-0.037	0.004	0.05
42	22.78	26	0.6452	-0.014	-0.002	0.003	-0.008	0.009	-0.016
48	34.58	32	0.3456	0	0.026	-0.038	-0.016	0.049	-0.031
54	38.28	38	0.4568	0.028	-0.02	-0.006	0.005	-0.022	0.005
60	42.86	44	0.5205	-0.027	-0.014	0.008	0.021	0.018	0.022
66	60.76	50	0.1417	0.007	0.009	0.074	0.054	0.012	-0.008
72	64.75	56	0.1978	-0.015	0.03	-0.025	0.01	0.005	0.008
78	69.98	62	0.2275	-0.017	0.03	-0.007	0.001	-0.022	-0.028

4.3.3 "Prewhitening" the output series

Next, each of the prewhitening filter corresponding to the set of inputs was applied to the output (ACDC(t)) in turns and the residual series were stored. A standard assumption was made that the three input series are not inter-correlated to a significant level [32]. In the present case, the correlation coefficient between BC(t) and the CA(t) series was found to be 0.43 (Table 4.1). Nevertheless, in next chapter, this assumption was removed and an alternate model was identified by combining the three inputs into a single composite input. The results of both the models were then compared. But for the time being, it was assumed that the set of three input series is not inter-correlated to a significant level.

The set of resultant series, obtained after 'prewhitening' the ACDC(t) series in turns, mayn't be pure white noise; in fact if they were white noise, then that would indicate absence of any relationship between the corresponding input and the output series.

4.3.4 Computing Cross-correlation between the prewhitened inputs and output series

The cross-correlation coefficients between a prewhitened input and the corresponding "prewhitened" output series are a direct measure of the relationship present between the two. In this step, the set of three cross-correlation coefficients tables, one between each of the prewhitened inputs and the corresponding "prewhitened" output (output "prewhitened" by the prewhitening filter of the corresponding input), were computed and plotted.

The plots and the table of cross-correlation coefficients are presented next.

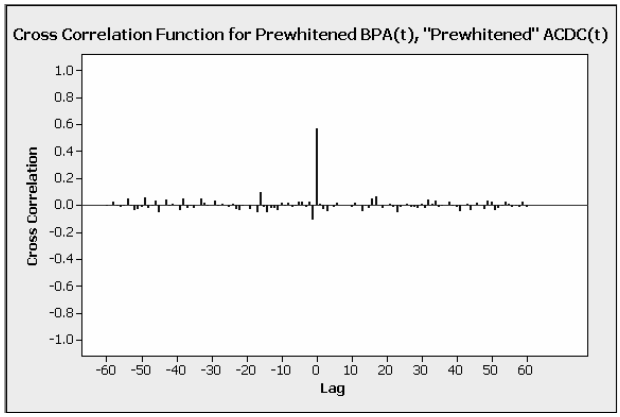

Figure 4-16: Cross-correlation plot for prewhitened BPA(t), "Prewhitened" ACDC(t)

Table 4-18: Cross-correlation coefficients for prewhitened BPA(t), "Prewhitened" ACDC(t)

Lag	CCF	Lag	CCF
-16	0.09462	**0**	**0.564951**
-15	-0.0155	1	0.008622
-14	-0.05499	2	-0.03094
-13	-0.02715	3	-0.04717
-12	-0.02263	4	-0.00256
-11	-0.03641	5	-0.01633
-10	0.016008	6	0.015393
-9	-0.01142	7	-0.00375
-8	0.016181	8	0.000149
-7	-0.01805	9	-0.00392
-6	-0.00821	10	-0.01896
-5	0.023636	11	0.01247
-4	0.026237	12	0.002916
-3	-0.01251	13	-0.04687
-2	0.027513	14	-0.00572
-1	-0.11196	15	-0.0267
		16	0.045595

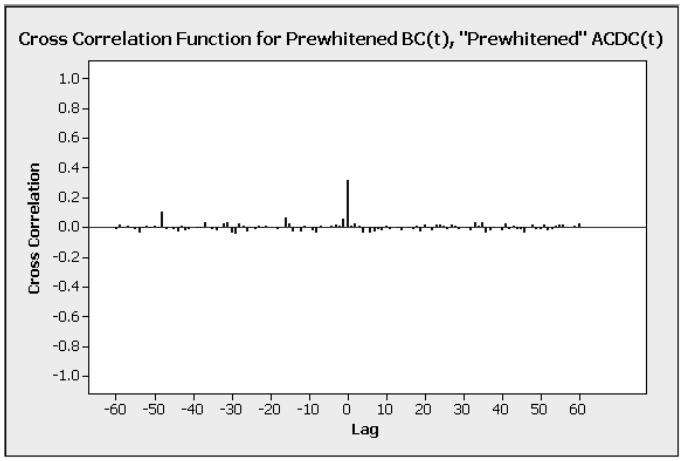

Figure 4-17: Cross-correlation plot for prewhitened BC(t), "Prewhitened" ACDC(t)

Table 4-19: Cross-correlation coefficients for prewhitened BC(t), "Prewhitened" ACDC(t)

Lag	CCF	Lag	CCF
-16	0.064754	**0**	**0.312409**
-15	0.024522	1	0.006988
-14	-0.03272	2	0.026673
-13	-0.00664	3	0.006425
-12	-0.03248	4	-0.0382
-11	0.005531	5	-0.00069
-10	0.001886	6	-0.04039
-9	-0.02119	7	-0.03463
-8	-0.03879	8	-0.0191
-7	0.009557	9	-0.02223
-6	-0.00324	10	0.007171
-5	-0.00224	11	-0.01395
-4	0.004073	12	-0.0012
-3	0.016874	13	-0.01075
-2	0.01179	14	-0.02222
-1	0.051689	15	0.002459
		16	-0.00505

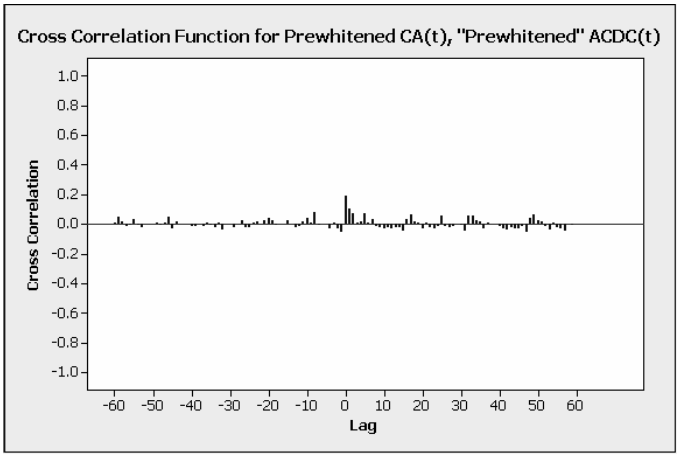

Figure 4-18: Cross-correlation plot for prewhitened CA(t), "Prewhitened" ACDC(t)

Table 4-20: Cross-correlation coefficients for prewhitened CA(t), "Prewhitened" ACDC(t)

Lag	CCF	Lag	CCF
-16	-0.00127	**0**	**0.189628**
-15	0.027346	**1**	**0.106012**
-14	0.001998	2	0.072262
-13	-0.02491	3	0.01075
-12	-0.0151	4	0.017857
-11	0.019679	5	0.068959
-10	0.036116	6	0.009884
-9	0.010794	7	0.031905
-8	0.082207	8	-0.01765
-7	-0.00573	9	-0.02691
-6	-0.00239	10	-0.02816
-5	0.000428	11	-0.0272
-4	-0.03483	12	-0.02872
-3	0.010106	13	-0.02529
-2	-0.03112	14	-0.02283
-1	-0.05912	15	-0.04581
		16	0.035482

4.3.5 Identifying the input transfer functions of the model

The cross-correlation plot obtained between the prewhitened BPA(t) input series and the "prewhitened" ACDC(t) output series using the prewhitening filter corresponding to BPA(t), is shown in Fig (4.16).

There were no significant coefficients at negative lags, a large significant coefficient existed at lag 0 while the coefficients at other positive lags were non-significant. Thus, I deduced that the relationship between the BPA(t) and ACDC(t) series is first-order linear function. (Note that the small spikes at negative lag were neglected. Even when they were considered, the coefficients corresponding to those lags turned out to be non-significant at the estimation stage) Likewise, the final relationship obtained between the BC(t) and ACDC(t) is linear (Fig 4.17) and that between the CA(t) and ACDC(t) (Fig 4.18) is a first-order rational transfer function of the form, $\frac{\omega_2}{\left(1-\delta_2 B\right)}$.

Table 4.21 provides the final estimates of coefficients for the transfer functions obtained using non-linear estimation (Section 3.3, chapter 3).

Table 4-21: Transfer function coefficients

Series	Series name	Coefficients
Input	BPA(t)	ω_0= 1.203
Input	BC(t)	ω_1= 6.84
Input	CA(t)	ω_2= 0.0004322, δ_2= 0.656

4.3.6 Extraction of Preliminary noise series and fitting an ARIMA model

Once the coefficients of the input transfer functions were available, the preliminary noise series, n_t, was extracted using Eqn. 3.21,Chapter-3.

The n_t values were analyzed using the three-step Box-Jenkins methodology and an appropriate ARMA (p_n ,0 , q_n) model was fitted. The autocorrelation and partial autocorrelation plots of the preliminary noise series were useful in determining the autoregressive and moving average orders of the ARMA model.

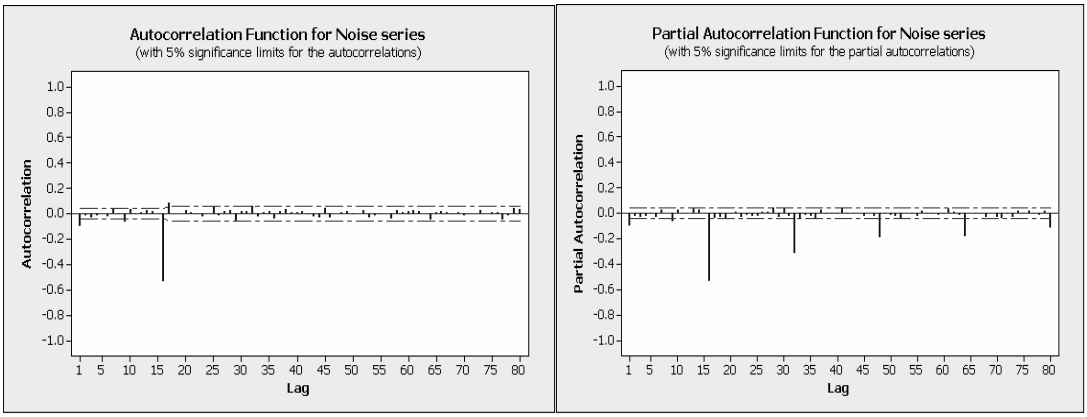

Figure 4-19: ACF and PACF plots for Preliminary Noise Series

The following ARMA model form was obtained for the preliminary noise series of the model by using the three-step Box-Jenkins iterative strategy and choosing the best fit from the AIC and SBC measures (Table 4.25):

$$\frac{1-\theta_{16}B^{16}}{1-\varphi_1 B^1-\varphi_2 B^2-\varphi_3 B^3-\varphi_9 B^9-\varphi_{15} B^{15}-\varphi_{32} B^{32}} \tag{4.4}$$

or in compact notation,

P= (1 2 3 9 15 32) Q=(16).

The parameter estimates are provided in Table 4.22 and the correlation matrix of parameter estimates is provided in Table 4.23.

Table 4-22: Parameter estimates of Preliminary Noise Series

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.93260	0.0089001	104.79	<.0001	16	ACDC	0
AR1,1	-0.14954	0.02267	-6.60	<.0001	1	ACDC	0
AR1,2	-0.05441	0.02253	-2.41	0.0158	2	ACDC	0
AR1,3	-0.07394	0.02229	-3.32	0.0009	3	ACDC	0
AR1,4	-0.06671	0.02204	-3.03	0.0025	9	ACDC	0
AR1,5	-0.10528	0.02206	-4.77	<.0001	15	ACDC	0
AR1,6	0.09578	0.02392	4.00	<.0001	32	ACDC	0
NUM1	1.20331	0.03177	37.88	<.0001	0	BPA	0
NUM2	6.84011	0.27151	25.19	<.0001	0	BC	0
NUM3	0.0004322	0.00009518	4.54	<.0001	0	CA	0
DEN1,1	0.65646	0.09432	6.96	<.0001	1	CA	0

Table 4-23: Correlation matrix of Parameter estimates of Preliminary Noise Series

Correlations of Parameter Estimates											
Variable Parameter	ACDC MA1,1	ACDC AR1,1	ACDC AR1,2	ACDC AR1,3	ACDC AR1,4	ACDC AR1,5	ACDC AR1,6	BPA NUM1	BC NUM2	CA NUM3	CA DEN1,1
ACDC MA1,1	1.000	-0.033	-0.020	-0.005	0.003	-0.019	0.363	0.065	0.090	-0.098	0.109
ACDC AR1,1	-0.033	1.000	0.159	0.046	0.017	-0.013	0.035	0.179	0.005	-0.021	0.020
ACDC AR1,2	-0.020	0.159	1.000	0.152	0.022	0.034	-0.004	0.052	-0.001	0.021	0.000
ACDC AR1,3	-0.005	0.046	0.152	1.000	0.031	-0.007	0.057	0.028	-0.014	0.006	0.018
ACDC AR1,4	0.003	0.017	0.022	0.031	1.000	0.028	0.035	-0.005	0.014	-0.026	0.011
ACDC AR1,5	-0.019	-0.013	0.034	-0.007	0.028	1.000	-0.016	-0.022	0.013	-0.008	-0.014
ACDC AR1,6	0.363	0.035	-0.004	0.057	0.035	-0.016	1.000	-0.007	0.006	0.025	-0.009
BPA NUM1	0.065	0.179	0.052	0.028	-0.005	-0.022	-0.007	1.000	0.162	-0.393	0.285
BC NUM2	0.090	0.005	-0.001	-0.014	0.014	0.013	0.006	0.162	1.000	-0.203	0.072
CA NUM3	-0.098	-0.021	0.021	0.006	-0.026	-0.008	0.025	-0.393	-0.203	1.000	-0.650
CA DEN1,1	0.109	0.020	0.000	0.018	0.011	-0.014	-0.009	0.285	0.072	-0.650	1.000

4.3.7 Identifying the final forecasting model

Thus, the final forecasting model becomes:

$$(1-B)\left(1-B^{16}\right)ACDC^{bxcx}(t)=\omega_0(1-B)\left(1-B^{16}\right)BPA^{bxcx}(t)+\omega_1(1-B)\left(1-B^{16}\right)BC^{bxcx}(t)+\frac{\omega_2}{1-\delta_2 B}(1-B)\left(1-B^{16}\right)CA^{bxcx}(t)+\frac{\left(1-\theta_{16}B^{16}\right)a_t}{1-\varphi_1 B^1-\varphi_2 B^2-\varphi_3 B^3-\varphi_9 B^9-\varphi_{15}B^{15}-\varphi_{32}B^{32}} \quad (4.5)$$

where $ACDC^{bxcx}(t)$ denotes the Box-Cox transformed series of standardized ACDC(t) series, and so on. The coefficients of the input transfer functions are provided in Table 4.21 and the parameters of the transfer function for noise series are provided in Table 4.22.

4.4 Diagnostic checking of the model

The following statistical tests were implemented to check the validity of the model:

4.4.1 Analysis of the final noise series

Presented in Fig 4-20 are the ACF and PACF plots of the final noise series, a_t. The absence of statistically significant spikes at all the lags beyond lag 0 in the autocorrelation and partial autocorrelation plots signify the absence of any defined pattern in the final noise series. Thus, the final noise series corresponds to a realization of white noise, confirming the validity of the model.

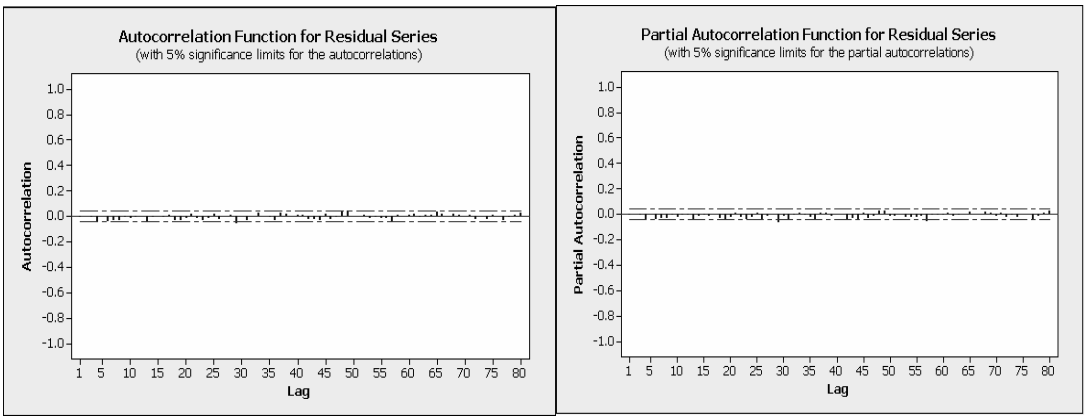

Figure 4-20: ACF and PACF plots for Final Noise Series

4.4.2 Ljung-Box Chi-Square statistical test

The Ljung-Box Chi-Square test can be used to determine if the set of autocorrelations in the ACF plot of Fig 4.20 are significantly different from zero (Eqn 3.14, Chapter 3).

Table 4.24 provides the Ljung-Box Chi-Square test results for the final noise series. At 95% significance level, Table 4.24 confirms that the residual series is a realization of white noise.

4.4.3 AIC and SBC tests

Three models for the preliminary noise series were entertained at the identification, estimation and diagnostic checking stages, based on the analysis of the series data. The AIC and SBC criteria was used to select the best model at the final stage.

The three competing models and the corresponding AIC and SBC values are shown in Table 4.25. As already mentioned in section 4.3.7, the model number 1, which

Table 4-24: Ljung-Box Chi-square test results for Final Noise Series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.007	0.000	-0.011	-0.045	-0.001	-0.043
12	13.13	5	0.0222	-0.032	-0.033	-0.008	-0.016	-0.006	-0.004
18	18.76	11	0.0655	-0.041	-0.008	-0.002	-0.011	0.006	-0.031
24	25.69	17	0.0802	-0.035	-0.018	0.019	-0.017	-0.032	-0.015
30	35.11	23	0.0507	0.019	-0.025	-0.002	0.005	-0.061	-0.003
36	41.79	29	0.0586	-0.038	0.003	0.024	0.003	-0.006	-0.036
42	46.04	35	0.1003	0.026	0.020	-0.000	0.011	0.008	-0.028
48	55.34	41	0.0666	-0.024	-0.034	0.016	-0.023	-0.005	0.045
54	59.43	47	0.1055	0.041	0.003	-0.004	0.008	-0.013	-0.010
60	64.68	53	0.1304	-0.014	-0.015	-0.044	0.006	0.003	0.012
66	68.37	59	0.1891	0.019	-0.004	0.006	0.009	0.034	0.014
72	70.55	65	0.2974	0.002	0.015	0.009	-0.002	0.008	-0.026
78	74.90	71	0.3531	0.003	-0.023	0.010	0.004	-0.036	-0.013

has the lowest AIC and SBC values, was chosen as the final model for the preliminary noise series.

Table 4-25: AIC and SBC values for competing models of preliminary noise series

Sr. No.	ARMA model	AIC	SBC	Variance estimate	Std error estimate
1	**P=(1 2 3 9 15 32) Q=(16)**	**7595.82**	**7657.33**	**2.688**	**1.639**
2	P=(1 2 3 9 15 32) Q=(16 32)	7602.57	7669.67	2.696	1.642
3	P=(1 2 3 9 15 16 32) Q=0	7932.16	7993.67	3.185	1.784

4.5 Results of fitting

The fitting plots (also called one-step ahead forecast plots) are provided next for the duration, Jan 2006 through May 2008.

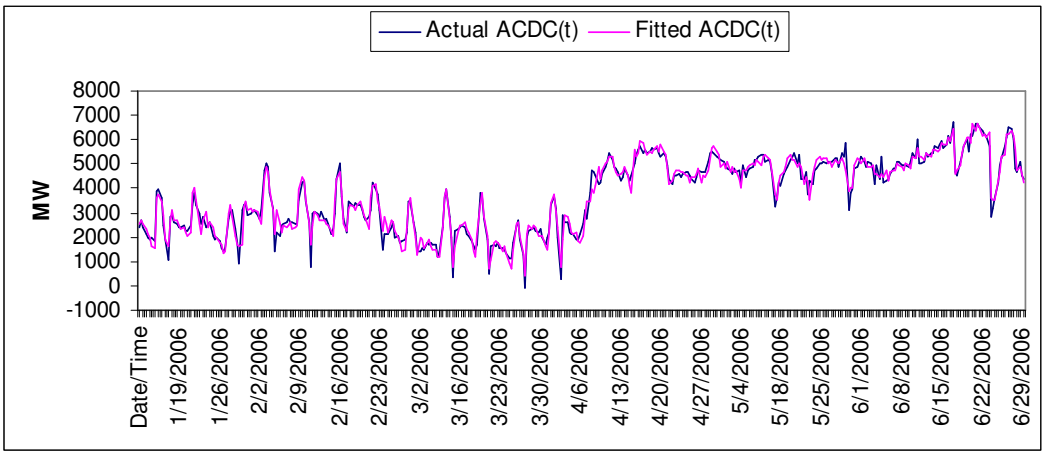

Figure 4-21: Actual and Fitted ACDC(t), Jan -June 2006

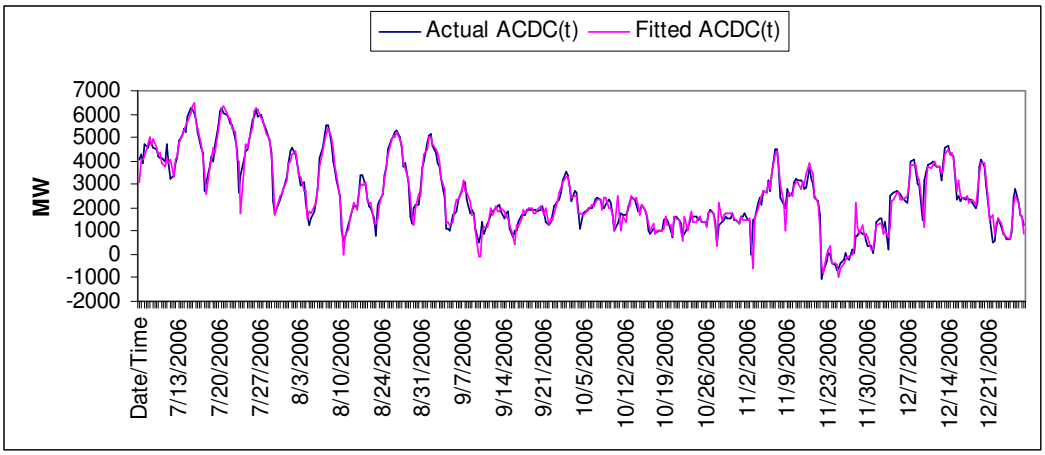

Figure 4-22: Actual and Fitted ACDC(t), July -Dec 2006

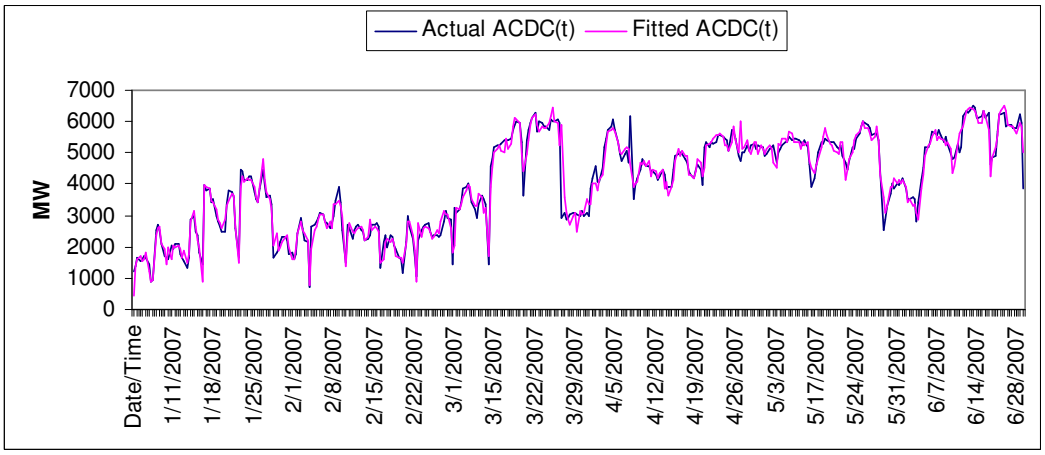

Figure 4-23: Actual and Fitted ACDC(t), Jan -June 2007

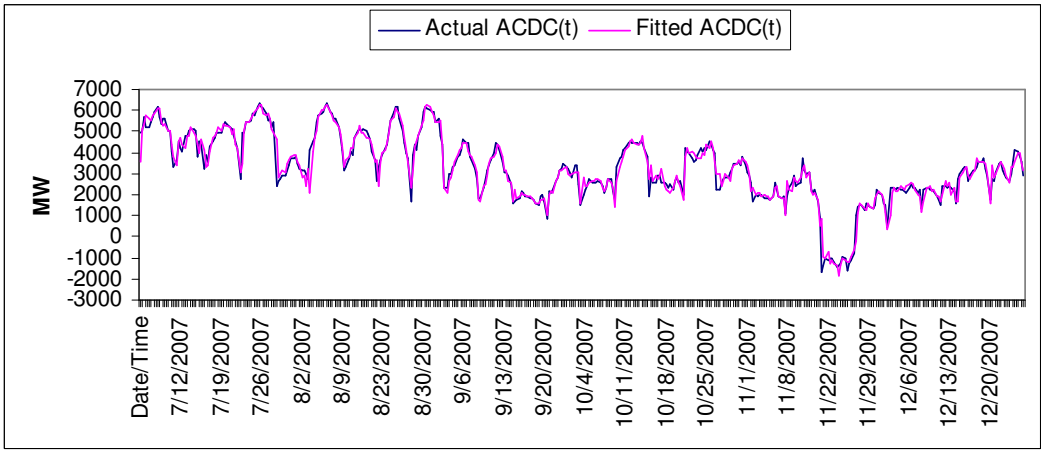

Figure 4-24: Actual and Fitted ACDC(t), July -Dec 2007

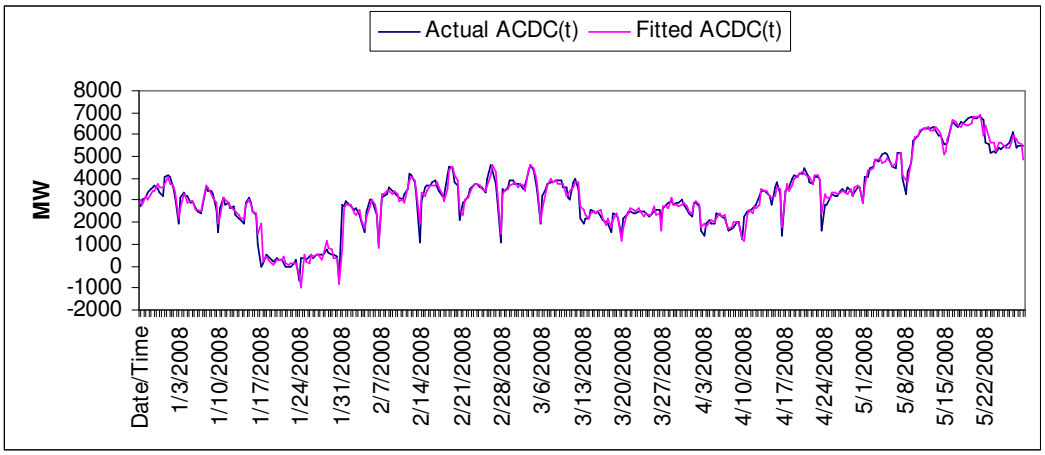

Figure 4-25: Actual and Fitted ACDC(t), Jan -May 2008

Several measures were computed to gauge the accuracy of fitting:

1) R-square

2) R-square (Adjusted)

3) Mean Error (ME)

4) Root Mean Square Error (RMSE)

5) Mean Absolute Deviation (MAD)

6) Mean Absolute Percentage Error (MAPE)

Table 4.26 presents the results of fitting accuracy measurements.

Table 4-26: Fitting accuracy measures

Criterion	Jan 2006-May2008
R-Sqr	0.961
R-Sqr Adj	0.960
ME	-2.788
RMSE	321.086
MAD	211.218
MAPE	14.964

4.6 Results of forecasting

The final model (Eqn 4.5) was used for forecasting the ACDC(t) values for the16 on-peak hours of the subsequent Thursday (5th June 2008). The forecasts for the 16 on-peak hours of 5th June 2008 and the 95% confidence limits, along with the original data, are shown in Fig. 4.26.

The various measures of forecasting accuracy were computed and the results are presented in Table 4.27.

As with all other statistical models, the form of the forecasting model (Eqn. 4.5) is likely to be the most accurate for the immediate forecast and may need revision as more and more time elapses. Thus, it is helpful to study the changes in forecasting accuracies as the same form of the model is used to fit and forecast for extended duration of time. Since the data was not available beyond 5th June into the future, alternate models with the same form (Eqn 4.5) were fitted to backforecast for the months of May and April 2008 and the forecasting accuracy was computed. For instance, to forecast the 16 on-peak hours of 22nd May 2008 (Fig 4.27), the model was fitted using historical data for each week's Thursday from Jan 1, 2006 through the previous Thursday, May 15th 2008.

Similarly, more models were identified and the forecasts were generated for 8th May, 24th April and 10th April (at 15 days interval). Various measures to computed to gauge the accuracy of forecasting. Table 4-27 presents the results.

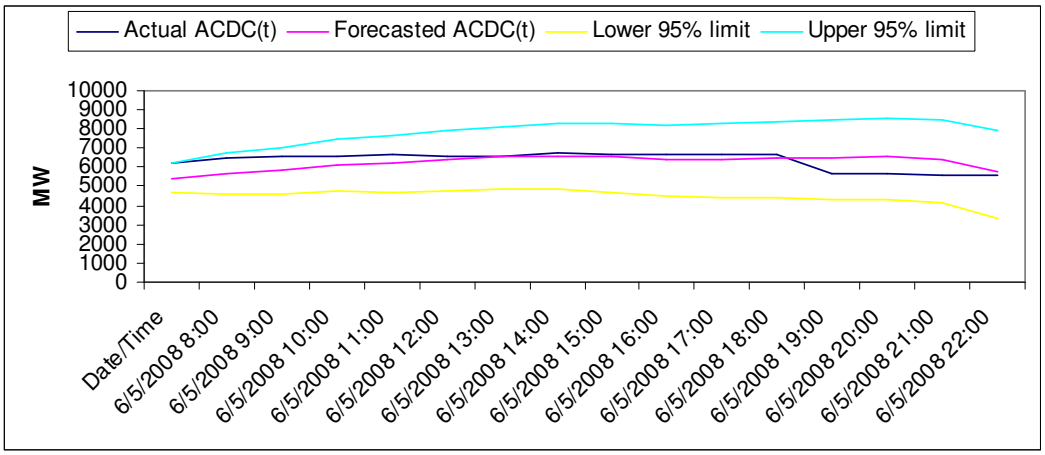

Figure 4-26: Actual and Forecasted ACDC(t), 5th June 2008

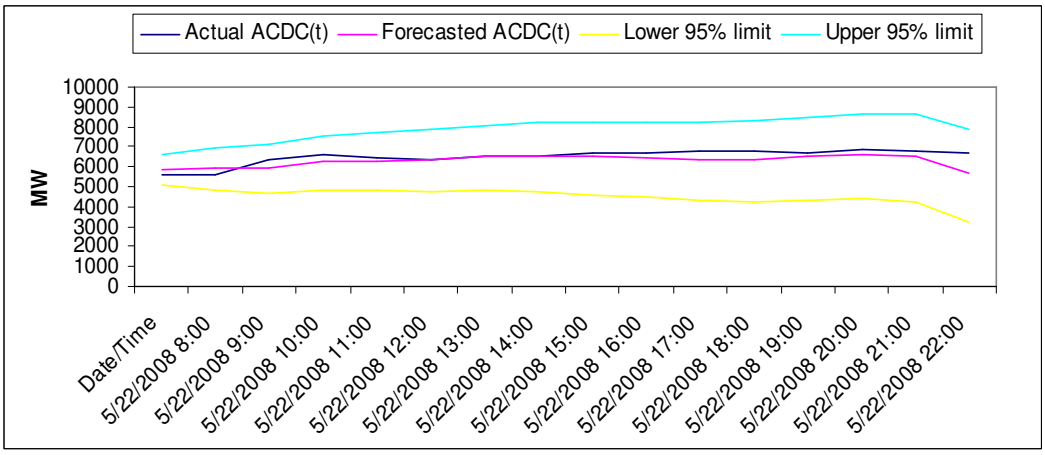

Figure 4-27: Actual and Forecasted ACDC(t), 22nd May 2008

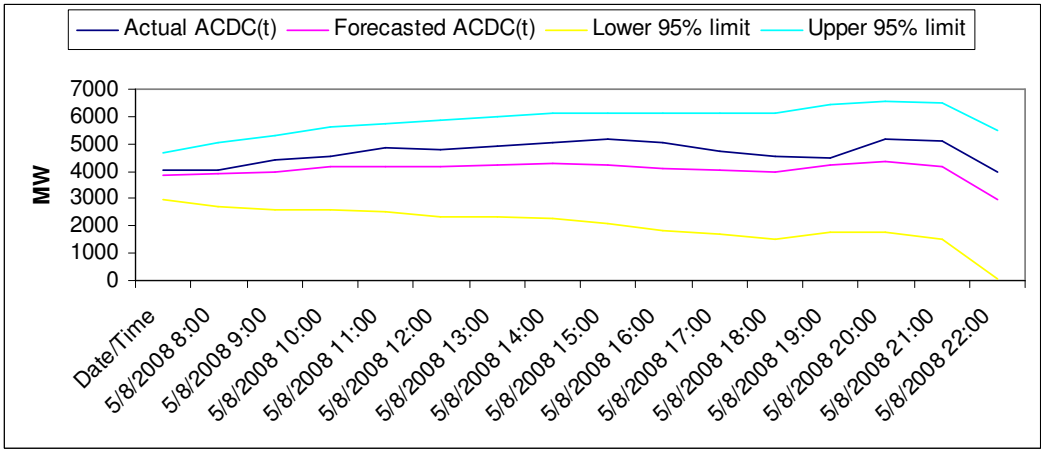

Figure 4-28: Actual and Forecasted ACDC(t), 8th May 2008

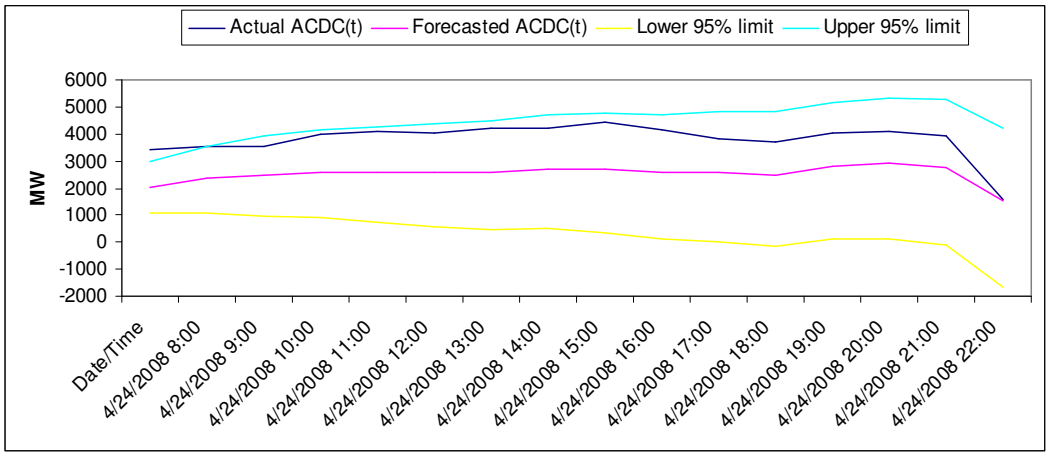

Figure 4-29: Actual and Forecasted ACDC(t), 24th April 2008

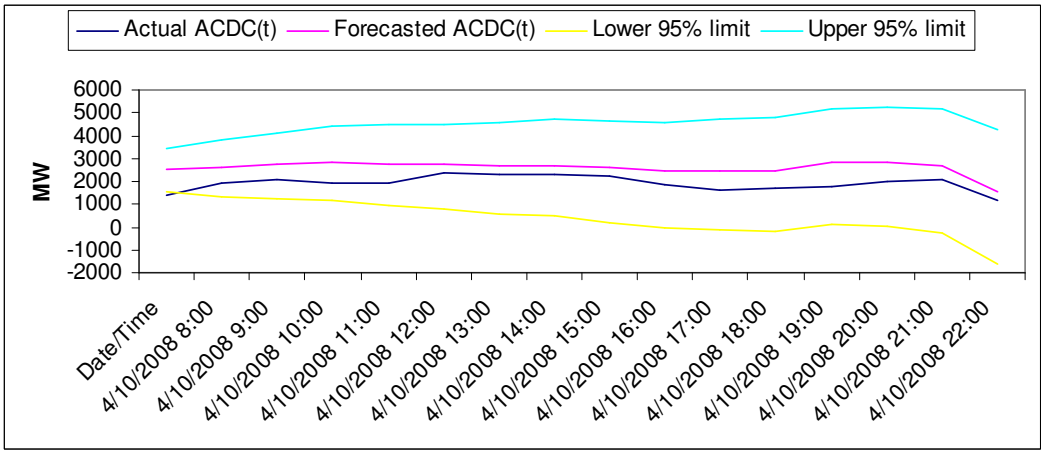

Figure 4-30: Actual and Forecasted ACDC(t), 10th April 2008

Table 4-27: Forecasting accuracy measures

Criterion/Day	5th June	22nd May	8th May	24th April	10th April
ME(forecast)	85.386	201.201	653.755	1288.655	-679.897
RMSE(forecast)	530.467	363.959	714.073	1340.334	720.399
MAD(forecast)	430.134	285.990	653.755	1288.655	679.897
MAPE(forecast)	7.026	4.399	13.696	32.637	37.360

4.7 Conclusions

The fitted ARIMA transfer function resulted in a large adjusted R-square (0.960) and low MAPE (14.964) as shown in Table 4-26. From Table 4.27 and the forecasting plots (Figs. 4.27 – 4.30), it can be deduced that the forecasting accuracy decreased as more time elapsed from the period of immediate forecast. In particular, the Mean Absolute Percentage Error (MAPE) increased from 7.02% to 37.3% as the same model form was used to forecast for 5th June and 10th April, 2008, respectively. However, the decrease in forecasting accuracy was not found to be rapid and several other factors, such as the quality of input series, was also found to affect the final forecasting accuracy. Thus, we concluded that the same transfer function form may be used for forecasting the series for up to 2 months period with reasonable accuracies; after which the transfer function form may need revision depending on the level of accuracy sought. The parameters of the transfer function model may need revision more often than every 2 months to improve the accuracy further.

Chapter 5 – Forecasting the Pacific Northwest Exports to California -An Alternate Model

The forecasting model built in chapter 4 was based on one approximation: It assumed a non-significant correlation between the three input series. In this chapter, this assumption is removed and an alternate model is identified by combining the three inputs into one single composite input. The results are then compared with the earlier model and conclusions are presented (Section 5.5). The SAS programs can be found in Appendix A.

5.1 Alternate model for forecasting the Exports to California

Specifically, before the three inputs series (Section 4.2, Chapter 4) are used in identifying the model, they are combined together to create one composite input series, denoted by X(t) in this chapter. The rest of the procedure is same as in section 4.3. It is briefly described below and the results are presented.

5.1.1 Transformation of the input and output series

The three input series, viz., BPA(t), BC(t) and CA(t) and the output series ACDC(t) are first converted to a standardized metric by carrying out the following transformation: $\left(\frac{x_t - \mu}{\sigma}\right)$,

where x_t = observation at time t, μ = mean of the series, σ = standard deviation of the series (standardization is required for comparing results with the model of chapter 4).The three standardized input series are then added together to form the X(t) input series.

For illustration purpose, the plot of the X(t) series is shown in Fig. 5.1 for the months of Jan- May 2008.

In order to make the input and the output series stationary in variance, a positive constant, 10, was added to observations of the series and Box-Cox transformation was applied. Table 5.3 provides the optimal coefficients that were obtained for the set of series. Tables 5.1 and 5.2 provides the results of related computations over a range of values of λ.

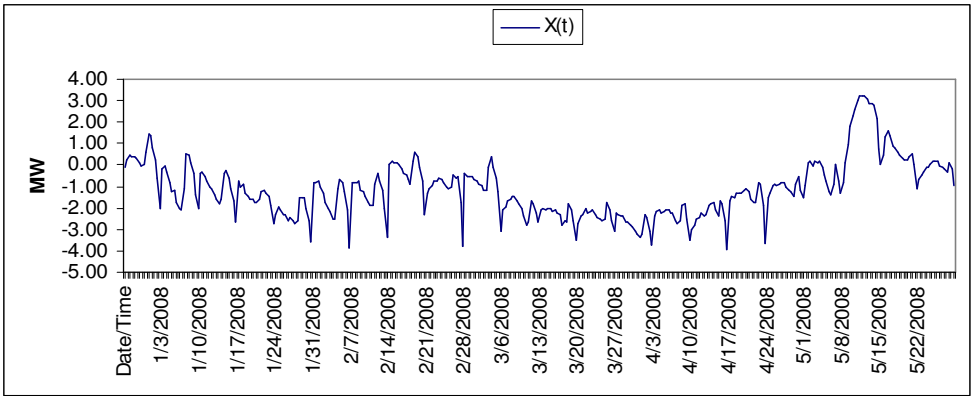

Figure 5-1: Plot of X(t), Jan-May 2008

Table 5-1: Optimal transformation for ACDC(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
2.50	-731.78	0.14	1475.56	1509.12
2.45	-730.63	0.14	1473.26	1506.82
2.39	-729.59	0.13	1471.17	1504.73
2.34	-728.65	0.13	1469.30	1502.85
2.29	-727.82	0.13	1467.64	1501.20
2.24	-727.10	0.13	1466.21	1499.76
2.18	-726.49	0.13	1464.99	1498.54
2.13	-726.00	0.13	1463.99	1497.55
2.08	-725.61	0.13	1463.22	1496.78
2.03	-725.34	0.13	1462.68	1496.23
1.97	-725.18	0.13	1462.36	1495.91
1.92	**-725.13**	**0.13**	**1462.27**	**1495.82**
1.87	-725.21	0.13	1462.41	1495.97
1.82	-725.39	0.13	1462.79	1496.34
1.76	-725.70	0.13	1463.39	1496.95
1.71	-726.12	0.13	1464.24	1497.79
1.66	-726.66	0.13	1465.32	1498.88
1.61	-727.32	0.13	1466.64	1500.20
1.55	-728.10	0.13	1468.21	1501.76
1.50	-729.01	0.13	1470.02	1503.57

Table 5-2: Optimal transformation for X(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
2.50	-1658.33	11.01	3328.66	3362.21
2.45	-1652.70	0.68	3317.41	3350.96
2.39	-1647.52	0.48	3307.04	3340.59
2.34	-1642.78	0.43	3297.56	3331.11
2.29	-1638.48	0.41	3288.97	3322.52
2.24	-1634.64	198957472.04	3281.28	3314.83
2.18	-1631.24	1.97	3274.49	3308.04
2.13	-1628.30	0.59	3268.61	3302.16
2.08	-1625.82	0.44	3263.64	3297.20
2.03	-1623.80	0.39	3259.60	3293.15
1.97	-1622.24	0.37	3256.47	3290.03
1.92	-1621.14	0.36	3254.28	3287.83
1.87	-1620.51	0.35	3253.01	3286.57
1.82	**-1620.34**	**0.34**	**3252.69**	**3286.24**
1.76	-1620.65	0.34	3253.31	3286.86
1.71	-1621.44	0.33	3254.87	3288.42
1.66	-1622.69	0.33	3257.39	3290.94
1.61	-1624.43	0.33	3260.86	3294.41
1.55	-1626.64	0.33	3265.29	3298.84
1.50	-1629.34	0.32	3270.68	3304.23

Table 5-3: Optimal transformation coefficients for X(t) and ACDC(t)

Series	Series name	Box-Cox coefficient (precise value)
Input	X(t)	1.81579
Output	ACDC(t)	1.92105

The original series was then transformed using the optimal Box-Cox transformation coefficients (Eqn 3.7, Chapter 3) to correct the non-stationarity in variance. Next, the autocorrelation and partial autocorrelation plots (Section 3.3, Chapter 3) of the set of series were analyzed to check for mean-stationarity. These are presented in Figs. 5.2 and 5.3.

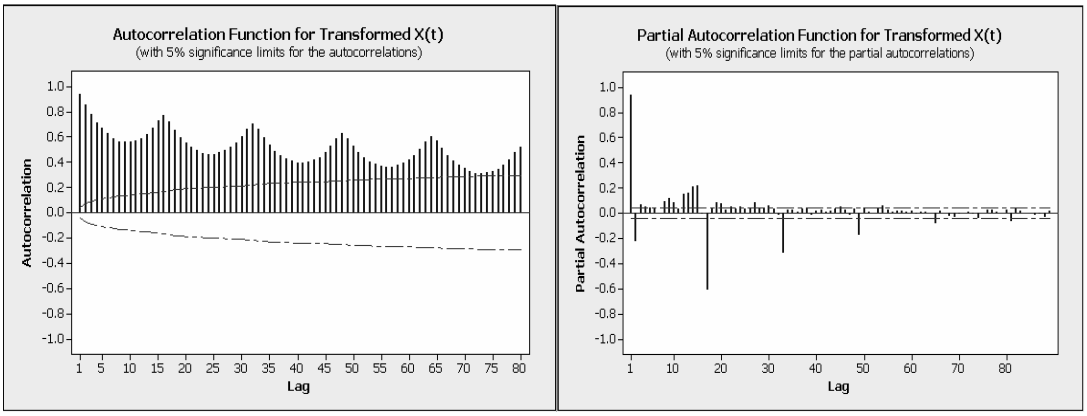

Figure 5-2: ACF and PACF plots of transformed X(t)

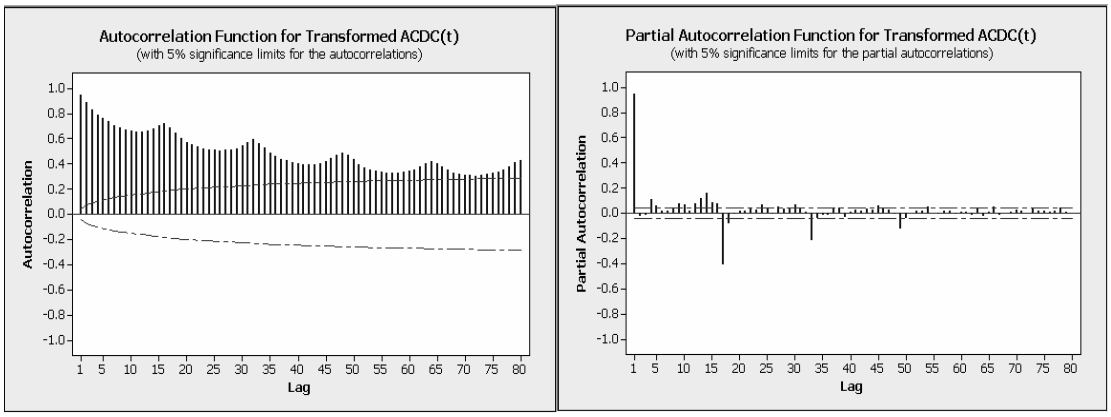

Figure 5-3: ACF and PACF plots of transformed ACDC(t)

Table 5-4 presents the orders of non-seasonal and seasonal differencing that were obtained for the set of series to render them mean-stationary:

The ACF and PACF plots for the differenced series are shown in Figs 5.4 and 5.5. The resultant series is stationary in mean.

Table 5-4: Optimal Differencing orders for ACDC(t) and X(t)

Type	**Series name**	**Differencing order (d= nonseasonal order; D= seasonal order)**
Input	X(t)	d=1, D=16
Output	ACDC(t)	d=1, D=16

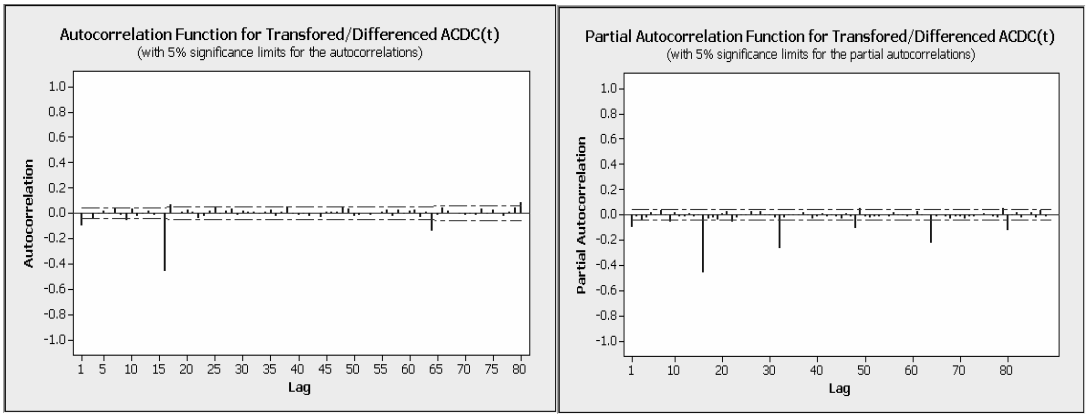

Figure 5-4: ACF and PACF plots of transformed/differenced ACDC(t)

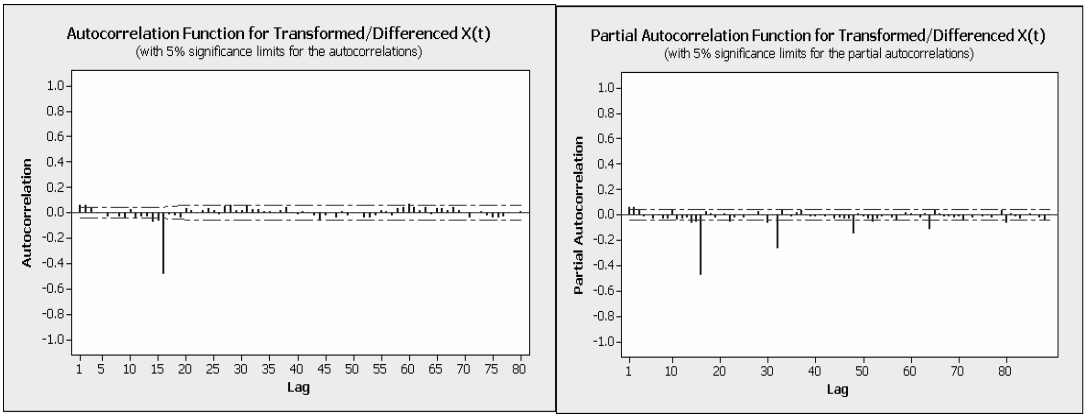

Figure 5-5: ACF and PACF plots of transformed/differenced X(t)

5.1.2 Prewhitening the input series

The (transformed and differenced) input series was prewhitened in this step. The three-stage analysis procedure of ARMA model fitting, viz., the identification , estimation and diagnostic checking (Section 3.3, Chapter 3), was carried for the X(t) series and suitable ARMA filter was built. Two ARMA model forms were selected after final analysis; the comparison chart is shown in Table 5.5.

Table 5-5: Prewhitening filter forms for the set of input series

No.	Prewhitening filter form	AIC	SBC	Variance estimate	Std error estimate
1	P=(16) Q=(1 6 7 8 9 14 19)	9874.09	9918.83	8.4779	2.9116
2	**P=0** **Q=(4 5 6 7 8 9 11 12 14 16) (16)**	**9851.02**	**9912.53**	**8.3672**	**2.8926**

The second filter was chosen because it has lower AIC and SBC values. The parameter estimates for the second filter are provided in Table 5.6.

The autocorrelation and partial autocorrelation plots of the residual series and the tables of related computations are shown next:

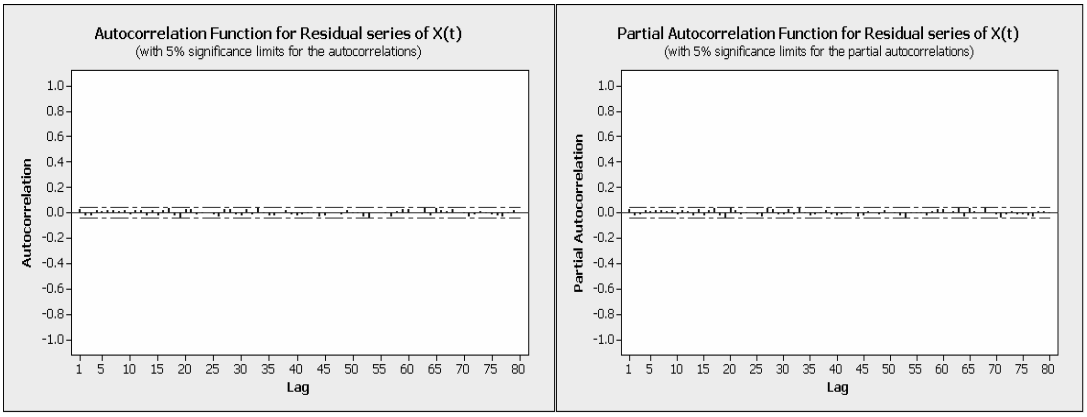

Figure 5-6: ACF and PACF plots of residual series of X(t)

Table 5-6: Parameter estimates of X(t) series

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.06286	0.02185	2.88	0.0041	4
MA1,2	0.04729	0.02196	2.15	0.0314	5
MA1,3	0.08264	0.02164	3.82	0.0001	6
MA1,4	0.09641	0.02185	4.41	<.0001	7
MA1,5	0.10027	0.02178	4.60	<.0001	8
MA1,6	0.10909	0.02167	5.03	<.0001	9
MA1,7	0.06980	0.02183	3.20	0.0014	11
MA1,8	0.05159	0.02201	2.34	0.0192	12
MA1,9	0.11004	0.02218	4.96	<.0001	14
MA1,10	0.10640	0.03925	2.71	0.0068	16
MA2,1	0.58583	0.03327	17.61	<.0001	16

Table 5-7: Correlation matrix of Parameter estimates of X(t) series

Correlations of Parameter Estimates											
Parameter	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5	MA1,6	MA1,7	MA1,8	MA1,9	MA1,10	MA2,1
MA1,1	1.000	-0.043	-0.032	-0.010	-0.104	-0.105	-0.147	-0.136	-0.002	-0.048	0.008
MA1,2	-0.043	1.000	-0.049	-0.003	-0.001	-0.099	-0.111	-0.139	-0.149	-0.050	-0.002
MA1,3	-0.032	-0.049	1.000	-0.054	-0.030	-0.031	-0.107	-0.118	-0.143	-0.051	0.063
MA1,4	-0.010	-0.003	-0.054	1.000	-0.045	-0.023	-0.105	-0.093	-0.144	-0.127	0.057
MA1,5	-0.104	-0.001	-0.030	-0.045	1.000	-0.053	-0.012	-0.095	-0.113	-0.143	0.081
MA1,6	-0.105	-0.099	-0.031	-0.023	-0.053	1.000	-0.022	-0.005	-0.095	-0.133	0.067
MA1,7	-0.147	-0.111	-0.107	-0.105	-0.012	-0.022	1.000	-0.031	-0.000	-0.100	0.065
MA1,8	-0.136	-0.139	-0.118	-0.093	-0.095	-0.005	-0.031	1.000	0.001	-0.068	0.031
MA1,9	-0.002	-0.149	-0.143	-0.144	-0.113	-0.095	-0.000	0.001	1.000	-0.091	0.111
MA1,10	-0.048	-0.050	-0.051	-0.127	-0.143	-0.133	-0.100	-0.068	-0.091	1.000	-0.826
MA2,1	0.008	-0.002	0.063	0.057	0.081	0.067	0.065	0.031	0.111	-0.826	1.000

Table 5-8: Ljung-Box Chi-square test results for X(t) series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.026	-0.025	-0.021	0.014	0.010	0.013
12	7.21	1	0.0072	0.016	0.011	0.014	-0.019	0.016	0.014
18	14.07	7	0.0500	-0.024	0.020	-0.022	0.017	0.033	-0.025
24	21.39	13	0.0656	-0.043	0.028	0.024	-0.017	-0.010	0.001
30	28.73	19	0.0704	-0.014	-0.033	0.025	0.029	-0.015	-0.026
36	35.52	25	0.0792	0.025	-0.014	0.038	-0.006	-0.024	-0.022
42	39.16	31	0.1492	-0.001	0.019	-0.019	-0.025	-0.020	-0.008
48	42.81	37	0.2358	-0.006	-0.032	-0.024	-0.003	0.001	-0.013
54	49.57	43	0.2278	0.016	-0.005	-0.008	-0.033	-0.044	-0.002
60	55.37	49	0.2470	-0.003	-0.011	-0.035	0.012	0.022	0.029
66	62.88	55	0.2173	-0.000	0.001	0.040	-0.028	0.032	0.015
72	67.41	61	0.2674	0.006	0.028	-0.002	-0.009	-0.033	-0.015
78	71.38	67	0.3345	0.004	-0.005	-0.019	-0.023	-0.031	-0.005

5.1.3 "Prewhitening" the output series and computing Cross-Correlations between the prewhitened input and output series

In this step, the ARMA filter, which was fitted to the input series in Section 5.1.2, was applied to the output (ACDC(t)) and the residual series was stored.

The cross-correlation coefficients between the prewhitened input and the "prewhitened" output series were computed next.

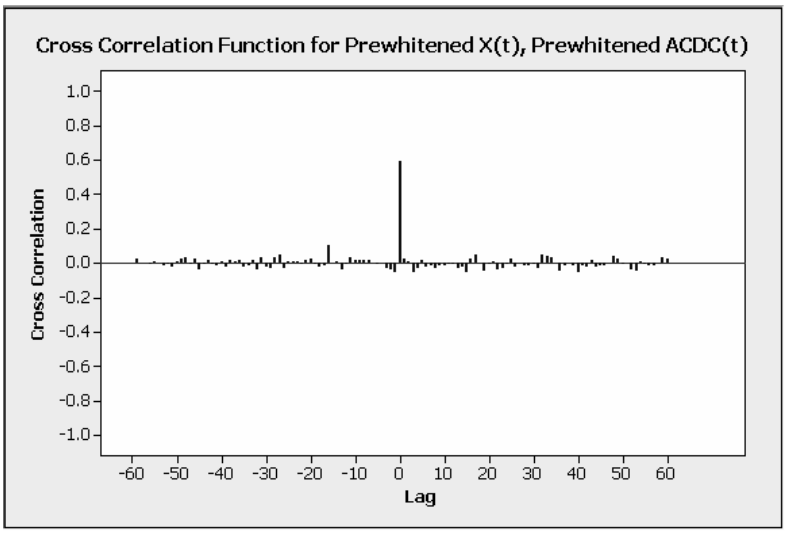

Figure 5-7: Cross-correlation plot for prewhitened X(t), "Prewhitened" ACDC(t)

Table 5-9: Cross-correlation coefficients for prewhitened X(t), "Prewhitened" ACDC(t)

Lag	CCF	Lag	CCF
-16	0.104004	**0**	**0.593233**
-15	0.002086	1	0.020136
-14	0.010159	2	0.011545
-13	-0.037	3	-0.05772
-12	-0.01037	4	-0.03129
-11	0.035355	5	0.014328
-10	0.017299	6	-0.02076
-9	0.016163	7	-0.01274
-8	0.012607	8	-0.03284
-7	0.01446	9	-0.01412
-6	0.002689	10	-0.01909
-5	-0.00717	11	-0.00874
-4	-0.00072	12	-0.00523
-3	-0.02847	13	-0.03379
-2	-0.03739	14	-0.02608
-1	-0.05227	15	-0.05717
		16	0.020646

5.1.4 Identifying the input transfer function of the model

From Fig. 5.7 and Table 5.9, it is clear that the relationship between the X(t) and ACDC(t) set of series is first-order linear.

The final estimated coefficient (Section 3.3, chapter 3) is presented in Table 5.10.

Table 5-10: Input Transfer function coefficient for X(t) series

Series	Series name	Coefficients
Input	X(t)	$\omega_0 = 0.4663$

5.1.5 Extraction of Preliminary Noise Series and fitting an ARIMA model

The Preliminary Noise Series, n_t, was extracted next (Equation 3.21,Chapter-3). After estimating the noise series in the above step, the n_t series was analyzed using the three-step Box-Jenkins methodology (Section 3.3, chapter 3) and an appropriate ARMA $(p_n, 0, q_n)$ model was fitted.

The ACF and PACF plots of the preliminary noise series are shown in Fig 5-8.

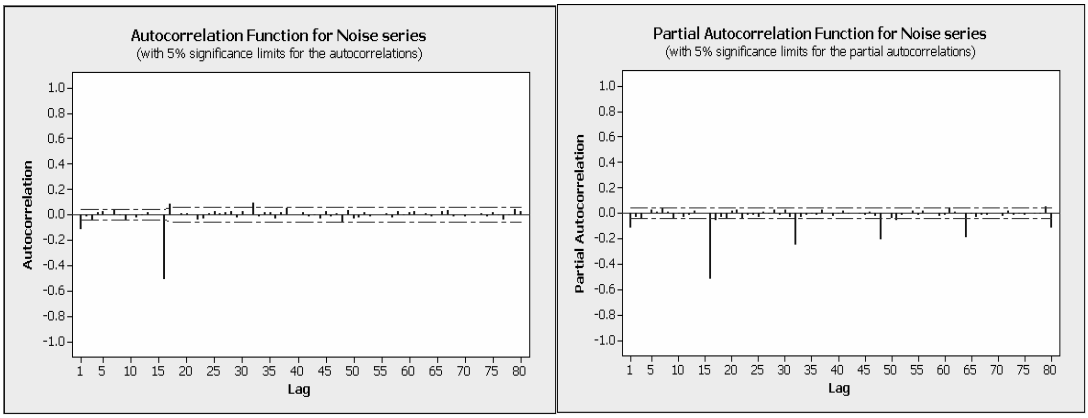

Figure 5-8: ACF and PACF plots for Preliminary Noise Series

The following ARMA model form was selected for the noise series (Table 5.14):

$$\left(1-\theta_1 B^1-\theta_2 B^2-\theta_3 B^3-\theta_9 B^9-\theta_{13} B^{13}-\theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16}-\theta_{48} B^{48}\right) \qquad (5.1)$$

or, in compact form:

P=(0) Q=(1 2 3 9 13 15) (16 48)

The parameter estimates are provided in Table 5.11.

Table 5-11: Parameter estimates of Preliminary Noise Series

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.19476	0.02228	8.74	<.0001	1	ACDC	0
MA1,2	0.05170	0.02261	2.29	0.0224	2	ACDC	0
MA1,3	0.10066	0.02207	4.56	<.0001	3	ACDC	0
MA1,4	0.08702	0.02159	4.03	<.0001	9	ACDC	0
MA1,5	0.07502	0.02176	3.45	0.0006	13	ACDC	0
MA1,6	0.11605	0.02205	5.26	<.0001	15	ACDC	0
MA2,1	0.80889	0.01795	45.07	<.0001	16	ACDC	0
MA2,2	0.11902	0.01789	6.65	<.0001	48	ACDC	0
NUM1	0.46631	0.01236	37.73	<.0001	0	Xt	0

Table 5-12: Correlation matrix of Parameter estimates of Preliminary Noise Series

Correlations of Parameter Estimates									
Variable Parameter	ACDC MA1,1	ACDC MA1,2	ACDC MA1,3	ACDC MA1,4	ACDC MA1,5	ACDC MA1,6	ACDC MA2,1	ACDC MA2,2	Xt NUM1
ACDC MA1,1	1.000	-0.193	-0.075	-0.010	-0.042	0.002	-0.081	0.058	-0.122
ACDC MA1,2	-0.193	1.000	-0.200	0.016	0.008	-0.102	0.013	-0.003	-0.082
ACDC MA1,3	-0.075	-0.200	1.000	-0.021	-0.015	-0.017	-0.043	0.024	-0.028
ACDC MA1,4	-0.010	0.016	-0.021	1.000	-0.077	-0.012	0.060	-0.050	0.089
ACDC MA1,5	-0.042	0.008	-0.015	-0.077	1.000	-0.117	-0.068	0.078	0.012
ACDC MA1,6	0.002	-0.102	-0.017	-0.012	-0.117	1.000	-0.152	0.105	0.035
ACDC MA2,1	-0.081	0.013	-0.043	0.060	-0.068	-0.152	1.000	-0.853	0.033
ACDC MA2,2	0.058	-0.003	0.024	-0.050	0.078	0.105	-0.853	1.000	-0.018
Xt NUM1	-0.122	-0.082	-0.028	0.089	0.012	0.035	0.033	-0.018	1.000

5.1.6 Identifying the final forecasting model

Thus, the final forecasting model becomes:

$$(1-B)\left(1-B^{16}\right) ACDC^{bxcx}(t) = \omega_0 (1-B)\left(1-B^{16}\right) X^{bxcx}(t) + \left(1-\theta_1 B^1 - \theta_2 B^2 - \theta_3 B^3 - \theta_9 B^9 - \theta_{13} B^{13} - \theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16} - \theta_{48} B^{48}\right) a_t \quad (5.2)$$

where $ACDC^{bxcx}(t)$ denotes the Box-Cox transformation of standardized ACDC(t) series, and so on. The coefficients of the input transfer functions are provided in Table 5.10 and the parameters of the transfer function for noise series are provided in Table 5.11.

5.2 Diagnostic checking of the model

The following statistical tests were implemented to check the validity of model:

5.2.1 Analysis of the final residuals

Presented in Fig 5-9 are the Autocorrelation and Partial Autocorrelation plots of the final noise series, a_t:

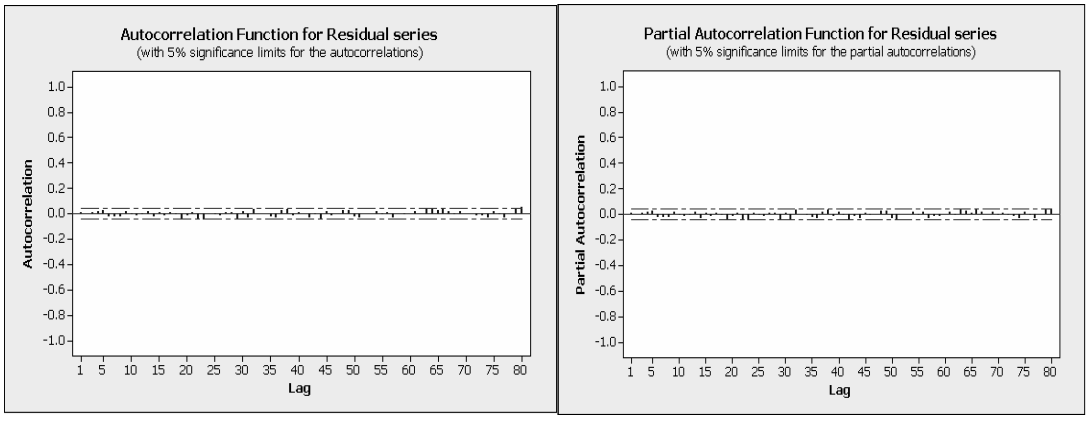

Figure 5-9: ACF and PACF plots for Final Noise Series

The absence of statistically significant spikes at all the lags beyond lag 0 in the autocorrelation and partial autocorrelation plots signify the absence of any defined pattern in the final noise series. Thus, it was concluded that the final noise series is a realization of white noise, confirming the validity of the model.

5.2.2 Box-Pierce chi-square statistical test

Table 5.13 provides the Ljung-Box Chi-Square test results for the final noise series, which confirms that the final noise series is a realization of white noise.

Table 5-13: Ljung-Box Chi-square test results for Final Noise Series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.009	0.002	0.009	0.016	0.022	-0.026
12	6.77	4	0.1487	-0.029	-0.021	0.013	-0.011	-0.015	-0.007
18	10.05	10	0.4364	0.017	-0.027	0.008	-0.019	0.011	-0.007
24	22.25	16	0.1353	-0.043	-0.013	0.012	-0.038	-0.050	0.004
30	26.15	22	0.2451	-0.008	-0.012	0.006	0.007	-0.038	0.015
36	34.44	28	0.1867	-0.036	0.032	0.004	-0.004	-0.026	-0.033
42	41.52	34	0.1756	0.024	0.036	-0.015	0.009	0.000	-0.036
48	47.15	40	0.2034	-0.010	-0.041	0.015	-0.014	-0.002	0.024
54	52.62	46	0.2333	0.025	-0.021	-0.036	-0.001	0.004	0.018
60	55.56	52	0.3421	-0.006	0.011	-0.034	-0.003	-0.012	-0.001
66	64.60	58	0.2570	0.018	-0.002	0.032	0.033	0.021	0.039
72	66.24	64	0.3996	0.016	-0.004	0.017	-0.002	0.003	-0.015
78	72.56	70	0.3935	-0.015	-0.034	0.016	-0.011	-0.036	-0.003

5.2.3 AIC and BIC tests

Two competing models were selected for the preliminary noise series after final analysis. As mentioned in section 5.1.7, model 2 was chosen as the final model for forecasting because of lower AIC and SBC values (Table 5.14).

Table 5-14: AIC and SBC values for competing models

No.	ARMA model	AIC	SBC	Variance estimate	Std error estimate
1	P=(1 2 3 9 15 16 23 32) Q=(16)	8031.08	8087	3.343	1.828
2	**P=0 Q=(1 2 3 9 13 15) (16 48)**	**8022.79**	**8073.12**	**3.33**	**1.825**

5.3 Results of fitting

The fitting plots for Jan 2006 through May 2008 are shown in Figs 5.10 – 5.14.

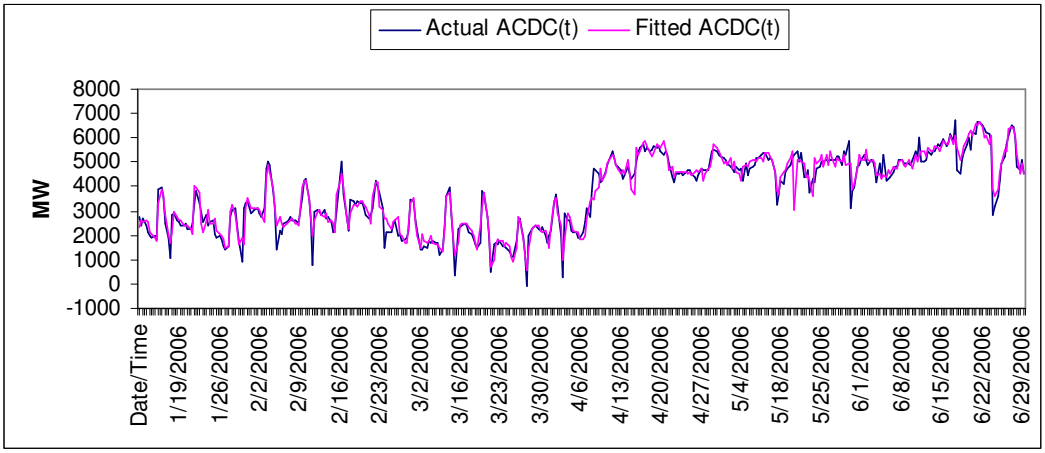

Figure 5-10: Actual and Fitted ACDC(t), Jan -June 2006

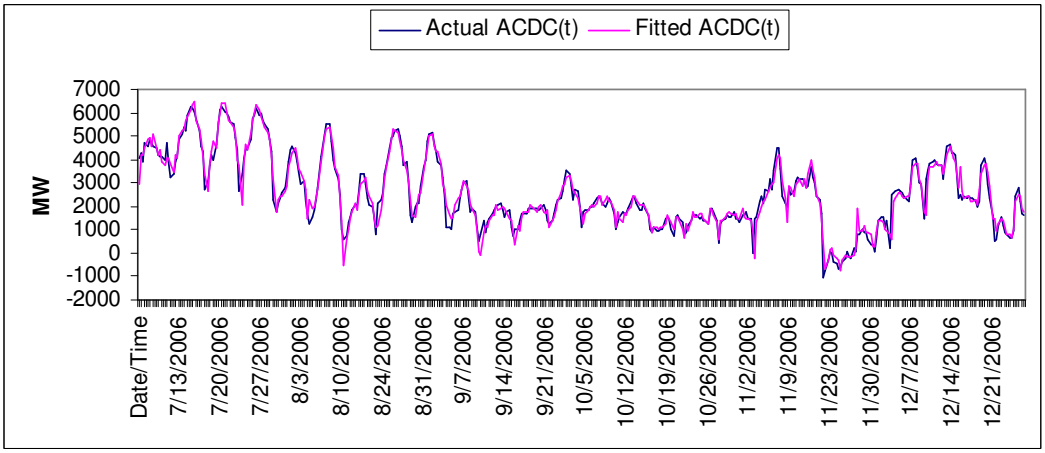

Figure 5-11: Actual and Fitted ACDC(t), July -Dec 2006

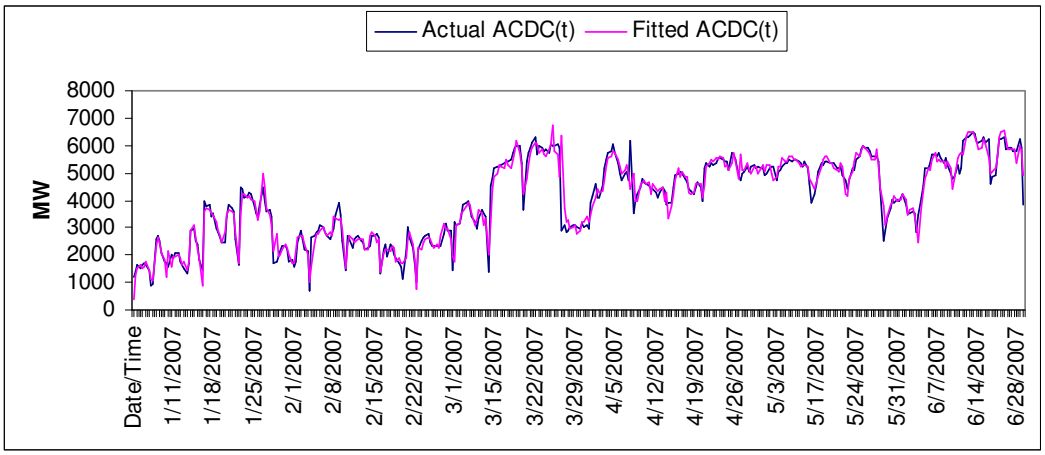

Figure 5-12: Actual and Fitted ACDC(t), Jan -June 2007

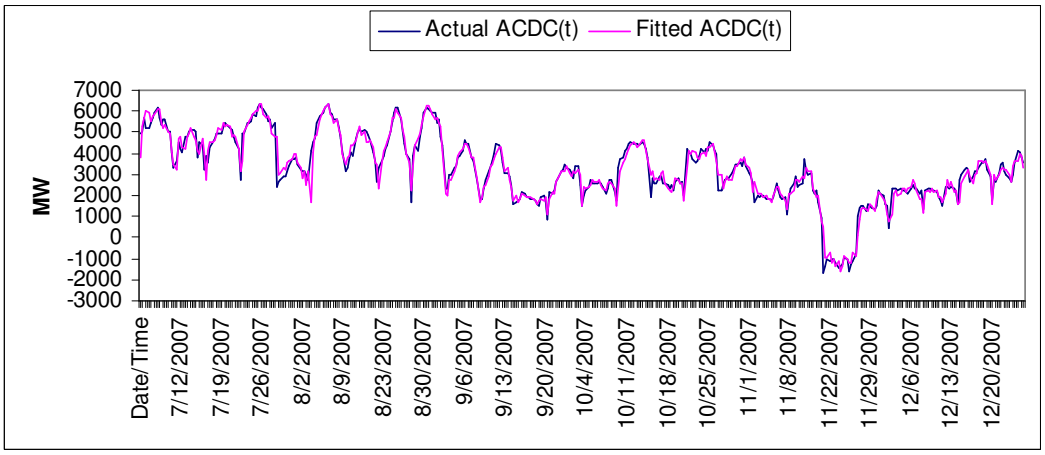

Figure 5-13: Actual and Fitted ACDC(t), July -Dec 2007

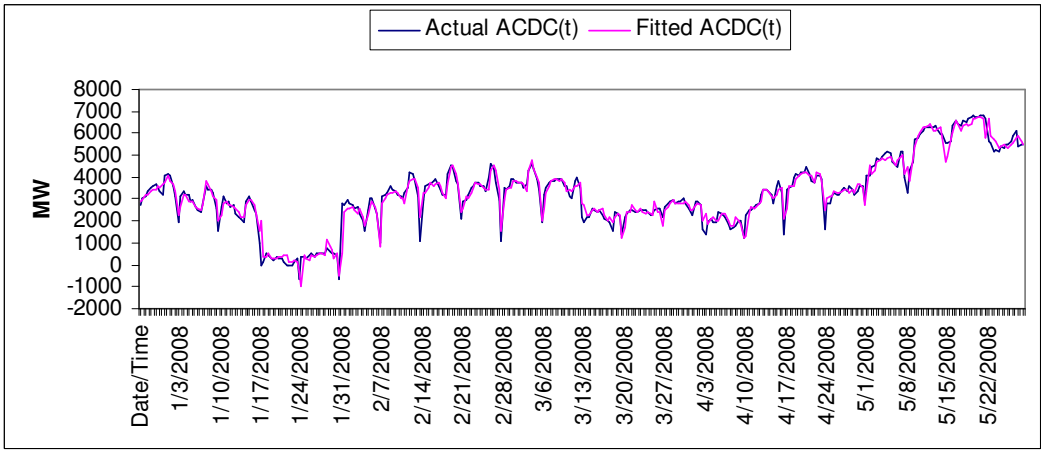

Figure 5-14: Actual and Fitted ACDC(t), Jan -May 2008

Several fitting accuracy measures were computed and the results are presented in Table 5-15.

Table 5-15: Fitting accuracy measures

Criterion	Jan 2006-May2008
R-Sqr	0.952
R-Sqr Adj	0.951
ME	-7.300
RMSE	353.921
MAD	231.270
MAPE	17.356

5.4 Results of forecasting

The final model (Eqn 5.2) was used to forecast for the 16 on-peak hours of subsequent Thursday (5th June, 2008). Also, alternate models with the same form (Eqn 5.2) were fitted to forecast for the months of May and April 2008 (Refer section 4.6, Chapter 4).

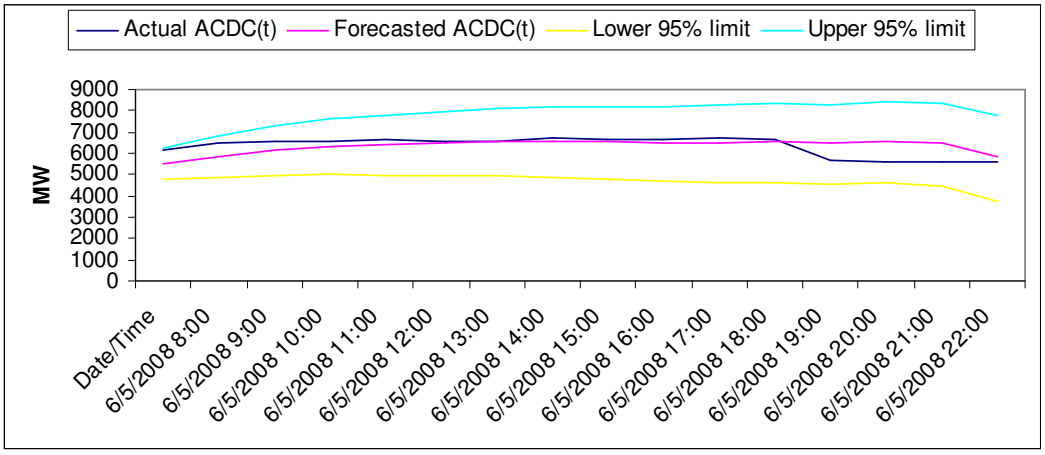

Figure 5-15: Actual and Forecasted ACDC(t), 5th June, 2008

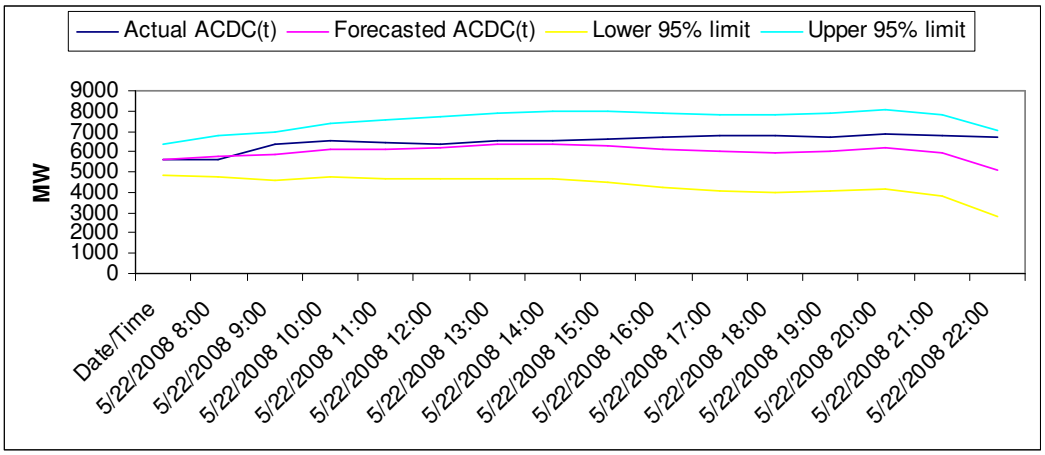

Figure 5-16: Actual and Forecasted ACDC(t), 22nd may, 2008

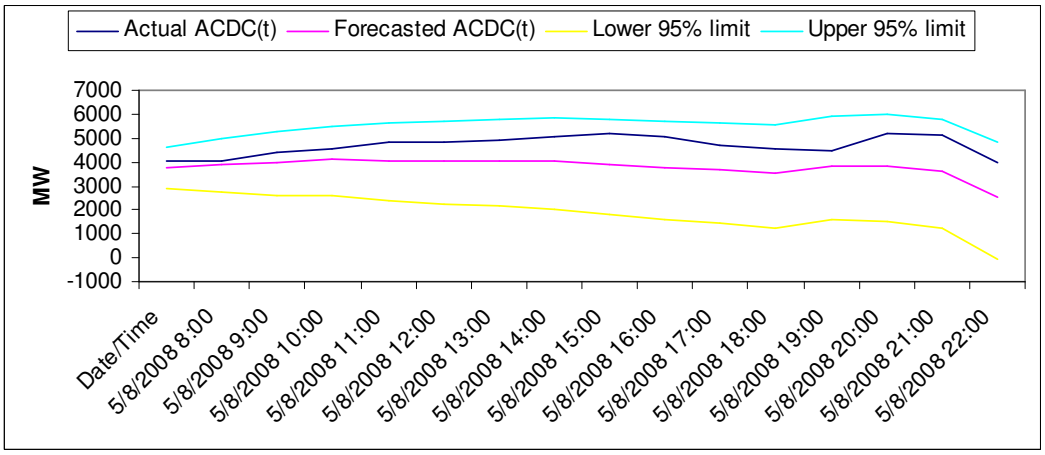

Figure 5-17: Actual and Forecasted ACDC(t), 8th May, 2008

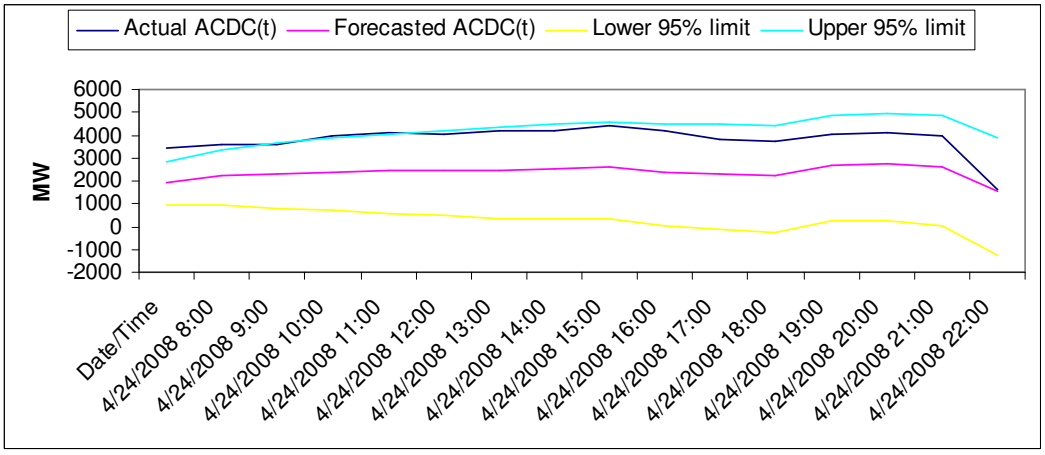

Figure 5-18: Actual and Forecasted ACDC(t), 24th April, 2008

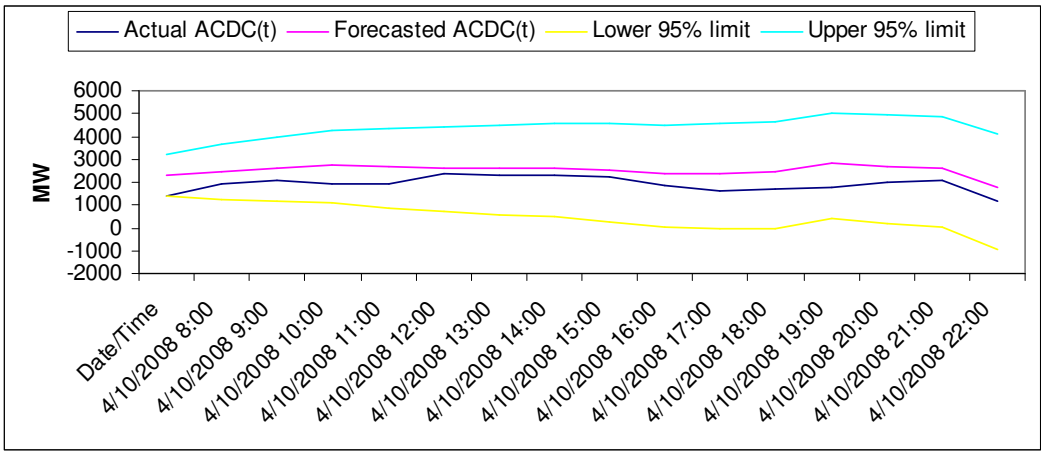

Figure 5-19: Actual and Forecasted ACDC(t), 10th April, 2008

The forecasting accuracy was measured by various statistical measures. Table 5.16 provides the results.

Table 5-16: Forecasting accuracy measures

Day/Criterion	5th June	22nd May	8th May	24th April	10th April
ME(forecast)	-4.018	489.944	1800.813	1446.505	-612.917
RMSE(forecast)	481.470	642.841	1832.468	1499.749	650.844
MAD(forecast)	361.096	514.871	1800.813	1446.505	612.917
MAPE(forecast)	5.999	7.743	27.698	36.702	34.149

5.5 Conclusions

It is helpful to compare the performance of the model developed in chapter 4 and the alternate model developed in this chapter (Tables 5.17 and 5.18). It was found that the alternate model developed produced slightly better results than the earlier model, in the case of immediate forecast period, 5th June 2008. For forecasts in other periods, however, this was not generally true and the performance of both the models was either comparable or the earlier model worked slightly superior (Table 5.18). This may be due to the fact

that some information is lost when the three input series are combined into one composite series. Thus, it can be concluded that the assumption of insignificant correlation between the three input series was a better trade-off than combining the three inputs into one and using it as the composite input to the forecasting model of Exports to California.

Table 5-17: Fitting accuracy comparison chart

Criterion	Model developed in this chapter	Model developed in chapter 4
R-Sqr	0.952	0.961
R-Sqr Adj	0.951	0.960
ME	-7.300	-2.788
RMSE	353.921	321.086
MAD	231.270	211.218
MAPE	17.356	14.964

Table 5-18: Forecasting accuracy comparison chart

	Model developed in this chapter				Model developed in chapter 4			
Day/Criterion	**ME**	**RMSE**	**MAD**	**MAPE**	**ME**	**RMSE**	**MAD**	**MAPE**
5th June	-4.018	481.47	361.096	5.999	85.386	530.467	430.134	7.026
22nd May	489.944	642.841	514.871	7.743	201.201	363.959	285.99	4.399
8th May	1800.813	1832.468	1800.813	27.698	653.755	714.073	653.755	13.696
24th April	1446.505	1499.749	1446.505	36.702	1288.655	1340.334	1288.655	32.637
10th April	-612.917	650.844	612.917	34.149	-679.897	720.399	679.897	37.36

Chapter 6 – Forecasting the Power Exports to California ISO market

6.1 Definition of problem

In Section 2.2 of Chapter 2, it was mentioned that the California ISO (CAISO) market has contracts to share only 3233 MW out of the 4800 MW north-to-south import capacity of the California-Oregon Intertie (COI), also called Path 66. Specifically, the Path 66 consists of three 500 kV AC transmission lines which enter the California Borders from the North side. Out of its 4800 MW North-to-South transfer capacity, 1,600 MW is in the COTP, also known as the Third AC line (Section 2.2, Chapter 2). The CAISO has contracts to share only 33 MW out of 1,600 MW of COTP transmission capacity. Thus, the total import capacity contracts available to CAISO market via path 66 are approximately 3233 MW. Interest in forecasting the CAISO portion of path 66 MW exports is of crucial importance to traders and other market players who trade in or are associated with CAISO markets.

In chapters 4 and 5, models were built to forecast the total power exports to California via the COI and the PDCI interties. In this chapter, the forecasts from those models are used as inputs to specifically forecast the CAISO contracted portion of path 66 intertie flows (hereafter referred as path66(t) series in this chapter for simplicity since CAISO market refers to it that way). It should be clarified that the path66(t) forecasts thus generated are "actual real-time scheduled flows" and not the "actual metered electrical flows", as was the case with ACDC(t) flows forecasted in earlier chapters. The reason for this is that the CAISO market makes available the actual real-time scheduled

flows and not the actual electrical flows on its market website; consequently market analysts and traders are more interested in forecasting these real- time scheduled flows since they play a direct role in affecting the market prices and other indices.

The inputs to this model are the forecasts of ACDC(t) flows, obtained from the models developed in chapter 4 and 5. The procedure involved in identifying the forecasting model for path66(t) parallels the procedure outlined in chapter 3 for identifying ARIMA transfer function models; the main difference is that the inputs used in these models are forecasts and not actual data. Thus, the accuracy of forecasts produced by these models also depend on the quality of forecasts produced by the ACDC(t) forecasting model. Another difference is the inherent nature of the actual real-time scheduled flow data of path66(t) series. Since it does not represent the actual electrical flows, there is always some level of extraneous behavior present in the data. The forecasting results presented later in the chapter demonstrate that the Box-Jenkins ARIMA modeling was able to capture these trends.

Section 6.2 presents the steps of identification of the model. In section 6.3, the fitted model is validated by relevant statistical diagnostic tests. Sections 6.4 and 6.5 present the results of fitting and forecasting, respectively. Section 6.6 presents the conclusions.

6.2 Identifying the path66(t) forecasting model

The procedure to identify the path66(t) model is similar to the one followed in Section 4.3 of Chapter 4. Only the major steps of model identification and results are presented in brief in the sections that follow.

6.2.1 Transformation of the input and output series

For illustration purposes, the snapshots of ACDC(t) and Path66(t) data is shown in Figures 6-1 and 6-2 for Jan-May 2008.

In order to make the input and the output series stationary in variance, a positive constant, 5000, was added to observations of the original series and several Box-Cox transformation coefficients were analyzed to detect the optimal transformation coefficient (Table 6.3). Box-Cox computational results for ACDC(t) and Path66(t) series are provided in Tables 6.1 and 6.2, respectively.

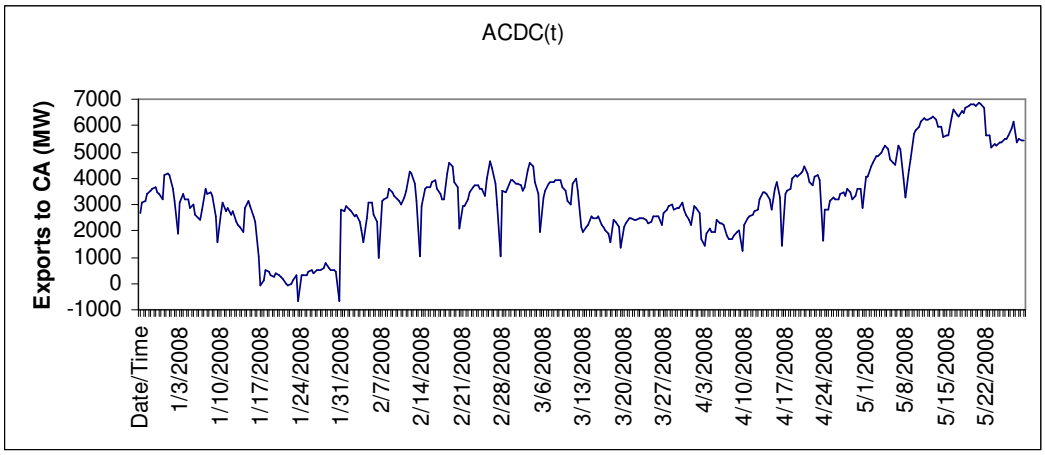

Figure 6-1: Plot of ACDC(t), Jan-May 2008

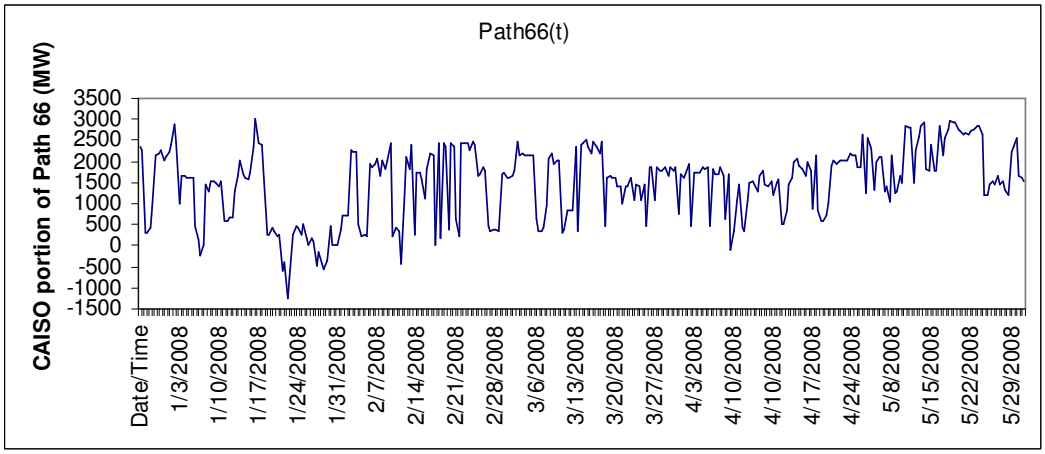

Figure 6-2: Plot of Path66(t), Jan-May 2008

Table 6-1: Optimal transformation for ACDC(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
2.00	-15263.15	402920.59	30538.30	30571.81
1.95	-15259.19	389869.62	30530.37	30563.88
1.89	-15255.60	380810.08	30523.20	30556.70
1.84	-15252.40	373954.70	30516.79	30550.30
1.79	-15249.59	368489.58	30511.18	30544.69
1.74	-15247.19	363983.51	30506.38	30539.88
1.68	-15245.20	429031.89	30502.40	30535.91
1.63	-15243.63	363350.75	30499.26	30532.77
1.58	-15242.49	352500.35	30496.99	30530.49
1.53	-15241.80	347055.90	30495.59	30529.10
1.47	**-15241.55**	**343299.26**	**30495.10**	**30528.60**
1.42	-15241.76	351623.23	30495.52	30529.02
1.37	-15242.44	337625.99	30496.88	30530.38
1.32	-15243.60	332659.06	30499.19	30532.70
1.26	-15245.25	329463.81	30502.49	30536.00
1.21	-15247.39	327040.01	30506.79	30540.29
1.16	-15250.05	325087.99	30512.10	30545.61
1.11	-15253.23	323479.68	30518.46	30551.97
1.05	-15256.94	322147.67	30525.88	30559.39
1.00	-15261.20	321051.50	30534.39	30567.90

Table 6-2: Optimal transformation for Path66(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
3.25	-15477.10	34600936.60	30966.19	30999.70
3.20	-15476.29	7426276.73	30964.59	30998.09
3.14	-15475.58	39564347.30	30963.16	30996.66
3.09	-15474.96	9049930.06	30961.91	30995.42
3.04	-15474.42	76976436.99	30960.84	30994.35
2.99	-15473.98	8808395.64	30959.95	30993.46
2.93	-15473.62	4207153.60	30959.25	30992.75
2.88	-15473.36	1161977.94	30958.72	30992.23
2.83	-15473.19	1165171853.06	30958.38	30991.89
2.78	**-15473.11**	**1848943.01**	**30958.22**	**30991.73**
2.72	-15473.13	473468011.17	30958.25	30991.76
2.67	-15473.23	79315718.44	30958.47	30991.97
2.62	-15473.43	1278850.82	30958.87	30992.37
2.57	-15473.73	679243.38	30959.45	30992.96
2.51	-15474.11	572281.46	30960.23	30993.73
2.46	-15474.60	531760.09	30961.19	30994.70
2.41	-15475.17	510043.38	30962.35	30995.85
2.36	-15475.85	496000.50	30963.69	30997.20
2.30	-15476.62	485865.45	30965.23	30998.74
2.25	-15477.48	478035.84	30966.96	31000.47

The final optimal transformation coefficients are provided in Table 6.3.

Table 6-3: Optimal transformation coefficients

Series	Series name	Box-Cox coefficient (precise value)
Input	ACDC(t)	1.47368
Output	Path66(t)	2.77632

The set of series is then transformed using the optimal Box-Cox transformation coefficients. To remove the non-stationarity present in the mean, suitable non-seasonal and seasonal differencing is carried out.

Figs 6.3 and 6.4 present the autocorrelation and partial autocorrelation plots of the transformed series.

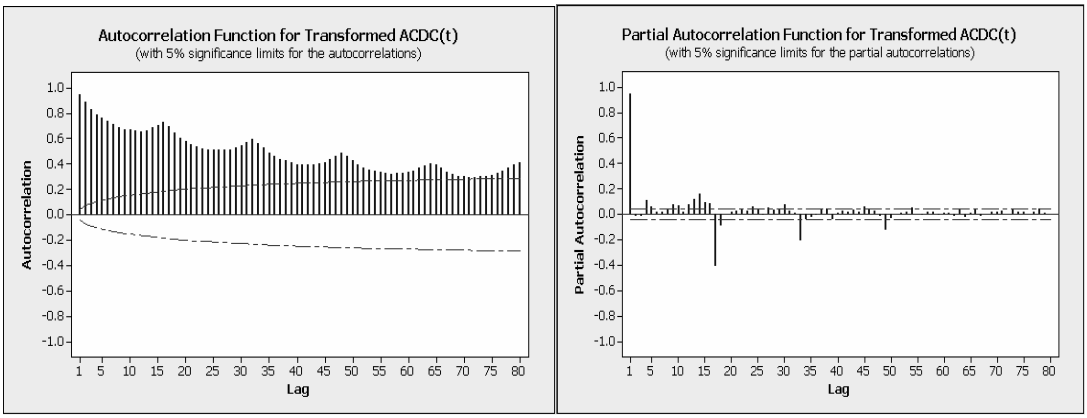

Figure 6-3: ACF and PACF plots of transformed ACDC(t)

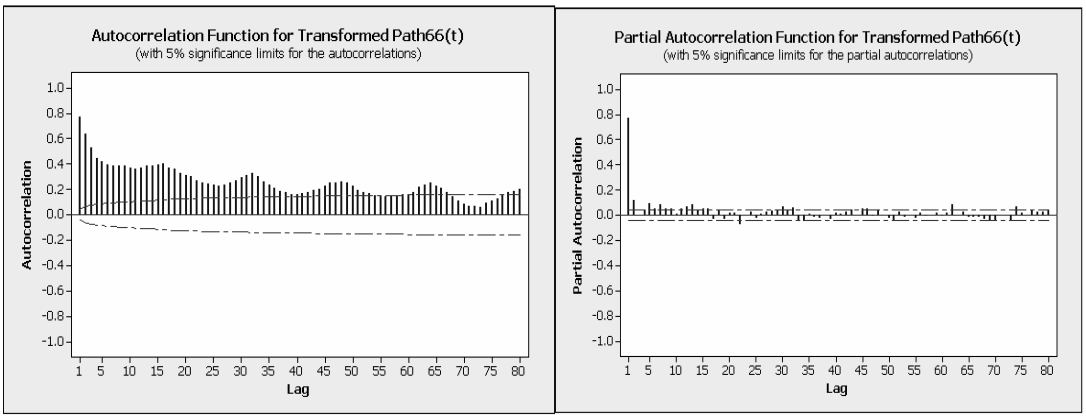

Figure 6-4: ACF and PACF plots of transformed Path66(t)

Table 6.4 presents the optimal orders of non-seasonal and seasonal differencing that were obtained for both the series.

The corresponding plots of the differenced series are shown in Figs 6-5 and 6-6.

Table 6-4: Optimal Differencing orders for ACDC(t) and Path66(t)

Series	**Series name**	**Differencing order** **(d= nonseasonal order; D= seasonal order)**
Input	ACDC(t)	d=1, D=16
Output	Path66(t)	d=1, D=16

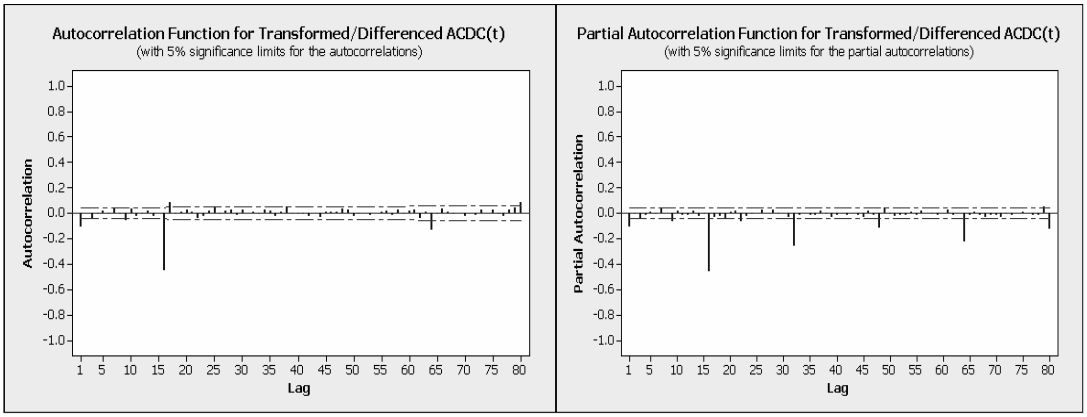

Figure 6-5: ACF and PACF plots of transformed/differenced ACDC(t)

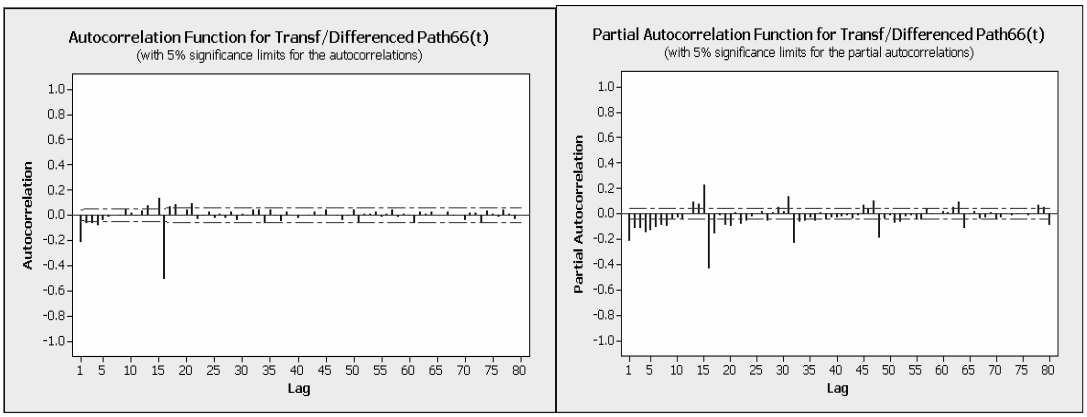

Figure 6-6: ACF and PACF plots of transformed/differenced Path66(t)

6.2.2 Prewhitening the input series

Two competing prewhitening filter forms for the input series (ACDC(t)) were selected and their AIC and SBC values were compared to decide on the final model.

Since SBC criterion penalizes over-fitting more vigorously over the AIC criterion, it was given more importance and the first filter form was chosen as the final model for forecasting. The parameter estimates for the first filter form are provided in Table 6.6.

Table 6-5: Prewhitening filter forms for the set of input series

No.	Prewhitening filter form	AIC	SBC	Variance estimate	Standard error est.
1	**P=0** **Q=(1 3 6 8 9 13 14 15 22) (16 64 80)**	**46451.1**	**46518.1**	**1.049E9**	**32390.45**
2	P=0 Q=(1 2 3 6 8 9 13 14 15 22) (16 64 80)	46450.14	46522.73	1.0482E9	32374.41

The corresponding ACF and PACF plots after applying the first filter are shown in Fig 6-7.

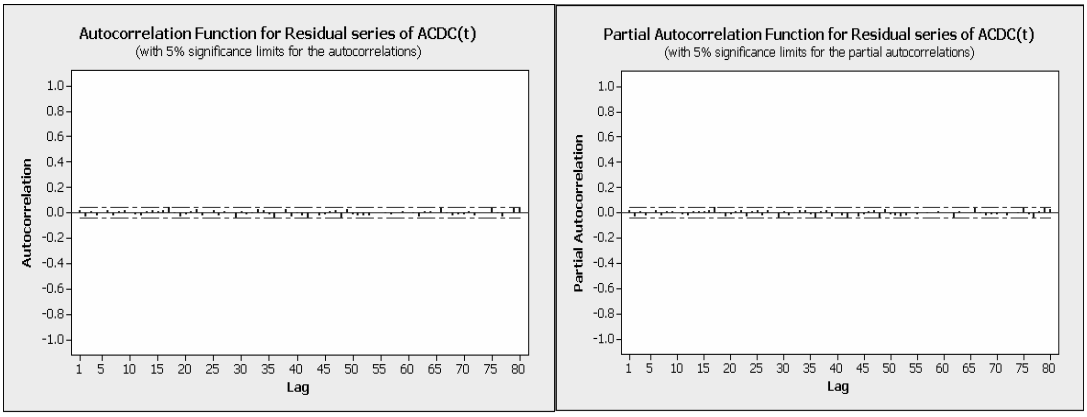

Figure 6-7: ACF and PACF plots of residual series of ACDC(t)

Table 6-6: Parameter estimates of ACDC(t) series

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.16822	0.02179	7.72	<.0001	1
MA1,2	0.11057	0.02171	5.09	<.0001	3
MA1,3	0.06333	0.02205	2.87	0.0041	6
MA1,4	0.04841	0.02212	2.19	0.0288	8
MA1,5	0.07271	0.02242	3.24	0.0012	9
MA1,6	0.07555	0.02208	3.42	0.0006	13
MA1,7	0.05769	0.02267	2.55	0.0110	14
MA1,8	0.07655	0.02259	3.39	0.0007	15
MA1,9	0.07846	0.02206	3.56	0.0004	22
MA2,1	0.69595	0.01768	39.36	<.0001	16
MA2,2	0.17007	0.02344	7.26	<.0001	64
MA2,3	-0.10965	0.02310	-4.75	<.0001	80

Table 6-7: Correlation matrix of Parameter estimates of ACDC(t) series

Correlations of Parameter Estimates												
Parameter	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5	MA1,6	MA1,7	MA1,8	MA1,9	MA2,1	MA2,2	MA2,3
MA1,1	1.000	-0.033	-0.011	-0.014	-0.078	-0.009	-0.093	-0.085	-0.030	-0.098	0.028	0.003
MA1,2	-0.033	1.000	-0.134	-0.022	-0.102	-0.047	0.000	-0.006	0.018	-0.036	0.015	0.023
MA1,3	-0.011	-0.134	1.000	-0.003	-0.106	-0.021	-0.063	-0.115	-0.067	0.029	-0.005	0.003
MA1,4	-0.014	-0.022	-0.003	1.000	-0.190	-0.004	-0.107	0.004	-0.105	-0.048	0.019	0.020
MA1,5	-0.078	-0.102	-0.106	-0.190	1.000	-0.009	0.024	-0.095	-0.085	0.026	-0.011	-0.010
MA1,6	-0.009	-0.047	-0.021	-0.004	-0.009	1.000	-0.193	0.010	-0.126	-0.069	0.033	0.004
MA1,7	-0.093	0.000	-0.063	-0.107	0.024	-0.193	1.000	-0.186	-0.054	0.052	-0.006	-0.037
MA1,8	-0.085	-0.006	-0.115	0.004	-0.095	0.010	-0.186	1.000	-0.017	-0.169	0.009	0.038
MA1,9	-0.030	0.018	-0.067	-0.105	-0.085	-0.126	-0.054	-0.017	1.000	0.009	0.036	-0.021
MA2,1	-0.098	-0.036	0.029	-0.048	0.026	-0.069	0.052	-0.169	0.009	1.000	-0.225	-0.135
MA2,2	0.028	0.015	-0.005	0.019	-0.011	0.033	-0.006	0.009	0.036	-0.225	1.000	-0.669
MA2,3	0.003	0.023	0.003	0.020	-0.010	0.004	-0.037	0.038	-0.021	-0.135	-0.669	1.000

Table 6-8: Ljung-Box Chi-square test results for ACDC(t) series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.017	-0.034	0.011	-0.024	-0.002	0.016
12	.	0	.	-0.023	0.006	0.013	-0.001	-0.013	-0.026
18	12.65	6	0.0489	0.011	0.015	0.011	0.017	0.039	-0.010
24	17.97	12	0.1165	-0.033	-0.014	0.007	0.022	-0.029	0.002
30	24.24	18	0.1474	0.019	-0.023	0.013	-0.003	-0.044	0.012
36	30.32	24	0.1742	-0.018	-0.008	0.025	0.016	-0.016	-0.039
42	39.75	30	0.1098	0.002	0.025	-0.031	-0.012	-0.023	-0.050
48	46.02	36	0.1223	0.001	-0.027	-0.020	0.008	0.020	-0.039
54	51.68	42	0.1455	0.025	-0.015	-0.025	-0.024	-0.026	0.002
60	52.35	48	0.3090	-0.002	-0.010	-0.014	-0.000	0.005	0.003
66	58.00	54	0.3300	-0.007	-0.037	0.011	0.009	0.001	0.034
72	61.44	60	0.4244	-0.006	-0.022	-0.018	-0.016	0.012	-0.021
78	66.57	66	0.4573	-0.000	-0.002	0.032	-0.009	-0.037	0.002

6.2.3 "Prewhitening" the output series and computing Cross-Correlations between the prewhitened input and output series

In this step, the ARMA filter corresponding to the input was applied to the output (Path66(t)) and the residual series (called "prewhitened" output) was stored. The cross-correlation coefficients between the prewhitened input and the "prewhitened" output were computed and plotted next.

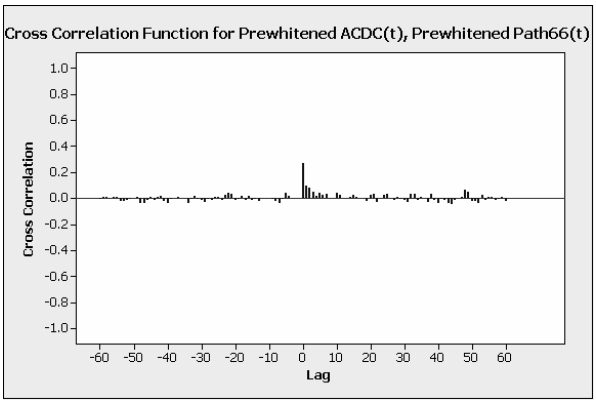

Figure 6-8: Cross-correlation plot for prewhitened ACDC(t), "Prewhitened" Path66(t)

Table 6-9: Cross-correlation coefficients for prewhitened ACDC(t), "Prewhitened" Path66(t)

Lag	CCF	Lag	CCF
-16	0.014252	**0**	**0.26942**
-15	-0.01968	1	0.092814
-14	-0.00825	2	0.075979
-13	-0.02167	3	0.047837
-12	8.37E-05	4	0.018481
-11	0.000653	5	0.035967
-10	-0.00232	6	0.021976
-9	-0.00463	7	0.030347
-8	-0.02408	8	0.002991
-7	-0.03809	9	0.00138
-6	-0.009	10	0.035967
-5	0.038194	11	0.024691
-4	0.016834	12	-0.00214
-3	-0.00133	13	-0.00381
-2	-0.00138	14	0.004026
-1	0.002321	15	0.0244
		16	0.004005

6.2.4 Identifying the input transfer function of the model

Referring to Fig 6.8, since the lag 0 coefficient is significant and all others are non-significant, it can be deduced that the relationship between the ACDC(t) and Path66(t) series is first-order linear. The table of estimated coefficient (Section 3.3, Chapter 3) is given in Table 6-10.

6.2.5 Estimation of Preliminary Noise Series and fitting an ARMA model

In this step, the Preliminary Noise Series, n_t, was extracted using Equation 3.21,Chapter-3.

Table 6-10: Input Transfer function coefficient for ACDC(t) series

Series	Series name	Coefficient
Input	ACDC(t)	$\omega_0 = 24816.1$

The autocorrelation and partial autocorrelation plots of the preliminary noise series is shown in Fig 6-9.

The following ARIMA model form was obtained for the preliminary noise series:

$$\frac{\left(1-\theta_1 B-\theta_2 B^2-\theta_3 B^3-\theta_4 B^4-\theta_{21} B^{21}-\theta_{22} B^{22}\right)\left(1-\theta_{16} B^{16}\right)}{\left(1-\varphi_9 B^9\right)}\ , \qquad (6.1)$$

or, in compact notation,

P=(9) , Q=(1 2 3 4 21 22) (16)

Table 6.11 provides the estimates of parameters.

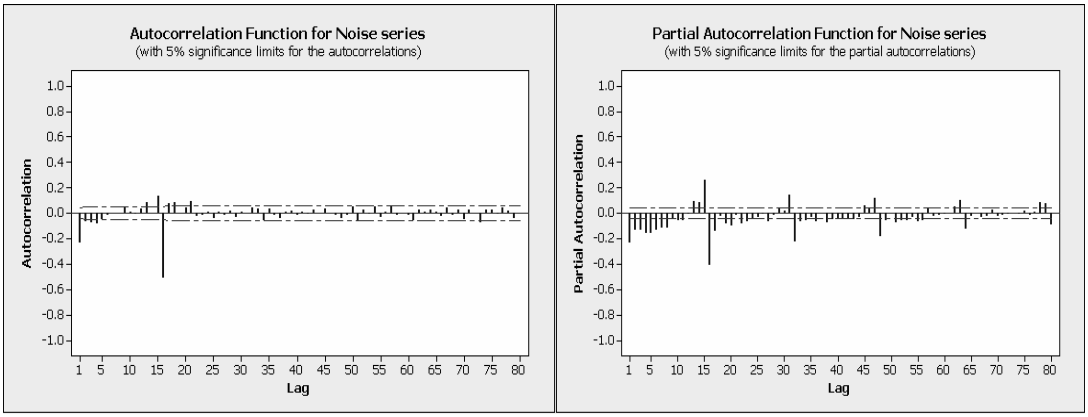

Figure 6-9: ACF and PACF plots for Preliminary Noise Series

Table 6-11: Parameter estimates of Preliminary Noise Series

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.43973	0.02237	19.66	<.0001	1	path66	0
MA1,2	0.16334	0.02420	6.75	<.0001	2	path66	0
MA1,3	0.16471	0.02424	6.80	<.0001	3	path66	0
MA1,4	0.13074	0.02263	5.78	<.0001	4	path66	0
MA1,5	-0.05621	0.02107	-2.67	0.0077	21	path66	0
MA1,6	0.07956	0.02092	3.80	0.0001	22	path66	0
MA2,1	0.91485	0.0094455	96.86	<.0001	16	path66	0
AR1,1	0.05255	0.02341	2.24	0.0249	9	path66	0
NUM1	24816.1	1434.3	17.30	<.0001	0	acdc	0

Table 6-12: Correlation matrix of Parameter estimates of Preliminary Noise Series

Correlations of Parameter Estimates									
Variable Parameter	path66 MA1,1	path66 MA1,2	path66 MA1,3	path66 MA1,4	path66 MA1,5	path66 MA1,6	path66 MA2,1	path66 AR1,1	acdc NUM1
path66 MA1,1	1.000	-0.430	-0.174	-0.217	-0.039	0.013	-0.018	0.059	0.004
path66 MA1,2	-0.430	1.000	-0.359	-0.166	0.010	-0.039	-0.048	0.036	0.003
path66 MA1,3	-0.174	-0.359	1.000	-0.424	-0.073	0.026	0.009	-0.015	-0.002
path66 MA1,4	-0.217	-0.166	-0.424	1.000	-0.007	-0.083	-0.020	0.087	-0.008
path66 MA1,5	-0.039	0.010	-0.073	-0.007	1.000	-0.754	-0.041	-0.063	0.023
path66 MA1,6	0.013	-0.039	0.026	-0.083	-0.754	1.000	0.003	-0.007	0.009
path66 MA2,1	-0.018	-0.048	0.009	-0.020	-0.041	0.003	1.000	-0.055	0.034
path66 AR1,1	0.059	0.036	-0.015	0.087	-0.063	-0.007	-0.055	1.000	0.001
acdc NUM1	0.004	0.003	-0.002	-0.008	0.023	0.009	0.034	0.001	1.000

6.2.6 Identifying the final forecasting model

Thus, the final forecasting model becomes:

$$(1-B)\left(1-B^{16}\right) Path66^{bxcx}(t) = \omega_0 (1-B)\left(1-B^{16}\right) ACDC^{bxcx}(t)$$

$$+ \frac{\left(1-\theta_1 B^1 - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_{21} B^{21} - \theta_{22} B^{22}\right)\left(1-\theta_{16} B^{16}\right)}{1-\varphi_9 B^9} a_t \tag{6.2}$$

where $ACDC^{bxcx}(t)$ denotes the Box-Cox transformation of original ACDC(t) series, and so on. The coefficients of the input transfer function are provided in Table 6.10 and the parameters of the transfer function for noise series are provided in Table 6.11.

6.3 Diagnostic checking of the path66(t) model

The following statistical tests were implemented to check the validity of the model:

6.3.1 Analysis of the final noise series

Presented in Fig 6-10 are the Autocorrelation and Partial Autocorrelation plots of the final noise series, a_t, of the model.

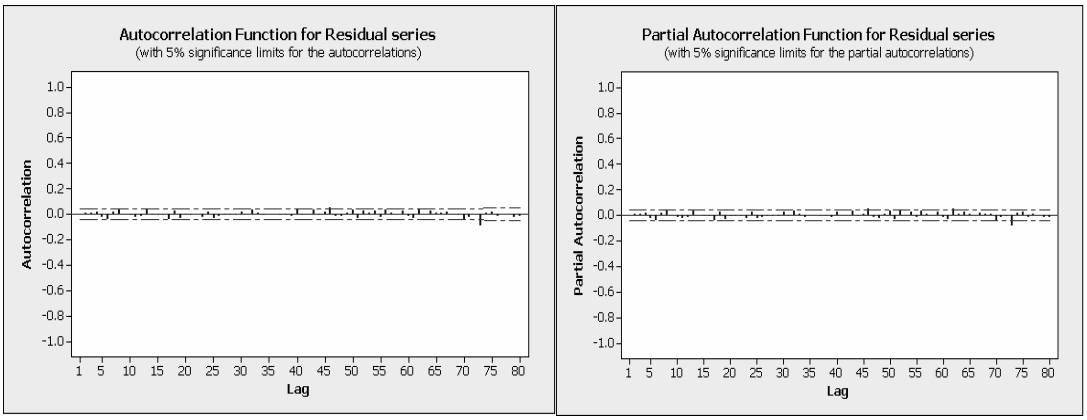

Figure 6-10: ACF and PACF plots for Final Noise Series

No significant spike exists beyond lag 0, signifying absence of any defined pattern in the final noise series. Thus, the final noise series may be a realization of white noise process.

6.3.2 Box-Pierce chi-square statistical test

Table 6.13 provides the Ljung-Box Chi-Square test results for the final noise series, which demonstrates that the final noise series may be a realization of white noise.

Table 6-13: Ljung-Box Chi-square test results for Final Noise Series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.005	0.008	0.013	0.015	-0.022	-0.040
12	10.61	4	0.0313	0.019	0.041	-0.001	-0.011	-0.021	-0.014
18	17.92	10	0.0562	0.036	-0.006	-0.006	-0.006	-0.042	0.023
24	22.28	16	0.1345	-0.035	-0.006	-0.004	-0.002	-0.021	0.021
30	25.75	22	0.2628	-0.029	-0.014	-0.011	-0.006	-0.010	0.021
36	28.75	28	0.4254	-0.003	0.036	0.009	-0.011	-0.001	-0.000
42	31.45	34	0.5930	-0.008	-0.009	-0.014	0.031	-0.006	0.001
48	39.85	40	0.4767	0.031	-0.005	0.017	0.049	-0.018	-0.015
54	48.43	46	0.3753	0.005	0.036	-0.036	0.026	0.006	0.029
60	53.73	52	0.4078	-0.025	0.032	0.009	0.000	0.025	-0.017
66	61.37	58	0.3561	-0.031	0.046	0.003	0.025	0.006	0.006
72	66.09	64	0.4045	0.020	0.003	-0.004	-0.038	-0.020	-0.002
78	83.15	70	0.1347	-0.087	0.008	0.022	-0.012	0.005	0.003

6.3.3 AIC and SBC tests

Two competing models were selected for the preliminary noise series after final analysis. As mentioned in section 6.2.7, model 1 was chosen as the final model for forecasting because of lower AIC and SBC values (Table 6.14).

Table 6-14: AIC and SBC values for competing models

No.	ARMA model	AIC	SBC	Variance estimate	Std error estimate
1	**P=(9) Q=(1 2 3 4 21 22) (16)**	**90813.14**	**90863.4**	**6.55E18**	**2.55E9**
2	P=0 Q=(1 2 3 4 9 21 22) (16)	90817.78	90868.04	6.565E18	2.56E19

6.4 Results of fitting

The fitting (one step-ahead forecasting) plots for Jan 2006 through May 2008 are provided in Figs 6-11 to 6-15.

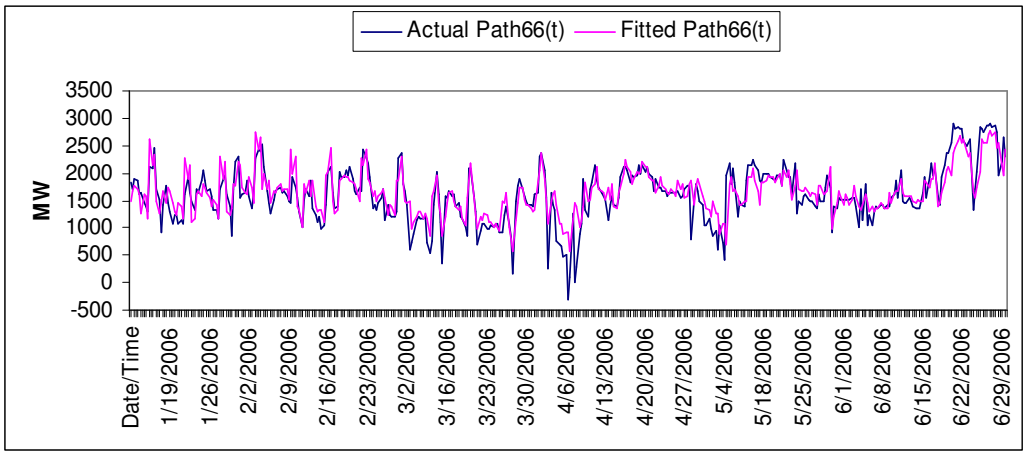

Figure 6-11: Actual and Fitted Path66(t), Jan -June 2006

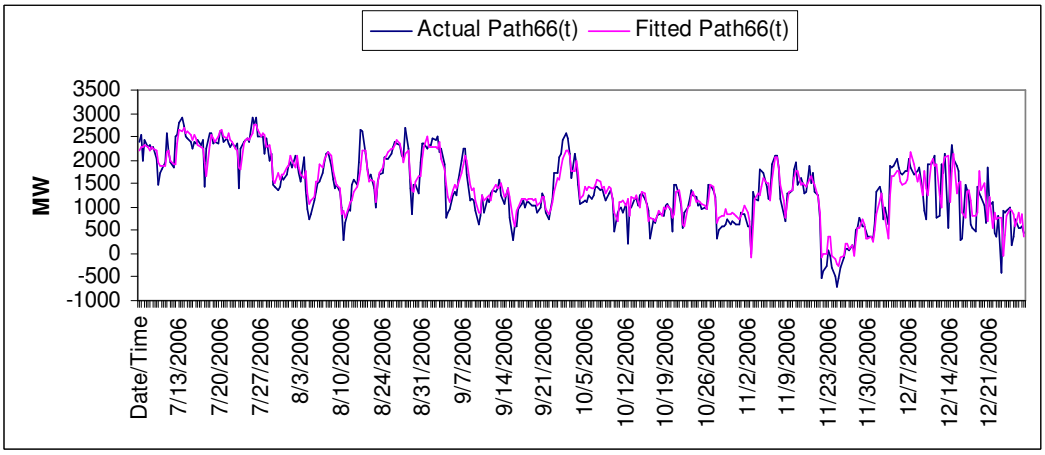

Figure 6-12: Actual and Fitted Path66(t), July -Dec 2006

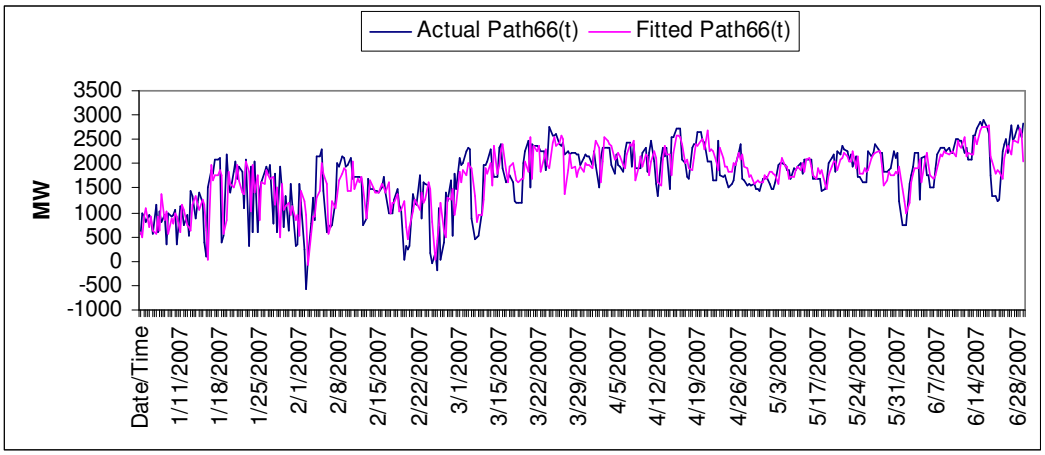

Figure 6-13: Actual and Fitted Path66(t), Jan -June 2007

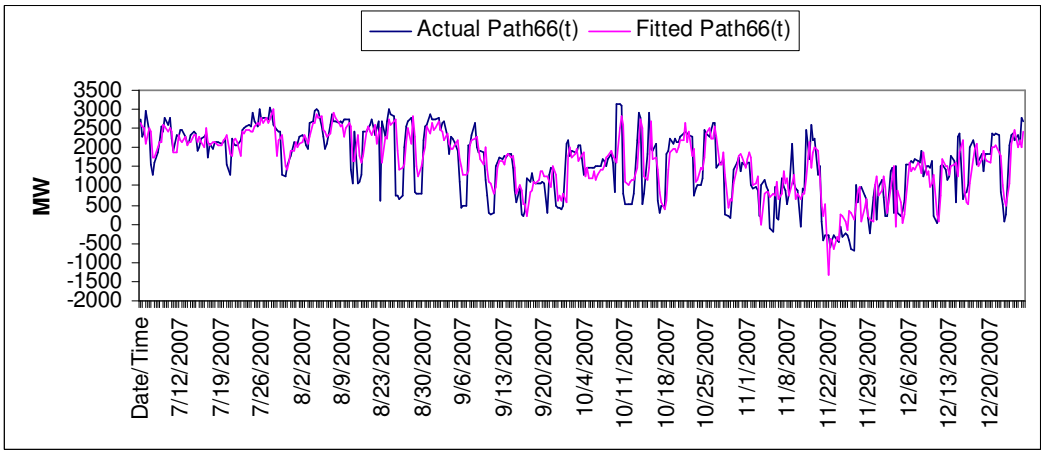

Figure 6-14: Actual and Fitted Path66(t), July -Dec 2007

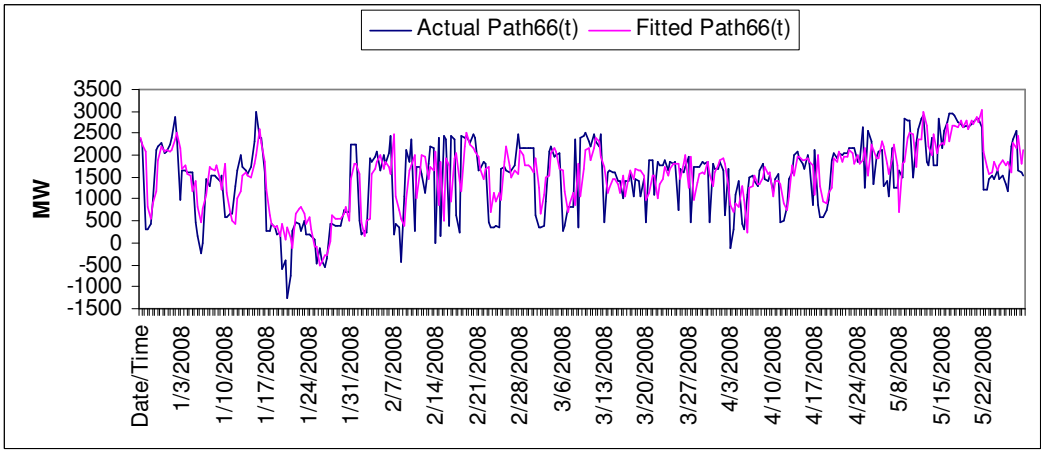

Figure 6-15: Actual and Fitted Path66(t), Jan -May 2008

The fitting accuracy was measured by various quantitative measures and the results are presented in Table 6-15.

Table 6-15: Fitting accuracy measures (Path66(t) model)

Duration/Criterion	Jan 2006-May2008
R-Sqr	0.731
R-Sqr Adj	0.730
FE	-12.608
RMSE	341.532
MAD	259.073
MAPE	38.462

6.5 Results of forecasting

The final model (Eqn 6.2) was used for forecasting the Path66(t) values for the16 on-peak hours of the subsequent Thursday (5th June 2008). The backforecasts for 22nd May, 8th May, 24th April and 10th April were also generated and compared (Table 6-16). The inputs used in the model are the ACDC(t) forecasts generated in Chapter 4 (Section 4.6) for the corresponding day. Thus, the inputs are ACDC(t) forecasts and not the actual future values. This was done to demonstrate the actual purpose of the Path66(t) forecasting model.

Figs. 6.16-6.20 presents the Path66(t) forecasting plots. Various forecasting accuracy measures were computed and the results are presented in Table 6-16.

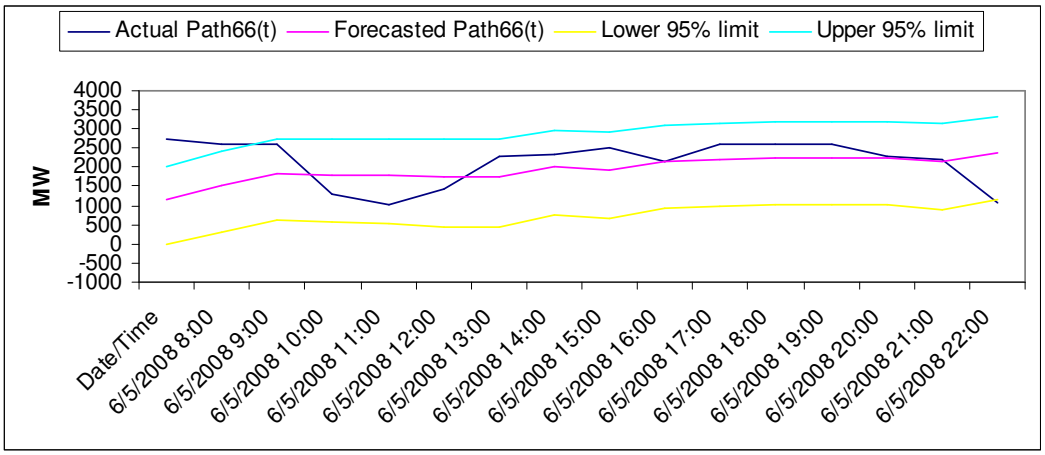

Figure 6-16: Actual and Forecasted Path66(t), 5th June 2008

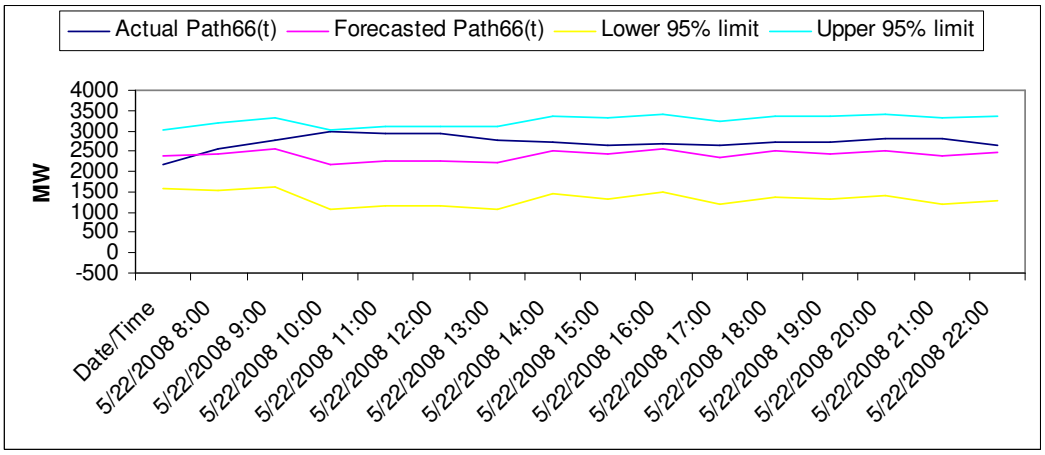

Figure 6-17: Actual and Forecasted Path66(t), 22nd May 2008

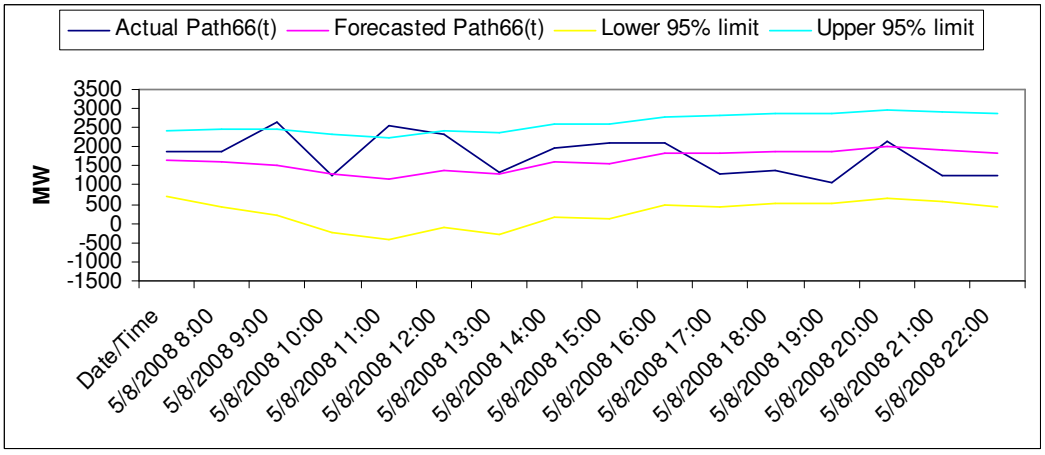

Figure 6-18: Actual and Forecasted Path66(t), 8th May 2008

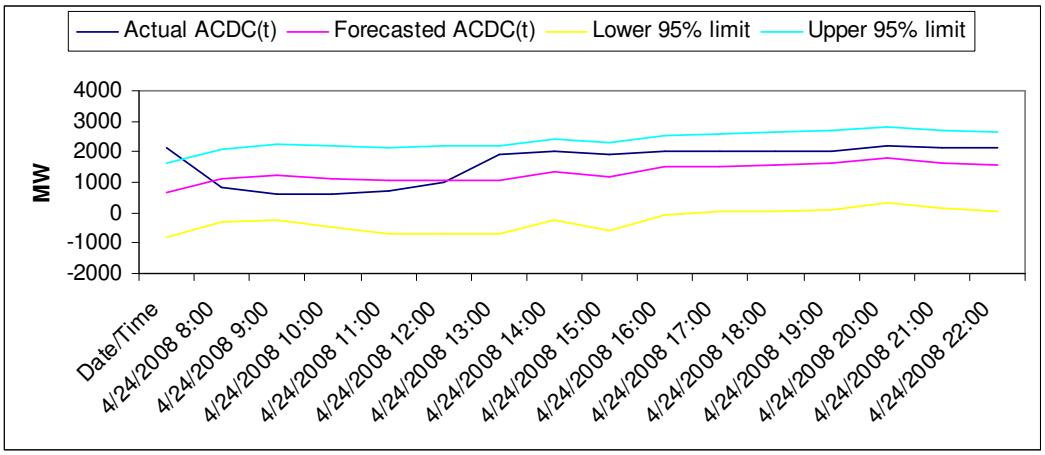

Figure 6-19: Actual and Forecasted Path66(t), 24th April 2008

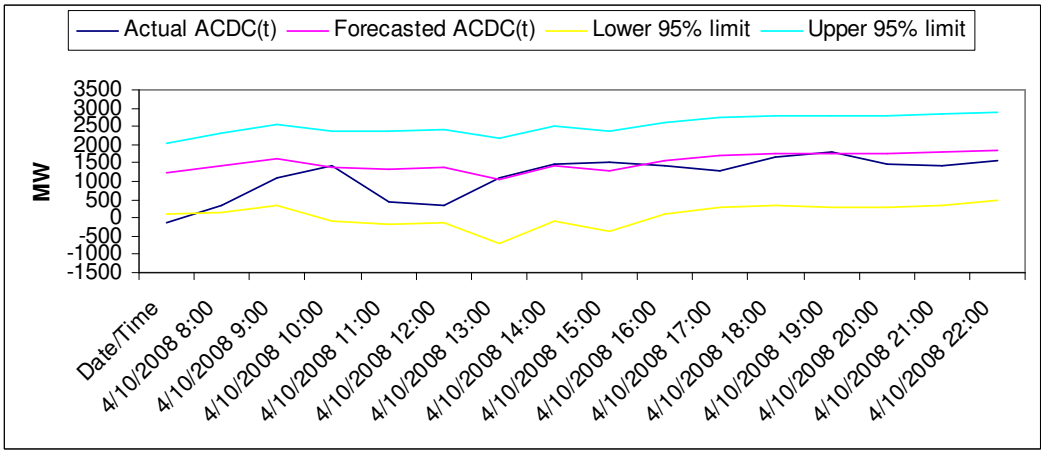

Figure 6-20: Actual and Forecasted Path66(t), 10th April 2008

Table 6-16: Forecasting accuracy measures (Path66(t) model)

Day/Criterion	5th June	22nd May	8th May	24th April	10th April
FE	199.941	313.763	133.929	337.652	-314.745
RMSE	712.794	401.896	640.992	642.200	477.056
MAD	559.970	341.251	524.373	565.691	371.452
MAPE	30.730	12.278	30.270	39.825	53.550

6.6 Conclusions

An adjusted R-square of 0.73 was obtained for Path66(t) series (Table 6-15), in contrast to 0.96 obtained for the ACDC(t) series (Section 4.5, Chapter 4). This is primarily due to the non-electrical format of the Path66(t) data, as described earlier in Section 6.1.

The forecasting accuracies of Path66(t) model (MAPE average over the five forecasting days: 33.3) is lower than that in the case of ACDC(t) model of Chapter 4 (MAPE average: 19.02). This was expected since the forecasts of Path66(t) were based directly on the corresponding forecasts of ACDC(t) series. Also, the lower R-Square fitting of the Path66(t) series, due to the extraneous behavior present in the series, contributed to lower forecasting accuracies.

Chapter 7 – Forecasting the Power Exports from British Columbia Hydro

7.1 Definition of the problem

One of the major source of power exports to California and one of the three inputs used in the forecasting models of chapters 4 and 5 is the exports of power from British Columbia Hydro (BCH) region to the Northwestern US (Section 4.2, Chapter 4). These exports are largely a result of trading contracts between the BCH and CAISO market and also between BCH and BPA territory. All of the power exchange between the BCH and BPA occurs via the Northern Intertie, a tie-line with total north-to-south capacity of 3100 MW (Section 2.2, Chapter 2). The power is then routed through the PACI and PDCI interties to CAISO. An insight into the future patterns of the power flow via the Northern Intertie is very helpful to the Planning departments and to the various traders and market players in the BCH, BPA and CAISO region. In this chapter, a suitable model is identified to forecast these flows on an hourly basis.

The exports from BCH to the BPA territory of US (i.e, the BC(t) series) are likely to depend mainly on the generation surplus (after the regional demand is met) in the BCH region, in addition to some additional factors. However, the surplus generation data was not available for this study due to security policies of Canadian Power Sector. For this reason, the BC(t) series is modeled as a stand-alone forecasting model, using seasonal ARIMA methodology (Section 3.5, Chapter 3).

Section 7.2 presents the steps in identifying the forecasting model. Section 7.3 and 7.4 present the fitting and forecasting results, respectively. In section 7.5, pole-zero analysis of the ARIMA filter of BC(t) series is presented.

Sometime, certain outlying observations, called outliers, may affect the performance of the model. In Section 7.6, an outlier detection technique is presented and is applied on the SAS platform to the BC(t) model built in Section 7.2. The gained improvement is demonstrated.

7.2 Identifying the BC(t) forecasting model

Fig. 4.2 of Chapter 4 presents a snapshot of the BC(t) series for months of Jan-May 2008, which is reproduced below in Fig. 7-1 for convenience.

The standardized series was analyzed by Box-Cox Transformation techniques to detect the non-stationarity in variance. The original series was transformed using a range of λ values and an autoregressive model was fit to the series each time. The likelihood of

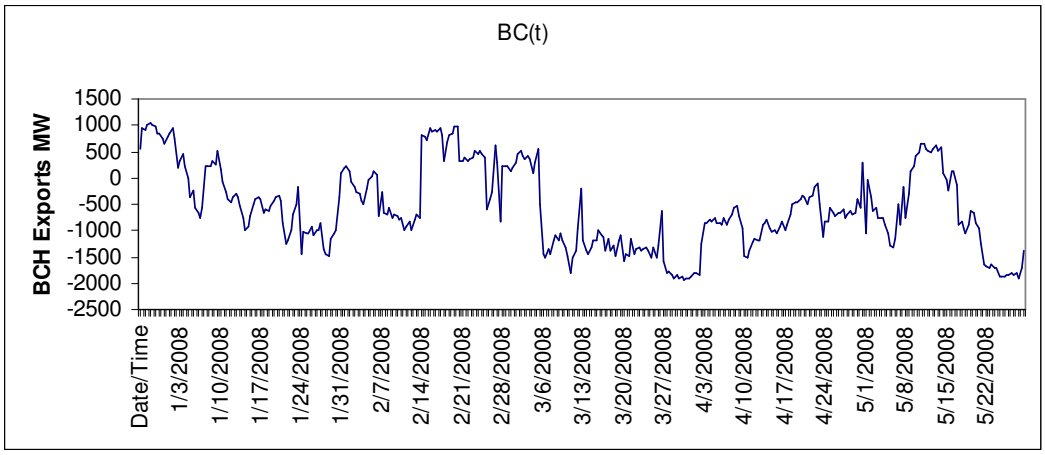

Figure 7-1: Plot of BC(t), Jan-May 2008

the data under each autoregressive model was computed and the value of λ producing the maximum likelihood was chosen as the optimal coefficient of transformation.

The optimum lambda was found to be 0.63. Table 7.1 provides the results for the range of λ values from 0 to 1.

The BC(t) series was then transformed using the optimal coefficient. To remove the non-stationarity present in the mean, suitable non-seasonal and seasonal differencing was required.

Table 7-1: Optimal transformation for BC(t) series

LAMBDA	LOGLIK	RMSE	AIC	SBC
1.00	-1124.91	0.18	2261.82	2295.38
0.95	-1124.39	0.18	2260.79	2294.34
0.89	-1123.95	0.18	2259.91	2293.46
0.84	-1123.59	0.18	2259.18	2292.73
0.79	-1123.30	0.18	2258.60	2292.15
0.74	-1123.08	0.18	2258.17	2291.72
0.68	-1122.94	0.18	2257.89	2291.44
0.63	**-1122.88**	**0.18**	**2257.76**	**2291.31**
0.58	-1122.89	0.18	2257.78	2291.34
0.53	-1122.98	0.18	2257.95	2291.51
0.47	-1123.14	0.18	2258.27	2291.82
0.42	-1123.37	0.18	2258.74	2292.29
0.37	-1123.68	0.18	2259.35	2292.91
0.32	-1124.06	0.18	2260.12	2293.67
0.26	-1124.51	0.18	2261.03	2294.58
0.21	-1125.04	0.18	2262.08	2295.64
0.16	-1125.64	0.18	2263.29	2296.84
0.11	-1126.32	0.18	2264.63	2298.19
0.05	-1127.06	0.18	2266.13	2299.68
0.00	-1127.88	0.18	2267.77	2301.32

Fig 7-2 presents the autocorrelation and partial autocorrelation plots of the transformed series.

First order non-seasonal and seasonal differencing (d=1, D=1) was found to be sufficient to render the series mean-stationary. The corresponding plots for the differenced series are shown in Fig 7-3.

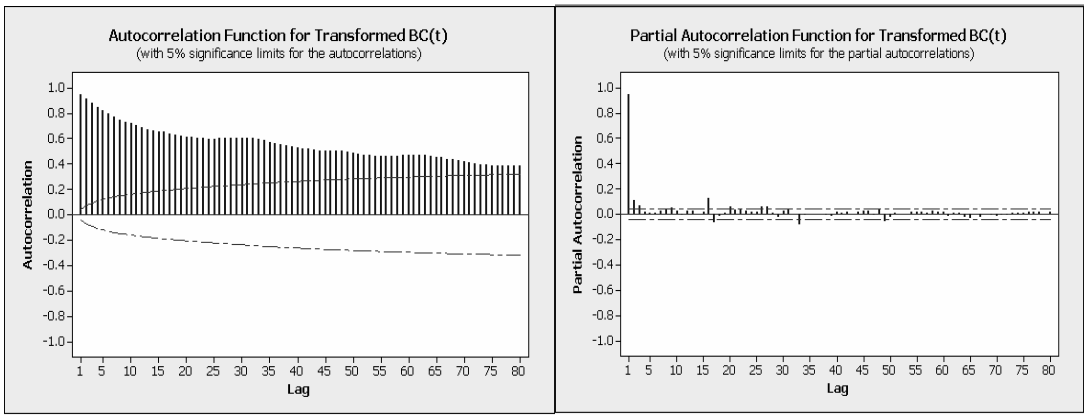

Figure 7-2: ACF and PACF plots for Transformed BC(t) series

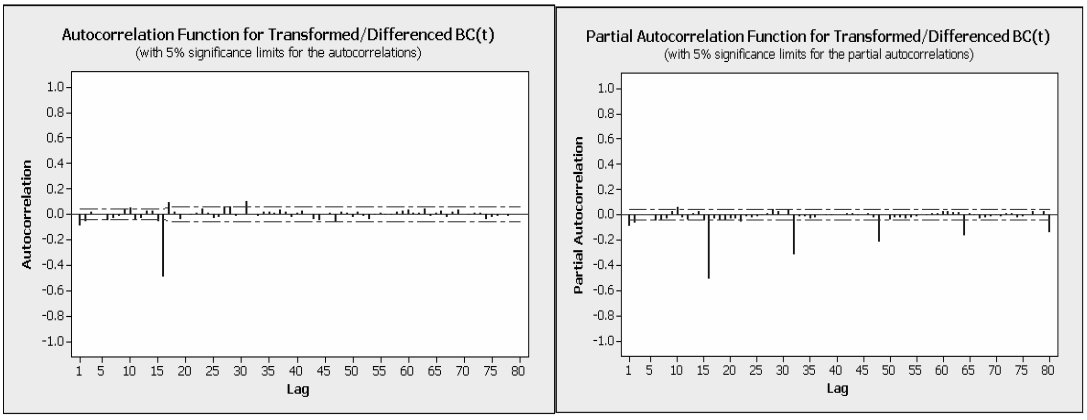

Figure 7-3: ACF and PACF plots for Transformed/Differenced BC(t) series

Next, the three-step iterative strategy of Box-Jenkins (Section 3.3, Chapter 3) was employed to fit a suitable ARMA model to the BC(t) series.

From the analysis of the autocorrelation and partial autocorrelations coefficients of the BC(t) series, three ARIMA models were finally selected (Table 7-2)

Table 7-2: AIC and SBC tests on selected ARIMA models

Sr. No.	ARMA model	AIC	SBC	Variance estimate	Std error estimate
1	**P=0 Q=(1 2 4 6 15) (16)**	**-2306.25**	**-2272.5**	**0.0189**	**0.137**
2	P=0 Q=(1 2 4 6 15) (16 32)	-2304.28	-2264.91	0.019	0.138
3	P=0 Q=(1 2 4 6 15 19) (16)	-2309.08	-2269.71	0.0189	0.137

Filter form 1 was selected on the basis of lower SBC value. It can be represented by:

$$\left(1-\theta_1 B^1-\theta_2 B^2-\theta_4 B^4-\theta_6 B^6-\theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16}\right) \tag{7.1}$$

or, in compact notation,

P=0 Q=(1 2 4 6 15) (16).

Thus, the final model form becomes:

$$\left(1-B^{16}\right)(1-B)\,BC^{bxcx}(t)=\left(1-\theta_1 B^1-\theta_2 B^2-\theta_4 B^4-\theta_6 B^6-\theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16}\right)a_t \tag{7.2}$$

Where $BC^{bxcx}(t)$ is the Box-Cox transformed series of standardized BC(t) series and a_t is the residual series. The parameter estimates are provided in Table 7.3.

In order to check the validity of the model, the ACF and PACF coefficients of the residual series were computed (Fig 7.4).

No significant spikes were present at all the lags beyond lag 0 in the autocorrelation and partial autocorrelation plots. Thus, the final noise series may correspond to a realization of white noise.

The correlation matrix of parameter estimates of BC(t) series is provided in Table 7.4. It was verified that none of the parameters were correlated with each other to a significant level.

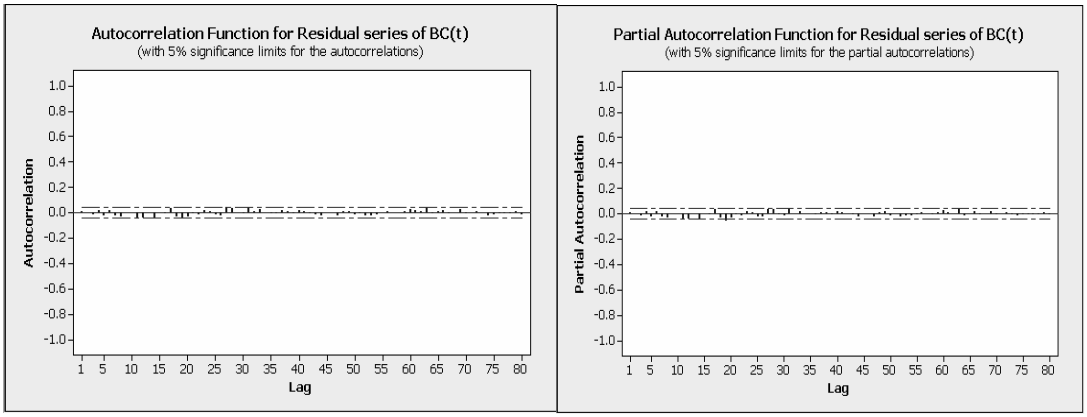

Figure 7-4: ACF and PACF plots for Residual Series of BC(t)

Table 7-3: Parameter estimates of BC(t) series

Conditional Least Squares Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	0.16165	0.02219	7.28	<.0001	1
MA1,2	0.08880	0.02238	3.97	<.0001	2
MA1,3	0.06137	0.02215	2.77	0.0057	4
MA1,4	0.06804	0.02202	3.09	0.0020	6
MA1,5	0.11591	0.02204	5.26	<.0001	15
MA2,1	0.83590	0.01252	66.77	<.0001	16

Table 7-4: Correlation matrix of parameter estimates of BC(t) series

Correlations of Parameter Estimates						
Parameter	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5	MA2,1
MA1,1	1.000	-0.178	0.001	0.013	0.042	-0.077
MA1,2	-0.178	1.000	-0.121	-0.067	-0.014	0.041
MA1,3	0.001	-0.121	1.000	-0.122	0.036	-0.002
MA1,4	0.013	-0.067	-0.122	1.000	-0.004	0.001
MA1,5	0.042	-0.014	0.036	-0.004	1.000	-0.138
MA2,1	-0.077	0.041	-0.002	0.001	-0.138	1.000

The Ljung-Box Chi-square test results are provided in Table 7-5. The residual series was confirmed to be a realization of white noise process, confirming the validity of the model.

Table 7-5: Ljung-Box Chi-square test results for BC(t) series

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.007	0.004	-0.014	0.016	-0.028	0.013
12	12.57	6	0.0503	-0.028	-0.030	-0.002	0.004	-0.040	-0.039
18	20.56	12	0.0572	-0.001	-0.045	0.001	-0.005	0.033	-0.029
24	28.63	18	0.0531	-0.048	-0.033	-0.006	-0.014	0.019	0.007
30	35.42	24	0.0624	-0.018	-0.021	0.039	0.032	-0.003	-0.007
36	40.28	30	0.0996	0.041	0.008	0.022	0.004	-0.006	-0.009
42	41.93	36	0.2292	0.015	0.006	-0.008	0.021	0.006	-0.005
48	45.00	42	0.3474	-0.017	-0.022	-0.003	-0.003	-0.026	0.005
54	48.99	48	0.4332	0.012	-0.017	-0.005	-0.024	-0.026	-0.016
60	50.75	54	0.6005	0.004	0.012	-0.002	0.001	0.011	0.024
66	54.55	60	0.6745	0.014	0.006	0.034	-0.007	0.009	0.018
72	56.29	66	0.7973	-0.010	-0.000	0.023	-0.012	-0.003	0.009
78	58.89	72	0.8665	-0.003	-0.027	-0.015	-0.010	-0.011	-0.010

7.3 Results of fitting

The fitting (one step-ahead forecasting) plots for Jan 2006 through May 2008 are provided in Figs 7-5 to 7-9.

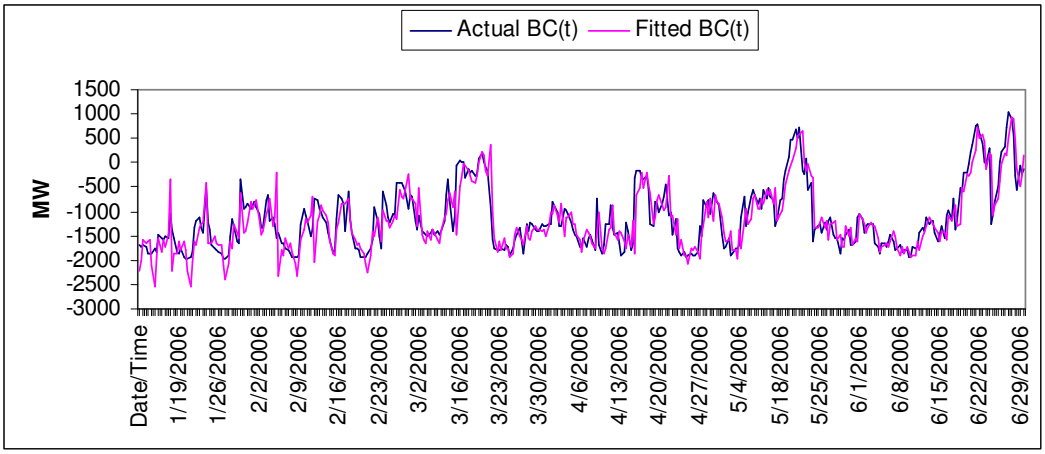

Figure 7-5: Actual and Fitted BC(t), Jan -June 2006

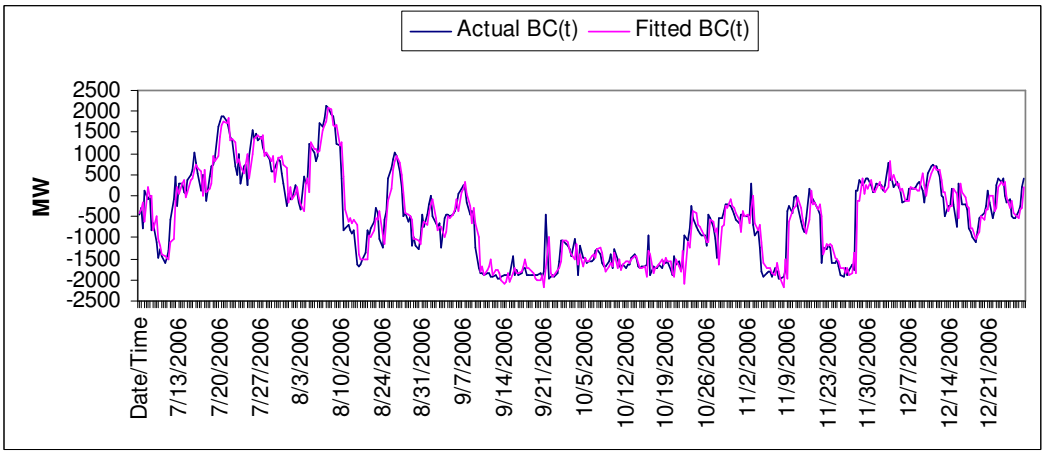

Figure 7-6: Actual and Fitted BC(t), July -Dec 2006

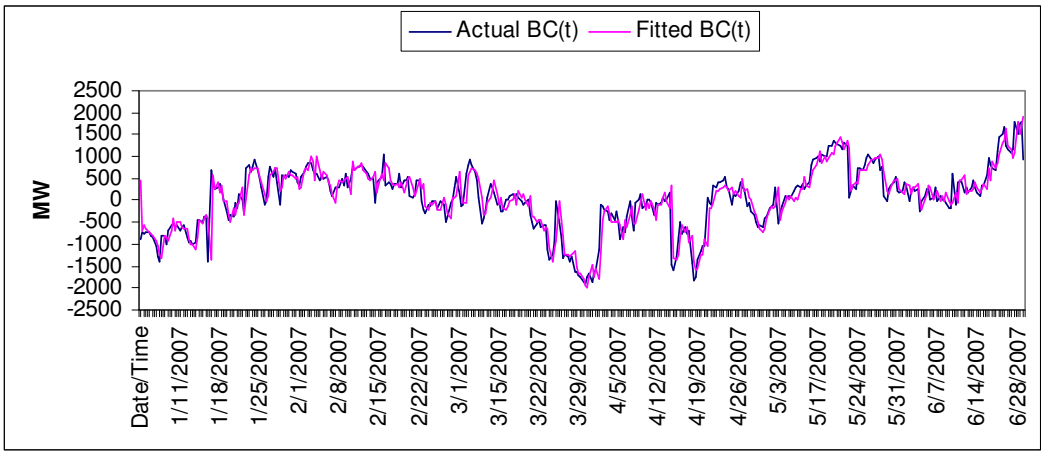

Figure 7-7: Actual and Fitted BC(t), Jan -June 2007

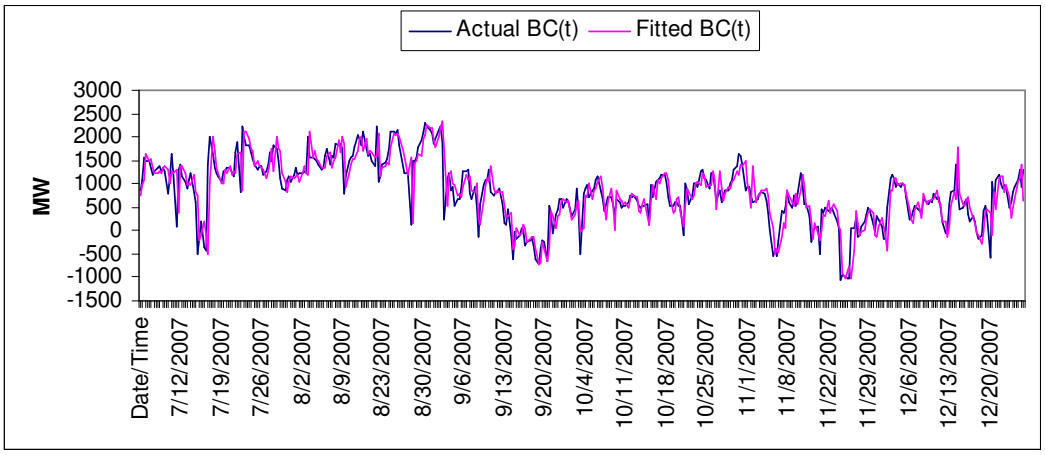

Figure 7-8: Actual and Fitted BC(t), July -Dec 2007

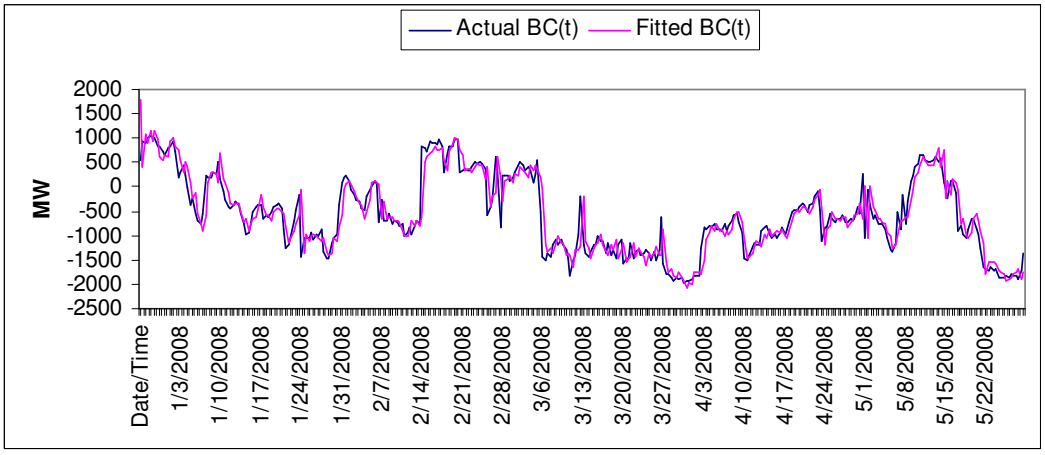

Figure 7-9: Actual and Fitted BC(t), Jan -May 2008

Several fitting accuracy measures were computed to gauge the performance of the model (Table 7-6). The R-square of fitting was 0.92 and the MAPE was 32.778. The slightly lower performance of the model, as compared to models developed in chapters 4 and 5, is due to its stand-alone nature and absence of any inputs.

Table 7-6: Fitting accuracy measures for BC(t) series

Duration/Criterion	Jan 2006-May2008
R-Sqr	0.920
R-Sqr Adj	0.919
FE	-0.678
RMSE	284.244
MAD	198.984
MAPE	32.778

7.4 Results of forecasting

The final model was used to forecast for the 16 on-peak hours of subsequent Thursday (5th June, 2008) (Fig. 7-10). Backforecasts for 29th May, 2008 are also presented (Fig. 7-11). The forecasting accuracy measures are presented in Table 7-7.

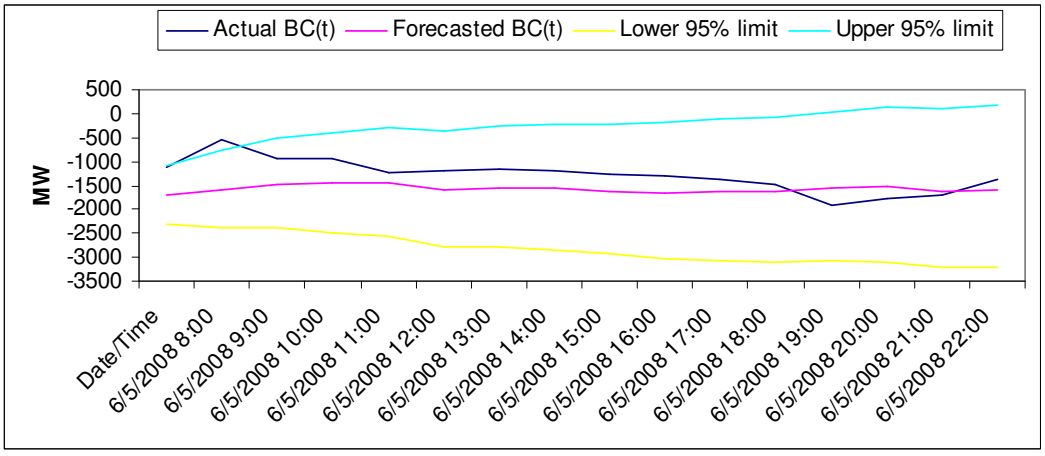

Figure 7-10: Actual and Forecasted BC(t), 5th June 2008

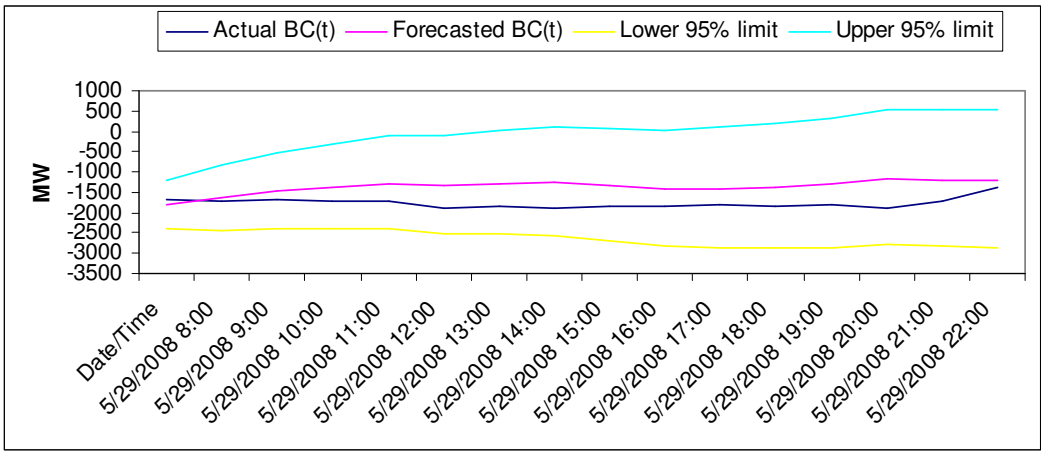

Figure 7-11: Actual and Forecasted BC(t), 29th May 2008

Table 7-7: Forecasting accuracy measures for BC(t) series

Criterion/Day	5th June	29th May
ME(forecast)	294.612	-389.430
RMSE(forecast)	439.761	443.018
MAD(forecast)	383.132	403.794
MAPE(forecast)	38.515	22.433

7.5 Pole and Zero analysis of the ARIMA transfer function

Consider the final forecasting model form of the BC(t) series (Eqn. 7.2):

$$\left(1-B^{16}\right)(1-B)\,BC^{bxcx}(t)=\left(1-\theta_1 B^1-\theta_2 B^2-\theta_4 B^4-\theta_6 B^6-\theta_{15} B^{15}\right)\left(1-\theta_{16} B^{16}\right)a_t$$

It is helpful to analyze the nature of the ARIMA filter to study its stability and frequency response spectrum. In this section, a brief pole-zero analysis of the ARIMA filter of the forecasting model of BC(t) is carried out to detect the location of poles and zeros and confirm the stability of the ARIMA filter.

Let,

$$H(B) = \left(1 - \theta_1 B^1 - \theta_2 B^2 - \theta_4 B^4 - \theta_6 B^6 - \theta_{15} B^{15}\right)\left(1 - \theta_{16} B^{16}\right),$$

the ARIMA filter (Moving Average Filter) of the final model (Eqn. 7.2).

Therefore, in z-domain,

$$H(z) = \left(1 - \theta_1 z^1 - \theta_2 z^2 - \theta_4 z^4 - \theta_6 z^6 - \theta_{15} z^{15}\right)\left(1 - \theta_{16} z^{16}\right)$$

Thus, it is a 31 order all-zero filter. The parameter estimates are provided in Table 7-3. Using MATLAB, the location of Zeros were computed (Table 7-8)

The Pole-Zero graph is provided in Fig 7.12 and the frequency response plot in Fig 7.13. Thus, all the zeros of the filter were found to lie outside the unit circle. The ARIMA filter is stable.

Table 7-8: Location of Zeros of the ARIMA filter of BC(t) series

Zeros	Zeros
-0.9521 + 0.6942i	0.0000 + 1.0113i
-0.9521 - 0.6942i	0.0000 - 1.0113i
-1.1300 + 0.2213i	0.3781 + 1.0894i
-1.1300 - 0.2213i	0.3781 - 1.0894i
-1.0113	0.3870 + 0.9343i
-0.9343 + 0.3870i	0.3870 - 0.9343i
-0.9343 - 0.3870i	0.7749 + 0.8506i
-0.5696 + 1.0086i	0.7749 - 0.8506i
-0.5696 - 1.0086i	0.7151 + 0.7151i
-0.7151 + 0.7151i	0.7151 - 0.7151i
-0.7151 - 0.7151i	1.0622 + 0.4385i
-0.3870 + 0.9343i	1.0622 - 0.4385i
-0.3870 - 0.9343i	1.1025
-0.1147 + 1.1617i	1.0113
-0.1147 - 1.1617i	0.9343 + 0.3870i
	0.9343 - 0.3870i

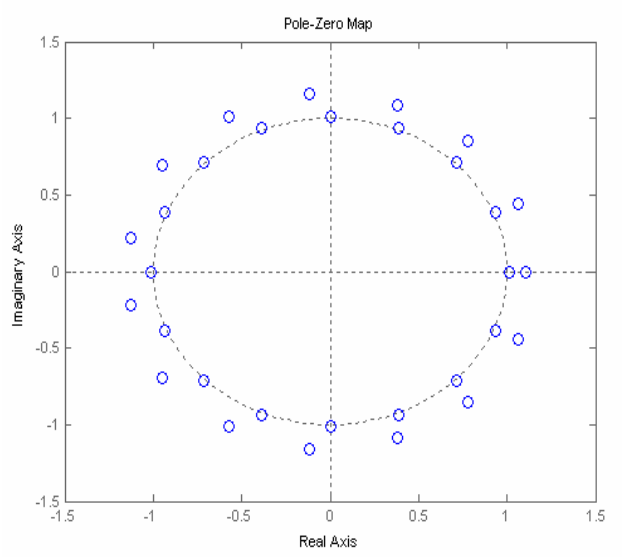

Figure 7-12: Pole-Zero Map of Moving Average filter (BC(t) series)

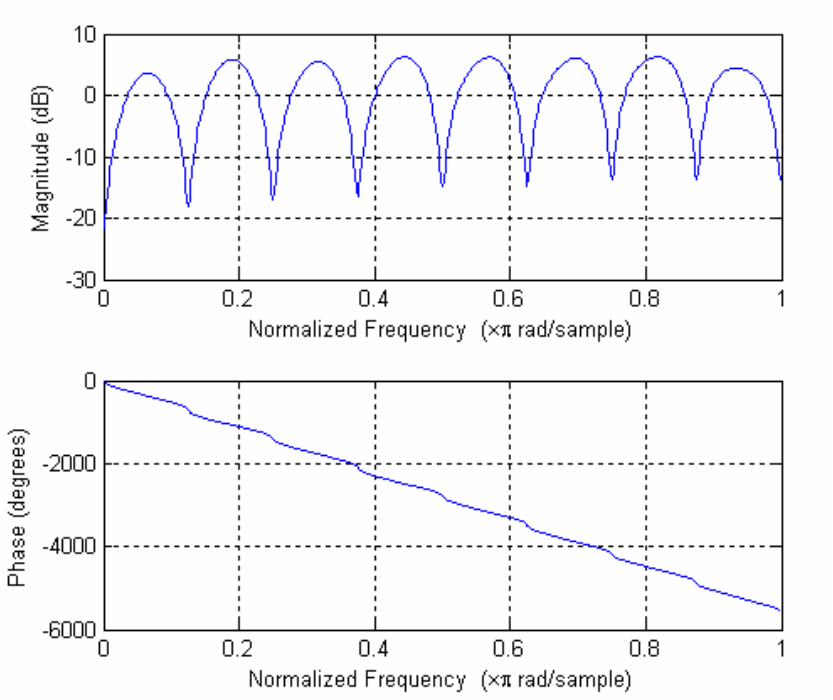

Figure 7-13: Frequency plot for Moving Average filter (BC(t) series)

7.6 Outlier Detection in ARIMA filter of the forecasting model

In the context of this chapter, outliers are defined as those observations of the time series of interest for which there exists a large difference between the actual observation value and the response predicted by the fitted model. Outliers may either be due to erroneous measurements or they maybe perfectly legitimate observations. One of the ways to deal with outliers in the BC(t) series is to remove them and refit the model for better accuracy. However, in the situation of uncertainty regarding the legitimacy of the outliers, an alternate method is proposed below which retains the outliers but introduces additional parameters (called shock signatures) to remove the influence of the outliers.

Consider a general ARIMA model (Equation 3.18, Chapter 3):

$$\varphi\left(B^s\right)\varphi(B)\left(1-B^s\right)^D(1-B)^d\, y_t = \varphi\left(B^s\right)\varphi(B)\, a_t$$

or, $$\left(1-B^s\right)^D(1-B)^d\, y_t = \frac{\theta\left(B^s\right)\theta(B)}{\varphi\left(B^s\right)\varphi(B)}\, a_t$$

or, $$D(B)\, y_t = \frac{\Theta(B)}{\Phi(B)}\, a_t$$

Let the shock signature α_t be a regression variable (with coefficient β) in the model describing some type of change in the mean response [33],

$$D(B)\left(y_t - \beta\alpha_t\right) = \frac{\Theta(B)}{\Phi(B)}\, a_t, \tag{7.3}$$

Three common types of outliers are:

1) Additive Outliers: An additive outlier at some time t_1 is defined as:

$\alpha_t = 1$ for $t = t_1$

$\alpha_t = 0$ for all other t .

2) Level Shifts Outliers: A Level shift at some time t_1 is defined as:

$\alpha_t = 1$ for $t < t_1$

$\alpha_t = 0$ for $t \geq t_1$.

3) Temporary Changes: A Temporary change originating at some time t_1 and with duration d is defined as:

$\alpha_t = 1$ for $t_1 \leq t \leq t_1 + d$

$\alpha_t = 0$ for all other t .

The procedure to detect the presence of shock signature involves testing:

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$, in the model,

$$D(B)\left(y_t - \beta \alpha_t\right) = \frac{\Theta(B)}{\Phi(B)} a_t ,$$

for each of the shock signatures present in the model.

Eqn (6.5) can be written as:

$$N_t = \beta \zeta_t + \frac{\Theta(B)}{\Phi(B)} a_t \qquad (7.4)$$

Where $N_t = D(B)\, y_t$ and

$\zeta_t = D(B)\alpha_t$, the 'effective' shock signature.

In this setting, under H_0 ,

$N = \left(N_1, N_2, \ldots \cdots, N_n\right)^T$ is a mean zero vector with variance covariance matrix $\sigma^2 \Sigma$, where σ^2 is the variance of the white noise a_t and Σ is the variance covariance matrix associated with the ARMA model.

Under H_1, N has $\beta\zeta$ as the mean vector where $\zeta = \left(\zeta_1, \zeta_2, \ldots, \zeta_n\right)^T$

The least squares estimate of β is then given by [33]:

$\bar{\beta} = \frac{\delta}{\kappa}$, and

$Var\left(\bar{\beta}\right) = \frac{\sigma^2}{\kappa}$,

where $\delta = \zeta^T \Sigma^{-1} N$ and $\kappa = \zeta^T \Sigma^{-1} \zeta$.

The test statistic $\tau^2 = \frac{\delta^2}{\sigma^2 \kappa}$ has an approximate distribution with 1 degree of freedom under H_0 and is used to test the significance of β. The robust estimate of σ^2 is computed as:

$\bar{\sigma}^2 = \left(1.49 \times median\left(\left|\bar{a}_t\right|\right)\right)^2$,

Where $\bar{a}_t$ are the standardized residuals.

7.6.1 Detecting outliers in the BC(t) series

The procedure described in the preceding section was demonstrated by detecting the dominant outliers in the BC(t) fitting model of Eqn 7.2. The five outliers found, their types and their approximate (standardized) impacts are shown in Table 7-9.

Table 7-9: Results of the outlier detection algorithm for BC(t) model

Outlier Details				
Obs	**Type**	**Estimate**	**Chi-Square**	**Approx Prob>ChiSq**
753	Shift	0.83969	101.16	<.0001
864	Additive	-0.62515	97.08	<.0001
1632	Additive	-0.53904	72.20	<.0001
1344	Additive	0.51321	65.65	<.0001
1073	Shift	-0.65472	64.45	<.0001

The residuals corresponding to the dominant outliers are presented in Fig 7.14.

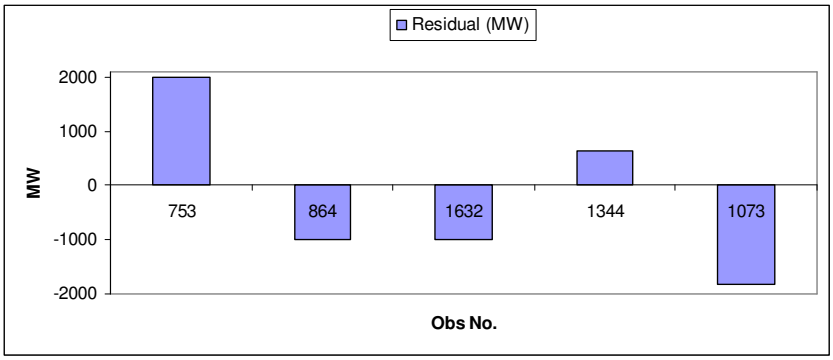

Figure 7-14: Dominant outliers in the BC(t) model

Five shock signatures as regression variables were included in the final model (Eqn. 7.2) and non-linear estimation was carried out. The parameter estimates and the significance test results are presented in Table 7.10.

Thus, all the parameters were found to be significant, including the regression coefficients corresponding to the shock signatures. The correlation matrix of parameters is provided in Table 7.11. None of the parameters were found to be significantly correlated with each other.

Table 7-10: Parameter estimates of the BC(t) model with shock signatures

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.11904	0.02186	5.44	<.0001	1	BC	0
MA1,2	0.10553	0.02202	4.79	<.0001	2	BC	0
MA1,3	0.06140	0.02196	2.80	0.0052	4	BC	0
MA1,4	0.07207	0.02184	3.30	0.0010	6	BC	0
MA1,5	0.12539	0.02178	5.76	<.0001	15	BC	0
MA2,1	0.82620	0.01267	65.20	<.0001	16	BC	0
NUM1	0.86370	0.12274	7.04	<.0001	0	LS753	0
NUM2	-0.62012	0.09290	-6.68	<.0001	0	AO864	0
NUM3	-0.55341	0.09288	-5.96	<.0001	0	AO1632	0
NUM4	0.52164	0.09333	5.59	<.0001	0	AO1344	0
NUM5	-0.66298	0.12275	-5.40	<.0001	0	LS1073	0

Table 7-11: Correlation matrix of parameter estimates of BC(t) model with shock signatures

Correlations of Parameter Estimates											
Variable Parameter	BC MA1,1	BC MA1,2	BC MA1,3	BC MA1,4	BC MA1,5	BC MA2,1	LS753 NUM1	AO864 NUM2	AO1632 NUM3	AO1344 NUM4	LS1073 NUM5
BC MA1,1	1.000	-0.135	0.002	0.022	0.028	-0.081	-0.036	-0.029	-0.006	0.007	0.028
BC MA1,2	-0.135	1.000	-0.127	-0.069	-0.016	0.032	0.010	0.004	0.005	-0.003	0.004
BC MA1,3	0.002	-0.127	1.000	-0.125	0.028	-0.006	-0.018	-0.008	-0.022	0.006	0.007
BC MA1,4	0.022	-0.069	-0.125	1.000	0.018	0.005	-0.020	-0.014	-0.005	0.028	-0.001
BC MA1,5	0.028	-0.016	0.028	0.018	1.000	-0.115	-0.002	-0.003	-0.022	-0.072	0.018
BC MA2,1	-0.081	0.032	-0.006	0.005	-0.115	1.000	0.008	0.021	0.022	0.076	0.032
LS753 NUM1	-0.036	0.010	-0.018	-0.020	-0.002	0.008	1.000	-0.014	0.001	-0.001	0.002
AO864 NUM2	-0.029	0.004	-0.008	-0.014	-0.003	0.021	-0.014	1.000	0.001	0.001	-0.006
AO1632 NUM3	-0.006	0.005	-0.022	-0.005	-0.022	0.022	0.001	0.001	1.000	0.005	-0.000
AO1344 NUM4	0.007	-0.003	0.006	0.028	-0.072	0.076	-0.001	0.001	0.005	1.000	-0.001
LS1073 NUM5	0.028	0.004	0.007	-0.001	0.018	0.032	0.002	-0.006	-0.000	-0.001	1.000

The ACF and PACF plots of the final noise series of the BC(t) model with shock signatures are presented in Fig 7.15.

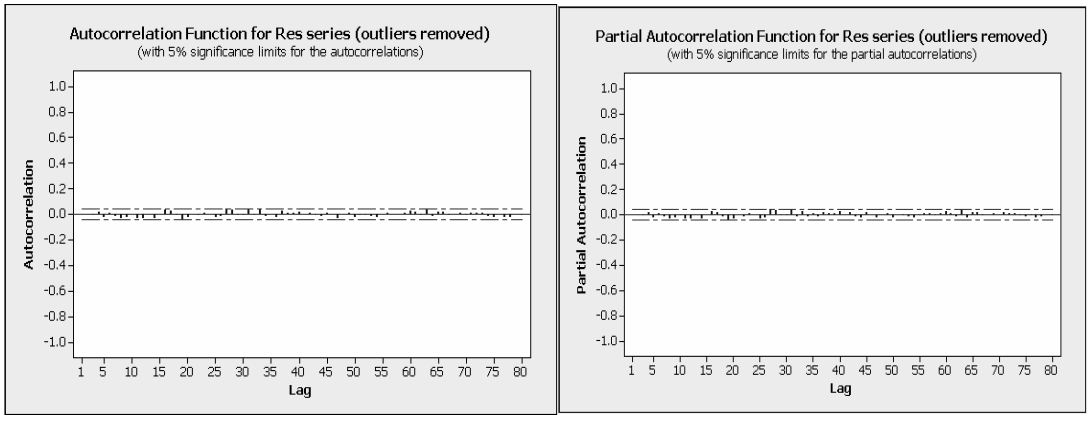

Figure 7-15: ACF and PACF plots for residual series of BC(t) model with shock signatures

Thus, the residual series may be a realization of white noise process. The Ljung-Box Chi-square test was carried out to further confirm the randomness of the residual series (Table 7-12)

Fig 7-16 presents the residuals corresponding to the five outliers after including the shock signatures in the model. It can be seen that the effect of outliers was greatly reduced by including shock signatures in the model.

This process may be extended to any number of outliers present in a general model; however, care must be taken not to over-fit the model since over-fitting may reduce the forecasting accuracies.

Table 7-12: Ljung-Box Chi-square test results for residual series of BC(t) model with shock signatures

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	0.003	0.003	-0.012	0.014	-0.028	0.010
12	11.08	6	0.0859	-0.020	-0.030	-0.027	-0.004	-0.033	-0.031
18	16.76	12	0.1587	-0.001	-0.032	-0.003	0.032	0.024	-0.012
24	22.82	18	0.1976	-0.045	-0.026	0.002	-0.011	0.009	0.004
30	31.79	24	0.1324	-0.029	-0.021	0.044	0.032	-0.010	-0.004
36	39.26	30	0.1201	0.041	-0.011	0.031	-0.013	0.002	-0.025
42	41.48	36	0.2440	0.023	0.006	0.005	0.021	-0.001	0.005
48	44.89	42	0.3516	-0.012	-0.019	0.012	-0.001	-0.031	0.001
54	48.15	48	0.4670	0.004	-0.025	-0.004	-0.010	-0.018	-0.021
60	49.80	54	0.6369	-0.000	0.009	0.004	-0.002	0.013	0.023
66	55.35	60	0.6460	0.013	-0.009	0.040	-0.018	0.015	0.016
72	56.11	66	0.8022	-0.011	0.002	0.007	-0.007	0.006	0.010
78	60.23	72	0.8373	0.011	-0.016	-0.022	-0.001	-0.025	-0.021

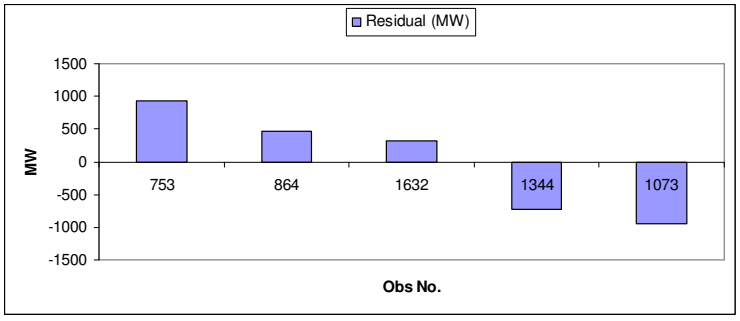

Figure 7-16: Outliers in the BC(t) model after including shock signatures

Chapter 8- Conclusions and Future Work

The problem of short-term forecasting of power flows on transmission interties with reasonable data requirements was successfully solved by employing the statistical technique of Box-Jenkins ARIMA methodology. Despite the non-availability of data or power flow analysis techniques, reasonable fitting and forecasting accuracies were achieved. The specific application of forecasting the power flows on some of the major interties of the Pacific Northwestern US was demonstrated with results in this thesis. In chapter 4, the statistical model identified to forecast the combined power flows on the Pacific AC Intertie and the Pacific DC Intertie was shown to achieve an R-Square value of 0.961 and MAPE (Mean Absolute Percentage Error) of 14.9 for fitting purpose, using 29 months of historical data. Reasonable forecasting accuracy was achieved from the model (average MAPE of 19.02 for the 5 selected days for forecasting). The alternate model identified in chapter 5 for forecasting the PACI and PDCI flows improved upon the procedure follower in chapter 4. However, in terms of forecasting accuracy over the longer term, the earlier model proved to be slightly more accurate. In chapter 6, a model was identified to forecast the CAISO contracted portion of COI intertie. The forecasted values were for real-time scheduled flows and their accuracy depended upon the quality of inputs used in the model, which were the forecasts from the model developed in chapter 4. Forecasts for several days were generated and reasonable accuracy was shown to have been achieved. In chapter 7, a model was identified to forecast the power exports from British Columbia to the Pacific Northwestern US via the Northern Intertie. The model was essentially a stand-alone model since no relevant inputs for the model were

available. Nevertheless, reasonable fitting accuracy (R square = 0.92, MAPE =32.7) was achieved for 29 months of historical data and the generated forecasts were adequately accurate.

Based on the accuracy of forecasts obtained from the various models built in this thesis, it was demonstrated that Box and Jenkins ARIMA model can be successfully employed for the purpose of forecasting power flows on transmission lines. The results of this work demonstrates the usefulness and motivates the need of employing the statistical technique of ARIMA methodology in forecasting power flow applications, either independently or in conjunction with the traditional methods, to result in a less computationally and data intensive method. Nevertheless, some work still needs to be done to validate further the use of such techniques in various other scenarios faced by the power systems in different power markets of the world.

As a future task to the work presented in this thesis, the forecasting model for BCH exports may be refitted if the relevant inputs (mainly the BCH surplus generation over demand) can be made available in the future. A transfer function model may be built and the results may be compared with the stand-alone model proposed in this thesis.

Another possible area of future work is the modeling of the highly volatile power flow pattern on the Pacific DC Intertie. The model built in chapter 4 was used to forecast the combined power flows over the PACI and PDCI interties; it may also be of interest to forecast the power flows on PDCI by itself. Figs 8-1 and 8-2 present some sample snapshots of the actual PDCI flows to demonstrate the high level of volatility in them.

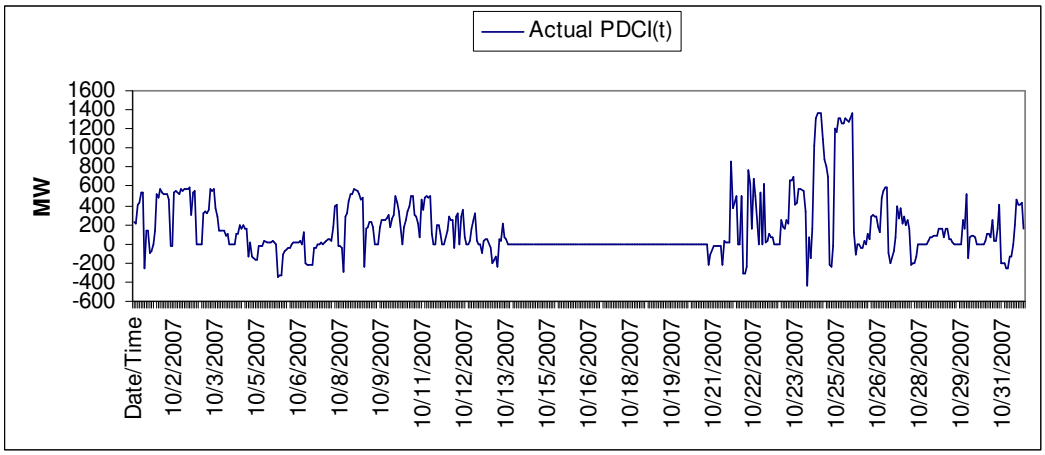

Figure 8-1: PDCI flows for month of October, 2007

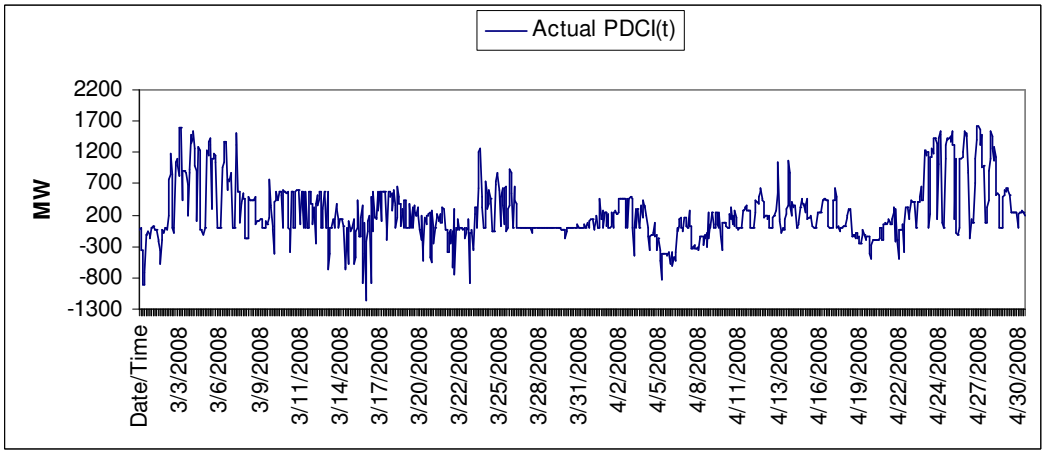

Figure 8-2: PDCI flows for months of March and April, 2008

Such high volatile flows are a result of the various complex operating procedures followed jointly by the BPA and CAISO in operating the PDCI line from hour-to-hour. It may first appear that the PDCI limits are zero when the PDCI flows are zero and vice versa. Investigation showed that though the vice versa may be generally true; the PDCI limits may not always be zero when the PDCI flows are zero. In other words, the PDCI flows may drop to zero even when the limits are non-zero, thus signifying the involvement of other factors in the operation of PDCI. Additionally, switching

mechanisms, inherent with the operation and control of DC transmission lines, may also contribute to the unpredictability of the power flows. Successful modeling of such highly volatile flows will require use of some additional techniques besides the ARIMA modeling. Nevertheless, for some simple purposes, one may only be interested in knowing whether the flows on PDCI will be zero or non-zero at a give time in future; in other words, one may only like to forecast whether the line will carry power in next few hours or not. A statistical model may be built to forecast the 1's and 0's, where 1's implies presence of flows on the PDCI line.

Another interesting topic of future work in this field could be the study of the impact on CAISO market prices due to the forecasts of power flows, available from the models developed in this work. Depending on seasonal conditions, the Pacific Intertie is capable of collectively transferring up to 7,900 MW of surplus power generation in PNW region to California markets, which is over 14% of the total installed MW capacity of California. As mentioned in Chapter 2, most of these exports are during the spring and summer seasons, when the demand is high in California. Consequently, these exports play some direct role in setting the electricity prices in California ISO markets, which runs a short-term wholesale power market. Forecasting these price impacts and their subsequent affects on the operation of power markets, thus, may attract substantial academic interest.

References

[1] "Federal Energy Regulatory Commission, Order No. 888, Docket Nos. RM95-8-000 and RM94-7-001 *Promoting Wholesale Competition Through Open Access Nondiscriminatory Transmission Services by public Utilities;Recovery of Stranded Costs by Public Utilities and Transmitting Utilities*," pp. 1, April 24, 1996.

[2] R. Weron, "Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach," 2006.

[3] K. B. H. Zareipour, C.A.Canizares, "Forecasting the Hourly Ontario Energy Price by Multivariate Adaptive Regression Splines," *IEEE*, 2006.

[4] S. M. Cottet Remy, "Bayesian Modeling and Forecasting of Intraday Electricity Load," 2006.

[5] K. T. Zoran Obradovic, "Time Series Methods of Forecasting Electricity Market Pricing," *IEEE*, 1999.

[6] M. Z. Chengjun Li, "Application of GARCH Model in the Forecasting of Day-Ahead Electricity Prices," *IEEE*, 2007.

[7] R. E. Javier Contreras, Francisco J. Nogales, Antonio J. Conejo, "ARIMA Models to Predict Next-Day Electricity Prices," 2003.

[8] Z. Y. Ming Zhou, Yixin Ni, Gengyin Li, "An ARIMA Approach to Forecasting Electricity Price with Accuracy Improvement by Predicted Errors," 2004.

[9] D.-X. N. Jian-Chang Lu, Zheng-Yuan Jia, "A Study of Short-Term Load Forecasting based on ARIMA-ANN " *IEEE*, 2004.

[10] S.-K. C. Arun Sekar, Wai-Ran Wu, "Base-Case Power Flow Model of an Interconnected Power System," 2002.

[11] D. D. R. Jaime Cerda Jacobo, "A Decentralized DC Optimal Power Flow Model," *IEEE*, 2008.

[12] B. Z. Hongbin Sun, "Distributed Power Flow Calculation for Whole Networks Including Transmission and Distribution," *IEEE*, 2008.

[13] "Northwest Power and Conservation Council, *Electricity Generation for the Pacific Northwest*," June, 2006.

[14] "Northwest River Partners, *Hydropower's importance to the Northwest*," 2006.

[15] "Bonneville Power Administration, *2007 BPA Facts*," 2007.

[16] US Bureau of Reclamation, BPA, US Army Corps of Engineers, "The Columbia River System Inside Story," *Federal Columbia River Power System*, April, 2001.

[17] "British Columbia Hydro website, *www.bchydro.com*," 2008.

[18] "Energy Almanac, *California's Major Sources of Energy*," *The California Energy Commission*, 2008.

[19] "Market Issues and Performance, 2007 Annual Report," *Department of Market Monitoring, California Independent System Operator Corporation*, 2007.

[20] "Oasis / Operations Information (OPI)," *Bonneville Power Administration Transmission Services*, 2008.

[21] "Pacific Intertie: The California Connection on the Electron Superhighway," *Northwest Power Planning Council*, May, 2001.

[22] B. W. Griess, “Transmission Agency of Northern California, *Implications of the CAISO's IBAA proposal on the California-Oregon Transmission Project*,” March 6, 2008.

[23] “Transmission Business Line (TBL),” *Bonneville Power Administration*, April 20, 2005.

[24] G. M. Jenkins, George E.P. Box, *Time Series Analysis forecasting and control*, 1983.

[25] A. Pankratz, *Forecasting with Univariate Box-Jenkins Models*: Wiley Series in Probability and Mathematical Statistics, 1983.

[26] L. S. Gareth Janacek, *Time Series: forecasting, simulations, applications*, 1993.

[27] N. T. Thomopoulos, *Applied Forecasting Methods*, 1980.

[28] Makridakas, *Forecasting: Methods and Applications*, 1983.

[29] O'Donovan, *Short Term Forecasting*, 1983.

[30] K. M. Regan, “An evaluation of ARIMA (Box-Jenkins) models for forecasting wastewater treatment process variables,” 1984.

[31] G. Kress, *Practical techniques of Business Forecasting*, 1985.

[32] W. W. S. Wei, *Time Series Analysis : Univariate and Multivariate Methods*, 2006.

[33] *SAS/ETS 9.1 User's Guide*: SAS Publishing, 2004.

Appendix A

A.1 SAS program for Identifying the ACDC(t) model:

```
Proc print data=work2.THURSacdcmodelORIGstd;
run;
%BOXCOXAR(work2.THURSacdcmodelORIGstd,acdc,lambdahi=2.5,lam
bdalo=1.5,nlambda=20,dif=(1,16),const=10,out=bcxxl);
data THURSacdcmodelORIG;
set work2.THURSacdcmodelORIGstd;
bxcxacdc=&BOXCOXAR;
acdc=(((acdc+10)**bxcxacdc)-1)/bxcxacdc;
run;
%BOXCOXAR(work2.THURSacdcmodelORIGstd,bpa,lambdahi=2,lambda
lo=1,nlambda=20,dif=(1,16),const=10,out=bcxxbpa);
data THURSacdcmodelORIG;
set THURSacdcmodelORIG;
bxcxbpa=&BOXCOXAR;
bpa=(((bpa+10)**bxcxbpa)-1)/bxcxbpa;
run;
%BOXCOXAR(work2.THURSacdcmodelORIGstd,bc,lambdahi=1,lambdal
o=0,nlambda=20,dif=(1,16),const=10,out=bcxxbc);
data THURSacdcmodelORIG;
set THURSacdcmodelORIG;
bxcxbc=&BOXCOXAR;
bc=(((bc+10)**bxcxbc)-1)/bxcxbc;
run;
```

```
%BOXCOXAR(work2.THURSacdcmodelORIGstd,ca,lambdahi=4.5,lambd
alo=3.5,nlambda=20,dif=(1,16),const=10,out=bcxxca);
data THURSacdcmodelORIG;
set THURSacdcmodelORIG;
bxcxca=&BOXCOXAR;
ca=(((ca+10)**bxcxca)-1)/bxcxca;
run;
proc print data=THURSacdcmodelORIG;
run;
ods html file = "c:/pppi.html";
proc arima data=THURSacdcmodelORIG;
identify var = bpa(1,16) nlag = 80;
estimate  q=(2 3 4 5 8 16 17)(16 64)noint plot;
forecast lead=0 out = ip1; /*contains the prewhitened input
bpa series */
identify var = bc(1,16) nlag = 80;
estimate  q=(1 2 4 6 15)(16)noint plot;
forecast lead=0 out = ip2; /*contains the prewhitened input
bc series */
identify var = ca(1,16) nlag = 80;
estimate q=(1 2 3 4 7 8 9 10 11 12 13 14 15 16)(16 32)noint
plot;
forecast lead=0 out = ip3; /*contains the prewhitened input
ca series */
identify var = acdc(1 16) nlag =80;
estimate  q=(2 3 4 5 8 16 17)(16 64)noint plot;
forecast lead=0 out = op1; /*contains the prewhitened acdc
corresponding bpa series */
identify var = acdc(1 16) nlag =80;
```

```
estimate  q=(1 2 4 6 15)(16)noint plot;
forecast lead=0 out = op2; /*contains the prewhitened acdc
corresponding bc series */
identify var = acdc(1 16) nlag =80;
estimate q=(1 2 3 4 7 8 9 10 11 12 13 14 15 16)(16 32)noint
plot;
forecast lead=0 out = op3; /*contains the prewhitened acdc
corresponding ca series */
identify var = acdc(1 16) nlag =80 crosscorr = (bpa(1 16)
bc(1 16) ca(1 16));
estimate input =(bpa bc /(1)ca) noint plot;
forecast lead=0 out=op4; /*contains the noise series */
estimate p=(1 2 3 9 15 32) q=(16)input =(bpa bc /(1)ca)
noint plot;
forecast lead = 64 out = opop;
run;
ods html file = off;
proc print data=opop;
run;
%BOXCOXAR(work2.THURSacdcmodelORIGstd,acdc,lambdahi=2.5,lam
bdalo=1.5,nlambda=20,dif=(1,16),const=10);
data opop;
set opop;
ACDC = ((exp(log(ACDC*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
l95=((exp(log(l95*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
u95=((exp(log(u95*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
```

```
forecast = ((exp(log(forecast*&BOXCOXAR+1)/(&BOXCOXAR)) -
10)*1599.371948)+3346.158;
run;

proc print data=opop;
run;
```

A.2 SAS program for Identifying the ACDC(t) alternate model:

```
Proc print data=work2.THURSacdcmodelstd;
run;
%BOXCOXAR(work2.THURSacdcmodelstd,acdc,lambdahi=2.5,lambdal
o=1.5,nlambda=20,dif=(1,16),const=10,out=bxcxl)
data THURSacdcmodel;
set work2.THURSacdcmodelstd;
bxcxacdc=&BOXCOXAR;
acdc=(((acdc+10)**bxcxacdc)-1)/bxcxacdc;
run;
%BOXCOXAR(work2.THURSacdcmodelstd,Xt,lambdahi=2.5,lambdalo=
1.5,nlambda=20,dif=(1,16),const=10 ,out=bxcxx)
data THURSacdcmodel;
set THURSacdcmodel;
bxcxXt=&BOXCOXAR;
Xt=(((Xt+10)**bxcxXt)-1)/bxcxXt;
run;
proc print data=THURSacdcmodel;
```

```
run;
ods html file = "c:/pppi.html";
proc arima data=THURSacdcmodel;
identify var = Xt(1,16) nlag = 80;
estimate  q=(4 5 6 7 8 9 11 12 14 16)(16)noint plot;
forecast lead=0 out = ip1; /*contains the prewhitened input
Xt series */
identify var =acdc(1,16) nlag = 80;
estimate  q=(4 5 6 7 8 9 11 12 14 16)(16)noint plot;
forecast lead =0 out=ip2;/*contains the prewhitened output
acdc series */
identify var = acdc(1 16) nlag =80 crosscorr = (Xt(1 16));
estimate input =(Xt) noint plot;
forecast lead=0 out=ip3; /*contains the noise series */
/*estimate p=(1 2 3 9 15 16 23 32) q=(16)input =(Xt) noint
plot;*/
estimate q=(1 2 3 9 13  15)(16 48)input =(Xt) noint plot;
forecast lead = 64 out = opop;
run;
ods html file = off;
proc print data=opop;
run;
%BOXCOXAR(work2.THURSacdcmodelstd,acdc,lambdahi=2.5,lambdal
o=1.5,nlambda=20,dif=(1,16),const=10);
data opop;
      set opop;
        ACDC    =    ((exp(log(ACDC*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
```

```
        l95=((exp(log(l95*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
        u95=((exp(log(u95*&BOXCOXAR+1)/(&BOXCOXAR))-
10)*1599.371948)+3346.158;
      forecast                                          =
((exp(log(forecast*&BOXCOXAR+1)/(&BOXCOXAR))            -
10)*1599.371948)+3346.158;
      run;


proc print data=opop;
run;
```

A.3 SAS program for Identifying the Path66(t) model:

```
Proc print data=work2.THURSpath66model;
run;
%BOXCOXAR(work2.THURSpath66model,acdc,lambdahi=2,lambdalo=1
,nlambda=20,dif=(1,16),const=5000,out=bxcxl)
data THURSpath66model;
set yoohoo.THURSpath66model;
bxcxacdc=&BOXCOXAR;
acdc=(((acdc+5000)**bxcxacdc)-1)/bxcxacdc;
run;
%BOXCOXAR(work2.THURSpath66model,path66,lambdahi=3.25,lambd
alo=2.25,nlambda=20,dif=(1,16),const=5000 ,out=bxcxx)
data THURSpath66model;
```

```
set THURSpath66model;
bxcxpath66=&BOXCOXAR;
path66=(((path66+5000)**bxcxpath66)-1)/bxcxpath66;
run;
proc print data=THURSpath66model;
run;
ods html file = "c:/pppi.html";
proc arima data=THURSpath66model;
identify var = acdc(1 16) nlag = 80;
estimate  q = (1 3 6 8 9 13 14 15 22)(16 64 80) noint plot;
/*estimate  q = (1 2 3 6 8 9 13 14 15 22)(16 64 80) noint
plot;*/
forecast lead=0 out = ip1; /*contains the prewhitened input
acdc series */
identify var =path66(1 16) nlag = 80;
estimate  q = (1 3 6 9 13 15 22)(16 64 80) noint plot;
forecast lead =0 out=ip2;/*contains the prewhitened output
path66 series */
identify var = path66(1 16) nlag =80 crosscorr = (acdc(1
16));
estimate input =(acdc) noint plot;
forecast lead=0 out=ip3; /*contains the noise series */
/*estimate p=(1 2 3 9 15 16 23 32) q=(16)input =(acdc)
noint plot;*/
estimate p=(9) q=(1 2 3 4 21 22)(16)input =(acdc) noint
plot;
forecast lead = 10 out = opop;
run;
ods html file = off;
```

```
proc print data=opop;
run;
%BOXCOXAR(work2.THURSpath66model,path66,lambdahi=3.25,lambd
alo=2.25,nlambda=20,dif=(1,16),const=5000);
data opop;
      set opop;
        path66 = (exp(log(path66*&BOXCOXAR+1)/(&BOXCOXAR))-
5000);
        l95=(exp(log(l95*&BOXCOXAR+1)/(&BOXCOXAR))-5000);
        u95=(exp(log(u95*&BOXCOXAR+1)/(&BOXCOXAR))-5000);
      forecast =
(exp(log(forecast*&BOXCOXAR+1)/(&BOXCOXAR)) -5000);
   run;

proc print data=opop;
run;
```

A.4 SAS program for Identifying the BC(t) model:

```
Proc print data=work2.THURSbcmodel;
run;
%BOXCOXAR(work2.THURSbcmodel,bc,lambdahi=1,lambdalo=0,nlambda=20,dif=(1
,16),const=5000,out=bxcxbc)
data THURSbcmodel;
set work2.THURSbcmodel;
bxcxbc=&BOXCOXAR;
```

```
bc=(((bc+5000)**bxcxbc)-1)/bxcxbc;
run;
proc print data=THURSbcmodel;
run;
ods html file = "c:/pppi.html";
proc arima data=THURSbcmodel;
identify var = bc(1 16) nlag = 80;
estimate  q = (1 2 4 6 11 15 19)(16) noint plot;
forecast lead = 10 out = opop;
run;
ods html file = off;
proc print data=opop;
run;
%BOXCOXAR(work2.THURSbcmodel,bc,lambdahi=1,lambdalo=0,nlambda=20,dif=(1
,16),const=5000);
data opop;
      set opop;
        bc = (exp(log(bc*&BOXCOXAR+1)/(&BOXCOXAR))-5000);
        l95=(exp(log(l95*&BOXCOXAR+1)/(&BOXCOXAR))-5000);
        u95=(exp(log(u95*&BOXCOXAR+1)/(&BOXCOXAR))-5000);
      forecast = (exp(log(forecast*&BOXCOXAR+1)/(&BOXCOXAR)) -5000);
   run;

proc print data=opop;
run;
```

Appendix B

List of BPA generators used in the study and their name plate capacities:

Table C-1 BPA generators and name plate capacities

Generator	Nameplate Rating (MW)
Bonneville	1077
Chief Joseph	2458
Grand Coulee	6795
Big Cliff	18
Ice Harbor	603
John Day	2160
Little Goose	810
Lost Creek	49
Lower Granite	810
Lower Monumental	810
McNary	980
The Dalles	1808
Lookout Point	120
Hills Creek	30
Green Peter	80
Foster	20
Dexter	15
Detroit	100
Cougar	25