To Kenny

Title: Time Series Forecasting Methodologies for Electricity Supply Systems

Candidate: Fiona Murray, Dip. App. Sci., B.Sc., M.Sc.

Ph.D. Thesis

University Name: Dublin City University

Supervisor: Dr. John Ringwood

School: School of Electronic Engineering

Month and Year of Submission: February 1996

I hereby certify that the material, which I now submit for assessment on the program of study leading to the award of Ph.D. is entirely my own work and has not been taken from work of others save and to the extent that such work has been cited and acknowledged with the text of my work

Signed: Fiona Murray ID No.: 92700632

Date: 07/02/1997

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. John Ringwood, for his encouragement, patience and guidance over the past four years. I would also like to thank the Electricity Supply Board, Ireland for the use of their data and their helpful advice. Thank-you control systems group. I would also like to acknowledge the School of Electronic Engineering for funding this work. Finally, thanks to my parents for their never ending support and encouragement.

Contents

Title: Time Series Forecasting Methodologies for Electricity Supply Systems

Author: Fiona Murray

ABSTRACT

Forecasting is an essential function in the electricity supply industry. Electricity demand forecasting is performed on number of different time-scales depending on the function for which they are required. In the short term (hourly) forecasts of electricity demand are required for the safe and efficient operation of the power system. Medium term forecasts (weekly) are needed for economic planning and long term (yearly) forecasts are required for deciding on system generation and transmission expansion plans. In recent years the electricity supply industry in some countries has undergone significant changes mainly due to a levelling off in the growth of electricity demand and also due to technological advances. There has been a move toward the existence of a number of smaller generating companies and the emergence of a competitors market has resulted. These changes in the structure of the industry have led to new requirements in the area of forecasting, where forecasts are now required on a small time-scale over a longer forecasting horizon, for example, the production of hourly forecasts over a period of a month.

The thesis presents a novel approach to the solution of the production of short term forecasts over a relatively long term forecast horizon. The mathematical formulation of the technique is presented and an application procedure is developed. Two applications of the technique are given and the issues involved in the implementation investigated. In addition, the production of weekly electricity demand forecasts using the optimal form of the available weather variables is investigated. The value of using such a variable in cases where it is not a dominant influencing factor in the system is assessed. The application of neural networks to the problem of weekly electricity demand forecasting is examined. Neural networks are also applied to the problem of the production of both aggregate and disaggregate electricity sales forecasts for up to five years in advance. Conclusions regarding the methodologies presented in the thesis are drawn and directions for future works are considered.

CHAPTER 1

INTRODUCTION

1.1 Description of Research Area

Forecasting can be an important function in many areas of science, engineering and economics. The focus of this thesis is on forecasting problems in the electricity supply industry. Electricity demand forecasting is important for both power system operation and planning. Different areas within the power utility will have different forecast requirements that may depend on the characteristics of the power system itself, power utility policy and also on the market environment in which the power system operates. On the whole, electricity demand forecasting can be divided into three ranges depending on the forecast period. Short term forecasts are usually performed on an hourly basis and are intended to be valid for a few hours up to a few days in advance. They play an important role in the daily operation of the power system and are required for functions such as unit commitment, economic dispatch and load management. Medium term forecasts are performed on a weekly or monthly time-scale and are valid for up to one to two years in advance. This type of forecast is necessary for activities such as planning fuel procurement, scheduling unit maintenance, revenue assessment and financial and market planning. Long term forecasts are normally required for a forecast horizon of five to twenty years. They are essential for the long term planning of capacity requirements (system generation and transmission expansion), market, strategic and financial planning and, if in operation, the assessment of demand side management.

An important aspect of electricity demand forecasting on any time-scale is the analysis of the relationship between electricity demand and various influencing factors. Electricity demand is highly influenced by several variables, such as weather, socio-economic and demographic variables. Moreover, different variables are pertinent on different time-scales. For example, an economic factor such as GDP is not going to influence short term electricity demand but does influence long term electricity demand, whereas in contrast weather is an influencing factor on short and medium term demand only. In addition, electricity demand is characterised by temporal factors, for example the demand during the day is different to the night-time demand and weekday and weekend demand also differ. Public holidays and special events, such as industrial disputes, elections and major television events, can also have a significant effect on electricity demand on an hourly and daily time-scale. A common practice in this

case is the use of a 'standard day profile', where previous knowledge based on the behaviour of such days is employed to cater for the occurrence of such an event. Table 1.1 summarises the type of influencing variables relevant on each time-scale.

Table 1.1 Electricity demand influencing variables on different time-scales

Time-Scale	Influencing variables
hourly	weather and time of day
daily	weather, day of week and special events
weekly	weather and week of year
yearly	socio-economic and demographic

It is often of benefit to break the electricity demand system down into individual sectors, where this is usually performed at the yearly time-scale. The most common disaggregation involves splitting the system into an industrial sector, a domestic sector and a commercial sector. The advantage of performing this disaggregation is that different factors influence the different sectors and thus analysis at the disaggregate level can provide a greater understating of the system. For example, in some electricity systems the behaviour of the price of electricity would be have a significant effect on the behaviour of the industrial and domestic sectors but would have a less significant effect on the commercial sector.

There are three main agents in the electricity supply industry: the generators, the distributors and the consumers. In some markets there is a middle man between the distributors and the consumers. Depending on the characteristics of the market environment in which the power system operates there may be a number of distributors in the market, however there is traditionally only one generating company. In some cases the same power utility performs both the generation and the distribution, where this is often the case in a small isolated island network. The main reason for the generators monopoly is caused by increasing returns to scale, that is as output increases unit costs fall and there is room for only one producer in the market. However, in recent years, the overall structure of the electricity supply industry of some countries in the developed world has undergone dramatic changes. Due to the relative slowing down of the growth of electricity demand there has been a move away from the construction of large generating units that have become uneconomical. Instead there has been a move toward smaller installations thus leading to a number of producers and the emergence of a competitor's market. In some cases this has led to operation of a power Pool system. In this system a number of distributors or large consumers, such as chemical companies or metal

producers buy, or bid for, electricity from the generators through the medium of a Pool. The Pool is often classified as a "day-ahead spot market" where transactors specify supply or demand for every half hour of the day 24 hours in advance. In such a market there is an onus on the generators to bid their plant in such a manner to provide an economic return and the ability of a generating company, or supplier, to accurately forecast the balance between potential income and direct energy cost is crucial. To achieve this a forecast of demand *and* of Pool Price are required. These forecasts are required on a half-hourly basis for intervals of one-day-ahead so as to participate in bidding and on a half-hourly basis for intervals of one-year-ahead for planning annual budgetary requirements. Therefore, in contrast to the short term forecasting case described above in this situation a short term forecast of electricity demand is required over a long term forecast horizon. In some cases, for example in Norway, there is a one-day-ahead spot market and also a weekly bidding market. In the weekly case the market participants submit bids for loads covering one week, where these can cover weekly periods one year into the future.

1.2 Motivation for Research

A significant motivating factor for the production of accurate electricity demand forecasts on any time-scale is to save costs. The ultimate aim of an electricity utility is to supply the customer entirely while ensuring economic operation of the power system. On a short time-scale (hourly) this involves matching generation supply with customer demand as closely as possible. The achievement of such an objective is crucially dependent on the forecast of the consumers electricity requirement, where even a moderate degree of forecast variance can be costly. A good example is that quoted by Bunn and Farmer (1985), of the U.K. Central Electricity Generating Board, where in 1984 an improvement in forecasting accuracy of 1% was estimated to yield a saving in operating costs of approximately £10 million per year. On a medium and long-time-scale accurate forecasts of electricity demand allows for the optimal economic planning of the power system, an example of this would be in the area of fuel purchasing policies.

The recent and ongoing changes in the market structures of the electricity supply industry of the developed world has resulted in the emergence of new forecasting problems in the area of electricity demand forecasting. Specifically, the production of medium and long term forecasts have become increasingly critical and while the production of short term forecasts are still required they are now needed over a much longer forecast horizon. The focus of attention in the past was on the production of accurate short term forecasts and although

medium and long term forecasting were important functions the substantial bulk of research was in the area of short term forecasting. This thesis addresses the problem of medium and long term electricity demand forecasting. In addition, it presents a proposal of a novel technique for the solution the production of short term forecasts over a long forecast horizon. The work is based on the use of linear time series forecasting analysis and also on time series neural network analysis, where the latter has received little attention in the area of medium and long term forecasting.

1.3 Main Thesis Contributions

The main contributions of this thesis may be summarised as follows:

1. Proposal of a novel multi-time-scale integration technique for the production of long range electricity demand forecasts based on a short sampling interval (Chapter 5).

2. Development of an application procedure for the implementation of the multi-time-scale integration technique (Chapter 5).

3. Application of neural networks to the weekly electricity demand forecasting problem (Chapter 4).

4. Applications of neural networks to forecast yearly aggregate and yearly disaggregate electricity sales (Chapter 4).

A portion of the work conducted in the thesis has been published or has been submitted for publication, where details of the publications are as follows:

- Murray, F and Ringwood, J., "Improvement of Electricity Demand Forecasts Using Temperature Inputs", *Simulation Practice and Theory*, **2** (1994), pp. 121- 139.
- Ringwood, J and Murray, F, "Forecasting of Weekly Electricity Demand Using Neural Networks", in Proc. *IDSPCC*, (Dublin,1996), pp. 349- 356.
- Murray, F. and Ringwood, J., "Integration of Multi-Time-Scale Models in Time Series Forecasting", *Journal of Forecasting*, submitted Sept. 1994.
- Murray, F, Ringwood, J., and O'Brien, K. "Yearly Electricity Sales Forecasting Using Neural Networks", *Neural Computing and Applications*, submitted Jan. 1996.

1.4 Thesis Structure

Initial development is in the area of weekly electricity demand forecasting using linear models with exogenous weather variables. The work then goes on to apply neural networks to the problem of weekly electricity demand and yearly electricity sales forecasting. Finally, the multi-time-scale integration technique is developed and then applied to a number of applications, where both the linear and nonlinear analysis of the previous chapters form a basis for these applications.

Chapter 2 provides an overview of the time series forecasting methods relevant to the work carried out in the thesis. It also simultaneously reviews the literature available on these forecasting techniques, with particular emphasis on works that deal with applications in the electricity supply industry. In addition, it examines a number of diverse standard time series, taken from the literature, with the objective of exhibiting the various characteristics that a time series may posses and discusses the parameters that the properties of a time series may depend upon.

The objective of the work carried out in Chapter 3 is twofold. Primarily it assesses the use of an exogenous weather variable when forecasting weekly electricity demand and evaluates the use an appropriate form of this weather variable. Secondly, it determines a suitable linear statistical model to represent weekly electricity demand for use in the multi-time-scale integration weekly/yearly application dealt with in Chapter 6.

The purpose of work conducted in Chapter 4 is to assess the capability of neural networks to forecast weekly electricity demand and yearly electricity sales. It involves the development of different neural network approaches based on the linear modelling techniques dealt with in Chapter 3. One approach involves the use of linear forecasting models to determine the input structure to the neural network and the second is a hybrid linear-neural network modelling approach. The yearly neural network analysis provides a basis for an analysis conducted in the weekly/yearly multi-time-scale integration application presented in Chapter 6.

Chapter 5 develops the proposed multi-time-scale integration technique. It presents the mathematical formulation of the technique and discusses a number of the solution parameters. Furthermore, it develops a multi-time-scale integration methodology and applies this procedure to a standard time series from the literature. The objective here is to discuss the issues involved in the use of the technique and to illustrate its application on an alternative time series.

Chapter 6 is purely applications based and deals with two multi-time-scale integration electricity demand applications. The first of these is a weekly/yearly application and the second an hourly/daily application. It employs the methodology developed in Chapter 5 and in addition relies on work carried out in Chapter 3 and Chapter 4.

Chapter 7, in conclusion, assesses and evaluates the proposed forecasting methodologies. Guidelines for the practical implementation of the multi-time-scale integration technique are given and directions for future work in this area are discussed.

CHAPTER 2

Overview of Existing Time Series Forecasting Methodologies

2.1 Introduction

This chapter provides an overview of the principal time series modelling techniques specifically employed for the work conducted in this thesis. The main areas of interest are statistical linear analysis and neural network analysis. In addition, the chapter provides a synopsis of the literature available on time series forecasting methods. There is a vast amount of published material on both the theoretical and practical aspects of time series analysis and it would be impractical to include a complete survey of the subject here. The main objective is to review works which specifically deal with applications in the electricity supply industry, and also to review works which employ forecasting techniques that are relevant to the electricity demand forecasting methods proposed in the thesis. However, in order to provide some form of general overview of time series methods, Section 2.2 reviews some of the works which deal with the comparison of a number of forecasting techniques, mainly through the realm of forecasting competitions.

2.2 Introduction to Time Series Forecasting Analysis

Forecasting techniques may be divided into two main classes: qualitative and quantitative. Qualitative techniques are conjectural in nature and are mainly based on expert judgement. Methods include the Dephi technique (Weber, 1990) and the Subjective Curve Fitting technique (Bowerman and O'Connell, 1987). Quantitative techniques usually involve the analysis of historical data in an attempt to make future predictions of a variable of interest. The majority of forecasting techniques are extrapolative in nature and are based on time series data. There are many different time series forecasting procedures available, ranging from simple techniques, such as moving average or exponential smoothing (Gardner, 1985) approaches, to relatively complex methods, such as the Box-Jenkins approach (1976) or the Bayesian forecasting approach (Harrison and Stevens, 1971). An extensive amount of literature has been published on the subject of time series analysis, however there are few works which provide a comprehensive guide to the available forecasting methods. A number of forecasting competitions have been carried out in an attempt to collate and assess the various time series methods. Fildes *et al* (1991) compares, contrasts and assesses the impact

time series forecasting of four such competitions which were conducted in the seventies and eighties. He concludes that the competition which had the greatest impact on the forecasting community appears to be the M-competition conducted by of Makridakis *et al* (1982). This competition distinguished 24 different forecasting procedures which were applied to 1001 different time series taken from the business, economic, finance and social sciences sectors and were based on yearly, quarterly and monthly time scales. These procedures may be broadly classified according to the following methods: naive method (Hibon, 1984), the Box-Jenkins approach (Box-Jenkins, 1976); adaptive estimation procedures (Carbone *et al*, 1984); Bayesian forecasting (Harrison and Stevens, 1971; Fildes, 1984); variants of exponential smoothing (Gardner and McKenzie, 1989); regression methods (Bowerman and O'Connell, 1987); Lewandowski's FORSYS method (1982); Parzen ARARMA method (1982) and combining forecasts (Winkler, 1983). The main conclusion of the competition is that different factors (time scale of data, type of time series, forecasting horizon) affect forecasting accuracy and that different methods perform differently under different conditions. Also, combining forecasts of a few methods improves overall forecasting accuracy over and above that of the individual forecasting methods. A detailed discussion of the M-competition results (Makridakis, 1984) examines the forecasting accuracy achieved and looks at the different methods individually but do not show how any of the various methods may be related. Harvey (1984) aims to address this problem when he presents a unified framework within which the different forecasting procedures may be placed. His paper deals with the forecasting procedures used within the M-competition and also with the state space structural time series models. A very recent competition carried out by Weigend and Gershenfeld (1992) highlights the emergence of neural networks as a valuable tool in time series forecasting applications. Six time series of varying attributes, such short/long, stationary/nonstationary, linear/nonlinear, clean/noisy and natural/synthetic were used in the competition, where the majority of contributions, and indeed the best predictions, were obtained using neural networks. The contrast between the mainly linear modelling techniques of the M-competition, carried out in the eighties, and the nonlinear modelling techniques of the Weigend and Gershenfeld competition is notable, where significantly greater computing power was utilised in the latter competition. This emphasises the effect of the widespread availability of powerful personal computers, for running simulations of neural networks, on time series forecasting techniques.

2.3 Time Series Examples

Before proceeding any further, examples of a variety of time series and the characteristics which they may exhibit are presented. The first example given in Figure 2.1 shows a time series of reported purse snatching in the Hyde Park area of Chicago (Janecek and Swift, 1993). The observations are recorded on days 28 days apart and run from January 1968 to September 1973. Graphical examination of the profile suggest that the level of purse snatching is fluctuating around some constant level.

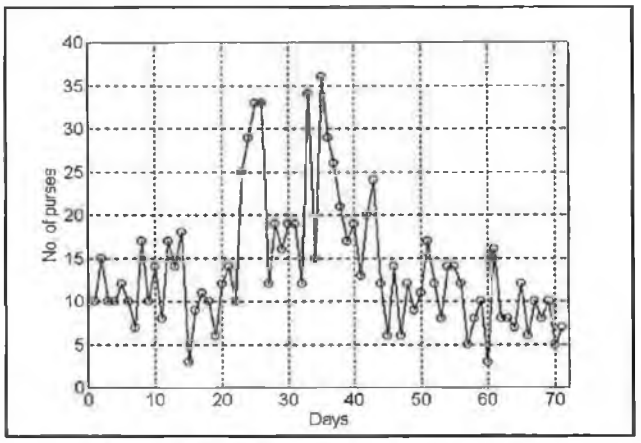

Figure 2.1 Purse snatching in Chicago

In contrast, Figure 2.2 shows the time series of the population of the USA, at ten-year intervals, from 1790 to 1980 (Brockwell and Davies, 1987).

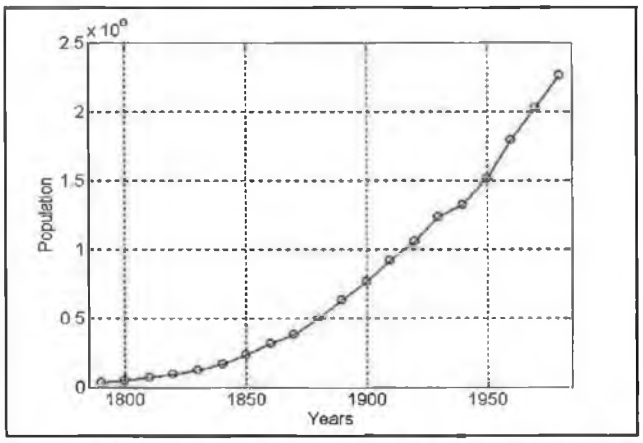

Figure 2.2 Population of the USA at ten-year intervals

It can be seen that this time series has a rising trend. To understand what is meant by the term *trend* of a time series consider the profile given in Figure 2.3 which shows monthly international airline passenger data, 1949 - 1960 (Box-Jenkins, 1976).

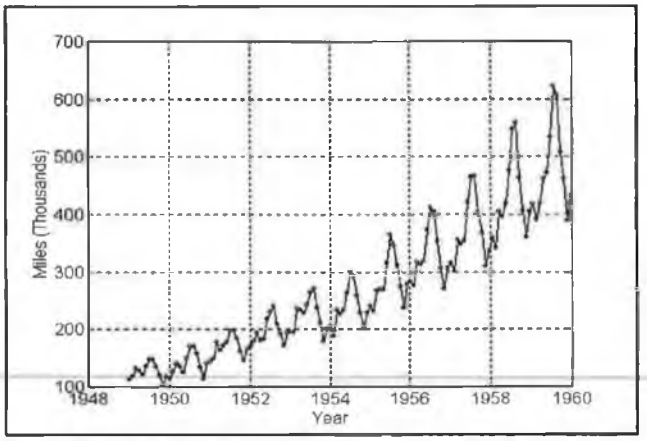

Figure 2.3 Monthly number of passenger air miles flown

This series has a rising trend and seasonality, with a seasonal length equal to 12 (corresponding to 12 months). The trend of the time series may be defined as the low frequency behaviour, or long term movement, of the series which changes relatively slowly over time. Conversely, the seasonal component may be defined as the relatively high frequency behaviour of the series which repeats every seasonal period *s*. The seasonal component may be examined further by over-plotting each year on the same graph as shown in Figure 2.4.

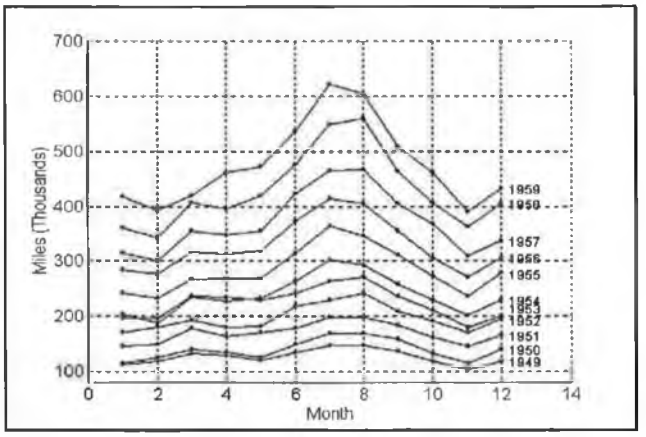

Figure 2.4 Over-plot of passenger air miles flown

This figure shows that the pattern of repetition is consistent and also that the amplitude of the seasonal pattern is steadily increasing each year along with the average level of the series determined by the trend. Many time series have cyclical patterns that do not repeat within a year; the series in Figure 2.5, which shows the Wolfer Annual Sunspot Number 1770 - 1869 (Box-Jenkins, 1976), is an example of this. The sunspot series has a strong cyclical component with a period of approximately equal to 11 years.

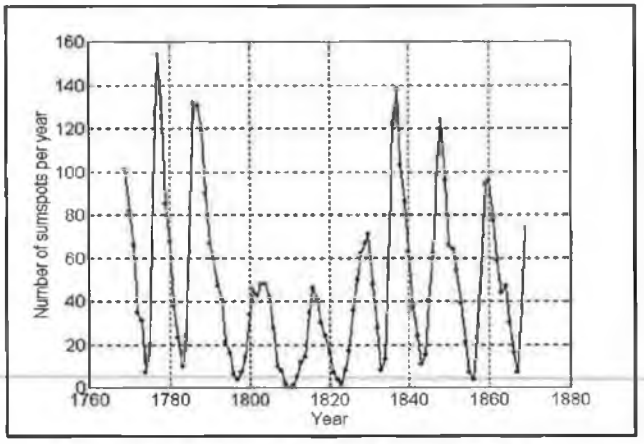

Figure 2.5 Wolfer Annual Sunspot Numbers

Each of the above time series exhibit different characteristics in response to the vastly different environments in which they exist. Different factors within the environment may influence the short and long term behaviour of the time series. An example of a short term effect on a series is the seasonal effect on the airline data (Figure 2.3), where the number of passenger miles flown increases during the summer months in each year, largely related to the seasonal holiday effect. In contrast, the long term effect of technological advances and changing social attitudes increases the number of people travelling abroad each year which results in a rising trend over time. Therefore, various factors, such as temporal, demographic economic and social, may affect how a series will evolve over time, where these factors are referred to as *exogenous* variables.

Mathematical models may be constructed to represent the series where the current value of the time series is based solely on past observations of the series; this is termed a *univariate* model. Therefore, the current value of the time series $y(k)$ may be represented by the following equation

$$y(k) = f\big(y(k-1), y(k-2)\ldots\ldots, y(k-na)\big) \tag{2.1}$$

where f is some function describing the time series and na is the number of past values of the time series that $y(k)$ depends upon. However, this type of model may be restrictive because of the exclusion of exogenous variables which have a significant influence on the time series of interest. A more appropriate model may be of the following form:

$$\begin{aligned} y(k) = g\big(&y(k-1), y(k-2)\ldots\ldots, y(k-na), u_1(k)\ldots\ldots u_1(k-nb_1), \\ &u_2(k)\ldots\ldots u_2(k-nb_2), \ldots\ldots\ldots, u_{nu}(k)\ldots\ldots u_{nu}(k-nb_{nu})\big) \end{aligned} \tag{2.2}$$

where g is some function describing the time series and u_1, u_2,....u_m are exogenous variables that have an effect on the time series of interest $y(k)$, with nb_1, nb_2,....nb_{nu} the number of past values of the exogenous variables u_1, u_2,....u_{nu} affecting $y(k)$ respectively. For example, Figure 2.6 show a plot of the seasonally adjusted quarterly UK employment figures from 1963 to 1979 with the corresponding output index (1980 = 100) in UK manufacturing from 1963 to 1979 also given on the same graph. It is suggested that employment can be regarded as being dependent on manufacturing firms' output expectations (Harvey, 1989).

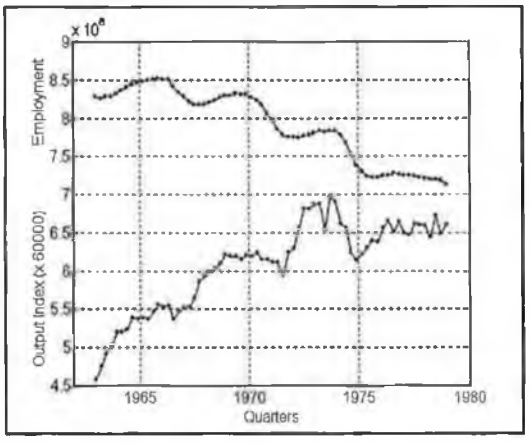

Figure 2.6 No. Employed in UK and output in UK manufacturing

Comparison of the short term movements (1 to 5 years) of the two series show that they are correlated. For example, the behaviour of the series from Quarter 1 in 1963 to Quarter 4 in 1966 is such that an increase in manufacturing output yields a corresponding increase in employment. A similar pattern can be seen from Quarter 1 in 1967 to Quarter 4 in 1971 and again from Quarter 1 in 1972 to Quarter 2 in 1976. From Quarter 3 in 1976 to the end of the series there appears to be a levelling off in the manufacturing output. However, overall the level of the manufacturing output does not account for the long term downward decrease in employment. This downward trend is in response to a dramatic improvement in technical progress over the period in question which resulted in a decrease in manpower requirements. A mathematical model used to represent the employment series could be based solely on previous observations of employment but could also include the effect of the manufacturing output and the rate of change of technology. The latter is difficult to measure and alternative methods are required to account for this (Harvey, 1989), however the short term movements may be accounted for by inclusion of the manufacturing output variable in the mathematical model.

The primary objective of the work conducted in the thesis is the construction of mathematical models to represent electricity demand time series, with the ultimate aim of producing a forecast of future values of the series. Analysis which is based solely on the electricity demand time series whose forecast is required is referred to as a *univariate* analysis, whereas a *multivariate* analysis refers to the construction of mathematical models based on the electricity demand times series of interest but also on associated exogenous variable time series. A bivariate analysis refers to the case where only a single exogenous variable is employed in the forecasting model.

The attributes which a time series possess may also depend on the time-scale on which it is based. For example, a variable which is based on a weekly or monthly time-scale may exhibit a trend *and* a seasonal component (due to repetitive behaviour), however the same variable based on a yearly time-scale may only exhibit the trend characteristic. It is possible to deal with the same time series variable on a number of different time scales, however, for practical reasons certain time series variables are only dealt with on particular time-scales, possibly due to measurement constraints. For example, economic variables, such as gross domestic product (GDP) or gross national product (GNP) tend to be based on a quarterly or yearly basis, whereas the share index is recorded on a daily basis. Analysis developed in Chapter 5 and Chapter 6 examine electricity demand time series which are based on different time scales and this analysis is referred to as *multi-time-scale* analysis.

2.4 Classical Linear Time Series Analysis

2.4.1 Background

The classical methods of linear time series analysis are based on two crucial assumptions (Priestly, 1981):

(i) all series are stationary, or can be reduced to stationarity by some simple transformation, such as detrending or differencing.

(ii) all models are linear, that is the observed series can be represented as a linear function of present and past values of an independent "white noise" process (linear stochastic process).

Before proceeding any further a number of terms, such as linear stochastic process, stationary and stationarity transformations, used in the above assumptions are now defined.

2.4.1.1 Stationary Time Series

A stochastic process is a family of random variables $\{y(k): k \in K\}$ defined on a probability space (Ω, F, P). In time series analysis the index set K is a set of time points, usually equal to $\{0, \pm1, \pm2,....\}$, $\{1, 2, 3,.....\}$, $[0,\infty]$ or $(-\infty, \infty)$. For each fixed $k \in K$, $y(k)$ is a function on the set Ω and for each fixed $\omega \in \Omega$, $y(k,\omega)$ is a function on K. The functions $\{y(k,\omega),\omega \in \Omega\}$ on the index set K are known as *realisations* or a *sample-paths* of the process $\{y(k): k \in K\}$. The collection of all realisations is called the *ensemble* of realisations. The process $\{y(k)\ k \in \mathbf{Z}\}$ with index set $\mathbf{Z} = \{0, \pm1, \pm2,....\}$ is *strictly stationary* if the joint distributions of

$$(y(1), y(2),.....,y(N))^T \text{ and } (y(1+\tau), y(2+\tau),.....,y(N+\tau))^T$$

are the same for all positive integers N and τ. Thus, it is required that the distributions over the ensemble are identical. Intuitively, strict stationarity means that the graphs over two equal lengths of a realisation of a time series should exhibit similar statistical characteristics. In most applications the form of the distribution on which the time series is defined is not known and thus it is common practice to deal with only the first two moments of the series. Based on this a time series may be defined to be *weakly stationary* if the following conditions are satisfied:

(i) $E(y(k)) = \mu$ is independent of k.

(ii) $E[\ (y(k) - \mu)^2\ ] = \text{var}(y(k))$ is independent of k. (2.3)

(iii) $E[\ (y(k) - \mu)(y(k - \tau) - \mu)\ ] = R(\tau)$ depends only on τ.

where $R(\tau)$ is the *autocovariance function* defined as follows:

$$R(\tau) = \text{cov}(y(k), y(k - \tau)) = E[\ (y(k) - \mu)\ (y(k - \tau)-\mu)\ ] \qquad \tau =...,-2,-1,0,1,2,... \qquad (2.4)$$

$R(\tau)$ measures the covariance between pairs of values of the process separated by an interval τ, where τ is usually termed the lag. The terms *stationary in the wide sense*, *covariance stationary*, *second order stationary* and *stationary* are also used to describe a weakly stationary time series (2.3), where the term "stationary" is used throughout the thesis. The fact that for a stationary process the mean, variance and covariance do not depend on τ means that they may be estimated from a single realisation as follows (Harvey, 1989):

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} y(k) \qquad (2.5)$$

$$\text{var}(y(k)) = \hat{R}(0) = \frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{\mu})^2 \tag{2.6}$$

$$\hat{R}(\tau) = \frac{1}{N}\sum_{k=\tau+1}^{N}(y(k) - \hat{\mu})(y(k-\tau) - \hat{\mu}) \qquad \tau = ...-2,-1,0,1,2.... \tag{2.7}$$

Equations (2.5), (2.6) and (2.7) are referred to as the sample mean, sample variance and sample autocovariance function respectively. The sample autocovariance function (2.7) can be used to describe the dynamic properties of a time series and it is a well employed tool in classical linear time series analysis. However, the normalised sample autocovariance function is usually used, where this is called the *sample autocorrelation function* (SACF) and is defined as follows:

$$\hat{\rho}(\tau) = \frac{\hat{R}(\tau)}{\hat{R}(0)} \qquad \tau = ...,-2,-1,0,1,2.... \tag{2.8}$$

Intuitively, $\rho(\tau)$ may be thought of as a measure of the "similarity" between a realisation of $y(k)$ and the same realisation shifted by τ units. For a real series $\rho(\tau) = \rho(-\tau)$ and thus the plot of the SACF as a function of τ is generally examined for positive values of $\tau = 1, 2,.....$ The SACF (2.8) may be used to calculate the *sample partial autocorrelation function* (SPACF) which represents the sample autocorrelations of the time series observations separated by a lag of τ with the effects of the intervening observations eliminated. The SPACF is defined as follows (Bowerman and O'Connell, 1987):

$$\begin{aligned} &\alpha(\tau,\tau) = \rho(\tau) \qquad \text{if } \tau = 1 \\ &\alpha(\tau,\tau) = \rho(\tau) - \sum_{j=1}^{\tau-1}\alpha(\tau-1,j)\alpha(\tau-j) \qquad \text{if } \tau = 2,\ 3,... \end{aligned} \tag{2.9}$$

where

$$\alpha(\tau,j) = \alpha(\tau-1,j) - \alpha(\tau,\tau)\alpha(\tau-1,\tau-j) \qquad j = 1,\ 2,.....\tau-1$$

The autocorrelation function defines the correlation structure *within* a process. However, when dealing with an exogenous variable time series that has an influence on the time series of interest it is useful to measure the correlation *between* two processes. Consider two stochastic processes, $y(k)$ and $u(k)$, k=1, ...,N, the two processes are a *stationary bivariate process* (or that $y(k)$ and $u(k)$ are jointly stationary) if the following hold:

(i) $y(k)$ and $u(k)$ are each stationary as defined in (2.1)

(ii) $E[\ (y(k) - \mu_1)(u(k-\tau) - \mu_2)\ ] = R_{uy}(\tau)$ depends only on τ, with μ_1 the mean of $y(k)$ and μ_2 the mean of $u(k)$

where $R_{uy}(\tau)$ is the *cross covariance function* which describes the correlation structure between the two processes and is defined as follows:

$$R_{uy}(\tau) = \text{cov}(y(k),u(k-\tau)) = E[\ (y(k) - \mu_1)\ (u(k-\tau)\ -\mu_2)\ ] \qquad \tau =...,-2,-1,0,1,2,... \qquad (2.10)$$

The sample cross covariance function may be calculated as follows:

$$\hat{R}_{uy}(\tau) = \frac{1}{N}\sum_{k=\tau+1}^{N}(y(k)-\hat{\mu}_1)(u(k-\tau)-\hat{\mu}_2) \qquad \tau =...-2,-1,0,1,2.... \qquad (2.11)$$

The normalised version of (2.11) is called the *sample cross correlation function* (SCCF) given by:

$$\hat{\rho}_{uy}(\tau) = \frac{\hat{R}_{uy}(\tau)}{\left(\hat{R}_{yy}(0)\hat{R}_{uu}(0)\right)^{1/2}} \qquad \tau =....,-2,-1,0,1,2.... \qquad (2.12)$$

2.4.1.2 Time Series Stationarity Transformations

Stationarity transformations that convert a nonstationary time series to an approximately stationary time series play an important role in classical linear time series analysis. The purpose of the transformation is to produce a new series which is more compatible with the assumption of stationarity. If the time series exhibits a trend or seasonal characteristic then these components are usually removed from the time series using an appropriate transformation.

There are a number of different approaches to the transformation of time series, where one approach is that followed by (Ljung, 1987) which involves the removal of the linear trend and bias from the series, that is the removal of the best straight-line fit. This may be described as follows:

$$\begin{aligned} y^*(k) &= \alpha_0 + \alpha_1 k \\ d(k) &= y(k) - y^*(k) \end{aligned} \qquad (2.13)$$

where α_0 and α_1 are the coefficients of the linear equation and $d(k)$ is the detrended time series.

Alternatively the trend of a time series may be removed using the following differencing transformation (Box and Jenkins, 1976):

$$z(k) = (1 - q^{-1})^d \, y(k) \tag{2.14}$$

where d represents the order of differencing. The transformation in (2.14) may be extended to the case where both trend and seasonality are removed from the time series through the following:

$$z(k) = (1 - q^{-1})^d (1 - q^{-s})^D \, y(k) \tag{2.15}$$

where

- d is the order of nonseasonal differencing
- D is the order of seasonal differencing.
- s is the seasonal length.

The application of the differencing transformations given by (2.14) and (2.15) require the determination of appropriate orders of d and D. The approach adopted in the thesis is the one suggested by Box and Jenkins (1976) that involves the application of the transformation and the subsequent examination of the SACF (2.8) of the differenced time series to check for the presence of nonstationarity. They consider the time series to be nonstationary if the SACF *dies down extremely slowly*, where the precise meaning of the term "dies down extremely slowly" is somewhat arbitrary and is best determined through experience. To illustrate what is meant by this term an example of the SACF of a time series that exhibits a trend and seasonal component, that is a weekly electricity demand time series, is shown in Figure 2.7, where a plot of the original series is given in Figure 2.8, note that this data is scaled for confidentiality reasons.

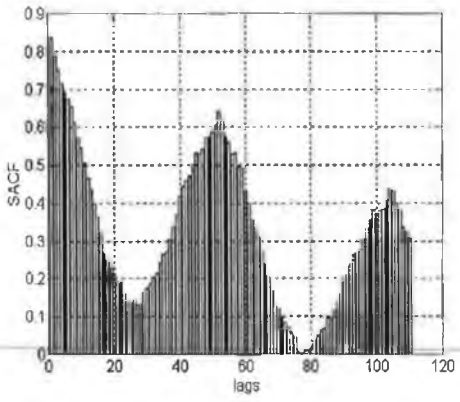

Figure 2.7 SACF of original time series

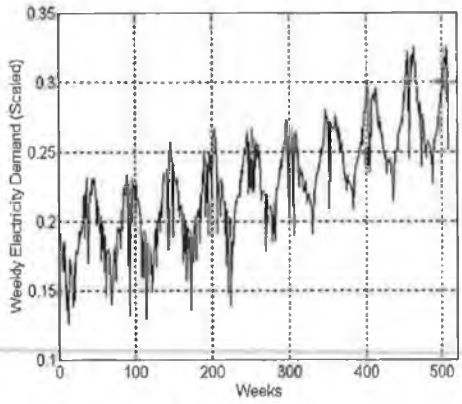

Figure 2.8 Original time series

The SACF exhibits a sinusoidal pattern due to the presence of the seasonal component in the data. Now consider the SACF of the transformed time series using (2.15), with $d = 1$ and $D = 1$. The SACF of the differenced data is given in Figure 2.9 and the differenced time series is shown in Figure 2.10.

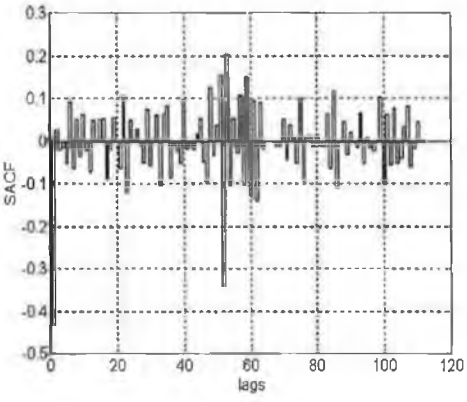

Figure 2.9 SACF of differenced time series

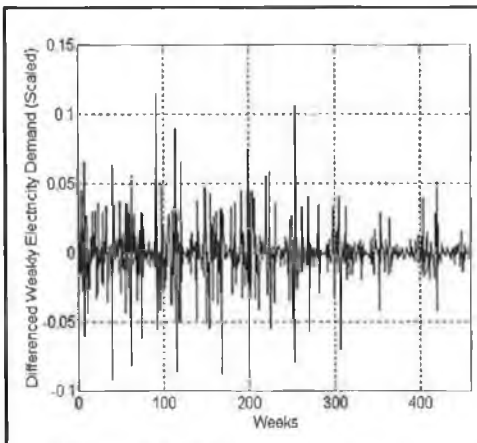

Figure 2.10 Differenced time series

Comparing the SACF of the original (Figure 2.7) and differenced time series (Figure 2.9) shows that the SACF of the differenced series is not dying down slowly and therefore the series is considered to be stationary and thus the differencing transformation with $d = 1$ and $D = 1$ is considered to be suitable transformation for the weekly electricity demand time series.

2.4.2 Classical Time Series Models

The classical time series models that are dealt with in this thesis may be derived from the following general model structure:

$$A(q)y(k) = \frac{B(q)}{F(q)}u(k-b) + \frac{C(q)}{D(q)}\varepsilon(k) \tag{2.16}$$

where

- $y(k)$, $t = 1, 2, ...N$ is the univariate time series.
- $u(k)$, $t = 1, 2, ...N$ is an exogenous variable time series.
- b is the delay between the exogenous variable and univariate time series.
- $\varepsilon(k)$ is a zero mean white noise disturbance term with variance σ_ε^2.
- A, B, C, D, F are polynomials in the delay operator q^{-1}.

2.4.2.1 Classical Univariate Models

The simplest univariate model is the Autoregressive (AR) model; derived from (2.16) and is given as follows:

$$A(q)y(k) = \varepsilon(k) \tag{2.17}$$

where $A(q^{-1}) = 1 + a_1 q^{-1} + ... + a_{na} q^{-na}$. This models represents the current value of the time series $y(k)$ as past values of the time series $y(k)$, $....y(k\text{-}na)$. The AR model may be extended to include a moving average term, where this model is referred to as an Autoregressive Moving Average (ARMA) model and is given as:

$$A(q)y(k) = C(q)\varepsilon(k) \tag{2.18}$$

where

- $A(q) = 1 + a_1 q^{-1} + ... + a_{na} q^{-na}$ is the autoregressive operator.
- $C(q) = 1 + c_1 q^{-1} + ... + c_{nc} q^{-nc}$ is the moving average operator.

If the ARMA model is used to represent a series that has been differenced using (2.14) then it is referred to as an Autoregressive Integrated Moving Average (ARIMA) model (Box and Jenkins, 1976). Box and Jenkins (1976) also developed a model to represent a series with both trend and seasonal characteristics, where the series is transformed using (2.15) and the autoregressive and moving average operators are split up into separate trend and seasonal component parts. The model is called a Seasonal Autoregressive Integrated Moving Average (SARIMA) model and is given as follows:

$$A(q)z(k) = C(q)\varepsilon(k) \tag{2.19}$$

where

- $z(k)$ is the differenced time series (2.15)
- $A(q) = (1 - a_1 q^{-1} - ..... - a_p q^{-p})(1 - a_{1,s} q^{-1s} - a_{2,s} q^{-2s} ..... + a_{P,s} q^{-Ps})$, with p the order of the nonseasonal autoregressor and P the order of the seasonal autoregressors.
- $C(q) = (1 - c_1 q^{-1} - ..... - c_m q^{-m})(1 - c_{1,s} q^{-1s} - c_{2,s} q^{-2s} ..... + c_{M,s} q^{-Ms})$, with m the order of the nonseasonal moving average regressor and M the order of the seasonal moving average regressor.

2.4.2.2 Classical Multivariate Models

The simplest multivariate model structure based on $y(k)$ and on an exogenous variable time series $u(k)$ is derived from (2.16) and is called an Autoregressive with Exogenous variable (ARX) model, where this is given as follows:

$$A(q)y(k) = B(q)u(k-b) + \varepsilon(k) \tag{2.20}$$

where

- $A(q) = 1 + a_1 q^{-1} + ... + a_{na} q^{-na}$
- $B(q) = b_0 + b_1 q^{-1} + ... + b_{nb} q^{-nb}$
- b is the delay between $u(k)$ and $y(k)$

The Autoregressive Moving Average with Exogenous variable (ARMAX) model derived from (2.16) is given as:

$$A(q)y(k) = B(q)u(k-b) + C(q)\varepsilon(k) \tag{2.21}$$

where

- $A(q) = 1 + a_1 q^{-1} + ... + a_{na} q^{-na}$
- $B(q) = b_0 + b_1 q^{-1} + ... + b_{nb} q^{-nb}$
- $C(q) = 1 + c_1 q^{-1} + ... + c_{nc} q^{-nc}$
- b is the delay between $u(k)$ and $y(k)$

Box and Jenkins (1976) developed a transfer function model that includes an explicit model to represent the noise or residuals of the system. The model is referred to as the Box-Jenkins Transfer Function (BJTF) model and is given as follows:

$$y(k) = \frac{B(q)}{F(q)} u(k-b) + N(k) \tag{2.22}$$

where

- $B(q) = b_0 + b_1 q^{-1} + \ldots\ldots + b_{nb} q^{-nb}$
- $F(q) = 1 + f_1 q^{-1} + \ldots\ldots + f_{nf} q^{-nf}$
- b is the delay between $u(k)$ and $y(k)$
- $N(k) = C(q)/D(q)$ is represented by an ARMA model

The above models deal with a single exogenous variable, where it is straightforward to extend each model to the multivariate case. For example, the structure of the BJTF model (2.22) with *nu* exogenous variables is as follows:

$$y(k) = \frac{B_1(q)}{F_1(q)} u_1(k-b_1) + \ldots\ldots + \frac{B_{nu}(q)}{F_{nu}(q)} u_{nu}(k-b_{nu}) + N(k) \tag{2.23}$$

where $B_j(q)$ and $F_j(q)$, $j = 1,\ldots nu$ each have a similar polynomial structure to the $B(q)$ and $F(q)$ polynomials in the BJTF model given by equation (2.22) and b_j, $j = 1,\ldots nu$ are the corresponding delay parameters. The ARX and ARMAX model structures may be expanded in a similar manner (Ljung, 1987).

2.4.3 Classical Time Series Modelling Methodology

The first step involved in any time series analysis is the examination of the plot of the series to identify the characteristics of the profile and also to check for the presence of outliers in the data. In classical linear analysis it is necessary to determine if a transformation is required to convert the time series to a stationary series. The presence of trend or seasonal characteristics may require the use of a stationarity transformation, where the application of such transformations have already been discussed in Section 2.4.1.2. Given a stationary time series the development of a model to represent the time series generally requires the following three basic steps:

- model structure identification
- parameter estimation
- model adequacy checking

The identification of the model structure involves the selection of suitable orders of the polynomials operators. It is desirable to identify a parsimonious model, that is a model with

the smallest possible number of parameters required for an adequate representation of the time series. Once a suitable model structure has been identified it is necessary to estimate the coefficients of the polynomial operators. Finally, the estimated model is checked to determine if it can adequately represent the time series. There are two main approaches to the selection of the model structure for the models given by equations (2.17) to (2.23).

The first approach was developed by Box and Jenkins (1976), where the main tools employed are the SACF (2.8), SPACF (2.9) and SCCF (2.12). This approach involves identifying a model structure using the autocorrelation and cross correlations functions, estimating the parameters of this model structure and testing the adequacy of the model. If the model is found to be inadequate then a new model structure is identified, the parameters estimated and the model checked, where the process is repeated until a parsimonious model is determined.

The alternative approach is to estimate the parameters of a variety of different model structures and compare them according to some chosen criterion, such as Akaike's Information Theoretic Criterion (AIC) (Akaike, 1973) or Rissanen's Minimum Description Length (MDL) criterion (Rissanen, 1983). The adequacy of the selected model structure is then checked, where if this model is found to be inadequate a new set of model structures are estimated and compared.

The difference between the two approaches is that the Box-Jenkins method requires a considerable amount of user interaction, where it may be necessary to repeat the three modelling steps a number of times before a suitable model structure is found. Each of the two approaches for identifying the model structure were employed in the work conducted in the thesis and they are now discussed. The parameter estimation techniques and model adequacy checking are dealt with in Section 2.4.3.3 and Section 2.4.3.4 respectively.

2.4.3.1 Model Structure Identification Using Box-Jenkins Methods

In the case of the univariate models Box and Jenkins (1976) suggest the examination of the SACF (2.8) and SPACF (2.9) to *tentatively* identify the orders of the moving average and autoregressive polynomials respectively. Here "tentative" refers to the use of the functions to aid in the selection of the model structure as part of the iterative modelling procedure described above, where this procedure may have to be repeated a number of times to determine a final parsimonious model. The SACF and SPACF functions are said to *cut off* at the appropriate orders of the polynomials. For example, in the case of the AR model (2.17) the SPACF is said to cut off at lag *na*. In the case of the ARMA model (2.18) the SPACF

cuts off at lag *na* and the SACF at lag *nc*. In the case of the SARIMA model (2.19) it is necessary to examine the function at the nonseasonal ($\tau = 1, \ldots \tau$-3) and also at the seasonal level ($\tau = 1s, 2s \ldots\ldots$) to determine the orders of the polynomials. The SACF is said to cut offs at the following lags:

- *m* at the nonseasonal level
- *M* at the seasonal level

and the SPACF at lags

- *p* at the nonseasonal level
- *P* at the seasonal level

The precise meaning of the term 'cut off ' is somewhat arbitrary, however it is possible to calculate a test statistic for $\rho(\tau)$ at each lag in the SACF and for $\alpha(\tau, \tau)$ at each lag of the SPACF. This test statistic for $\rho(\tau)$ at lags τ=1, 2.....(Bowerman and O'Connell, 1987) is given as follows:

$$t_{\rho(\tau)} = \frac{N^{\frac{1}{2}}\rho(\tau)}{\left(1+2\sum_{j=1}^{\tau-1}\rho^2(j)\right)^{\frac{1}{2}}} \tag{2.24}$$

where *N* is the number of sample points in the time series for which the SACF was calculated. Bowerman and O'Connell (1987) provide the guideline that if the absolute value of the *t*-statistic (2.24) is greater than 1.6 at the nonseasonal level and greater than 2 at the seasonal level then the $\rho(\tau)$ is statistically different from zero and thus there is a cut off at lag τ. In the case of the SPACF the test statistic for $\alpha(\tau, \tau)$ is given as:

$$t_{\alpha(\tau,\tau)} = N^{\frac{1}{2}}\alpha(\tau,\tau) \tag{2.25}$$

where the same guideline used for the SACF is also used for $t_{\alpha(\tau,\tau)}$.

The identification of the BJTF model (2.23) involves the determination of *nb, nf* and *b* through the use of the SCCF (2.12) and the determination of $N(k)$ using the univariate model identification techniques discussed above. The procedure to determine the BJTF model structure may be split into the following two steps:

i. Use the SCCF to identify *b, nb* and *nf* to obtain a *preliminary transfer function model.*

ii. Identify $N(k)$ by examining the SACF and SPACF of the residuals of the estimated preliminary transfer function determined in (i).

Step (i)

The SCCF is calculated between the time series of interest and each of exogenous variable time series. However, since there may be autocorrelation in the time series the true relationship between the two series may not be reflected in the SCCF and it is thus necessary to account for the possible presence of this autocorrelation. This is achieved through 'prewhitening' which involves constructing a univariate SARIMA (2.19) for the exogenous variable time series and using this to filter (or prewhiten) the time series $y(k)$. The behaviour of the SCCF is examined, where the following guidelines may be used to determine the model structure (Bowerman and O'Connell, 1987):

1. The lag at which the first cut off appears is equal to the parameter b.
2. The first cut off in the SCCF will be followed by a dying down pattern, where nb is set equal to the number of lags that reside between lag b and the lags at which this decay begins.
3. If the SCCF decays in a damped exponential fashion, then $nf = 1$ or if it decays in a damped sine wave fashion then $nf = 2$.

Step (ii)

The preliminary transfer function model identified in Step (i) is estimated and the SACF and SPACF of the residuals of this model are examined in order to identify a SARIMA model for $N(k)$.

In the applications dealt with in the thesis it was found that the guidelines to determine the nb and nf orders for the preliminary transfer function models were difficult to apply in practice and it was necessary to test a number of different model structures before appropriate orders could be found. This involved a significant amount of development effort as it was necessary to determine a suitable noise model $N(k)$ for each preliminary transfer function model structure.

Although it is possible to use the SCCF to determine the delay parameter b in the ARX (2.20) and ARMAX (2.21) model structures, there are no equivalent guidelines available for the use of the function to determine the order of the $A(q)$, $B(q)$ and $C(q)$ in these models. The model

structure selection technique described in the next section is usually employed for such models (Ljung, 1987; Soderstrom and Stocia, 1989).

2.4.3.2 Model Structure Identification Using Criterion Functions

This approach involves comparing different model structure performances according to a suitable criterion function. One such criterion function, referred to as the *loss function* is given as follows (Soderstrom and Stocia, 1989):

$$V = \frac{1}{N}\sum_{k=1}^{N}\left(\varepsilon(k)\right)^2 \tag{2.26}$$

where $\varepsilon(k) = y(k) - \hat{y}(k)$ and N is the number of sample points over which the loss function is calculated. The procedure involves selecting a data set on which to estimate the parameters of the model and a different data set on which to compute the loss function. The model structure that yields the minimum value of the loss function is selected. It is possible to calculate alternative criterion based on the loss function value given by (2.26). The AIC (Akaike's, 1973) is a criterion that penalises high order (in terms of parsimony) model structures, where this is calculated as follows:

$$AIC = V * \left(1 + \frac{2np}{\mathrm{N}}\right) \tag{2.27}$$

where V is the loss function (2.26), N is the number of sample points over which the AIC is calculated and np is the number of parameters estimated in the model. The model structure that yields the smallest AIC is selected. Alternatively the MDL (Rissanen, 1983) may be calculated as follows:

$$MDL = V * \left(1 + \log(N)\frac{np}{N}\right) \tag{2.28}$$

where V is the loss function (2.26), N is the number of sample points over which the MDL is calculated and np is the number of parameters estimated in the model. The MDL also penalises model complexity.

2.4.3.3 Parameter Estimation

Once a structure has been selected for the model it is necessary to estimate the coefficients of the polynomial operators in the model. Consider the AR model given by equation (2.17); $y(k)$ may be rewritten as follows:

$$y(k) = \varphi^T(k)\theta + \varepsilon(k) \tag{2.29}$$

where $\varphi^T(k) = (-y(k\text{-}1),.....,-y(k\text{-}na))$ and $\theta = (a_1.....a_{na})^T$. The ordinary least squares estimator of θ is given as follows (Soderstrom and Stoicia, 1989; Fuller, 1996):

$$\hat{\theta} = \left[\frac{1}{N}\sum_{k=1}^{N}\varphi(k)\varphi^T(k)\right]^{-1}\left[\frac{1}{N}\sum_{k=1}^{N}\varphi(k)y(k)\right] \tag{2.30}$$

The error in this estimator is given as:

$$\hat{\theta} - \theta_0 = \left[\frac{1}{N}\sum_{k=1}^{N}\varphi(k)\varphi^T(k)\right]^{-1}\left[\frac{1}{N}\sum_{k=1}^{N}\varphi(k)v(k)\right] \tag{2.31}$$

where θ_0 represents the true parameter vector and $y(k) = \varphi(k)\theta_0 + v(k)$.

Similarly in the ARX model given by (2.20) $y(k)$ may be rewritten as:

$$y(k) = \varphi^T(k)\theta + \varepsilon(k) \tag{2.32}$$

where $\varphi^T(k) = (-y(k\text{-}1)\ .....-y(k\text{-}na)\ u(k\text{-}1).....u(k\text{-}nc))$ and $\theta = (a_1.....a_{na}\ b_1.....b_{nb})^T$ and the ordinary least squares estimator of θ is given by (2.30), with the error in the estimator given by (2.31).

Ordinary least squares is a simple method that can be used to obtain efficient estimators of an autoregressive time series. However, the estimation of a model with a moving average component is less straightforward, where the most common approach used is maximum likelihood estimation (Box and Jenkins, 1976; Brockwell and Davies, 1987; Fuller, 1996). A description of maximum likelihood estimation for time series is given in Appendix A and details pertaining to the use of the technique in relation to the classical linear models are now given.

Consider the ARMA model given by (2.18), if $y(k)$ is a stationary normal series then the log of the likelihood is given by the following (Fuller, 1996)

$$\log L(\psi) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\left|\log \Sigma_{YY}(\psi)\right| - \frac{1}{2} Y^T \Sigma_{YY}^{-1}(\psi) Y \qquad (2.33)$$

where $\psi = (\theta, \sigma_\varepsilon^2)$, $\theta = (a_1 \ldots\ldots\ldots a_{na}\ c_1 \ldots\ldots\ldots c_{nc})^T$, $\mathbf{Y}^T = (y(1), y(2), \ldots\ldots, y(N))$ and $\Sigma_{YY} = \Sigma_{YY}(\psi) = E\{\mathbf{Y}\mathbf{Y}^T\}$. The maximum likelihood estimator of θ can be found by minimising the following:

$$l(\theta) = \frac{1}{N}\left|\mathbf{M}_{YY}(\theta)\right|^{1/N} \mathbf{Y}^T \mathbf{M}_{YY}^T(\theta)\mathbf{Y} \qquad (2.34)$$

where $|\mathbf{M}_{YY}(\theta)| = \sigma_\varepsilon^{-2}\, \Sigma_{YY}(\theta)$.

Maximum likelihood estimation may be also be used to estimate the parameters of a SARIMA model (2.19), where the differences $z(k) = (1-q^{-1})^d (1-q^{-s})^D y(k)$ constitute an ARMA process of order $na = p+sP$ and $nc = m+sM$ in which some of the coefficients are zero and the rest are functions of the $(p+P+m+M)$-dimensional vector $\beta = (a_1 \ldots\ldots a_p\ a_{1,s} \ldots\ldots a_{P,s}\ c_1 \ldots\ldots c_m\ c_{1,s} \ldots\ldots c_{M,s})^T$. The likelihood of $l(\beta)$ of the differences of $y(k)$, that is $z(k)$, can be computed using (2.34) and the maximum likelihood estimator of β is the value that minimises $l(\beta)$. In the case of an ARMAX model (2.21) equation (2.34) can be used to compute the likelihood $l(\alpha)$ for the vector of unknown parameters $\alpha = (a_1 \ldots\ldots a_{na}\ b_1 \ldots\ldots b_{nb}\ c_1 \ldots\ldots c_{nc})^T$, where the maximum likelihood estimator of α is the value that minimises $l(\alpha)$. Similarly, maximum likelihood estimation can be used to estimate the unknown parameters of the BJTF model (2.23), where the vector of unknown parameters in this case also contains the parameters of the ARMA noise model.

2.4.3.4 Model Adequacy Checking

The model adequacy of the univariate and multivariate classical models may be tested through the examination of the residuals of the model. In a well specified model, the residuals should be approximately random. The simplest way to test the randomness of the residuals is to examine them graphically. To illustrate this Figure 2.11 shows a plot of the residuals for the USA population data shown in Figure 2.2, where an AR(2) model was fitted to the detrended data.

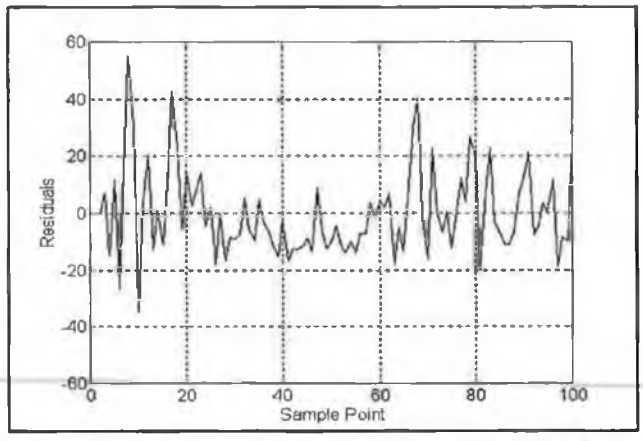

Figure 2.11 Plot of residuals of AR(2) model for population data

The plot shows that the residuals fluctuate around a constant mean and have no recognisable or repeating pattern and they are thus assumed to be random. If the residuals are random then there should be no serial correlation between them, where the SACF (2.8) of the residuals can be used to check for correlation. A method of testing the overall adequacy of a model is to examine a statistic that determines whether the first T sample autocorrelation, considered together, indicate adequacy of the model. This statistic is called the Ljung-Box test statistic and is given as (Bowerman and O'Connell, 1987)

$$Q(\mathrm{T}) = \frac{N}{N+2}\sum_{j=1}^{\mathrm{T}}(N-j)^{-1}\,\rho^{2}(j) \tag{2.35}$$

where N is the number of samples in the time series and $\rho(j)$ is the SACF (2.8) at lag j. The number of lags for which Q is calculated is $T = (N)^{1/2}$. If the model accounts adequately for the time series observations then the residuals should be unrelated and the autocorrelations of the residuals should be small. Thus, Q(T) (2.35) should be small. The larger Q(T) is, the larger are the autocorrelations of the residuals and the more related are the residuals. The distribution of Q(T) is assumed to be approximately chi-squared (Morrison, 1976) with (T - np) degrees of freedom, where np is the number of parameters estimated in the model under consideration. The adequacy of the model is therefore rejected at level α if

$$Q(\mathrm{T}) > \chi_{\alpha}^{2}(\mathrm{T}-np) \tag{2.36}$$

For a multivariate model there should be no serial correlation in the residuals of the model and in addition there should be independence between the residuals and past values of the exogenous variable time series. This may be checked by examining a plot of the SCCF (2.12)

calculated between the residuals and the exogenous variables, where if there are no cut offs in the SCCF then independence is assumed and thus the model is deemed adequate.

2.4.5 Literature Overview

2.4.5.1 Univariate Models

In the literature the ARMA and ARIMA modelling approach has been applied to a wide variety of time series, including forecasting hospital patient movements (Lin, 1989) to forecasting personnel computer product sales (Wu *et al*, 1991). Many books deal with both the theory and implementation of the univariate classical modelling procedure (Pankratz, 1983; Abraham and Ledolter, 1983; Janecek and Swift, 1993).

There are numerous examples of the use of the classical forecasting models for electricity demand time series forecasting (Di Caprio, 1983; Schneider *et al*, 1985; Ackerman, 1985; Demirovic, 1988; Moghrqm and Rahman, 1989; Bakarat *et al*; 1990; Komprej and Dipling, 1990; McCafferty *et al*, 1990; Mbamalu and El-Hawary, 1993). An advantage of the ARIMA modelling approach is its capability of handling seasonal time series through the use of the SARIMA model (2.19). As an example, Komprej and Dipling (1991) show that a SARIMA model produces forecasting results which are 25% more accurate than an exponential smoothing model when predicting the one o'clock peak load of the day in Slovenia. On weekly and monthly time scales seasonal factors have a significant effect on the characteristics of the electricity demand time series. These may be due to weather effects or due to other cyclic events such as heightened industrial activity or regular holidays. Barakat *et al* (1990) use a SARIMA model to predict a two-monthly forecast of peak electricity demand. This time series is interesting in that it is partially influenced by weather factors, mainly through air conditioning, but there are other cyclic and dynamic events that contribute significantly to the peak demand. These cyclic variable loads are due to major religious events, Ramadem, Eid and Hajj, which are related to the lunar calendar and thus move in cycles every Gregorian calendar year. The forecasting accuracy of the results produced by the SARIMA model are significantly improved through the adjustment of the original series to allow for these cyclic events. In this thesis SARI models are applied to two different seasonal weekly electricity demand time series. The main seasonal effects are attributable to different factors, where in one case the effects are due to the weather and in the other case the effects are due weather and also to the cyclical nature of the dairy industry.

2.4.5.2 Multivariate Modelling of Electricity Demand Systems

In the case of electricity demand there are a number of exogenous variables which have an external influence on demand, where different variables are pertinent on different time-scales (see Section 1.1). For electricity demand modelling the most common multivariate methods are based on classical regression models (Train *et al*, 1984; Baker, 1985; Demovric, 1988; Stetson and Stark, 1988; Engle *et al*, 1988, 1992; Haido and Muto, 1994; Al-Zayer, 1996). Examples of the application of BJTF models (2.23) can also be found (Pigott, 1985; McCafferty *et* al; 1990; Harris and Liu, 1993). The availability of significant exogenous variable data is a crucial factor in the development of the multivariate analysis. Another factor is that predictions of future values of the exogenous variables may be required and if this is the case then the accuracy of the forecast of the dependent variable will depend on the accuracy of these forecasts.

The external influences most commonly accounted for in electricity demand modelling are weather factors, particularly in hourly, weekly and monthly applications due to temporal (e.g. time of day) and seasonal effects respectively. Typical temporal factors are time of day, day of week and week of year Some of the more commonly used weather variables are:

- average temperature (Baker, 1985; Schneider *et al*, 1985; Ramanthan *et al*, 1985; Demirovic, 1988).
- humidity (Haida, 1994; Ramanthan *et al* , 1985; Hyde and Hodnett, 1991).
- effective temperature (Baker, 1985; Schneider, 1985).
- wind chill factor (Baker, 1985, Ramanthan, 1985, Hyde and Hodnett, 1991).

Two widely utilised weather variables are heating degree day (HDD) and cooling degree day (CDD) variables which take account of the effect of weather on electricity demand at selected points of the relevant temperature range (Train *et al*, 1984; Engle *et al* , 1992, 1988; Harris and Liu, 1993; Al-Zayer, 1996). Economic and demographic variables such as price of electricity; average income or number of customers (Ramanthan *et al*, 1985; Engel *et al*, 1988; Harris and Liu, 1993) may also be relevant, especially in the case of domestic demand forecasting. If future values of the exogenous variables are required then forecasting models may be built to predict these values or in the case of weather forecasts they are often obtained from meteorological stations and economic forecasts from government agencies.

A problem may exist in that the required exogenous variables may be unavailable on the time scale under consideration. McCafferty *et al* (1990) consider the use of mean monthly average

temperature when developing models to forecast disaggregate monthly industrial electricity demand for the Northern Ireland power system. They compare univariate (ARIMA) and multivariate models (BJTF) where they found that the effect of temperature on industrial monthly demand was negligible. They propose that economic factors are more likely to be relevant but since they are usually only available on a quarterly basis data for these variables are unobtainable on a monthly time scale.

The use of the external variables may not only provide more accurate forecast than the univariate case but also provide a greater understanding of the system. Harris and Liu (1993) provide an in-depth analysis of modelling monthly residential (domestic) electricity demand in the Southeast of the United States, where weather, economic and demographic influences are taken into account through the use of BJTF (2.23) models. They compare the forecasting results produced by a univariate (ARIMA) and to those produced by the BJTF model, with cooling degree day (CDD), heating degree day (HDD) and average price of electricity exogenous variables. The transfer function model produced more accurate results with the added advantage of providing a greater understanding of consumer behaviour in this residential sector. They found that residential electricity consumption is substantially influenced by price and weather factors, where weather is particularly influential in winter. Also, contrary to the economic reasoning average disposable income did not have a statistically significant effect on consumption. The most interesting revelation was the existence of a conservation effect among consumers, which was manifested in the time series through the presence of a downward trend from 1974 to 1989. This downward trend was exposed through the examination of temperature (CDD and HDD) adjusted data and also seasonally adjusted data. Two theories were considered to account for the conservation effect: the first postulates a heightened energy awareness following the world-wide energy demand crisis in 1973-1974; the second postulates that the effect reflects traditional consumer economics, where with rising inflation, the cost of electricity is trending upwards over time. They showed that the latter theory explained the effect better and that there was a damping out of this behaviour in the late 1980s.

Weather is a significant modelling factor in the area of short term load forecasting (STLF), for example, in daily peak or hourly electricity demand forecasting. Bunn and Farmer (1985) provide a review of STLF methods used in the electricity supply industry. A widespread approach is to split the load up into separate components, such as, a base load component and a weather sensitive load component, where a separate model is then used to represent each

component (Ackerman, 1985; Baker, 1985; Gupta, 1985; Schneidor *at al*; Demirovic, 1988; Hyde and Hodnett, 1991).

2.4.6 Drawbacks of Classical Linear Analysis

A main assumption in classical linear analysis is that all time series can be reduced to stationarity by a simple transformation such as detrending or differencing. This is a very well discussed and somewhat disputed topic in the literature. For example, Harvey (1989) accepts that some economic and social time series may be made approximately stationary through differencing but warns that if nonstationarity arises through a structural change in the series the model will be likely to break down in forecasting. In the case of the ARIMA model, he highlights the problem associated with the identification of the order of the autoregressive and moving average components of the model. The problem is that if the time series has a linear trend component, then differencing will yield a stationary time series but the invertibility conditions required for the ARIMA representation (Box and Jenkins, 1976) will be broken. To address the problem of the presence of a linear trend in the time series two different types of stationarity were defined by Kang (1990): a time series may be trend stationary, where the trend is modelled as a deterministic function of time, or difference stationary, where the trend component is modelled as stochastic in nature. He also suggests the use of a composite transfer function model which accounts for the theoretical possibility of the presence of both deterministic and stochastic trends. Both Morrison (1976) and Dickey and Fuller (1979) develop techniques that employ statistical hypothesis testing to determine if a time series is trend or difference stationary. The null hypothesis that a time series is difference stationary versus trend stationary is tested. In contrast Arellano and Pantulla (1995) test the null hypothesis that a time series is trend versus difference stationary.

The question over the stationarity transformations led some forecasters to develop modelling techniques that could handle nonstationary time series but that did not require the prior conversion to a stationary time series (Kitagawa and Gersch, 1984; Harvey and Durbin, 1986). These models are called *structural* time series models, also referred to as *unobserved components* models (Young, 1994), and the development of these models is dealt in Section 2.5.2. Alternatively, other forecasters rather than criticising the application of the transformation acknowledge the ambiguity associated with the selection of an appropriate differencing order and present techniques for its determination. For example, Chen and Jarrett (1991) present a hypothesis based technique that involves checking for the presence of unit roots to check the order of differencing. Their test also indicates if a linear trend should be

included in the model, thus addressing the problem that Harvey (1989) and Kang (1990) highlighted. Alternatively, Chang and Dickey (1994) develop a technique to recognise when overdifferencing has occurred, where this method is based on the examination of the inverted autocorrelation function (IACF). Frances (1991) also examines the problem of overdifferencing in the case of a seasonal monthly time series. He presents a test to check for seasonal unit roots and proposes an alternative to the SARIMA model (2.19) which uses first order differencing with the seasonality modelled with a constant and 11 seasonal dummy variables.

In Chapter 3 of this thesis a nonstationary seasonal weekly electricity demand time series is examined where, rather than using statistical tests to check for the type of stationary, three different types of modelling techniques were employed to handle the nonstationary. In the first case the trend component is removed from the time series , through detrending (Ljung, 1987), and a high order (to account for seasonality) AR (2.17) and ARMA (2.18) model is then fitted to the data. The second case uses the traditional Box-Jenkins methodology to model the series, where differencing is applied at the nonseasonal and seasonal level of the series to convert it to a stationary one and then a relatively low order SARI model is fitted to data. The third approach does not involve the use of transformations to induce stationary but uses a structural model in which the trend and seasonal components are allowed to slowly vary over time.

2.5 State Space Time Series Methods

2.5.1 Background

The motivation for the development of the structural time series models has been discussed in Section 2.4.6. The Box-Jenkins SARIMA model represents the differenced observed time series as a stationary stochastic process, whereas in contrast, the structural model does not attempt to represent the underlying data generation process but uses individual models to represent salient features within the time series, such as trend, seasonal or cyclical characteristics. Therefore, both the SARIMA and structural models deal with the presence of trend and seasonality in a time series but do so in a different manner. The SARIMA model removes the trend and seasonality through differencing, whereas the structural model uses a separate model to estimate the trend and seasonality. Harvey (1989) shows that the simplest form of structural model (linear and time-invariant) has a corresponding SARIMA model representation, where they are equivalent in that the SARIMA model will produce identical

forecasts to the structural form. Box *et al* (1987) present an SARIMA model based approach for estimating the trend of a seasonal time series, where they derive an equivalence relationship between this trend estimate and that of the trend component in the structural model. Espasa and Pena (1995) also carry out similar work to Box *et al* in which the forecast function of the SARIMA model is broken down into a permanent term (produced by the model's nonstationary operators) and a transitory term (produced by the model's stationary operators). The permanent (nonstationary) term can be further broken down into trend and seasonal components and the transitory (stationary) term into a seasonal and nonseasonal term.

The key to handling the structural time series models is their formulation in state space form, where the state of the system represents the various characteristics of the time series. The Kalman filter (1960) is used to estimate the states, with future predictions obtained through the extrapolation of the components into the future. It also possible to formulate the SARIMA model in state space form (Annsley and Kohn, 1985) but the classical representation is that of a stochastic difference equation. Before going on to describe the structural state space models the equations for the general state space model for a univariate time series are given.

The *observation* equation of the general state space model for a time series $y(k)$, k=1,.....,N is given by

$$y(k) = \mathbf{H}\mathbf{x}(k) + \boldsymbol{\varepsilon}(k) \tag{2.37}$$

where $\boldsymbol{H}$ is an $1 \times n$ matrix and $\varepsilon(k)$ is a serially uncorrelated disturbance term with zero mean and variance σ_ε^2. In general the elements of $\mathbf{x}(k)$, the state vector, are not observable but are known to be generated by a first order Markov process (Harvey, 1989) and are described by the *state* equation which is given by:

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}(k-1) + \mathbf{G}\boldsymbol{\eta}(k) \tag{2.38}$$

where

- $\mathbf{F} \in R^{n \times n}$ and is called the state transition matrix.
- $\mathbf{G} \in R^{n \times g}$
- $\eta(k) \in R^{g \times 1}$ is a vector of serially uncorrelated disturbances with a mean of zero and a covariance matrix $\mathbf{Q}$, that is, var($\boldsymbol{\eta}(k)$) = $\mathbf{Q}$.

The matrices **H**, **F**, **G** and **Q** are called the system matrices which are assumed to be non-stochastic and if they do change with time it is assumed that they do so in a deterministic way. If they do not change with time the model is said to be time-invariant. Two assumptions are made regarding the state space system. The first is that the disturbances and the initial state vector are normally distributed, thus ensuring that the model is linear. Therefore, based on this assumption the initial state vector $\mathbf{x}(0)$ has a mean $\mathbf{x_0}$ and a covariance matrix $\mathbf{P_0}$, that is $E(\mathbf{x}(0))=\mathbf{x_0}$ and $\text{Var}(\mathbf{x}(0))=\mathbf{P_0}$. The second assumption is that the disturbances $\varepsilon(k)$ and $\eta(k)$ are uncorrelated with each other for all k and are uncorrelated with the initial state. The Kalman filter is a recursive algorithm which may be used to compute the optimal estimate of the state vector at time k, based on observations up to and including $y(k)$. The system matrices and $\mathbf{x_0}$ and $\mathbf{P_0}$. are assumed to be known at all k. The Kalman filter equations for a time-invariant model are presented in Appendix 1 and derivations of the algorithm can be found in numerous works such as Gelb (1974), Anderson and Moore (1979) and in Kalman's (1960) first paper on the subject.

Given the optimal estimate of $\hat{x}(k)$ at sample k, an l-step-ahead forecast of the series may be obtained through the following:

$$\hat{\mathbf{x}}(k+l/k) = \mathbf{F}^{l}\hat{\mathbf{x}}(k) \tag{2.39}$$

$$\hat{\mathrm{y}}(\mathrm{k}+1/\mathrm{k}) = \mathbf{H}\hat{\mathbf{x}}(k+l/k) \tag{2.40}$$

where

- k represents the forecasting origin
- l represents the forecasting lead time
- $\hat{\mathbf{x}}(k+l/k)$ is the estimate of the state vector at sample $k+l$ given the state estimate at sample k
- $\hat{y}(k+l/k)$ represents the forecast of the time series at sample $k+l$.

2.5.2 State Space Structural Time Series Models

2.5.2.1 Univariate Structural Models

The univariate structural model (Harvey and Durbin, 1986; Harvey, 1989; Harvey, 1984) employed for the work carried out in this thesis is given by the following equation:

$$y(k) = \mu(k) + \gamma(k) + \varepsilon(k) \tag{2.41}$$

where

- $\mu(k)$ represents trend component
- $\gamma(k)$ represents the seasonal component
- $\varepsilon(k)$ is a zero mean random disturbance term with variance σ_ε^2.

The disturbance term may be interpreted as an irregular component in the time series or as a forecast error.

The trend component may be modelled by the following random walk model:

$$\begin{aligned} \mu(k) &= \mu(k-1) + \beta(k-1) + \eta(k) \\ \beta(k) &= \beta(k-1) + \zeta(k) \end{aligned} \tag{2.42}$$

where $\eta(k)$ and $\zeta(k)$ are mutually uncorrelated white-noise disturbance terms with zero mean and variances given by σ_η^2 and σ_ζ^2 respectively. The variances affect the stochastic movement of the trend component, where the larger the variance the greater the stochastic movement. The disturbance term $\eta(k)$ allows the level of the trend to move up and down, whereas $\zeta(k)$ allows the slope of the trend component to change. In formulating a model for the seasonal component the assumption is made that the sum of the seasonal effects is zero, this may be achieved using the following model which is referred to as a *dummy* seasonal model:

$$\gamma(k) = -\sum_{j=1}^{s-1} \gamma(k-j) + \omega(k) \tag{2.43}$$

where $\omega(k)$ is a zero mean random disturbance term which allows the seasonal effects to change over time, with the variance of $\omega(k)$ given by σ_ω^2. The larger the value of σ_ω^2 relative to σ_ε^2 the greater the discounting of past observations in the construction of the seasonal pattern for the model. The equations given by (2.42) and (2.43) may be combined and formulated in state space form to yield the Basic Structural Model (BSM) with a dummy seasonal component given by:

$$\begin{bmatrix} \mu(k) \\ \beta(k) \\ \hline \gamma(k) \\ \gamma(k-1) \\ . \\ . \\ \gamma(k-(s-1)) \end{bmatrix} = \left[\begin{array}{cc|cccccc} 1 & 1 & 0 & 0 & . & . & 0 \\ 0 & 1 & 0 & 0 & . & . & 0 \\ \hline 0 & 0 & -1 & -1 & . & . & -1 \\ 0 & 0 & 1 & 0 & . & . & 0 \\ . & . & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & 1 & 0 \end{array}\right] \begin{bmatrix} \mu(k-1) \\ \beta(k-1) \\ \hline \gamma(k-1) \\ \gamma(k-2) \\ . \\ . \\ \gamma(k-(s-2)) \end{bmatrix} + \begin{bmatrix} \eta(k) \\ \zeta(k) \\ \hline \omega(k) \\ 0 \\ . \\ . \\ 0 \end{bmatrix} \tag{2.44}$$

$$y(k) = \begin{bmatrix} 1 & 0 & 1 & 0 & . & . & 0 & 0 & 0 \end{bmatrix} \mathrm{x}(k) + \varepsilon(k)$$

Alternatively, the seasonal component may be represented by a set of trigonometric terms at seasonal frequencies $\lambda_j = {}^{2\pi j}/_{s}$, $j = 1,\ldots..{}^{s}/_{2}$, defined as:

$$\gamma(k) = \sum_{j=1}^{[s/2]} \gamma_j(k) \tag{2.45}$$

$$\left.\begin{aligned} \gamma_j(k) &= \gamma_j(k-1)\cos\lambda_j + \gamma_j^*(k-1)\sin\lambda_j + \omega_j(k) \\ \gamma_j^*(k) &= \gamma_j(k-1)\sin\lambda_j + \gamma_j^*(k-1)\cos\lambda_j + \omega_j^*(k) \end{aligned}\right\} \qquad j = 1,\ldots..,[s/2] \tag{2.46}$$

where $\omega(k)$ and $\omega_j^*(k)$ are zero mean white-noise disturbance terms which are uncorrelated. Equations (2.45) and (2.46) make the *trigonometric* seasonal model. Note that $\gamma_j^*(k)$ appears as a matter of construction in order to form $\gamma_j(k)$ and therefore does not have any real physical meaning. In addition, when s is even the $j = s/2$ component of λ_j becomes π and $\gamma_j^*(k)$ is not required for this frequency. It is possible to assign different variances to $\omega(k)$ and $\omega_j^*(k)$ which would allow each harmonic to evolve at varying rates. However, in practice, the number of parameters can be greatly reduced without affecting the goodness of fit of the model if all the variances are assumed to be equal; that is

$$\mathrm{var}(\omega_j(k)) = \mathrm{var}(\omega_j^*(k)) = \sigma_\omega^2 \qquad j=1,\ldots..,[s/2] \tag{2.47}$$

As in the case of the dummy seasonal component the larger the value of σ_ω^2 the more past values are discounted in constructing the seasonal pattern. Equations (2.42), (2.45) and (2.46) may be combined and formulated in state space form to yield the BSM with a trigonometric seasonal component. The full state space is given as follows:

$$\begin{bmatrix} \mu(k) \\ \beta(k) \\ \hline \gamma_1(k) \\ \gamma_1^*(k) \\ . \\ \gamma_{(j/2-1)}(k) \\ \gamma^*_{(j/2-1)}(k) \\ \gamma_{j/2}(k) \\ \gamma^*_{j/2}(k) \end{bmatrix} = \left[\begin{array}{cc|ccccccccc} 1 & 1 & 0 & 0 & 0 & . & . & . & 0 \\ 0 & 1 & 0 & 0 & 0 & . & . & . & 0 \\ \hline 0 & 0 & \cos\lambda_1 & \sin\lambda_1 & 0 & . & . & . & 0 \\ 0 & 0 & -\sin\lambda_1 & \cos\lambda_1 & 0 & . & . & . & 0 \\ 0 & 0 & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & \cos\lambda_{(j/2-1)} & \sin\lambda_{(j/2-1)} & 0 & 0 \\ 0 & 0 & 0 & 0 & . & -\sin\lambda_{(j/2-1)} & \cos\lambda_{(j/2-1)} & 0 & 0 \\ 0 & 0 & 0 & 0 & . & . & 0 & \cos\lambda_{j/2} & \sin\lambda_{j/2} \\ 0 & 0 & 0 & 0 & . & . & 0 & -\sin\lambda_{j/2} & \cos\lambda_{j/2} \end{array}\right] \begin{bmatrix} \mu(k-1) \\ \beta(k-1) \\ \hline \gamma_1(k-1) \\ \gamma_1^*(k-1) \\ . \\ \gamma_{(j/2-1)}(k-1) \\ \gamma^*_{(j/2-1)}(k-1) \\ \gamma_{j/2}(k) \\ \gamma^*_{j/2}(k) \end{bmatrix} + \begin{bmatrix} \eta(k) \\ \zeta(k) \\ \hline \omega_1(k) \\ \omega_1^*(k) \\ . \\ \omega_{(j/2-1)}(k) \\ \omega^*_{(j/2-1)}(k) \\ \omega_{j/2}(k) \\ \omega^*_{j/2}(k) \end{bmatrix}$$

$$y(k) = \begin{bmatrix} 1 & 0 & 1 & 0 & . & . & . & 1 & 0 \end{bmatrix} \mathrm{x}(k) + \varepsilon(k) \tag{2.48}$$

2.5.2.2 Multivariate Structural Model

The structural model of (2.41) may be expanded to include exogenous variables, referred to as a Basic Structural Model with Exogenous input (BSMX) model, where this model is written as follows (Harvey, 1989):

$$y(k) = \mu(k) + \gamma(k) + \mathbf{u}^{\mathrm{T}}(k)\mathbf{g} + \varepsilon(k) \tag{2.49}$$

where the vector $\mathbf{u}(k)$ is an nu x 1 vector containing the exogenous variables, while the nu x 1 vector $\mathbf{g}$ contains the unknown parameters associated with them. Thus a linear combination of $\mathbf{u}^{\mathrm{T}}(k)\mathbf{g}$ has been added to the model. The assumption is made that the relationship between $y(k)$ and $u(k)$ is a linear one. The trend and seasonal component may be represented by (2.44) or (2.48). The state equation given by (2.37) may be rewritten in terms of (2.49) as

$$y(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{u}^{T}(k)\mathbf{g} + \boldsymbol{\varepsilon}(k) \tag{2.50}$$

for $k = 1,.....,N$. If $\mathbf{g}$ were known, the addition of $\mathbf{u}^{\mathrm{T}}(k)\mathbf{g}$ would make no real difference to the operation of the Kalman filter. However, in the majority of cases $\mathbf{g}$ will be unknown and for the purposes of estimating it $\mathbf{g}$ is augmented into the state vector $\boldsymbol{x}(k)$. Therefore, the state space form is given by:

$$y(k) = \begin{bmatrix} \mathrm{H}^{T} & \mathrm{u}^{T} \end{bmatrix} \begin{bmatrix} \mathrm{x}(k) \\ \mathrm{g}(k) \end{bmatrix} + \boldsymbol{\varepsilon}(k) \tag{2.51}$$

$$\begin{bmatrix} \mathbf{x}(k) \\ \mathbf{g}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k-1) \\ \mathbf{g}(k-1) \end{bmatrix} + \mathbf{G} \begin{bmatrix} \boldsymbol{\eta}(k) \\ 0 \end{bmatrix} \tag{2.52}$$

where $k = 1,.....,N$ and **F**, **G** and **H** are of the form given in (2.44) or (2.48) depending on which model is being employed to represent the trend and seasonal component. The vector $\eta(k) = [\eta(k)\ \zeta(k)\ \omega(k)\ 0\ 0\ .....0]^T$ if (2.44) is used to represent the trend and seasonal component and $\eta(k) = [\eta(k)\ \zeta(k)\ \omega_1(k)\ \omega_1^*(k)\ .....\ \omega_{s/2}(k)]^T$ if (2.48) is used to represent the trend and seasonal component. Lagged exogenous variables can be introduced, where the state equation given by (2.50) becomes

$$y(k) = \mu(k) + \gamma(k) + \sum_{b=0}^{h} \mathbf{u}^T(k-b)\mathbf{g}_b + \varepsilon(k) \tag{2.53}$$

where b is the lag between the $y(k)$ and the variables in $\mathbf{u}(k)$ and $\mathbf{g}_b$ represents the unknown parameter vector associated with $\mathbf{u}(k - b)$.

2.5.3 Structural Model Estimation

The estimation of the state space structural model given by (2.44), (2.48) and (2.50) involve the estimation of the variances of each of the stochastic disturbance terms in the vector $\psi = \{\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2\}$, where the parameters in ψ are referred to as the hyperparameters of the model. The estimation of the hyperparameters may be carried out using maximum likelihood estimation via the prediction error decomposition. The prediction errors are computed using the Kalman filter which is run as a function of ψ and then the likelihood function is maximised with respect to ψ, where this is usually carried out using a suitable numerical optimisation routine, such as the Gill-Murray-Pithfield quasi newton algorithm (Harvey, 1981). The estimation procedure is an iterative procedure which may be described by the following steps:

1. Run the Kalman filter as a function of $\psi = \{\sigma_\varepsilon^2, \sigma_\varepsilon^2, \sigma_\varepsilon^2, \sigma_\varepsilon^2\}$ and compute prediction errors required to compute the likelihood function.
2. Compute the likelihood function, $L(\psi)$.
3. Maximise the likelihood function $L(\psi)$ with respect to ψ.

Steps 1 to Step 3 are iterated upon until the optimal values of $\psi = \{\sigma_\varepsilon^2, \sigma_\varepsilon^2, \sigma_\varepsilon^2, \sigma_\varepsilon^2\}$ are determined. Appendix A presents the Kalman filter equations for a general state space model and also describes the maximum likelihood estimation and the prediction error decomposition methods.

2.5.4 Literature Overview

There are a number of indicative examples in the literature of the application of structural state space models. An early example is that of Harvey and Durbin (1986), in which they present a case study of structural time series modelling applied to monthly road casualty rates in Great Britain. The paper deals with the full formulation of their proposed structural time series model with a trend, seasonal and irregular component. The Kalman filter is employed to compute the maximum likelihood estimates of the hyperparameters of the model. Only brief details of the estimation procedure are given. The multivariate case is dealt with where the BSM model is extended to include explanatory variables. For comparative purposes they also develop ARIMA models and obtain comparable results in terms of forecasting accuracy. However, they argue that the structural time series modelling procedure offers a more direct and transparent approach than the Box-Jenkins methodology and question the Box-Jenkins assertion that a nonstationary time series can be made stationary through differencing. In addition, they suggest that in contrast to the majority of cases the use of an ARIMA model in this particular application was relatively straightforward to deal with since an "Airline" model (Box-Jenkins, 1976) was fitted to data, where the identification of this model is considered a less complex procedure that alternative models.

In the work carried out in the thesis a SARI model (2.19) and a structural model were fitted to weekly electricity demand data. The structural model produced more accurate forecasting accuracy results, with the added advantage that the development effort involved was less complex in comparison with the development of the SARI model. However, the Box-Jenkins and structural models developed in the weekly electricity demand application were also employed in the work involving a neural network analysis. The results showed that the neural networks based on the SARI (2.19) and BJTF (2.22) models produced considerably more accurate results than those based on the structural models. It is the opinion of the author that the development effort required to identify a linear Box-Jenkins model is worthwhile if the model is used in conjunction with a nonlinear modelling tool such as neural networks.

The paper by Harvey and Durbin (1986) does not include details of the estimation techniques for the structural models. Harvey (1989), however, provides full details of the estimation methods in both the time and frequency domain. In this book he also deals with the development of the BSM models for the univariate time series; the extension of the BSM model to the multivariate case and model adequacy testing. Harvey and Peters (1990) discuss different estimation techniques for the structural time series models in the both the time

domain and the frequency domain. They compare the estimation procedures using six macroeconomic time series and also using the well known airline data series (Box-Jenkins, 1976). The procedures are compared according to forecasting accuracy results; computation time and reproducibility of estimates. They conclude that the results obtained by the time domain and the frequency domain methods are comparable, however the frequency domain methods appear computationally faster.

An even earlier example of the structural state space modelling approach was developed by Kitagawa and Gersch (1984), where the origins of this work can be found in Kitagawa (1981). In the approach given in this paper they decompose the observed time series into the following components:

- Additive local polynomial trend component.
- Seasonal component.
- Globally stationary autoregressive component.
- Observation error component.

Each component is characterised by a stochastic perturbed difference equation constraint. Prior knowledge of the time series is used to construct the constraint equation which is formulated in state space form,. A number of different models, characterised by alternative constraint equations, are constructed and the Kalman filter is used to compute the likelihood for each of the different models. Akaike's (1973, 1974) minimum AIC (Akaike Information Theoretic Criterion) procedure is used to determine the best of the alternative models fitted to the data. A fixed interval smoothing algorithm is applied to obtain smoothed estimates of the components and an l-step ahead forecast is obtained by repeated application of the Kalman filter recursion equations l times. The authors compare their approach to the Box-Jenkins modelling procedure and argue that their approach is semiautomatic whereas the Box-Jenkins approach requires extensive human intervention to achieve satisfactory modelling. It has also been the experience of the author that in comparison to the structural modelling approach that the Box-Jenkins modelling procedure requires a considerable amount of user interaction in order to develop a parsimonious forecasting model.

Ng and Young (1990) in their approach to structural time series modelling use similar models to those developed by Harvey (1986, 1989), however they use a different method of identification and estimation of the models. The method involves a spectral decomposition procedure which is based on the exploitation of recursive smoothing algorithms. The

technique involves the decomposition of the time series into various quasi-orthogonal components, where typical components are trend, signal, stochastic perturbation, seasonal and irregular. Subsequently, it appears to be straightforward to identify and estimate models for these components using recursive time-variable parameter estimation (Young, 1984, 1989; Ng *et al*, 1988). Examples of the models used include:

- IRW (integrated random walk) for the trend component.
- GTF (general transfer function) model for the stochastic component.
- PRW (periodic random walk), DPRW (differenced periodic random walk) or DHR (dynamic harmonic regression) model for the seasonal component.

The papers by Young and his co-workers (1984, 1988, 1989) present the full spectral decomposition technique and illustrates the use of the technique on a number of examples. A UK electricity demand (quarterly 1960 - 1985) example with outliers is given, where it is shown that the model performs well in the presence of the outlier. This technique is also dealt with in Young (1988); Young (1994); Young *et al* (1989); Young *et al* (1991). It has been the experience of the author that although the structural models developed by Young *et al* produce comparable results to those developed by Harvey, the estimation of the model using spectral decomposition techniques is computationally less demanding but requires more user interaction and development effort. Therefore, the Harvey structural models (2.44, 2.48, 2.50) were employed for the applications dealt with in the thesis since they required a reasonable amount of computation effort and were less complex to develop.

2.5.5 State Space ARMA and ARIMA Models

It is also possible to formulate the ARMA (2.18) and SARIMA (2.19) models in state space form. Jones (1990) presents the formulation of an ARMA process in state space form. The full procedure for evaluating the likelihood function using the Kalman filter recursion equations is given. The likelihood function can be used to obtain maximum likelihood estimates of the autoregressive and moving average coefficients. The variance of the observation error and random input process can also be computed from the likelihood function. Annsley and Kohn (1985a, 1985b); Kohn and Annsley (1986, 1989) also develop the state space formulation of the SARIMA model. They develop efficient modified Kalman filtering and smoothing algorithms which take advantage of the SARIMA state space form. The modified filtering and smoothing algorithms may be used to calculate the likelihood

function for a state space model with incompletely specified initial conditions, for example, when the time series is subject to missing observations. Harvey and Pierce (1984) obtain a state space representation for an SARIMA model and show how to obtain the maximum likelihood estimates of the parameters of the model.

2.6 Neural Networks

2.6.1 Background

The study of neural networks originates from work in the area of artificial intelligence where attempts were made to define a computational model to simulate the brains thinking process (Rosenblat, 1962). A neural network consists of a number of simple node (neuron) elements that are connected together to form either a single or multiple layers. The basic node element employed in a neural network differ according to network type. However a commonly used model is the form of the McCulloch and Pitts neuron (Aleksander, 1991) shown in Figure 2.12.

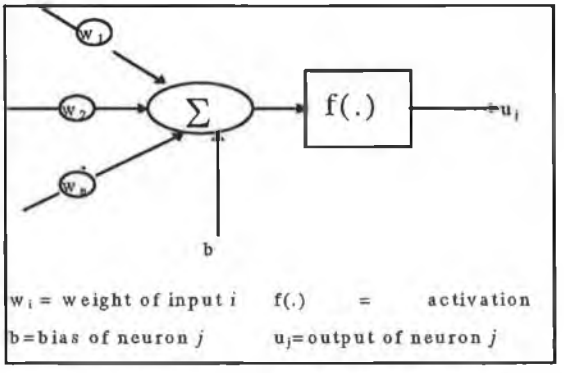

Figure 2.12 Outline of a basic McCulloch Pitts neuron

Associated with each of the interconnecting input links in a neuron is a *weight* value w_i. Each neuron computes the weighted sum of its inputs and adds a bias value, b, where inputs come in the form of data either from outside of the network, for example a representation of what the network is required to model, or from other network elements, possibly from outputs of neurons in previous layers. This sum is then passed through a nonlinear function called an activation function $f(.)$ to yield the output of the neuron u_j. The strengths of the weights between the nodes and the bias value are determined through network *learning*. The learning procedure can make use of an individual data set, after which the network weights are fixed, or learning can continue through the network's lifetime. In certain types of networks learning is directed, referred to as supervised learning, whereas other networks learn in an unsupervised manner. There are many different types of neural network that vary according

to factors such as: type of input, network structure and the type of learning procedure adopted. Some examples are the Hopfield network (Hertz *et al*, 1991); the Hamming network (Lippmann, 1987), the Multi-Layer-Perceptron (MLP) (Hecht-Neilsen, 1990), the Radial Basis Function (RBF), (Hush and Horne, 1993) and Kohonen Self Organising Feature Maps (Anderson and Rosenfeld, 1988).

2.6.2 Neural Networks for Time Series Forecasting

For time series forecasting purposes a neural network may be considered as a data processing technique that maps some type of input stream of information to an output stream of data (Azoff, 1994). In linear time series forecasting a linear relationship is formed between past and future values of the time series. Nonlinear time series forecasting usually involves forming a constrained nonlinear relationship (Chen and Billings, 1989; Gooijer and Kumar, 1992). Neural networks provide an unconstrained nonlinear modelling technique where a general nonlinear mapping is formed between some subset of past time series values and a future time series value. Temporal information may be presented to the network by a time-lagged vector of time series data at the input, often referred to as a tapped delay line, with the current value of the time series at the output as shown in Figure 2.13.

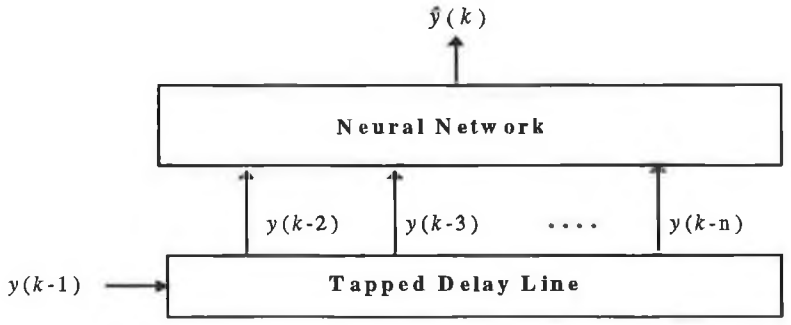

Figure 2.13 Time series neural network

The diagram in Figure 2.13 describes the univariate case; it is straightforward to extend this to the multivariate case. In the multivariate case the input is made up of past values of the time series of which the forecast is required but also on present and past values of the other exogenous variable time series. In contrast to the linear case where significant development effort is required at the identification stage, particularly in the case of the BJTF models (2.22), it is the relative ease at which the multivariate case may be dealt with using neural networks that makes them an attractive modelling tool for time series forecasting. It is also possible to introduce feedback into the network (Hush and Horne, 1993) by including a second tapped delay line through which the output of the network is fed, where this is shown in Figure 2.14.

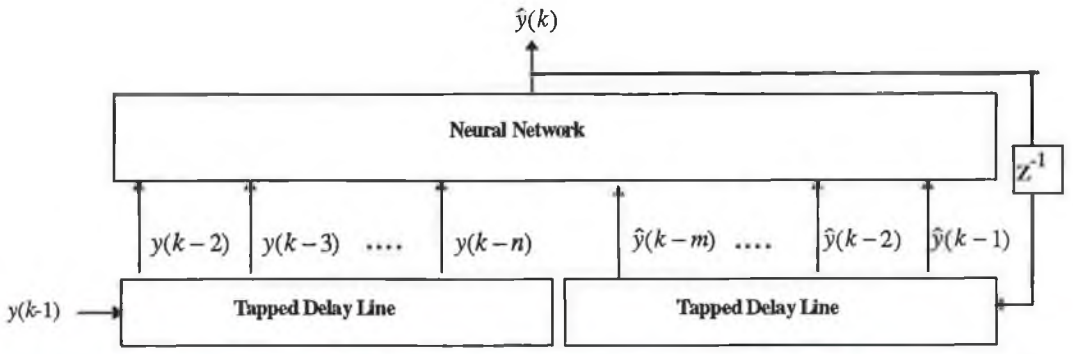

Figure 2.14 Time series neural network with output feedback

A widely applied neural network structure in the area of time series forecasting is the MLP, where the configuration given in Figure 2.13 is the most common approach (Park *et al*, 1991; Tang *et al*, 1991; Wong, 1991; Chakraborty *et al*, 1992; Chen *et al*, 1992; Weigend and Gershenfeld, 1992; Lu *et al*, 1993; Rogers and Vemuri, 1994; Papalexopoulos *et al*, 1994; Azoff, 1994; Lachtermacher and Fuller, 1995; Beltratti *et al*, 1996). An MLP is a network that consists of a layer input nodes; a layer of output neurons and one or more layers of hidden neurons sandwiched in between, with a description given in Figure 2.15

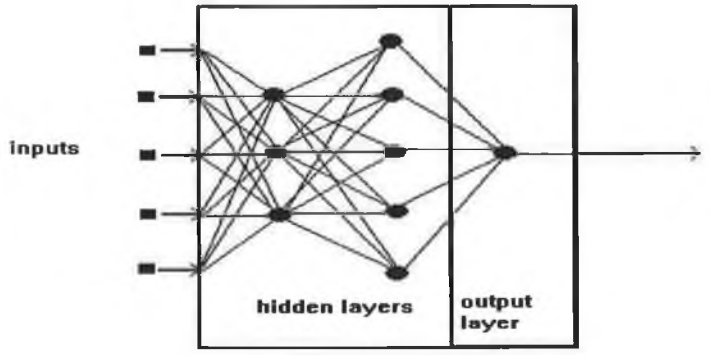

Figure 2.15 Structure of an MLP with two hidden layers

The type of neurons used in the MLP are of the form described in Figure 2.12. The MLP is known as a *feedforward* network, where a set of data is presented at the input layer; the outputs from this layer are fed to the first hidden layer and subsequently the outputs from the first hidden layer are fed to the second hidden layer, where this process continues until the output layer is reached. Rummelhart *et al* (1986) developed a learning algorithm for MLP's, called the *backpropagation* algorithm, to determine the weights and biases of the network. The MLP is a supervised learning network where each output layer value is compared with a desired or specific value and the weights and biases of the network are adjusted so as to

minimise the difference between the two. There are a number of different factors involved in the specification of an MLP, where these are as follows:

- the number and type of input nodes.
- the number of output neurons.
- the number of hidden layers.
- the number of neurons in each hidden layer.
- the type of neuron activation functions.

2.6.3 Literature Overview

2.6.3.1 Neural Networks for Time Series Forecasting

There are many works that deal with the specification of MLP networks, one such paper is the tutorial paper published by Lippman (1987). In this he discusses the MLP and also several other different neural networks, such as, the Hopfield network, the Hamming net and the Carpenter/Grossenberg classifier. Regarding the MLP he addresses issues such as the number of layers required and the conditions under which good generalisation can be achieved. He recommends that MLPs with one hidden layer are sufficient to solve arbitrary complex input-output mappings with typically three times as many hidden neurons as input nodes. Hecht-Neilsen (1990) also considers the same problem and suggests that a network with one hidden layer that has $2N+1$ hidden neurons, and with continuously increasing nonlinear transfer functions, can compute any continuous input function with N input nodes. However, in practice it has been found that these rules may not be easily implemented (Hunt and Sbarbaro, 1992). To date, there is still no firm theory or even a set of heuristic guidelines for the design of MLPs and it is very much an area of active research (Hertz *et al*, 1991; Villiers and Barnard, 1992; Hush and Horne, 1993; Cottrell *et al*, 1995).

A number of works suggest the use of linear statistical techniques to determine the input structure to the MLP, examples are Chen *et al* (1992), Peng *et al* (1993), Papalexopoulos *et al* (1994), Hegazy and Salama (1994) and Lachtermacher and Fuller (1995). The latter three works are applications based papers, whereas the former attempt to present a procedure for the general case. A similar approach is also used in the work carried out in this thesis, where MLP network are used when forecasting weekly and yearly electricity demand. In contrast to the work developed here, the work carried out by Hegazy and Salama (1994) and Lachtermacher and Fuller (1995) only deal with the use of autoregressive variables of the time

series in the input structure and do not address the case where exogenous variables are employed. Chen *et al* (1992), Peng *et al* (1993) and Papalexopoulos *et al* (1994) do use exogenous weather variables in the input structure but unlike the applications in this thesis they deal with STLF. Hegazy and Salama (1994) give a brief and somewhat unclear description of a procedure to determine the input variables (input nodes) for a neural network. It involves the identification and estimation of AR and ARI models, where the minimum AIC criterion (Akaike, 1973) is used to choose the model order that determines the number of input nodes for the neural network. Lachtermacher and Fuller give a much more detailed description of their proposed procedure that involves the following:

Step 1. Exploratory phase

1.1 Plot time series look for trend, seasonalities, outliers, plot SACF (2.8).

1.2 Box-Jenkins modelling.

Step 2. *Modelling Phase*

2.1 Use Box-Jenkins model to determine the input variables to the network.

2.2 Use heuristic guidelines from literature to set number of units in hidden layer.

2.3 Set weights by backpropagation; stop when forecasting performance is achieved.

2.4 Overfitting analysis: repeat 2.3 for networks with slightly different numbers of inputs and hidden units.

2.5 Select model from 2.3 and 2.4 having best performance.

The paper discusses the various MLP design criteria that are involved in the modelling phase (Step 2), such as: the learning rule; training and validation procedure; network architecture; performance criterion (RMSE) and overfitting analysis . They demonstrate the application of the procedure on two types of time series: stationary nonseasonal time series and nonstationary nonseasonal time series. Similar to the work developed in the current study, the nonstationary nonseasonal time series are yearly electricity demand, where four applications from different countries are examined, that is United States, Canada, Brazil and Australia. They suggest that their future work should involve the analysis of seasonal stationary and seasonal nonstationary time series. The work dealt with in this thesis specifically deals with a neural network application for modelling a nonstationary seasonal time series.

Network structures other than the MLP are also applied to the area of time series forecasting. For example, Peng *et al* (1993) use a neural network, referred to as a *Widrow* model, where this network is similar to an MLP but has different network architecture and a different

learning rule. Other networks are the *recurrent* neural network (Figure 2.10), this is a feedforward network with feedback (Connor *et al*, 1994; Jin *et al*, 1994) and the radial basis function network (Zaknich and Attikiouzel, 1994; Hudson *et al*, 1994). Note that Giles *et al* (1994) present a summary of the published literature in the area of recurrent neural networks.

The comparison of the accuracy of the forecasting results obtained using statistical linear models with those obtained using neural networks is dealt with in some works (Sharda and Patil, 1990; Tang *et al*, 1994; Blake *el al*, 1995). Sharda and Patil (1990) present results of a comparison between forecasts produced by neural networks and Box-Jenkins models. The results from seventy five series, (selected from the M competition data set (Makridakis, 1982)) where single-step-ahead forecasting of 8 yearly, 18 quarterly and 49 monthly series is conducted, are compared. The neural network used was a three layer peceptron with one hidden layer, and the same number of neurons in the hidden layer as input nodes. Using the mean absolute percentage error as a criterion the neural networks performed at least as well as the Box-Jenkins models. They suggest that as only simple neural network models were employed that the case for the use of neural networks in time series forecasting looks promising. They plan to carry out a similar analysis but where multi-step-ahead forecasting is performed. Note that multi-step-ahead forecasting is carried out using Box-Jenkins analysis and neural networks in the work conducted here.

Blake *et al* (1995) compare the forecasting ability of different neural network models with genetically identified univariate SARIMA Box-Jenkins models. The construction of the SARIMA models involve the use of a genetic algorithm to carry out the identification stage of the Box-Jenkins methodology. Three different neural network designs are used:

1. traditional feedforward network.
2. feedforward network with jump connections.
3. recurrent network.

They conduct the analysis using the 7 time series that Box and Jenkins (1976) deal with in their book. Short, medium and long term forecasts are compared using the mean square error and the average relative variance as a criteria. The analysis shows that the optimal network design in terms of forecasting accuracy depends on whether the time series is stationary. Preprocessing, although not essential to outperform the Box-Jenkins models, may enhance the neural networks performance ability. The neural networks Blake *et al* (1995) outperform the Box-Jenkins models for all forecast horizons and all time series.

Liu *et al* (1996) compare neural network, fuzzy logic and AR modelling applied to a STLF problem. The performances are evaluated through a simulation study where a 30-minute-ahead prediction is carried out. The fuzzy logic (RMS ≈ 0.1%) and neural network (RMS ≈ 1.0%) models outperform the AR model (RMS ≈ 7%). The AR model is only accurate for up to a few minutes ahead and from then on the forecasting accuracy deteriorates very significantly.

Apart from neural networks, other intelligent techniques that have been applied to time series forecasting problems are genetic algorithms (Beltratti *et al* , 1996; Yang *et al* , 1996); fuzzy logic (O'Sullivan, 1994) and knowledge based systems (Banim and Hodnett, 1991; Raham and Hazim, 1993). There are also examples of works that adopt a hybrid approach, for example, Jones (1993), Bornholdt and Graudenz (1992) and Beltratti (1996) all apply a hybrid genetic algorithms/neural networks approach, where the learning algorithm for the neural network is based on a genetic algorithm. Srinivasan *et al* (1995) apply a hybrid fuzzy/neural networks approach; here fuzzy logic is used to model the qualitative and quantitative knowledge about the system and its input parameters, where the neural network capture the relationship between the fuzzy inputs and outputs.

2.6.3.2 Neural Networks for Electricity Demand Forecasting

In the case of electricity demand prediction the most commonly dealt with applications are in the area of STLF. The general practice is the use of the MLP network, where examples of this are Bacha and Meyer (1992), Caire *et al* (1992), Tamura *et al* (1993) and Lu *et al* (1996). An advantage of neural networks over statistical models is the ability to model a multivariate problem without having to make complex dependency assumptions concerning the input variables. In the STLF problem it is well known that the load is dependent on weather factors (Bunn and Farmer, 1985). The majority of the neural network based works include weather and temporal input variables in the MLPs (Park *et al*, 1991; Chen *et al* , 1992; Kermanshahi *et al*, 1993; Lu *et al*, 1993; Morioka *et al*, 1993; Onoda, 1993; Azzam-ul-Asar *et al* , 1994; Bakirtzis *et al*, 1996). In contrast to the linear analysis some of the works employ a relatively large set of weather and temporal influencing variables, examples of this are Kiernan *et al* (1994) and Papalexopoulos *et al* (1994).

The use of neural networks as a modelling tool is a relatively new approach in the STLF domain, some of the earlier works are El-Sharkawi *et al* (1989) and Sobajic and Pao (1989). Consequently, a number of practical considerations regarding the design of neural networks

for this problem need to be addressed and as a result, there are a large number of applications based works. In this thesis neural networks are not applied to the STLF problem but to the problem of weekly and yearly demand forecasting. To the authors knowledge there are no published applications based works in these particular areas, with the exception in the yearly case of Lachtermacher and Fuller (1995). It is difficult to ascertain if the specific optimal neural network design specifications in the STLF problem are analogous to those in the weekly and yearly demand problem. However, it is assumed that some general guidelines pertaining to the use of neural networks in the electricity demand forecasting area may be acquired from the STLF based works. Some of the contributions relevant to the work carried out in this thesis, for example, input data normalisation, input structure and the type of exogenous variables employed are now discussed.

Azzam-ul-Asar *et al* (1994) examine different input data normalisation and input patterns when forecasting the peak load of the day. They found that the normalisation procedures based on the maximum and minimum values from the given data set yielded forecasts with relatively high errors and could be difficult to train. Note that this normalisation approach is suggested by Azoff (1994). The network that produced the most accurate forecast used a normalisation procedure that involved dividing the data by one number greater than the order of magnitude of the largest value. In relation to the different input patterns they find that the ratio of free weights to training vectors is high. This ratio represents the extent to which the network can generalise from the training set as opposed to fitting the noise. They attribute this problem to the fact that a limited data set was available for the work. They propose that a larger and more representative data set would alleviate this particular problem. An interesting point to note in this work is that the most accurate forecasting result (1.96 % in univariate case versus 2.39 % in multivariate case) was achieved with a neural network that did not use weather variables in its input structure.

Lu *et al* (1993) examine if the neural network model that they use is system dependent and/or case dependent. They do so by predicting hourly load using data from two notably different power boards, one in Asia and the second in the South East region of the United States. The main differences between the power boards are the following:

- the maximum (9875 MWh versus 2043 MWh) and minimum (2893 MWh versus 633 MWh) values in daily load;
- the maximum (95°F versus 93°F) and minimum (46.4°F versus -23°F) hourly temperatures;

- the typical yearly peak (August versus August and February).

They investigate if the neural network is case dependent by testing its ability to cope with weekends and holidays, seasonal changes and outliers. The input variables include weather and temporal factors that were selected based on prior knowledge of the system. Correlation analysis was used to determine the correlation between temperature and current / past load and the SACF (2.6) is used to determine the correlation between the current load and past load. The input data was normalised using the maximum and minimum values of load and temperature in the data. Note that this is a technique that Azzam-ul-Asar *et al* (1994) found unsatisfactory. The main conclusions of the work are as follows:

- the criteria used to select suitable network inputs is system dependent;
- once the network is trained it does not need to be modified frequently;
- the network does not depend on the season of the year, except in cases where a change in the season results in an abrupt change in the weather pattern;
- the neural network does not cope well with weekends and holidays;
- the networks are sensitive to outliers in the data (suggest filtering techniques as a solution).

In Chapter 4 of this thesis neural networks are applied to weekly time series from two very different power boards where it will be shown that the neural networks are case dependent.

An important contribution of the work carried out by Papalexopoulos *et al* (1994) is the use of several nonlinear temperature input variables, such as HDD, CDD, when forecasting peak and hourly load system demand for the PG&E (Pacific Gas and Electric) utility. Their justification for using these inputs is that when using neural networks with linear transfer functions in the output neuron and given the fact that most hidden neurons operate in the linear state (as opposed to non-activated state and saturated state) the network performance was deteriorating at two ends of temperature range. Under these conditions the load changes with temperature at a quadratic rate, however the output of the neural network changes with temperature at almost a linear rate. The use of CDD and HDD variables in the input resolved this problem. The neural networks used in the weekly demand application, dealt with in Chapter 4, also use HDD variables in the input structure. The neural network results were compared to those obtained from an existing fully operational model employing complex regression based statistical modelling techniques. Extensive testing showed that the neural network consistently outperformed the existing system in terms of average error over a long

period of time and number of larger errors. They also presented a successful neural network model that accounted for the holidays periods.

A summary of the main conclusions from the works by Papalexopoulous *et al* (1994), Lu *et al* (1993) and Azzam-ul-Asar *et al* (1994) are as follows:

- the type of input data normalisation may have an effect on neural network learning capabilities,
- the use of exogenous weather variables may not always improve the short term load prediction,
- when weather variables are employed it may be advantage to use nonlinear variables such as CDD and/or HDD,
- the network structure is system dependent and
- neural network performance may be affected by weekend and holiday effects.

2.7 Multi-Time-Scale Modelling

A multi-time-scale modelling approach is developed in this thesis which may also be thought of in terms of the *combination of forecasts*. For example, the weekly /yearly multi-time-scale application of Chapter 6 may be considered as the *combination* of forecasts produced by a time series model based on weekly observations with forecasts produced by a time series model based on a yearly observations. The literature available on the subject of combining forecasts may be split into two groups, the substantial bulk of which are based on the first approach.

1. the combination of forecasts produced on the same sampling interval but obtained using different forecasting methods - single-time-scale combination of forecasts.
2. the combination of forecasts produced on different sampling intervals combined on a chosen sampling interval - multi-time-scale combination of forecasts.

2.7.1 Combining Single-Time-Scale Forecasts

The rationale here is that the combined forecast is generally more accurate than any of the individual forecasts as the combined forecasts takes more information into consideration (Mahmoud, 1984). In addition, it has been shown that the accuracy of the combined forecast improves as more methods are included in the combination (Winkler and Makridakis, 1983). The general approach is to form a restricted (Mills and Stephenson, 1985; Clemen, 1986;

Holden and Peel, 1986) or unrestricted (Granger and Ramanathan, 1984; Mills and Stephenson, 1985) linear combination of the alternative forecasts. In the former case the weights of the linear combination are restricted by some constraint, for example that the sum of the weights add up to unity; whereas in the unrestricted case there is no such constraint. The weights of the combination are usually determined using least squares regression. Bordley (1982, 1986) obtains similar results to the aforementioned authors where he approaches the problem from a Bayesian perspective. Batchelor and Dua (1995) substantiate the generally proposed theory that combining diverse forecasts yields more accurate results than combining forecasts produced by similar methods and they develop a measure for assessing the benefits of combining a set of forecasts. Clemen *et al* (1995) and Chandrasekharan *et al*, (1994) propose procedures to evaluate the worth of the combination of a particular set of forecasts prior to their combination. Donaldson and Kamshra (1996) and Shi and Liu (1993) present a neural network based approach to the nonlinear combination of forecasts and propose through test cases that the principle of nonlinear combination is superior to a linear combination.

2.7.2 Combining Multi-Time-Scale Forecasts

The motivation here is similar to the single-time-scale case but the advantage of combining forecasts produced by models based on different sampling intervals is that one forecast may consider information or variables which the other forecast may not consider. In addition, one forecast may make different assumptions about the form of the relationships between the variables and thus may capture information which the other forecast may not. A limited amount of work has been carried out in this particular area of time series forecasting. Exceptions are Corrado and Greene (1988); and Fuhrer and Haltmaier (1988) who independently address the problem of combining a set of monthly and quarterly US economic forecasts for the Federal Reserve Board. Howry *et al* (1991) deal with a different set of US monthly and quarterly economic forecasts.

Corrado and Greene (1988) combine forecasts produced by a monthly *econometric* model, aggregated to a quarterly level, with forecasts produced by a quarterly econometric model. The approach extracts innovations from the aggregated monthly forecasts and uses Kalman filtering techniques to pass the innovations to the quarterly model forecast. This combined forecast can be revised each time actual monthly data becomes available. They develop the technique and discuss problems which may be encountered during practical implementation, such as model nonlinearity and serial correlation between the monthly and quarterly forecast

error terms. A full application is presented where the data set consists of a large set of endogenous and exogenous economic variables. It is shown that the combined forecasts yields significantly improved RMSE forecasting accuracy values (≈25%) over the original quarterly forecast. The important difference between this approach and the one presented in the thesis is that the forecasts are combined at the lower frequency level (quarterly level). However, Fuhrer and Haltmaier (1988) approach the same problem where similar techniques are used to produce the combined forecast at the higher frequency level (monthly level). They show (theoretically) that the combined forecast at the higher frequency level when aggregated is the same as the forecast which would have been obtained had it been combined at the lower frequency level.

Howry *et al* (1991) also combine monthly and quarterly forecasts produced by econometric models at the quarterly frequency level. They use the conditional expectation formulae for the multivariate normal distribution (Morrison, 1976) and develop an aggregation condition through which the quarterly and monthly variables are directly related. The application of the technique involves two steps: (i) the adjustment of the predicted monthly and quarterly forecasts so that they satisfy the aggregation condition, where this involves the use of the standard formulas for the multivariate normal distribution (Morrison, 1976); (ii) monthly updates of the combined forecast are performed by conditioning successively on the observed monthly values as they become available. They present an application of the technique based on economic data, where in general there is reasonable improvement in the RMSE forecasting accuracy over the original uncombined quarterly forecast.

2.8 Forecasting Accuracy Measures

A significant factor in the assessment of a time series forecasting methodology is its ability to predict future values of the time series with reasonable accuracy. There are many forecasting accuracy measures available, for example, Mathews and Diamantopoulos (1994) identify 14 such measures. Unfortunately, there is an absence of a universally accepted measure of accuracy (Mahmoud and Pegels, 1989; Mathews and Diamantopoulos, 1994). Mathews and Diamantopoulos (1994) concentrate on the problem of the assessment of forecasting performance within a multi-product (multi time series) context, where they find that for this case the reliance on a single forecast error measure is dangerous.

This thesis is devoted solely to the prediction of electricity demand, with the exception of a UK fuel consumption *example* in Chapter 5. Consequently, the focus of attention is on the

comparison of techniques used to produce forecasts of the one time series and considerations such as those outlined in Mathews and Diamantopoulos (1994) are not relevant.

For the most part, when comparing multi-step-ahead forecasting results four forecasting accuracy measures are employed. These are the Mean Square Error (MSE), one of the most commonly used measures in time series forecasting analysis, the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Percentage Error (MPE). Each of these are given as:

$$MSE = \sum_{n} \frac{e^2}{n} \tag{2.54}$$

$$MAE = \sum_{n} \frac{|e|}{n} \tag{2.55}$$

$$MAPE = \sum_{n} \frac{\left| \frac{e}{a} \times 100 \right|}{n} \tag{2.56}$$

$$MPE = \sum_{n} \frac{\frac{e}{a} \times 100}{n} \tag{2.57}$$

where a is actual, e = *actual - forecast* and n is the number of predictions carried out. The MAE and MSE measures are based on absolute errors and are thus scale dependent and the larger the value of the forecast the higher the value the measure are likely to be. Note that, in contrast to the MAE the MSE penalises large errors since it squares the error. The MAPE and MPE combine and average the percentage errors from the different time periods and provide an estimate of the forecasting accuracy over time. The MPE also indicates bias, where a negative value indicates forecasts that are on average higher than actual and positive values indicate forecasts that are on average lower than actual.

In the majority of cases the number of predictions in the forecast horizon will be greater than one, however in the case where a single forecast value is produced the *percentage error* associated with this forecast is given.

The neural network analysis conducted in Chapter 4 require the production of a large set of forecasting results that must be examined over three separate data sets (training, validation and prediction data set). Consequently, for the applications dealt with in this chapter only a

single forecasting accuracy measure, the MAE, is employed to compare the forecasting results over the prediction horizon.

2.9 Conclusion

In the literature time series forecasting techniques have been widely applied to the problem of electricity demand forecasting. This is particularly evident in the case of short term forecasting, where both statistical and neural network modelling techniques are well represented. Relatively speaking, medium and long term electricity demand forecasting problems are not dealt with in nearly the same depth. However, at present there is a requirement to address the problems associated with medium and long term forecasting in a response to changes in the market structures of the electricity supply industry in the developed world.

Also in response to the change in the operation of the electricity supply markets the problem of the production of electricity demand forecasts on a fine time scale over a long forecast horizon has become increasingly important. A solution to this problem is proposed in the thesis, where the problem is dealt with in terms of multi-time-scale modelling. It is the authors experience that there is little evidence of works that deal with time series forecasting problems based on multi-time-scales, with the exception of the papers discussed in Section 2.7.2. The proposed multi-time-scale technique requires a model which is in state space form and there are numerous examples of the formulation of such models in the literature.

CHAPTER 3

Weekly Electricity Demand Forecasting Using Exogenous Weather Variables

3.1 Introduction

This chapter investigates the use of exogenous temperature variables to improve the accuracy of weekly electricity demand forecasts carried out one year in advance. Depending on the climate in which the electricity supply system operates weather is expected to have a significant effect on weekly electricity demand due to the seasonal temperature changes throughout a given year. However, although temperature may be an important influencing factor it is not always the dominant one. It is also the purpose of this chapter to determine the benefit of including a non-dominant temperature variable in a weekly electricity demand forecasting model. In addition, it investigates the optimal configuration of such a variable for both the dominant and non-dominant cases, where this involves the use of a heating degree day (HDD) variable calculated from an average temperature °C (AT) variable. To achieve the objectives, the analysis was carried out using data obtained from two notably different power boards operating in diverse global regions. The first of these electricity utilities is the Irish national power board operating an isolated network with peak load of 2500 MW, this system is referred to as System A. The second is a regional power board in Northern New Zealand (System B), operating a peak load of 14 MW. The Irish and Northern New Zealand climates are described broadly by the same climate region classification i.e. Temperate Oceanic, although due to the closer proximity of New Zealand to the equator the temperature here is moderately higher. Weather affects electricity demand in each system primarily through heating requirements, since cooling requirements (air conditioning etc.) are not a major factor. In the specific area of New Zealand in question, weather, although a major influencing factor, is not the dominant factor; electricity demand here is driven primarily by the agricultural and forestry industry. In Ireland on a national scale weather is the dominant factor. A comparison of the two power boards is summarised in Table 3.1;

Table 3.1 Comparison of System A and System B

System A	System B
Temperature dominant input	Temperature non-dominant input
Operating in Temperate Oceanic climate	Operating in Temperate Oceanic climate
Small Island National Power Board	Island Regional Power Board
Peak load 2000 MW	Peak load 14 MW

There are several studies devoted to the use of exogenous variables in forecasting electricity demand. Train *et* al (1984), when forecasting monthly electricity sales using econometric models includes both HDD and cooling degree days (CDD) calculated at different base temperatures within the same model. For the systems dealt with here, only heating effects are significant and HDD calculated at the one base temperature is included in the model. They concentrate on finding the best method of calculating the weather variables to associate with monthly sales billing data and highlight the importance of finding an appropriate form of an input variable from the data available.

Harris and Liu (1993) include weather, demographic and economic variables in the Box-Jenkins transfer function models that they use to forecast monthly residential electricity consumption in Carolina, USA. The weather variables, HDD and CDD, are adjusted through the use of weighting factors for the geographic distribution of the customer. This geographic adjustment is due to the varying weather conditions throughout North (temperate) and South (sub-tropical) Carolina. In contrast, the small size and geographical conditions, in Ireland and in the New Zealand region in question, are such that a single set of temperature measurements recorded in one central weather station is sufficient for forecasting purposes, and thus no geographical adjustments are required.

Schneider *et al* (1985) when forecasting 24-hour loads utilise a temperature deviation variable which reflects the change in the load due to the weather being different than expected, it is calculated as the expected minus the actual (or forecast) temperature. He observed that during the heating season there is a base temperature above which a change in temperature had a less significant effect on the load than when the temperature falls below this base temperature. A similar but opposite effect existed during the cooling season. An effective temperature deviation was thus calculated taking this into account; this same effect is catered for here through the use of HDD the variable.

Haida and Muto (1994) use a multivariate regression model to forecast daily peak load for the Tokyo Electric Power Company, where peak load depends on daytime temperature and humidity. They present a transformation technique (based on translation and reflection methods) to deal with the fact that the relationship between peak load and weather variables may be assumed to be linear in the summer and winter seasons but is nonlinear in the transitional seasons, i.e. spring and autumn. They use the basic forecasting model throughout the year and apply the transformation technique in the transitional seasons.

This chapter examines the relationship between electricity demand and HDD at different times of the year, specifically during the transitional months from spring to summer and summer to autumn and also during the summer months. In contrast to the work conducted in this chapter, the common factor in all of the above works is that the weather variables are considered to be dominant. However, only average temperature and HDD were considered here, whereas some of the above works consider the use of wind and illumination variables which may be significant in medium term forecasting, unfortunately this data was not available during the time of this study.

3.2. Description of Weekly Electricity Demand and Temperature Data

Weekly electricity demand (MWh) and AT, in degrees Celsius, are available from the from 4th April 1982 to 28th December 1991, a total of 508 points and from the 5th April 1980 to 25th August 1990, a total of 543 points for the System A and System B data sets respectively. The time series data for each system are presented below in Figure 3.1 to Figure 3.4; where the data has been scaled for confidentiality reasons. For each system, profiles for a single year only are also given to show more clearly the effect of temperature on electricity demand.

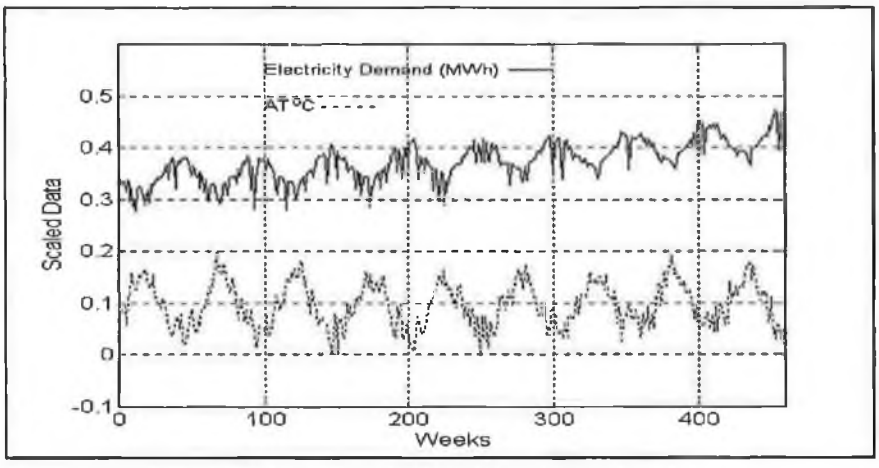

Figure 3.1 System A electricity demand and AT

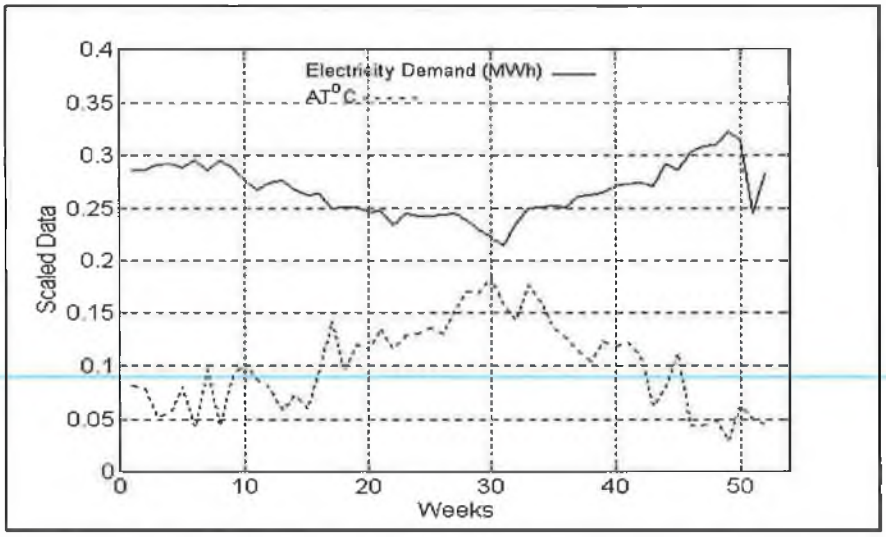

Figure 3.2 One year for System A

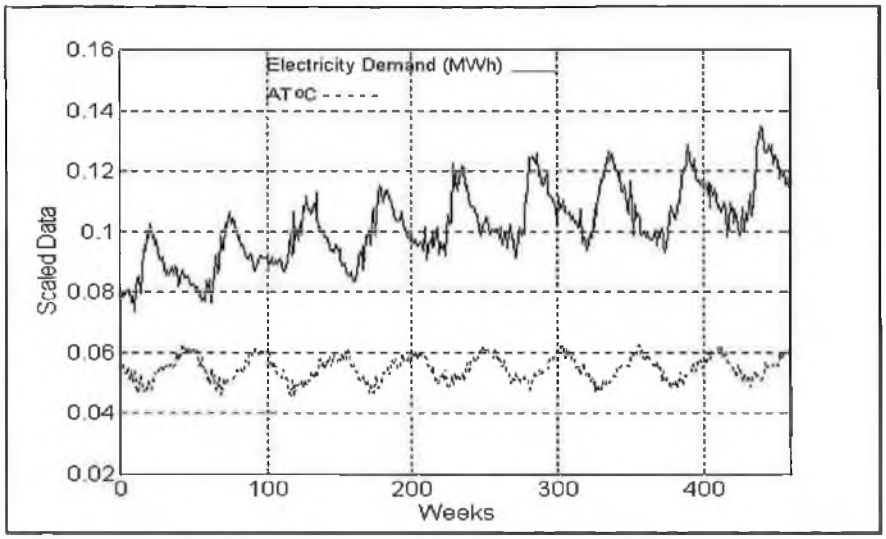

Figure 3.3 System B electricity demand and AT

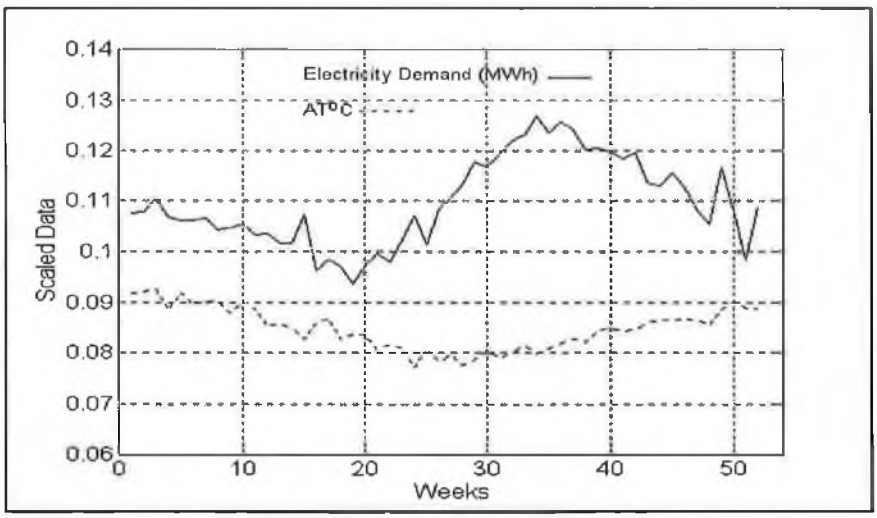

Figure 3.4 One year for System B

The plots of the electricity demand for System A (Figure 3.1) and System B (Figure 3.3), show that each time series exhibits a slowly rising trend coupled with strong seasonality of regular amplitude. For both systems, the rising trend is due to the annual growth in electricity consumption resulting from social, economic and demographic influences. For System B, the seasonal characteristics are due to annual temperature profile variations but are also due to the cyclical consumption patterns of the seasonal industries, primarily the dairy industry. Figure 3.4 shows that the temperature decrease in the winter months has the effect of increasing the electricity demand, however, demand continues to increase at a rate greater than the temperature decrease, which is due to the dominant influence of the dairy industry. The peak of electricity demand (August) is not co-incident with the trough in temperature (July), instead this peak corresponds to the peak in dairy production which occurs in Spring, i.e. August - October. Temperature is therefore a non-dominant exogenous variable in this system. In contrast, Figure 3.2 shows that the cyclical variations for System A are attributed to the seasonal temperature patterns throughout a given year. For this system, temperature is clearly a dominant exogenous variable, where a peak in temperature yields a trough in electricity demand.

3.3 System Temperature Exogenous Variables

For each system it is possible to use the AT data to evaluate a heating degree day at a particular base temperature ($HDD_{base\ temp}$). A $HDD_{base\ temp}$ is defined as follows:

If, AT < base temp **then**

$$HDD_{base\ temp} = \int_{day} \text{base temp} - AT$$

else

$$HDD_{base\ temp} = 0 \qquad (3.1)$$

Figure 3.5 depicts HDD_{18} for a typical summers day (System A) where the following are shown:

- AT on an hourly basis given by the solid line profile.
- the base temperature is given by the solid horizontal line at 18°C.
- HDD_{18} represented by the shaded area.

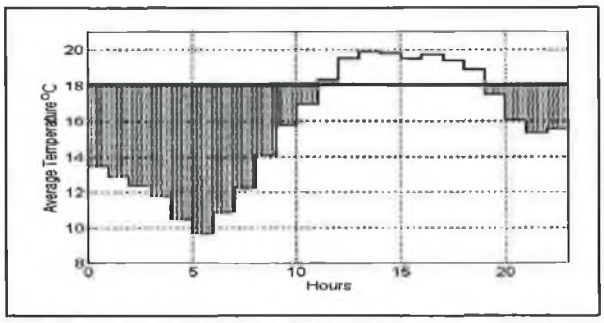

Figure 3.5 HDD_{18} for a typical summers day

3.3.1 Calculation of HDD

The AT data is available on different time-scales for each systems; hourly AT is available for System A and daily AT is available for System B. If HDD is to be used in the electricity demand forecasting models then it is required on a weekly basis. There are a number of ways in which to calculate *weekly* HDD depending on the time-scale on which the AT variable is based and it is necessary to choose the most appropriate AT time-scale to use in the calculation. In addition, the calculation of HDD involved the selection of the base temperature appropriate to the system. The appropriate base temperatures may be different in each system due basic differences between the systems but also due to the different comfort levels of the people in each country. Therefore, there are two factors involved in the calculation of the weekly HDD data:

- the AT time-scale
- the base temperature

3.3.1.1 HDD Base Temperature

It was necessary to determine for each system the base temperature that yields the weekly HDD values that are the most highly correlated with electricity demand. Therefore, the correlation between weekly electricity demand and weekly HDD using different base temperatures were compared. The correlation coefficient (Priestly, 1981) of electricity demand with each of the weekly HDD figures calculated using the base temperatures in the interval 14°C - 21°C were computed, where the objective was to select the base temperature which yielded highest correlation coefficient. The correlation coefficient results are given in Table 3.2.

Table 3.2. Correlation coefficients for electricity demand and HDD

System	Base Temperature							
	14°C	15 °C	16°C	17°C	18°C	19°C	20°C	21°C
A	0.85	0.86	0.86	0.86	0.87	0.85	0.85	0.85
B	0.49	0.50	0.50	0.50	0.51	0.50	0.50	0.50

The results show that the correlation coefficients do not vary greatly across the base temperature range. However, using these results it was possible to select the base temperature that yielded the highest correlation coefficient for each system, that is 18^{o}C. It can also be seen that the correlation coefficient values are higher for System A than for System B since temperature is a dominant exogenous variable in this system.

3.3.1.2 AT Time-Scale Used in HDD Calculation

Different methods may be used to calculate HDD depending on the AT time-scale available. For example, if AT is available on an hourly basis then weekly HDD may be calculated as follows:

Method 1: AT time-scale = hourly

If for each hour, hourly AT < base temp ***then***

$$HDD_{\text{hourly contribution}} = \text{base temp} - \text{hourly average temp}$$

else

$$HDD_{\text{hourly contribution}} = 0$$

endif

$$HDD_{\text{daily contribution}} = \frac{1}{24}\int_{00:00}^{23:00} HDD_{\text{hourly contribution}}$$

$$HDD_{\text{weekly contribution}} = \int_{\text{day 1}}^{\text{day 7}} HDD_{\text{daily contribution}}$$

(3.2)

If AT is available on a daily sampling period then HDD may calculated as follows:

Method 2: AT time-scale = daily

If for each day, daily AT < base temp ***then***

$$HDD_{\text{daily contribution}} = \text{base temp} - \text{daily average temp}$$

else

$$HDD_{\text{daily contribution}} = 0$$

endif

$$HDD_{\text{weekly}} = \int_{\text{day 1}}^{\text{day 7}} HDD_{\text{daily contribution}}$$

(3.3)

Alternatively, using AT on a weekly basis the HDD may be calculated using:

Method 3: AT time-scale = weekly

If for each week, weekly AT < base temp ***then***

$$HDD_{weekly} = (\text{base temp} - \text{weekly average temp}) * 7 \quad (3.4)$$

else

$$HDD_{weekly} = 0$$

endif

In the case of System A all three methods may be used. The first method may be employed directly using the hourly AT; the use of the second method involved the prior calculation of daily AT values using the hourly data and the use of Method 3 involved the prior calculation of weekly AT using the hourly data. In the case of System B Method 2 and Method 3 can be employed. The daily AT can be used directly in Method 2 and Method 3 can be employed following the calculation of weekly AT using the daily AT data.

Weekly HDD, using a base temperature of 18^0C, was produced using the above methods, (3.2) to (3.4) for System A and (3.3) to (3.4) for System B. It was found, however, that the figures varied during certain months of the year. Figure 3.6 shows an example of this where weekly HDD_{18} for the System B data set calculated using Method 2 and Method 3 are shown, from January 1983 to December 1985.

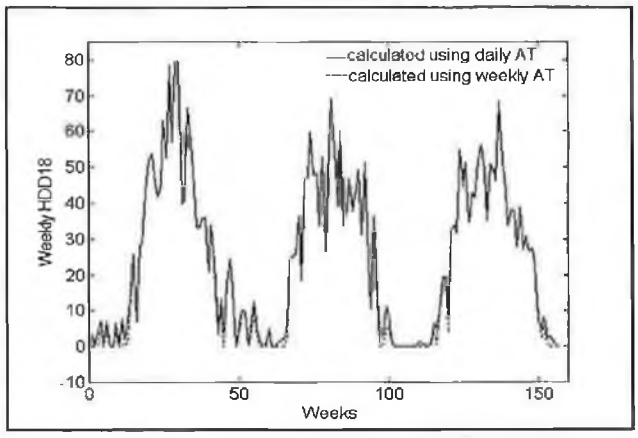

Figure 3.6 Weekly HDD_{18} 1983 - 1985 - System B

It can be seen from the graph that the weekly HDD_{18} figures differ during the spring and summer months only. Figure 3.7 provides a closer look at graph given in Figure 3.6 from October 1983 to April 1985.

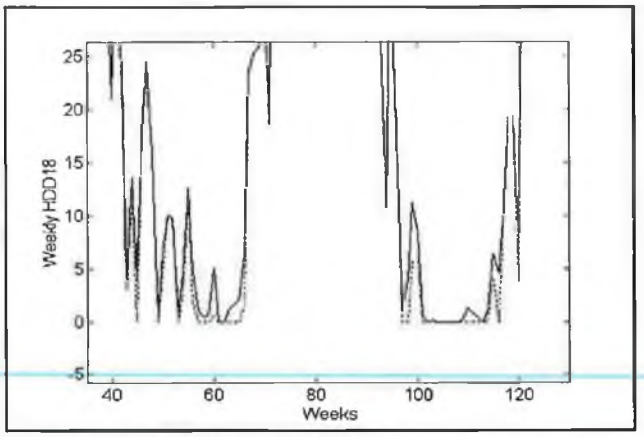

Figure 3.7 Weekly HDD_{18} Oct. 1983 - Apr. 1985 - System B

This graph indicates that the most significant differences occur during the seasonal transitional months at spring (November) and autumn (March), and also during the summer months, December - February. Figure 3.8 gives a plot of the magnitude of the differences between the HDD_{18} data shown in Figure 3.6 (1983 - 1985) to ascertain if the greatest variance between the HDD_{18} calculated using the different methods for these two years does indeed occur during the seasonal transitional and summer months.

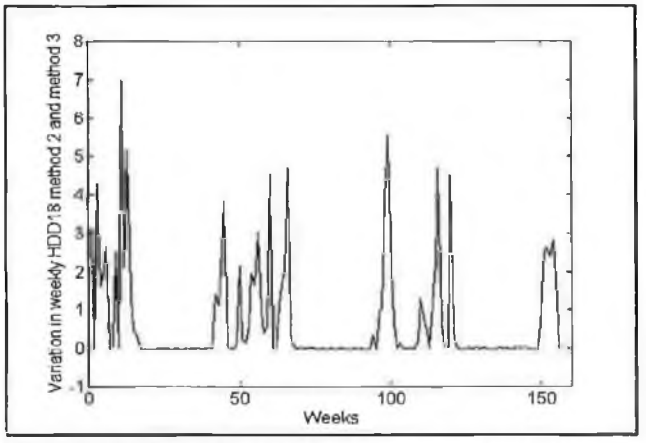

Figure 3.8. Magnitude of differences 1983/1985 - System B

It is clear from this graph where the main differences occur and it is also seen that the maximum variance is as high as 7 degree days. For System A, the weekly HDD_{18} figures calculated using Methods 1 to 3 exhibited a similar effect. In this case the differences occurred during the seasonal transitional months at May and September and also during the summer months June - August, where for the remaining months of the year the differences were minor.

The variance between weekly HDD_{18} produced using the different methods was calculated over the entire data set for each system, where Table 3.3 gives the maximum and minimum values of the differences.

Table 3.3 Magnitude of variance between HDD_{18} calculated using different methods

Difference Between Methods	System A		System B	
	Maximum	Minimum	Maximum	Minimum
1 and 2	8.34	-0.24	N/A	N/A
1 and 3	11.04	-0.46	N/A	N/A
2 and 3	4.7	-0.3	8.60	-0.03

Since there is a difference between the HDD_{18} figures computed by each of the different methods, it was necessary to determine the method that yielded the weekly HDD values that are the most highly correlated with weekly electricity demand. The correlation coefficient between electricity demand and HDD_{18} calculated using each of the different methods was computed for the data points for which the differences in the HDD_{18} values occurred. Use of the correlation coefficient as a criterion failed to distinguish the most appropriate method to use. Therefore, for each system a one-year-ahead prediction was carried out using an BJTF model (2.22) driven with weekly HDD_{18} produced by the different methods. The accuracy of the predictions were compared using the MAE, with the results presented in Table 3.4.

Table 3.4 Forecasting accuracy results using HDD_{18} calculated using different methods

Method for HDD	MAE - System A	MAE - System B
1	0.89×10^4	n/a
2	0.93×10^4	2.70×10^5
3	0.94×10^4	2.95×10^5

Using the MAE as a criterion, the calculation of weekly HDD_{18} using AT based on the finest time-scale available, hourly AT for System A and daily AT for System B, yielded the most accurate forecasting results. Thus, the HDD_{18} data was calculated using Method 1 (3.2) for System A and Method 2 (3.3) for System B.

3.4 Linear Forecasting Models

In order to determine if a gain in forecasting accuracy may be achieved through the use of an exogenous weather variable when predicting weekly electricity demand both univariate and bivariate forecasting models were built and their forecasting accuracy results compared. The forecasting accuracy of a bivariate model based on weekly electricity demand and weekly AT were compared to those obtained from a model based on weekly electricity demand and weekly HDD_{18} in an attempt to assess the advantage of using of HDD_{18} over AT in the forecasting

model. It was also necessary to develop univariate forecasting models to predict future values of the temperature variables for use when performing the bivariate model forecast of electricity demand.

In each system the weekly time series are nonstationary and exhibit trend and seasonal characteristics. In an effort to draw a broad conclusion two different classes of linear forecasting model that are suitable for representing a time series possessing the above characteristics were employed. The first class are the classical time series models (2.17 to 2.22) and the second class are the state space structural models (2.44 and 2.50).

For each system the data was divided into an identification and prediction set, where Table 3.5 described the data sets for each system.

Table 3.5 Data sets for each system

System	Identification Data Set	Prediction Data Set
A	04/04/1982 - 23/12/1990	30/12/1990 - 22/12/1991
B	30/03/1980 - 20/08/1989	27/08/1989 - 19/08/1990

Different software packages were employed to build the different forecasting models, where this software and the hardware used in each case are described in Table 3.6

Table 3.6 Hardware and software used to build time series forecasting models

Models	Software	Hardware
AR, ARMA, ARX and ARMAX	System Identification Toolbox for Use with MATLAB (The MATH WORKS Inc., 1991)	33 MHz IBM Compatible 486 PC with 16 MB of RAM.
SARI and BJTF	BMDP (BMDP, 1988)	DECVAX 6230
Structural	NAG Foundation Toolbox for Use with MATLAB (The MATH WORKS Inc., 1995)	160 MHz Pentium (586) PC with 16 MB of RAM.

3.4.1 Classical Models

The basic classical time series modelling methodology of identification, estimation and model adequacy checking described in Section 2.4.3 of Chapter 2 was adopted for the classical

models developed here. Although this is an iterative procedure where each step is dependent on the other steps of the procedure, for clarity each step of the procedure is described separately.

3.4.1.1 Transformation of Weekly Electricity Demand Time Series

Transformations were applied to the weekly electricity demand time series where two different approaches were adopted. The first approach involved the removal of the linear trend and bias from the weekly electricity demand time series using (2.13) and the second involved the removal of the trend and seasonality from the data through differencing (2.15). Figure 3.9 and Figure 3.10 show the detrended data for System A and System B respectively.

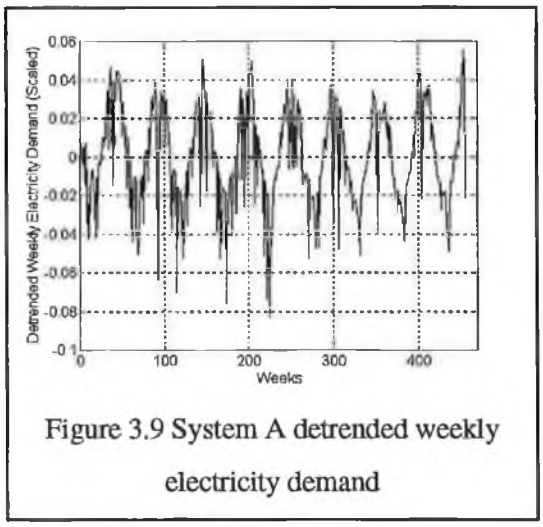

Figure 3.9 System A detrended weekly electricity demand

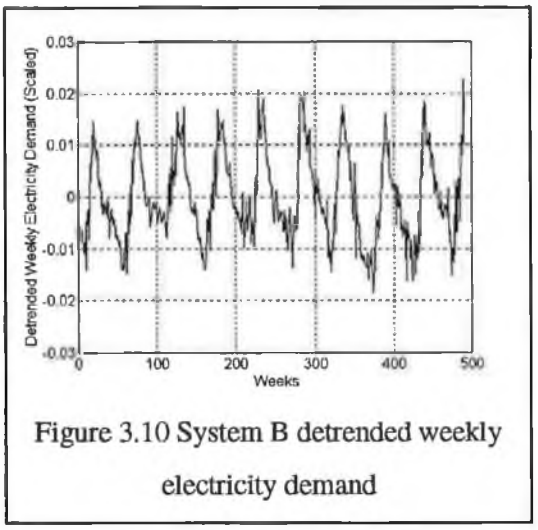

Figure 3.10 System B detrended weekly electricity demand

In the case of the differenced data it was necessary to determine an appropriate order for d and D in (2.15). The determination of these orders was carried out using the procedure described in Section 2.4.1.2. Figure 3.11 shows an example of the SACF, with the corresponding significance level (2.24), for the differenced System A electricity demand data with d=1 and D=1 and s=52.

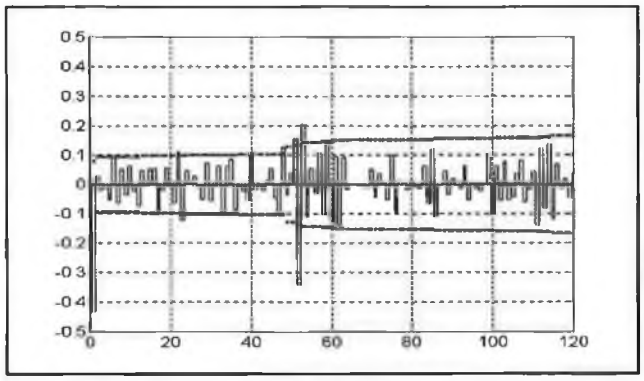

Figure 3.11 SACF for System A differenced electricity demand

The difference between the SACF shown in Figure 3.11 and the SACF of the original time series shown in Figure 2.7 of Chapter 2 is clear. In the differenced case the SACF does not posses a sinusoidal characteristic because the sesonality has been removed from the data. Instead the SACF cuts off at lag 1 at the nonseasonal level and at approximately lag s at the seasonal level, consequently the transformed time series was to be differenced appropriately. A similar analysis was performed on the System B data set, where the order of differencing used to difference the time series was d=1 and D=1 and s=52. Figure 2.10 of Chapter 2 shows the differenced data for System A and Figure 3.12 shows the differenced electricity demand data for System B.

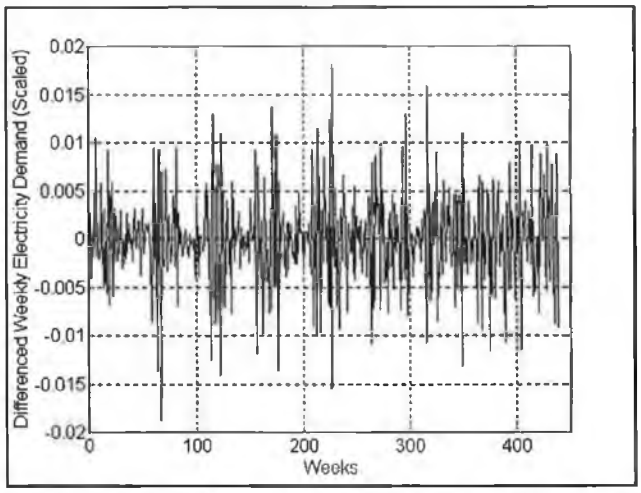

Figure 3.12 System B differenced weekly electricity demand

3.4.1.2 Weekly Electricity Demand Univariate Model Identification

It was necessary to determine a suitable univariate model structure to represent the detrended and differenced time series. The Box-Jenkins model structure identification techniques described in Section 2.4.3.1 were employed, where this involved the examination of the SACF(2.8) and SPACF (2.9) of the detrended and differenced data. Plots of these functions for the System A and System B time series are given in Figure 3.13 to Figure 3.20, where the significance levels (2.24 and 2.25) are shown by the dashed line.

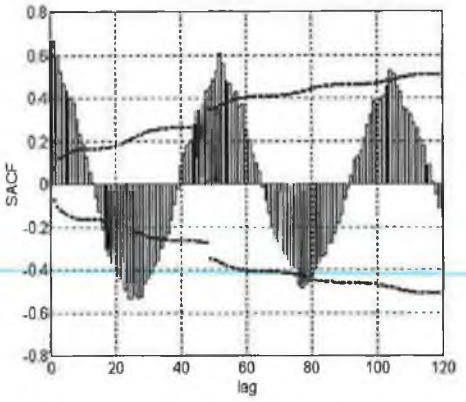

Figure 3.13 SACF for System A detrended electricity demand $d(k)$

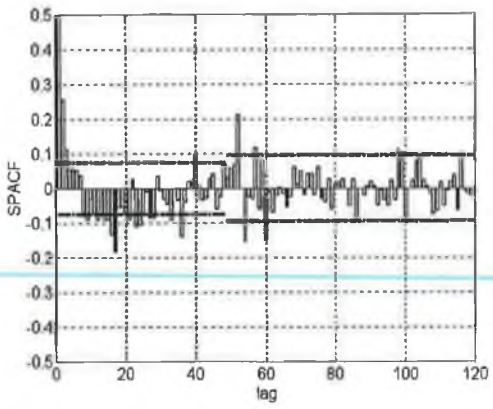

Figure 3.14 SPACF for System A detrended electricity demand $d(k)$

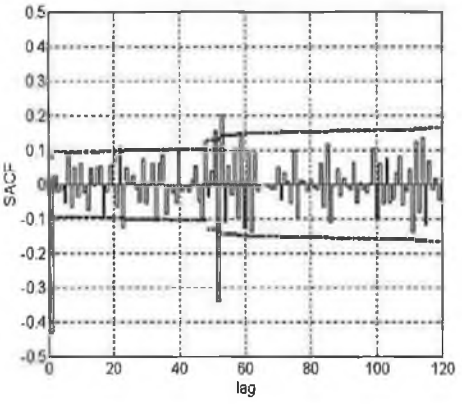

Figure 3.15 SACF for System A differenced electricity demand $z(k)$

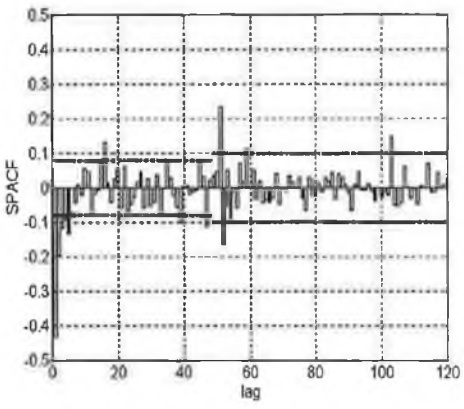

Figure 3.16 SPACF for System A differenced electricity demand $z(k)$

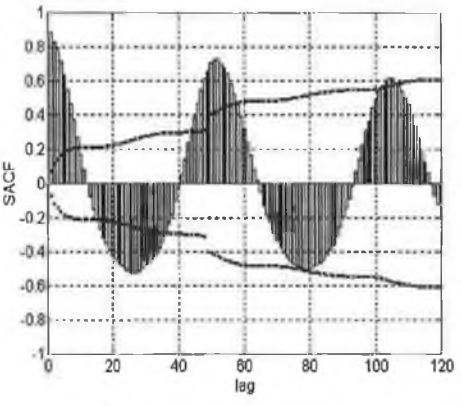

Figure 3.17 SACF for System B detrended electricity demand $d(k)$

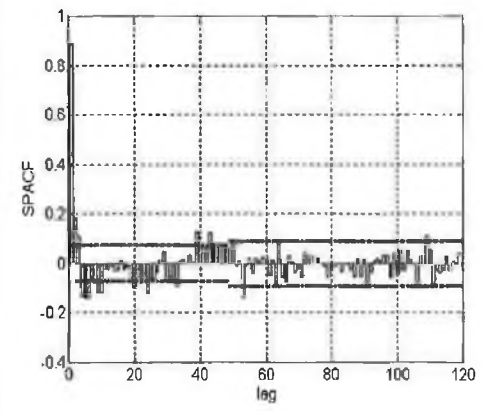

Figure 3.18 SPACF for System B detrended electricity demand $d(k)$

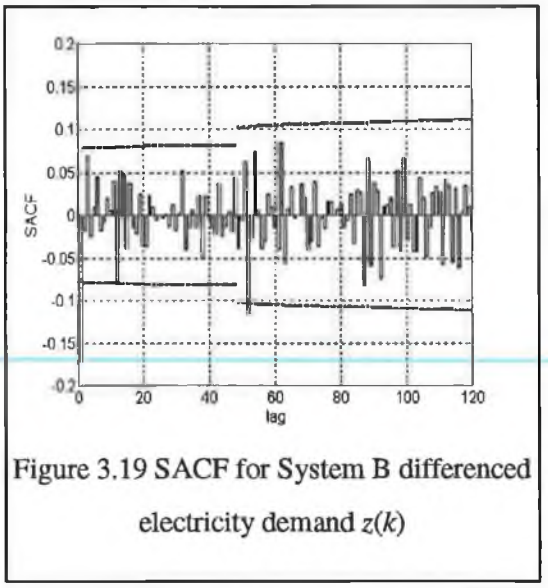

Figure 3.19 SACF for System B differenced electricity demand $z(k)$

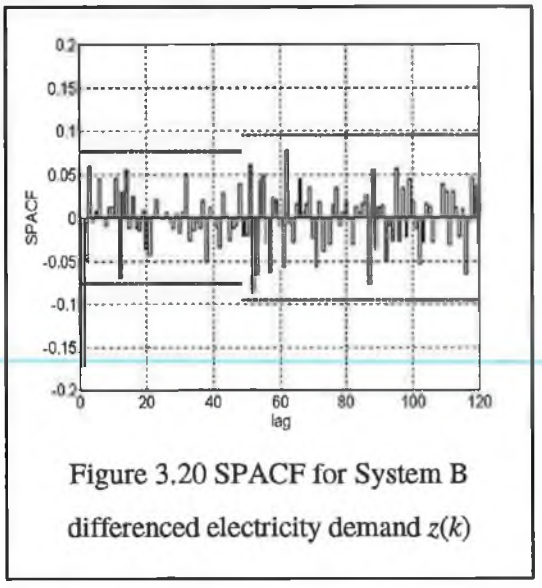

Figure 3.20 SPACF for System B differenced electricity demand $z(k)$

It can be seen that depending on the form of transformation used that the profile of the SACF and SPACF differ considerably. This is particularly prominent in the case of the SACF, where for the detrended data the function has a sinusoidal profile because the seasonality has not been removed from the time series. The sinusoidal pattern is not present in the SPACF of the data because this function calculates the correlation between $d(k)$ and $d(k+\tau)$ but were the effect of intervening observations are eliminated. Due to the presence of this sinusoidal pattern it was not possible to use the SACF to determine a suitable model structure for the detrended data and the identification techniques based on the use of criterion functions described in Section 2.4.3.2 of Chapter 2 were employed.

Examination of the SACF and SPACF of the System A differenced data, Figure 3.15 and Figure 3.16, show that both functions have cut offs and therefore a model with either an autoregressive or moving average operator is appropriate. In the case of the SACF there are cut offs at lag 1 at the nonseasonal level and approximately at lag 1s at the seasonal level and therefore a seasonal integrated moving average model (SIMA) derived form (2.19), with $m = 1$ and $M = 1$, was fitted to the differenced data. Examination of the SPACF shows that there are cut offs at the nonseasonal level at lag 5 and at the seasonal level at lag 2s, therefore a seasonal autoregressive integrated model (SARI) model also derived from equation (2.19), with $p = 5$ and $P = 2$, was tentatively identified for the series. The SIMA and SARI models were estimated using maximum likelihood estimation, see Section 2.4.3.3, and the adequacy of the models was checked using the techniques described in Section 2.4.3.4. The t-ratio test was used to determine the significance of an estimated parameter. This is defined as the ratio of the parameter estimate to the standard error of the parameter estimate. The guideline that was

used was that if the t-ratio was greater than 2 then the parameter was employed in the model (Bowerman and O'Connell, 1987). Based on the t-ratio test the SIMA model did not appear to be an appropriate model to represent the differenced series, where none of the parameters of the model were found to be significant. Consequently, the SARI model was used, where it was found using the t-ratio test that autoregressive parameters at lag 6 and at lag $2s$ were significant and thus the final model structure identified for the differenced time series was $p = 6$ and $P = 2$.

Similarly, both the SACF and SPACF of the differenced data for System B cut off; see Figure 3.19 and Figure 3.20. The cut offs at 1 and $1s$ in the SACF suggest an SIMA model with $m = 1$ and $M = 1$ and the cut off at 1 in the SPACF suggest an SARI model with $p = 1$. Using a similar analysis to that used in the System A case it was determined that a SARI model was the appropriate model structure. However, using the t-ratio test the final SARI model structure for the System B differenced data was determined to be $p = 3$ and $P = 2$.

This anomaly between the cut offs on the SPACF and the final SARI model structure identified highlights the unreliability of the sole use of the SPACF function to determine the model structure.

Due to the problem associated with the use of the SACF to identify a model for the detrended time series the approach based on criterion functions was employed. Different ARMA and AR model structure performances were compared according to the loss function criterion given by (2.26). Based on the Box-Jenkins analysis an MA model was not fitted to the data. The performance of the models were compared on a different data set to which they were estimated on. The identification data set was divided into an estimation data set and a validation data set. For both the System A and System B data sets at least one year of weekly data was used for the validation set, where the size of this set this depended on the model order, and the remainder of the data was used for the estimation set. In the case of the AR model the following model structures were compared:

- na =1 to 60 chosen to span the seasonal length

Figure 3.21 and Figure 3.22 show a plot of the loss function versus na for System A and System B respectively

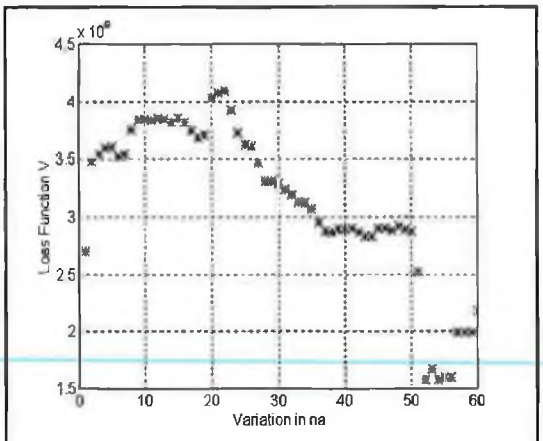

Figure 3.21 Different AR structures for System A

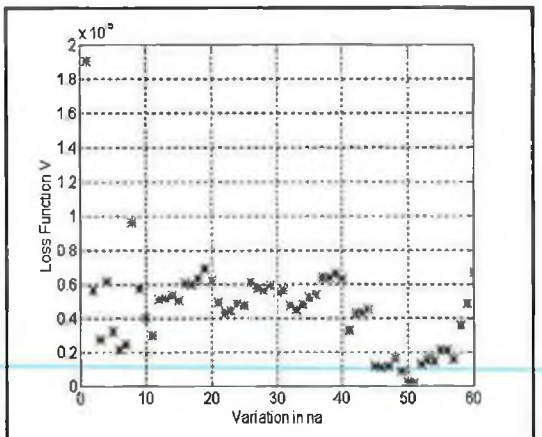

Figure 3.22 Different AR structures for System B

In the case of the System A data set the model order that yields the minimum loss function is equal to 52, with model orders 54, 55 and 56 also yielding low function values. Similarly, in the case of System B the minimum loss function was achieved for na = 52. Therefore, AR models with na = 52 were fitted to the detrended System A and System B electricity demand data.

In the case of the ARMA model the following model structures were compared:

- na = 52 chosen based on AR model analysis.
- nc = 1 to 20, chosen relative to na.

The plots of the loss function for different ARMA model structures for the detrended series $d(k)$ for both systems are given in Figure 3.23 and Figure 3.24, where the variation in the order of nc is shown on the x-axis.

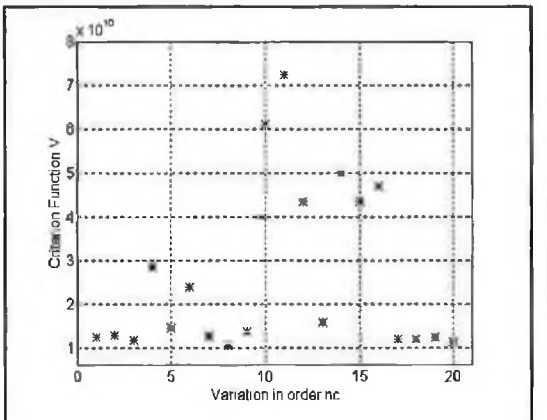

Figure 3.23 Different ARMA structures for System A

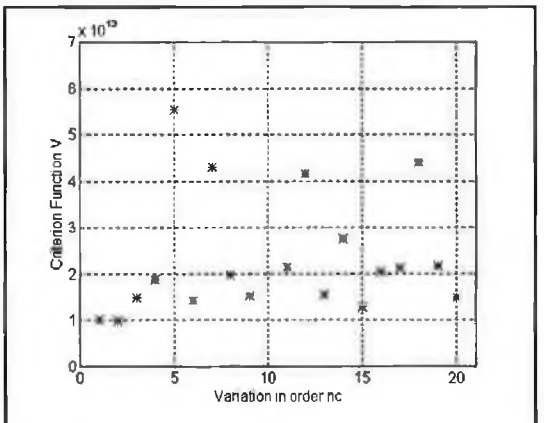

Figure 3.24 Different ARMA structures for System B

The analysis determined an ARMA model structure with na = 52 and nc = 8 for the System A data set and na = 52 and nc = 2 for the System B data set.

3.4.1.3 Temperature Variable Univariate Model Identification

Using a similar analysis to the one used to develop the weekly electricity demand models in Section 3.4.1.2 univariate models were developed for AT and HDD_{18}. Since the AT and HDD_{18} time series do not posses a trend characteristic it was not necessary to apply a detrending or differencing transformation to remove this from the data. An AR model was fitted to the original AT data, where the model order was determined using a loss function analysis to be na = 48 in the case of System A and na = 41 in the case of System B. In the case of the HDD_{18} time series for System A the AR model order was also determined to be equal to 48 but in the case System B the model order is na = 49. A SARI model was also fitted to the time series after it was differenced to remove the seasonal component from the data, however, it was found that this model produced poor forecasting results and thus was discarded from the model set.

3.4.1.4 Bivariate Model Structure and Transformation of Data

As in the case of the univariate models two different pre-processing approaches were adopted for the bivariate models. The first approach involved modelling the detrended weekly electricity demand time series and the original weekly temperature time series, where it was not necessary to detrend the AT or HDD_{18} data as these time series do not posses a trend characteristic. The second approach involved differencing both the weekly electricity demand and temperature time series to remove the trend and seasonal components with the subsequent fitting of a BJTF model (2.22) to the data. In Section 3.4.1.1 the order of the differencing operator for weekly electricity demand series was determined to be d=1, D=1 and s=52 for both System A and System B. Using a similar analysis the order of the differencing operators for the AT and HDD_{18} series for System A and System B were also determined to be d=1, D=1 and s=52.

3.4.1.5 Bivariate Model Identification

A BJTF model (2.22) was fitted to the differenced weekly electricity demand and differenced temperature data. The identification of the structure of this model was carried using the techniques described in Section 2.4.3.1. This involved the examination of the SCCF (2.12) calculated between the electricity demand and exogenous variable time series. Plots of this

function are Figure 3.25 to Figure 3.28, where the ninety five percent significance levels, given as 1.96 $\pm(N^{1/2})$ (Brockwell and Davies, 1987) are shown by the * symbol.

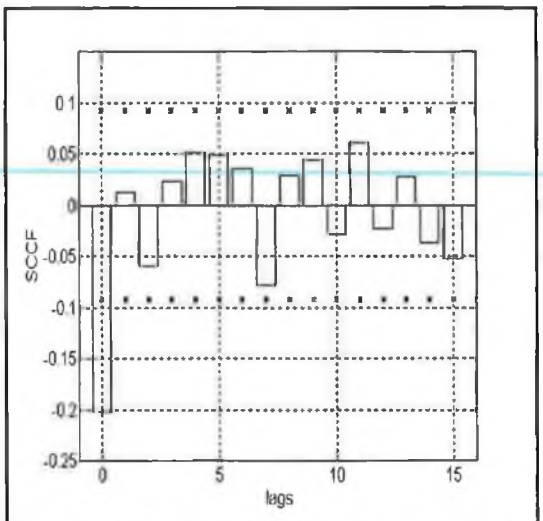

Figure 3.25 SCCF for electricity demand and AT - System A

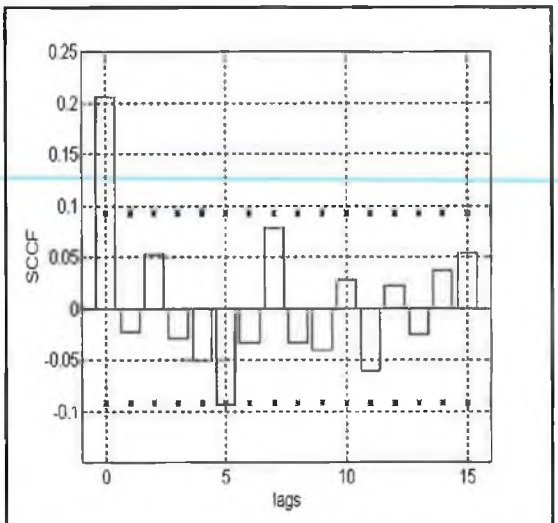

Figure 3.26 SCCF for electricity demand and HDD_{18} - System A

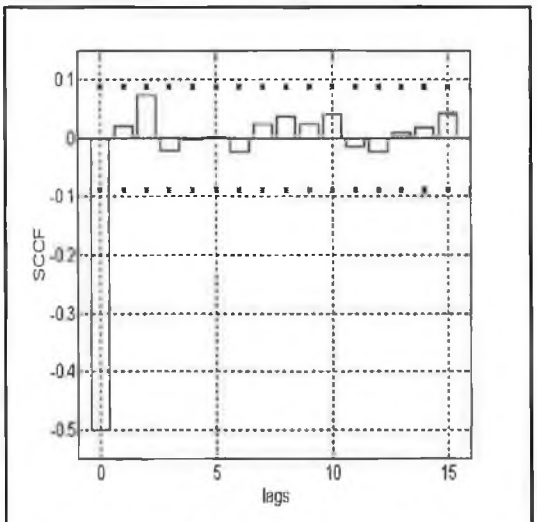

Figure 3.27 SCCF for electricity demand and AT - System B

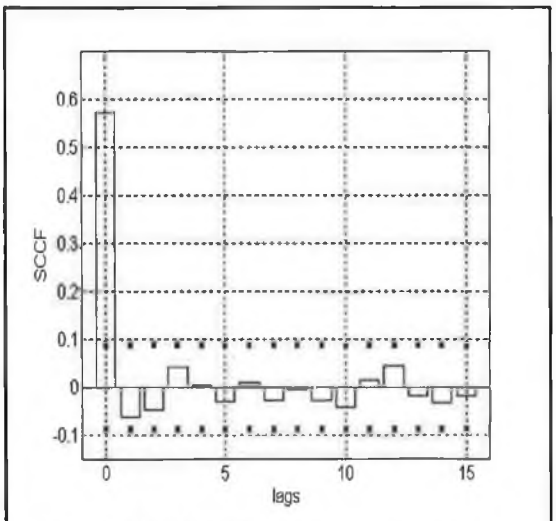

Figure 3.28 SCCF for electricity demand and HDD_{18} - System B

The SCCF was examined using the guidelines outlined in Section 2.4.3.1 of Chapter 2. Examination of the graphs show that for each system and for each exogenous variable that the order of the delay parameter *b* is equal to zero. For System A the SCCF for both AT and HDD_{18} starts to decay immediately after the first cut off at lag 0 and therefore *nb* was set equal

to zero in each model. In addition, in each case the SCCF dies down in a damped sine wave fashion and thus was set *nf* was set equal to 2. In the case of System B *nb* was set equal to zero for both the AT and HDD_{18} models since the SCCF starts to decay immediately after the first cut off at lag 0. In this case it is more difficult to establish how the SCCF is decaying and model structures with *nf* set equal to 0, 1, 2 were estimated and the *t*-ratio associated with the parameter examined for significance. Using this method the order for *nf* was determined to be equal to 0 for both AT and HDD_{18} models.

ARX (2.20) and ARMAX (2.21) models were fitted to the detrended weekly electricity demand and original temperature time series. The model structures were determined by comparing the performance of the loss function (2.26) as described in the univariate analysis. For each system the following ARX model structures were compared:

- *na* = 52 chosen based on AR model analysis.
- *nb* = 1 to 20, chosen relative to *na*.
- *b*= 0 chosen based on SCCF analysis.

Table 3.7 describes the model structures determined.

Table 3.7 Description of ARX model structures

Exogenous Variable	System A			System B		
	na	*nb*	*b*	*na*	*nb*	*b*
AT	52	3	0	52	2	0
HDD_{18}	52	1	0	52	10	0

The determination of the ARMAX model structure involved the comparison of the performance of models with the *na*, *nb* and *b* orders given in Table 3.7 and with different values of the *nc* parameter in the range from 1 to 20. The model structures determined for each case are described in Table 3.8.

Table 3.8 Description of ARMAX model structures

Exogenous Variable	System A				System B			
	na	*nb*	*nc*	*b*	*na*	*nb*	*nc*	*b*
AT	52	3	7	0	52	2	3	0
HDD_{18}	52	1	3	0	52	10	1	0

3.4.1.6 Estimation and Model Adequacy Checking of Classical Models

The techniques used by the software packages to estimate the univariate and bivariate models were previously described in Section 2.4.3.3. The adequacy of the estimated models was

checked using the methods outlined in Section 2.4.3.4. For example, Figure 3.29 shows a plot of the residuals for the SARI model for the System A data set.

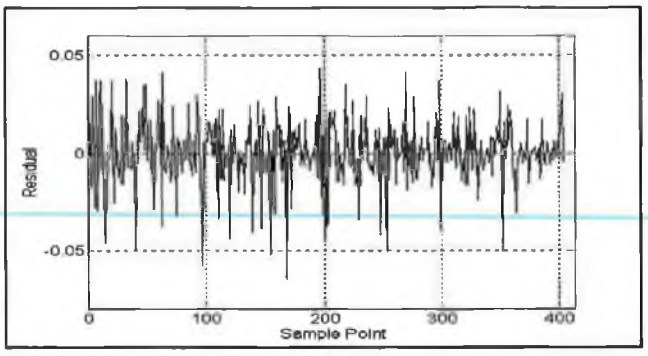

Figure 3.29 Plot of residuals of SARI model for System A

The plot shows that the residuals are random and show no distinguishable pattern. The Ljung-Box statistic for $T = 21$ (that is, $N^{1/2}$) was calculated as 21.34 and $\chi^2_{[.05]}(13) = 22.36$ and therefore the model was accepted as adequate. A similar analysis was performed on the other univariate models.

For the bivariate models independence between the residuals and past values of the exogenous variable of the model was tested. This was checked by examining a plot of the SCCF (2.12) calculated between the residual and the exogenous variable time series. For example, Figure 3.30 shows the SCCF calculated for the residuals of the System A ARX model and AT. Ninety five percent significance levels based on the hypothesis that the residuals are white and independent of the exogenous variable are given by the dashed line (Ljung, 1987). A similar plot is shown in Figure 3.31 for the System B data set for the ARMAX model with the HDD_{18} exogenous variable.

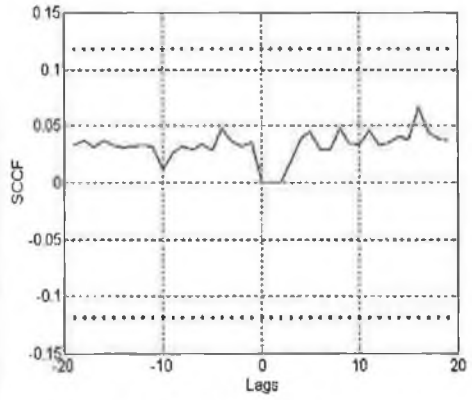

Figure 3.30 SCCF of residuals of ARX model and AT for System A

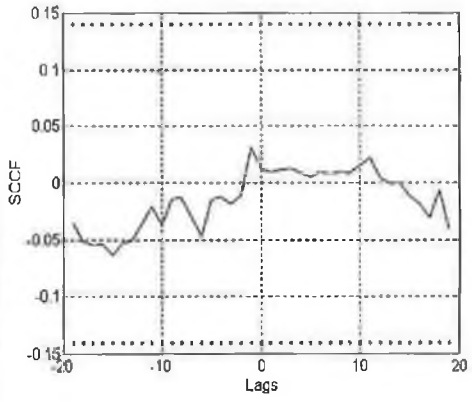

Figure 3.31 SCCF of residuals of ARMAX model and HDD_{18} for System B

In each case there are no cut offs and thus independence between the residuals and past values of the exogenous was assumed. A similar analysis was performed on all of the bivariate models.

3.4.2 State Space Structural Models

3.4.2.1 Univariate and Bivariate Model Structure

In the case of the structural models it was not necessary to transform the weekly time series data as this model can handle nonstationary series. A BSM with a dummy seasonal component was fitted to the System A and System B weekly electricity demand (2.44). From a complexity point of view a dummy seasonal component was employed rather than a trigonometric component (2.45 and 2.46) due to the relatively large seasonal length, s=52. A univariate BSM with dummy seasonal component was also used to model the AT and HDD_{18} time series.

A state space structural model with exogenous variable (2.49) was fitted to the weekly electricity demand and AT time series and also to the weekly electricity demand and HDD_{18} time series. The classical linear analysis carried out in Section 3.4.1.5 determined that the lag between weekly electricity demand and each of the weather variables is equal to zero. Therefore, the bivariate structural model may be derived from (2.49) and is given by the following equation:

$$y(k) = \mu(k) + \gamma(k) + u(k)g + \varepsilon(k) \tag{3.5}$$

where the trend component and seasonal components are represented by (2.44). The state space form of the (3.5) is given by equation (2.51) and (2.52), where the matrix **F** is given by the state transition matrix of the BSM with dummy seasonal component state space model given in (2.44)

3.4.2.2 Estimation of Structural Models

It was necessary to estimate the hyperparameters, $\{\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2\}$, of the BSM with dummy seasonal component and BSMX with dummy seasonal component models, where maximum likelihood estimation via the prediction error decomposition was used. A description of this was previously given in Section 2.5.3 of Chapter 2.

In order to reduce the computational time the number of parameters which it was required to estimate was reduced by reparameterising the structural models (Harvey, 1989). This was

achieved by choosing one of the variances of the disturbance terms as a scaling factor, denoted σ_*^2. In all cases σ_ε^2 was chosen as the scaling factor and thus the other variances given in terms of the scaling factor are:

$$\sigma_\eta^2 = \frac{Var(\eta(k))}{\sigma_\varepsilon^2}, \qquad \sigma_\zeta^2 = \frac{Var(\zeta(k))}{\sigma_\varepsilon^2} \quad \text{and} \quad \sigma_\omega^2 = \frac{Var(\omega(k))}{\sigma_\varepsilon^2}$$

with the scaling variance σ_*^2 is fixed to unity. Therefore, the number of hyperparameters required to be estimated was reduced to three, where the new vector of hyperparameters is given by $\psi=[\psi_* \ \sigma_*^2]$, where $\psi_*=[\sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2]$. It was necessary to rewrite the likelihood function given by (A.17) in terms of the newly defined parameters as, where this is given as:

$$\log L(\psi) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma_*^2 - \frac{1}{2}\sum_{k=1}^{N}\log\Sigma(k) - \frac{1}{2\sigma_*^2}\sum_{k=1}^{N}\frac{v(k)^2}{\Sigma(k)} \tag{3.6}$$

This expression was maximised with respect to σ_*^2 by setting the derivative to zero and solving to give

$$\sigma_*^2 = \frac{1}{N}\sum_{k=1}^{N}\frac{(v(k))^2}{\Sigma(k)} \tag{3.7}$$

Substitution of (3.7) into (3.6) yields the likelihood function in terms of the new parameters as:

$$\log L(\psi) = -\frac{N}{2}(\log 2\pi + 1) - \frac{1}{2}\sum_{k=1}^{N}\log\Sigma(k) - \frac{N}{2}\log\sigma_*^2 \tag{3.8}$$

The likelihood function (3.8) was maximised with respect to ψ, where a numerical optimisation routine was employed. The numerical routine employed was provided by the NAG Foundation Toolbox For Use with MATLAB (Matlab, 1994), where the routine which was used is E04UCF. This routine minimises a function of several variables using the sequential quadratic programming method with linear constraints and simple bounds on the variables (Taha, 1987). It was necessary to supply the function to be minimised, which in this case was -log $L(\psi)$, and an initial set of parameter values. In all cases, the initial parameters values were set equal to 0.01. Bounds were placed on the variables which in this case were that the σ_ζ^2 and σ_ω^2 are strictly positive (Harvey, 1989). The estimates of the hyperparameters for each model are given in Table 3.10 to Table 3.15 of Section 3.5.

3.4.2.3 Model Adequacy Checking of Structural Model

The techniques employed to check the adequacy of the classical linear models were also employed here. This involved checking that the residuals of the model were approximately random, that there was no serial correlation among in the residuals and that there was independence between the residuals and past inputs. The Ljung-Box test statistic (2.35) was also applied to the structural models where *np* in equation (2.36) (the number of parameters estimated) was in all cases equal to 3 since the models were reparameterised.

3.4.3 Summary of Classical and State Space Modelling Methods

Table 3.9 provides a summary of the linear modelling methods employed.

Table 3.9 Summary of modelling methods

Model	Transform	Identify	Estimation
AR	Detrend *y(k)*	*na*	Least Squares
ARMA	Detrend *y(k)*	*na, nc*	Maximum Likelihood
SARI	Difference *y(k)*	*d, D, s, p, P*	Least Squares
ARX	Detrend y(k)	*na, b*	Least Squares
ARMAX	Detrend *y(k)*	*na, nc, b*	Maximum Likelihood
BJTF	Difference *y(k)* Difference *u(k)*	*d, D, s, b, nb, nf, p, P*	Maximum Likelihood
BSM	None	*s*	Maximum Likelihood via PED
BSMX	None	*s*	Maximum Likelihood via PED

3.5 Forecasting Results

The forecasting accuracy results for the prediction set (see Table 3.5) produced by each of the models are given in Table 3.10 for System A and Table 3.13 for System B respectively. The results for the prediction of future values of the AT and HDD_{18} temperature variables are given in Table 3.11 and Table 3.12 for System A and Table 3.14 and Table 3.15 for System B. Note that, in the case of the HDD_{18} since there are zero values in the actual data over the prediction horizon the calculation of the MAPE and MPE results in a division by zero and thus these forecasting accuracy results are not calculated for these models. Associated with the forecasting accuracy result of all models is a summary of the model structure, where in the case of the structural models the values of the estimated hyperparameters are given. Note that, "diff" refers to differencing transformation and NF and NM transfer function model and noise model respectively in the BJTF.

Table 3.10 Electricity demand forecasting results for System A

No	Model	Input	Structure	MAE x 10^{-4}	MSE x 10^{-8}	MAPE	MPE
1	SARI		diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3,4,5,6), P=(52,104)	1.0691	1.8223	3.91	-2.64
2	AR		na=52	1.1024	2.1639	3.83	2.09
3	ARMA		na=52, nc=8	0.9721	1.9978	3.41	1.94
4	BSM		σ_ε=0.1253 x 10^{-3} σ_η=0.1643 x 10^{-7} σ_ζ=0.0001 x 10^{-7} σ_ω=0.0072 x 10^{-3}	0.78911	1.3022	2.78	0.35
5	BJTF	Actual HDD_{18}	*TF*: diff=$(1-q^{-1})(1-q^{-52})$, b=0, nb=0, nf=2 *N*(k): diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3,4,5,6), P=(52,104)	0.6777	1.0250	2.47	-1.15
6	BJTF	Pred HDD_{18}		0.8080	1.4151	2.90	-0.95
7	BJTF	Actual AT	*TF*: diff=$(1-q^{-1})(1-q^{-52})$, b=0, nb=0, nf=2 *NM*: diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3,4,5,6), P=(52,104)	0.8570	1.4426	3.09	-1.25
8	BJTF	Pred AT		0.9252	1.5815	3.34	-1.58
9	ARX	Actual HDD_{18}	na=52, nb=1	1.0714	2.0757	3.73	1.90
10	ARX	Pred HDD_{18}	na=52, nb=1	1.0736	2.0882	3.74	1.90
11	ARX	Actual AT	na=52, nb=3	1.1690	2.3058	4.06	2.42
12	ARX	Pred AT		1.1549	2.3416	4.01	2.42
13	ARMAX	Actual HDD_{18}	na= 52, nb=1, nc= 3	0.9966	1.9511	3.50	1.70
14	ARMAX	Pred HDD_{18}		0.9975	1.9791	3.50	1.70
15	ARMAX	Actual AT	na=52, nb=3, nc=7	1.0065	2.0708	3.53	1.93
16	ARMAX	Pred AT		1.0135	2.1234	3.55	1.93
17	BSMX	Actual HDD_{18}	σ_ε=0.1156 x 10^{-3} σ_η=0.5119 x 10^{-7} σ_ζ=0.0001 x 10^{-7} σ_ω=0.0079 x 10^{-3}	0.7111	1.0617	2.50	0.46
18	BSMX	Pred HDD_{18}		0.7667	1.2367	2.69	0.47
19	BSMX	Actual AT	σ_ε=0.1151 x 10^{-3} σ_η=0.1643 x 10^{-7} σ_ζ=0.0 σ_ω=0.0079 x 10^{-3}	0.7277	1.0958	2.55	0.84
20	BSMX	Pred AT		0.7959	1.2746	2.77	0.93

Table 3.11 AT forecasting results for System A

No	Model	Structure	MAE	MSE	MAPE	MPE	Used In
21	AR	na=48	1.52	3.46	27.39	-15.97	ARX, ARMAX and BJTF
22	BSM	σ_ε=0.3413 x 10^{-3} σ_η=0.0009 x 10^{-3} σ_ζ=0.0 σ_ω=0.0	1.47	3.09	24.08	-12.74	BSMX

Table 3.12 HDD_{18} forecasting results for System A

No.	Model	Structure	MAE	MSE	MAPE	MPE	Used in
23	AR	na=48	10.39	155.90	26.86	-16.11	ARX, ARMAX and BJTF
24	BSM	σ_ε=0.1617 x 10^{-3} σ_η=0.0003 x 10^{-3} σ_ζ=0.0 σ_ω=0.0	10.08	145.71	21.09	-11.01	BSMX

Table 3.13 Electricity demand forecasting results for System B

No	Model	Input	Structure	MAE x 10^{-5}	MSE x 10^{-11}	MAPE	MPE
25	SARI		diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3), P=(52,104)	3.2336	1.8817	2.62	-0.75
26	AR		na=52	3.3509	1.8553	2.69	1.13
27	ARMA		na=52, nc=2	3.3778	1.8872	2.72	1.27
28	BSM		σ_ε=0.5622 x 10^{-5} σ_η=0.1443 x 10^{-5} σ_ζ=0.0 σ_ω=0.0250 x 10^{-5}	2.9964	1.6162	2.43	-0.57
29	BJTF	Actual HDD_{18}	*TF*: diff=$(1-q^{-1})(1-q^{-52})$, b=0, nb=0, nf=0 *NM*: diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3,4,5,6,7), P=(52,104)	2.5543	1.1736	2.06	0.47
30	BJTF	Pred HDD_{18}		3.2288	1.7733	2.61	0.24
31	BJTF	Actual AT	*TF*: diff=$(1-q^{-1})(1-q^{-52})$, b=0, nb=0, nf=0 *N(k)*: diff=$(1-q^{-1})(1-q^{-52})$, p=(1,2,3), P=(52,104)	2.7767	1.3672	2.26	0.16
32	BJTF	Pred AT		3.2122	1.9133	2.61	-0.95
33	ARX	Actual HDD_{18}	na=52, nb=10	2.5147	1.1566	2.02	0.62

34	ARX	Pred HDD_{18}		3.0269	1.5974	2.44	0.41
35	ARX	Actual AT	*na*=52, *nb*=2	2.8764	1.4134	2.32	0.86
36	ARX	Pred AT		3.1850	1.6995	2.56	0.82
37	ARMAX	Actual HDD_{18}	*na*=52, *nb*=10, *nc*=1	2.2984	0.9035	1.85	0.05
38	ARMAX	Pred HDD_{18}		2.8791	1.3753	2.33	-0.15
39	ARMAX	Actual AT	*na*=52, *nb*=2, *nc*=3	2.5933	1.2718	2.01	0.11
40	ARMAX	Pred AT		3.3470	1.7274	2.69	0.18
41	BSMX	Actual HDD_{18}	σ_ε=0.4769 x 10^{-5} σ_η=0.0831 x 10^{-5} σ_ζ=0.0 σ_ω=0.008 x 10^{-5}	2.5606	1.1054	2.05	0.99
42	BSMX	Pred HDD_{18}		2.9324	1.4961	2.36	0.34
43	BSMX	Actual AT	σ_ε=0.5341 x 10^{-5} σ_η=0.0942 x 10^{-5} σ_ζ=0.0 σ_ω=0.0112 x 10^{-5}	2.7490	1.2916	2.22	0.81
45	BSMX	Pred AT		2.9474	1.5390	2.37	0.21

Table 3.14 AT forecasting results for System B

No	Model	Structure	MAE	MSE	MAPE	MPE	Used In
46	AR	*na*=41	1.42	2.91	9.25	2.04	ARX, ARMAX and BJTF
47	BSM	σ_ε=0.2117 x 10^{-3} σ_η=0.0008 x 10^{-3} σ_ζ=0.0 σ_ω=0.0006 x 10^{-5}	1.36	2.88	8.80	3.34	BSMX

Table 3.15 HDD_{18} forecasting results for System B

No	Model	Structure	MAE	MSE	MAPE	MPE	Used in
48	AR	*na*=49	7.25	90.95	N/A	N/A	ARX, ARMAX and BJTF
49	BSM	σ_ε=0.7856 x 10^{-2} σ_η=0.0028 x 10^{-2} σ_ζ=0.0 σ_ω=0.7240 x 10^{-5}	6.77	87.05	N/A	N/A	BSMX

3.5.1 Discussion of Classical Model Results

Comparison of the System A forecasting accuracy measures show that as expected the use of an exogenous weather variable significantly improves the one-year-ahead prediction. Examination of the forecasting accuracy of the predicted AT and HDD_{18} for this system shows that these weather variables are difficult to predict, where the MAPE is as high as

approximately 27 for both variables. However, in spite of this considerable improvement was obtained through the employment of these variables. It can also be seen that the improvement obtained is greater when HDD_{18} is employed rather than AT, where this improvement is approximately of the order of 11%. The improvement in the bivariate forecast using HDD_{18} over the univariate forecast is of the order of 16% in the MAE and 27% in the MSE, with a reduction in the MAPE from 3.41 to 2.90 and from 1.94 to -0.94 in the MPE.

In the case of System B the use of the exogenous weather variables also improves the accuracy of the prediction but it is can be seen that it is necessary to employ HDD_{18} if a significant improvement is to be obtained. This highlights the importance of the use of the optimal form of the available weather variable. The improvement obtained in the bivariate forecast over the univariate forecast for System B is of the order of 13% in the MAE, 20% in MSE and a reduction in MAPE from 2.62 to 2.33 and in MPE from -0.75 to -0.15. Relatively high forecasting errors were also experienced in the prediction of the weather variables in this case but the results appear better than those obtained in the System A case.

Figure 3.32 gives a plot of the actual versus predicted System A results, where the most accurate univariate model result (ARMA) and the most accurate bivariate model result (BJTF, with HDD_{18}) is given. Figure 3.33 gives similar results for System B, where the most accurate univariate model result is produced by the SARI model and the most accurate bivariate model result is produced by an ARMAX model with HDD_{18} as the exogenous variable.

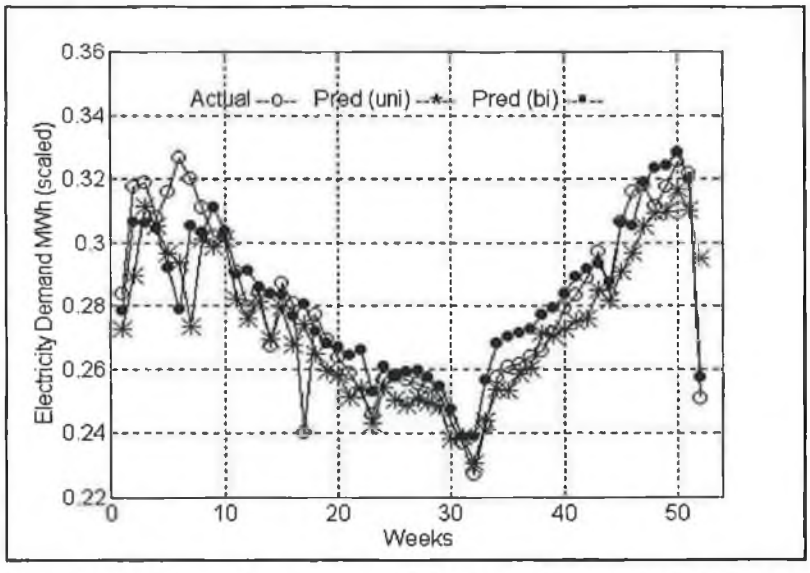

Figure 3.32 System A actual vs predicted - ARMA and BJTF(HDD_{18})

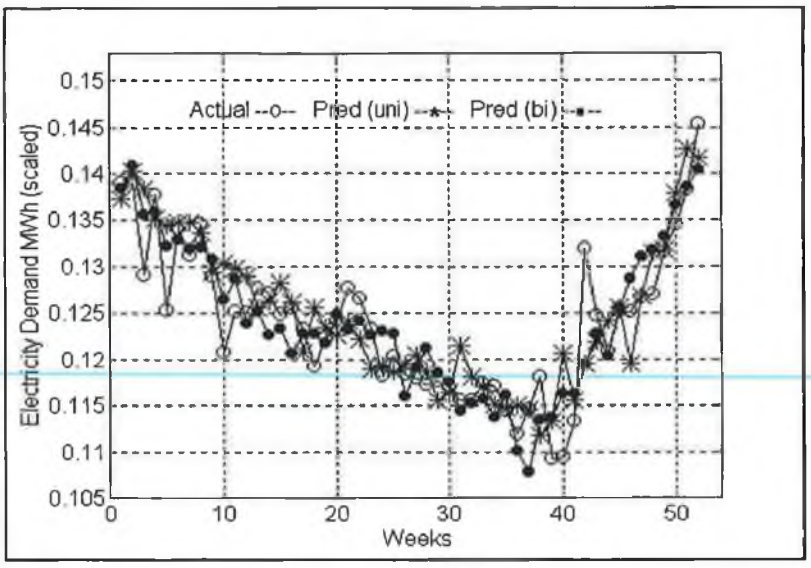

Figure 3.33 System B actual vs predicted - SARI and ARMAX(HDD_{18})

In the case of System A the most significant improvement in the bivariate forecast over the univariate forecast can be seen at Weeks 15 to 20 (April/May), Weeks 25 to 30 (July/August) and Weeks 40 to 47 (October/November). Each of these weekly periods contain a seasonal transitional month, that is at May, August and November, thus suggesting that during these months that the change in the weather pattern has a considerable effect on the demand which the univariate model is not capable of modelling. In the case of System B it is more difficult to establish the effect of the weather variable. It can be seen that at various weeks throughout the forecast horizon the presence of the weather variable in the bivariate improves the forecast, however it is difficult to establish any particular pattern. A possible reason for this is that weather is a non-dominant exogenous variable in this system.

A final note is made regarding the different transformations performed on the time series data prior to the fitting of the classical models. As expected the type of transformation appropriate is case dependent. For example, in the case of System A the univariate ARMA model which was used to represent the detrended electricity demand time series produced the most accurate result. In contrast, in the case of System B the SARI model which was used to represent the differenced the time series produced the most accurate result.

3.5.2 Discussion of Structural Model Results

The structural model results show that the bivariate structural model produces more accurate results than the univariate model. However the improvement obtained is significantly less than that attained in the equivalent results produced by the classical models. For example, in the System A case the improvement is only of the order of approximately 3% in the MAE and 5%

in the MSE, with similar results in case of System B. As in the case of the classical models it is necessary to employ HDD_{18} over AT in the bivariate model to achieve the greatest improvement over the univariate model results. The results also show that as in the case of the classical models the BSM univariate models employed to produce predictions of AT and HDD_{18} do not produce accurate results, where as before the weather data in System B appears to be relatively easier to predict. Figure 3.34 and Figure 3.35 compare the univariate and bivariate forecasts for System A and System B respectively.

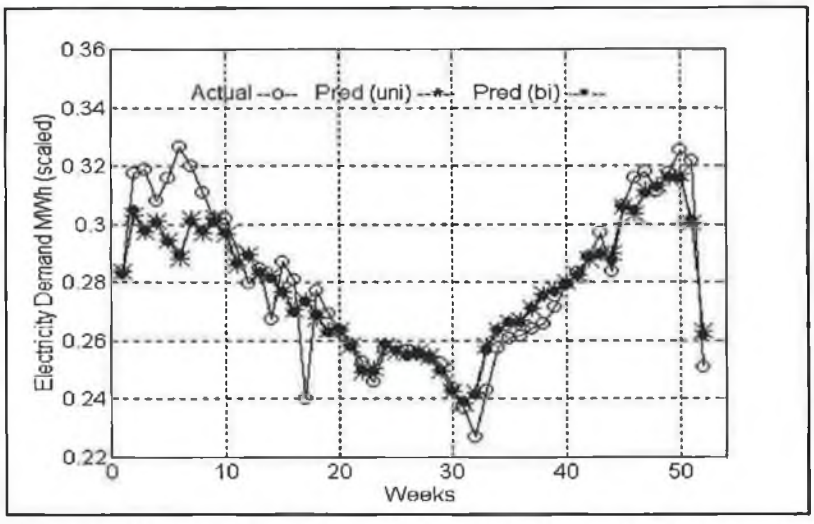

Figure 3.34 System A actual vs predicted - BSM and BSMX(HDD_{18})

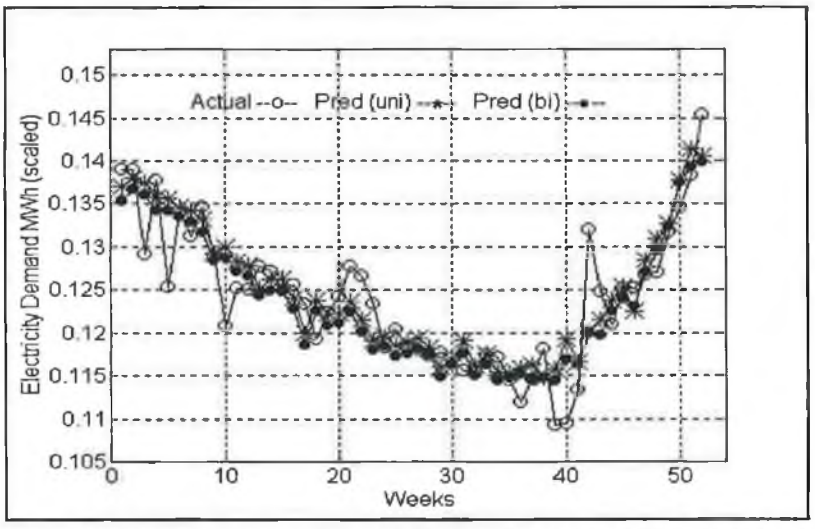

Figure 3.35 System B actual vs predicted - BSM and BSMX(HDD_{18})

In both systems it is difficult to select a week at which any significant improvement occurred, especially in the System A case, and it is thus assumed that minor improvements occurred at some weeks throughout the entire forecasting horizon.

3.5.3 Comparison of Classical and Structural Model Results

Comparing the forecasting results produced by the System A univariate and bivariate classical and structural models shows that the structural models outperform the classical models. In the univariate case the MAE and MSE of the structural model results are approximately 18% and 34% more accurate than those produced by the classical models. There is also substantial improvement in the MAPE and MPE, where the structural model MAPE is 2.78 as opposed to 3.83 for the classical model and for the MPE the difference is even more significant with the structural model MPE equal to 0.35 and the classical model MPE equal to 2.09. The latter results demonstrate that the structural model produces an overall more accurate prediction over the entire forecast horizon. The variance in the performance between the System A bivariate models is not as great, with the MAE of the structural approximately 5% more accurate and the MSE approximately 12% more accurate. The reason for this is that the inclusion of exogenous weather variables in the structural model does not have as notable an effect as in the classical model case.

Although a difference between the performance of the different classes of the univariate model can also seen in the System B results the variance is not as large, with the structural model MAE 8% more accurate than the classical model MAE and the MSE 12% more accurate. In contrast to the above in the bivariate case the classical ARMAX with HDD_{18} as an exogenous variable produces a more accurate result than the structural bivariate model. Figure 3.36 compares the most accurate System A univariate classical model result to the System A univariate structural model result, with the results for the bivariate models in Figure 3.37. Similar results are given for System B in Figure 3.38 and Figure 3.39.

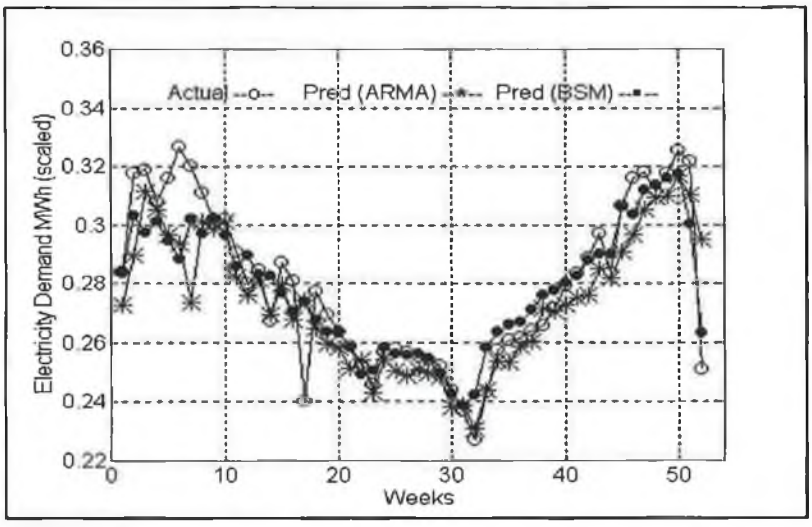

Figure 3.36 System A comparison of classical and structural univariate result

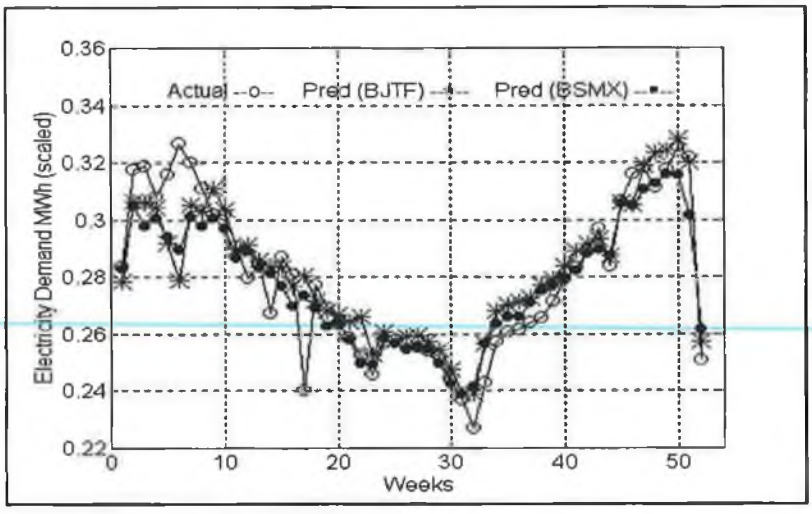

Figure 3.37 System A comparison of classical and structural bivariate result

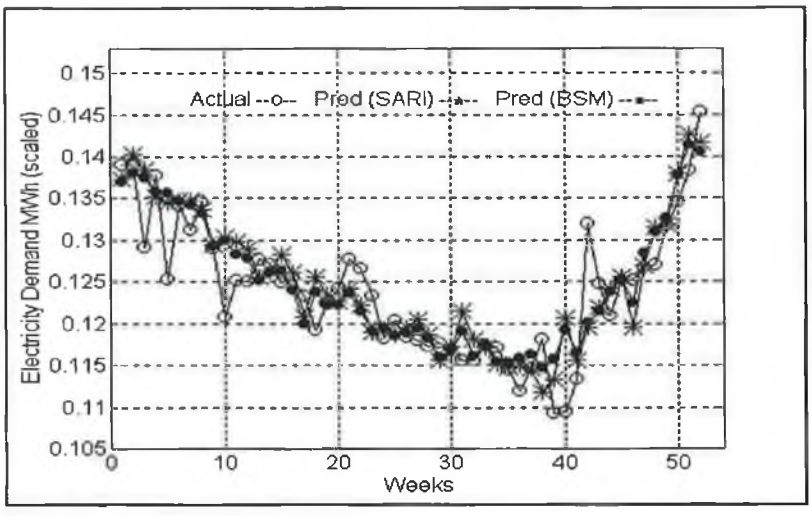

Figure 3.38 System B comparison of classical and structural univariate result

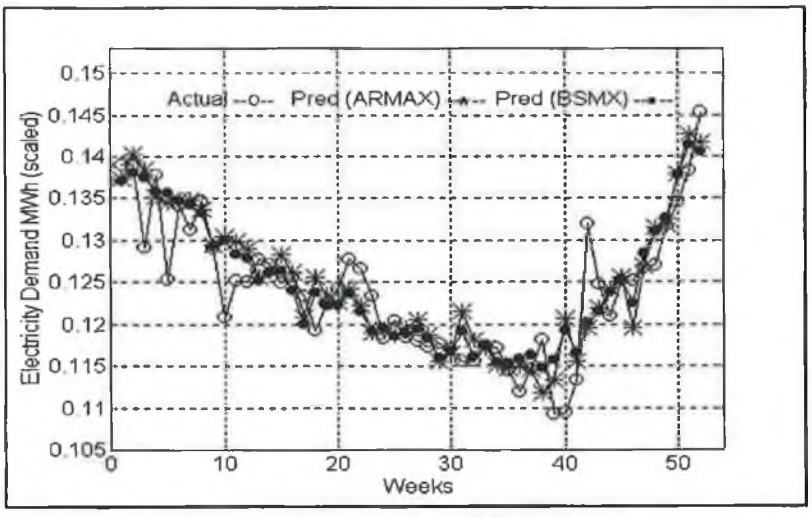

Figure 3.39 System B comparison of classical and structural bivariate result

Comparing the System A univariate results shows that from Weeks 19 to 30 that the ARMA model performance is inferior in comparison with the BSM model. Examination of the bivariate profiles show that the BJTF model performs better over this period but produces particularly poor predictions from Weeks 47 to 50. The reason for this is that the prediction of HDD_{18} was relatively inaccurate over this period. Therefore, since the inclusion of this exogenous weather variable has a greater influence on the prediction in the case of the classical model in comparison with the structural model the variance between the results of two models is notable during this period. In the case of System B the results show that the variance between the results produced by the two models is more subtle, with relatively minor differences between the performances of the models over the entire forecast horizon.

3.6 Conclusions

This chapter shows that an improvement in the accuracy of weekly electricity demand forecasting results may be achieved through the use of an appropriately selected exogenous weather variable; this was shown to be true for both dominant and non-dominant exogenous variables. The degree of improvement obtained is greater for the dominant case. This increased accuracy is attained for both cases even when future values of the exogenous variable are unknown and it is necessary to predict it through the use of a univariate forecasting model. However, it was found that the exogenous variables were difficult to predict using the linear forecasting models and thus it may be more appropriate to use an alternative modelling tool, where the use of neural networks is a possibility.

Methods of determining the optimal configuration of the exogenous temperature variable for the electricity demand systems is considered. For both systems, it was found that HDD was more effective than AT, where a base temperature of $18^{o}C$ was selected in each case. The evaluation of HDD was also studied with respect to the availability of temperature data on various time scales. It was not possible to use correlation analysis to ascertain the best time scale to use when calculating HDD, however, using the MAE as a criterion it was found for each system that the use of the finest AT time scale available yielded the HDD values which had the highest correlation with electricity demand.

The results demonstrated that the structural time series models produced more accurate forecasting results than the classical models. The variance between the results produced by the different class of models was considerably greater in the univariate case than in the bivariate case, thus it appears that the BSMX model is not a particularly effective model for representing the bivariate system.

CHAPTER 4

Application of Neural Networks to Weekly Electricity Demand and Yearly Electricity Sales

4.1 Introduction

This chapter examines the application of neural networks to modelling electricity demand time series data. A neural network may be considered as a data processing technique that formulates a relationship, or mapping, between some type of input stream of information and an output stream of information. The process of constructing this relationship is achieved through the use of *learning* algorithms. For the purposes of time series analysis, a neural network can be considered as a general nonlinear mapping between some subset of past time series values and a future time series value. It is this nonlinear mapping which makes neural networks attractive as a time series modelling tool. The traditional linear time series approaches (Box-Jenkins, 1976; Harvey, 1989) assume a linear relationship among the variables, which may be an inadequate representation of the system. Although nonlinear forecasting techniques have been developed (Priestly, 1988; Tong, 1990) they can often involve complex methods which result in unwieldy models (Gooijer and Kumar, 1992). Another advantages of using neural networks for time series forecasting is their excellent learning and generalisation capabilities, where the neural network has the capability to infer and deduce the system behaviour from the time series data.

Two applications are considered where the main objective is to address some of the issues involved in the design of neural networks for time series forecasting. The first application examines the weekly electricity demand data sets for System A and System B, which was modelled using linear forecasting techniques in Chapter 3, and the second application examines yearly electricity demand data.

4.2 Neural Networks for Time Series Forecasting

There are a number of works devoted to the subject of time series forecasting using neural networks (Wong, 1991; Weigend and Gershenfeld, 1992; Lu *et* al, 1993; Peng *et al*, 1993; Lachtermacher and Fuller, 1995). A widely applied network structure is the MLP (multi-layer perceptron) (Tang *et al*, 1991; Park *et al*, 1991; Chen *et al*, 1992; Chakraborty *et al*, 1992;

Papalexopoulos *et al*, 1994; Azoff, 1994). The MLP is an example of a *feedforward* network, where information is fed forward through the network during learning. Typically, temporal information is presented to the network by a time-lagged vector of time series data at the input, with the current value of the time series at the output. Following learning the MLP may be configured to operate in *recurrent* mode, where the production of a multi-step-ahead forecast involves the use of predicted values of the time series which are fed back to the input. In contrast, examples of works which use neural networks which operate recurrently during the learning stage, referred to as recurrent networks, are Blake *et al* (1995) and Connor *et al* (1994). Another network structure which may be configured to operate recurrently during the forecasting stage is the radial basis function network (Hudson *et al*, 1994; Zaknich and Attikiouzel, 1994), where the main difference between this network structure and the MLP is the training algorithm employed.

In the case of electricity demand forecasting there is a substantial amount of literature which deals with the application of short term and peak load forecasting (Park *et al*, 1991; Chen *et al*, 1992; Kermanshahi *et al*, 1993; Lu *et al*, 1993; Morioka *et al*, 1993; Onoda, 1993; Azzam-ul-Asar *et al*, 1994; Papalexopoulos *et al*, 1994; Bakirtzis *et al*, 1996 among others). However, the application of weekly and yearly forecasting are not dealt with in nearly the same depth, one example in the weekly case is Smith (1994) and Lachtermacher and Fuller (1995) in the yearly case.

4.2.1 Background

The study of neural networks can be considered as an attempt to understand a replicate human biological nervous system. Of particular interest is the formulation of an 'artificial' computational form which mimic the brain's thinking process. The basic structure of a neural network consists of processing elements, called nodes or neurons, which are based on the physiology and individual neurons in the human brain. These neurons are interconnected with one-way signal channels, called connections. The relative strengths and nature of the connecting links determine the overall operational characteristics of the network. Network *learning* (training) involves selecting and modifying the interconnecting links in an adaptive manner so as to execute the specific task(s) required of the network. An important property of a neural network is that it can learn from particular patterns of data presented to it, where it can recognise such patterns when they occur again, but can also recognise similar patterns through generalisation. It is this capability of learning and generalisation coupled with the

property of potentially faster operational speeds, realised through inherent parallel operation, which make neural network a very powerful computational tool.

There are many different types of neural networks, with variations in network topology, neuron characteristics and learning rules. One method of broadly classifying them is according to their learning techniques. They may be broadly classified into two learning classes: supervised learning and unsupervised learning. Supervised learning involves presenting the network with a set of example input and output data pairs and the network learns how to match the examples as closely as possible. In contrast, for unsupervised learning the networks are only presented with input samples and samples are grouped into classes which are self-similar; networks of this type are called self-organising networks. Examples of self-organising networks are the ART (adaptive resonance theory) network (Hertz *et al*, 1991) and Kohonen's Self Organising feature maps (Lipmann, 1987). Supervised learning neural networks may be further divided into static and dynamic networks (Hush and Horne, 1993). Static networks are memoryless, that is, the output of the network during learning is only a function of the current input and not on past or future inputs or outputs. Examples of static models are the MLP (multi layer perceptron) network (Rummelhart *et al*, 1986) and the RBF (radial basis function) network (Hush and Horne, 1993). Dynamic networks are systems with memory which may have feedforward dynamics, output feedback or state feedback. Examples of dynamic networks are Hopfield networks (Hertz *et al*, 1991) and recurrent networks (Giles *et al*, 1994).

4.2.2 Choice of Network Type and Architecture

The type of neural network used depends on the task for which it is being employed. In time series analysis applications the autoregressive nature of time series require a network structure that can operate recurrently and produce a continuous output. In addition, a network which adopts a supervised learning methodology is appropriate due to presence of recurring patterns within the time series data. An MLP network structure fulfils these criteria and is adopted for the time series applications dealt with in this neural network analysis.

An MLP is a feedforward network which consists of a layer of input nodes; a layer of output neurons and one or more additional layers of neurons sandwiched in between. The additional layers are referred to as the hidden layers. Each neuron computes the weighted sum of its inputs, adds a bias value and then passes this sum through a nonlinear function called an activation function, as in Figure 4.1.

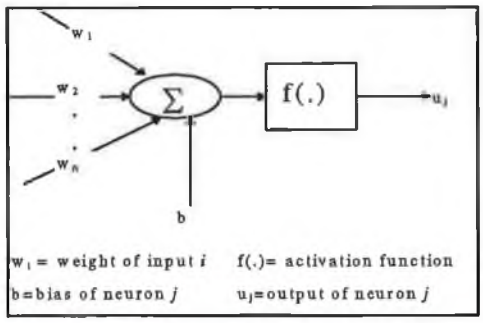

Figure 4.1 Outline of a basic neuron

An example of an MLP network with two input nodes, a single hidden layer with three hidden neurons and two output neurons is shown in Figure 4.2.

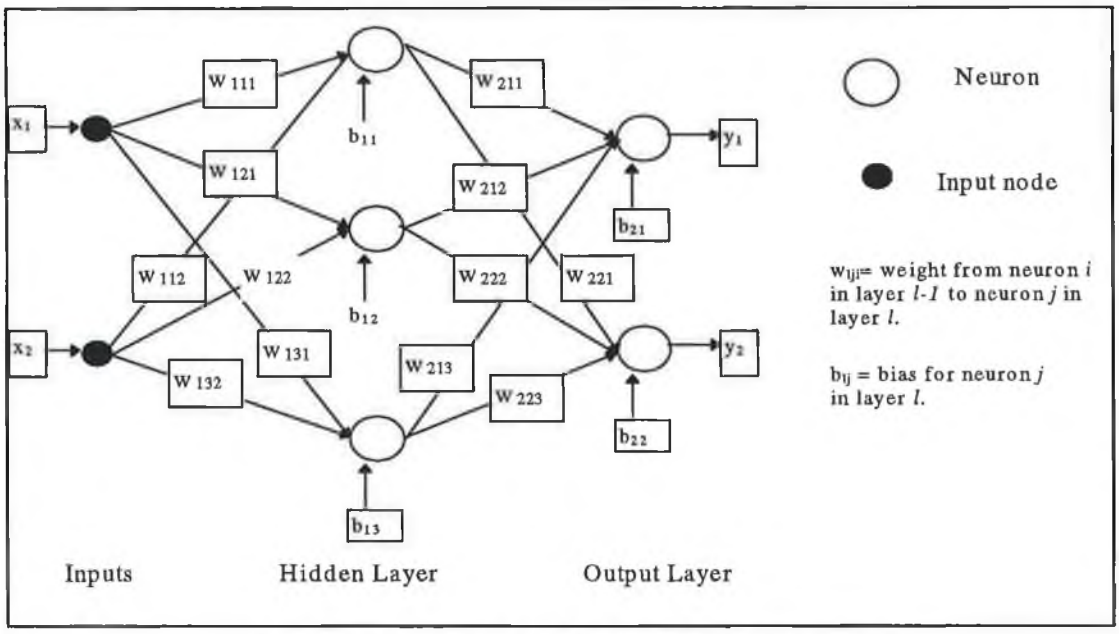

Figure 4.2 Example of an MLP with one hidden layer

The values of the connection weights w_{lji} and bias values b_{lj} of the network are determined through network training. A training algorithm developed by Rummelhart *et al* (1986) for MLP networks, referred to as the 'backpropagation' training algorithm, is described in Section 4.2.3. However, the basic framework of network training may be described as follows: the network is presented with a set of input/output data vectors which is representative of the process which the neural network is attempting to model. Upon each presentation, the weights are adjusted to decrease the difference between the networks's output and the target output. The error is minimised using the technique of gradient descent. The bias values b_{lj} serve as a threshold for the neurons to which they are connected, where these values are adjusted in a similar manner to the weights w_{lji}. . Network training is an iterative process which involves the following basic steps:

- the input and output data is presented to the network;
- the network output is calculated for a particular set of weight and bias values;
- the weights and bias values are adapted accordingly;
- repeat this procedure for a specified number of iterations or until a specified error criteria has been achieved.

Nonlinearity is introduced into the network through the activation functions. The type of activation function used in the neural network depends on the intended application. The simplest activation function is the linear function, whose output equals the input. Smooth nonlinearities, such as, tan-sigmoid or log-sigmoid are suitable for networks which are to model continuous variables, whereas hard-limiting nonlinearities are better for classifiers. Figure 4.3 and Figure 4.4 describe log-sigmoid and tan-sigmoid activation function respectively. These activation functions have a squashing role in restricting the possible neuron output, where a value lying in the range $(-\infty,\infty)$ is constrained to [0,1] in the log-sigmoid case and [-1,1] in the tan-sigmoid case.

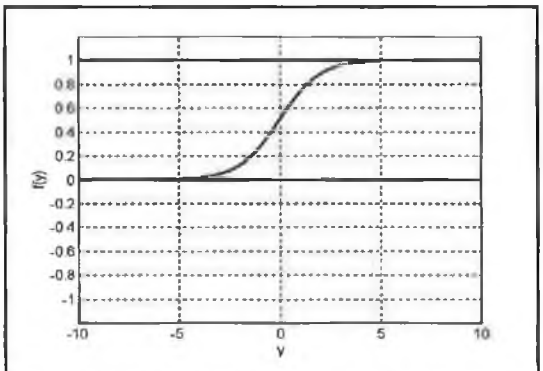

Figure 4.3 Log-sigmoid activation function

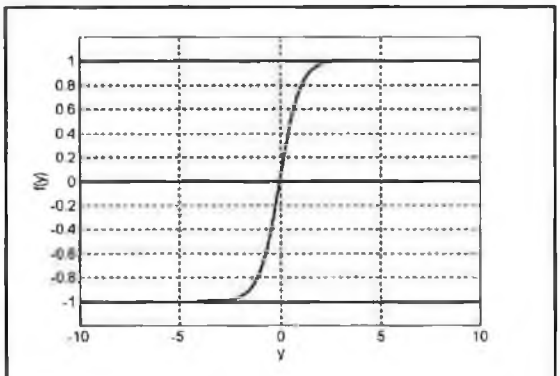

Figure 4.4 Tan-sigmoid activation function

The number of nodes in the input and output layer are determined by the nature of the task which the neural network is to perform. The input layer to the network does not perform processing but merely acts as a means by which data is introduced to the network. It is made up of the variables which are necessary for modelling the target output, where from a complexity and parsimony point of view it is desirable to select the absolute minimum number of essential input variables. The number of neurons in the output layer should also be the minimum that satisfy the problem requirements. A single output neuron has the advantage that the network is focused on one task and there are no conflicting outputs pulling the weights in opposing directions during training. For time series forecasting applications the output layer

usually consists of a single neuron, where this neuron represents the current value of the time series. The input variables often consists of past values of the time series and current and past values of exogenous variables which have an external influence on the time series.

In the majority of cases it will be necessary to normalise the input data to the network, where normalisation standardises the possible numerical range that the input data to the network can take. It also ensures that the weights of the network are of the same order of magnitude which allows the network to train more easily. An exception to the requirement for normalisation may be when the data concerned is in binary form, when it is, in a sense, already normalised. There are a number of methods available for normalisation (Azoff, 1994). The choice of method depends on the composition of the input vector but ultimately it is a matter of how the network performs, as different normalisation approaches may affect the convergence of the training process (Azzam-ul-Asar, 1994). A common approach in time series analysis is to divide all the values of the data by one number larger than the magnitude of the greatest value of the time series (Azzam-ul-Asar, 1994; Lachtermacher and Fuller, 1994; Smith, 1994), for the purposes of this analysis this number shall be referred to as the normalisation factor (NF). This approach is an external normalisation method, that is, the data is scaled prior to the construction of the input vectors. It may also be necessary to normalise the target output data. The main consideration is the output range of the output neuron's activation function. In time series forecasting applications the output neurons often have linear activation functions at the output layer, thus allowing an unrestricted output range. To aid the learning process it is useful to ensure that the weights of the output neurons are of the same order of magnitude as those in the input and hidden regions and therefore the target output data is scaled to the same order of magnitude as the input data. As a result of the target data normalisation the actual network output will correspond to the normalised scaling. Therefore, in order to interpret the network output on the original scale it will be necessary to re-scale it with the normalising parameters.

The characteristics of the MLP depend on the input and output structure but also depend on the hidden layer region. An MLP may contain one or more hidden layers. For classification problems Lippmann (1986) demonstrated that an MLP network with one hidden layer can implement arbitrary convex decision boundaries. However, for some problems, a network with two hidden layers with a relatively small number of neurons in each layer can be used where a network with one hidden layer would require an infinite number of neurons (Hush and Horne, 1993). Chester's (1990) work shows that networks with two hidden layers appear to provide

higher accuracy and better generalisation than a single hidden layer network with the added advantage that fewer neurons are required. The decision of the number of neurons in the hidden layers of the network is highly problem dependent and is best determined through experimentation (Azoff, 1994). Azoff also notes that some problems work well with a bottleneck structure, where there are fewer neurons in the hidden layer than nodes in the input layer. The experimentation exercise involves training a number of neural networks with different hidden layer architectures whose performance is then compared according to some chosen criterion. This may be computationally demanding due to the large number of possible architectures. It would be possible to search the set of all possible architectures using a genetic algorithm (Hertz *et al*, 1991; Dodd, 1992; Jones, 1993; Beltratti *et al* 1996) but this kind of search would frequently require an unreasonable amount of CPU time.

4.2.3 Training Algorithm

The aim of the training process is to choose values of the weights and biases of the network so as to realise the desired mapping from inputs to outputs. Each weight in a network represents a unique dimension in a multi-dimensional weight space. Rummelhart (1986) developed an algorithm, named the 'backpropagation algorithm' to train MLP networks. The algorithm is an iterative gradient algorithm which minimises a cost function which is a function of the weights of the network. The traditional MLP training algorithm cost function is the Sum Squared Error (SSE) cost function. The algorithm also requires an activation function which is differentiable, where the most common nonlinearity used is a sigmoid function. The algorithm is described for a network of L layers, with a sigmoid nonlinearity and a SSE cost function. The following notation is used:

- $u_{l,j}$ represents the output of the jth node on layer l.
- $w_{l,j,i}$ represents the weight which connects the ith node in layer l-1 to the jth node in layer l.

To account for the bias weights define the 0th component of the input vector to each layer l to be equal to 1, therefore, $u_{l,0} = 1$; which represents the bias weight $w_{l,j,0}$. Therefore, using this notation the output of a node j in layer l is given by:

$$u_{l,j} = f\left(\sum_{i=}^{N_{l-1}} w_{l,j,i} u_{l-1,i}\right) \tag{4.1}$$

where $f(.)$ represents a sigmoidal function and N_l is the number of nodes in layer l. The algorithm is now described in Steps 1 to 6.

Step 1. Define cost function - the SSE criterion function may be described as:

$$J(w) = \sum_{p=1}^{P} J_p(w) \tag{4.2}$$

where P = number of training patterns

$$J_p(w) = \sum_{q=1}^{N_L} \left(u_{L,q}(x_p) - d_q(x_p) \right)^2$$

$u_{L,q}$ = output of qth neuron in output layer.

N_L = number of neurons in output layer.

d_q = desired response for qth output neuron.

x_p = pth training sample.

Step 2. Initialise weights and biases to random values.

Step 3. Present input training vectors and desired output vector.

Step 4. Calculate actual output at each layer:

$$u_{l,j} = f\left(\sum_{i=0}^{N_{l-1}} w_{l,j,i} u_{l-1,i} \right) \tag{4.3}$$

where $u_{l,j}$ = output of jth neuron in layer l.

$w_{l,j,i}$ = weight which connects the ith neuron in layer $l-1$ to jth neuron in layer l.

N_l = number of neurons in layer l.

$f(.)$ = sigmoid nonlinearity.

Step 5. Adapt each weight iteratively using Equation (4.4) by starting at the output layer neurons and working back to the preceding layer of neurons until the first hidden layer is reached:

$$w_{l,j,i}(k+1) = w_{l,j,i}(k) - lr \sum_{p=1}^{P} \frac{\partial J_p(w)}{\partial w_{l,j,i}}\bigg|_{w(k)} \tag{4.4}$$

where lr = the learning rate.

$$\frac{\partial J_p(\mathbf{w})}{\partial w_{l,j,i}} = \frac{\partial J_p(\mathbf{w})}{\partial u_{l,j}} \frac{\partial u_{l,j}}{\partial w_{l,j,i}}$$

$$\frac{\partial u_{l,j}}{\partial w_{l,j,i}} = u_{l,j}\left(1 - u_{l,j}\right) u_{l-1,i}$$

From one hidden layer to the next hidden layer:

$$\frac{\partial J_p(\mathbf{w})}{\partial u_{l,j}} = \sum_{m=1}^{N_{l+1}} \frac{\partial J_p(\mathbf{w})}{\partial u_{l+1,m}} u_{l+1,m}\left(1 - u_{l+1,m}\right) w_{l+1,m,j}$$

At the output layer:

$$\frac{\partial J_p(\mathbf{w})}{\partial u_{L,j}} = u_{L,j}(x_p) - d_j(x_p)$$

Step 6. the process of computing the gradient and adjusting the weights is repeated by going to back to Step 3.

Note that Step 3 to Step 6 is referred to as a training epoch.

The learning rate *lr* defines the amount by which the weights and biases of a network are adjusted. If the learning rate is set too small the network may take a long time to train, while setting the learning rate too large may result in the error minimum being 'jumped over'. Since there is no way of calculating the optimum learning rate a solution to the problem is to adapt the learning rate during training, where the objective is to keep the learning rate as large as possible while ensuring stable learning.

A problem with the gradient descent algorithm is that learning may finish in a local, instead of a global minimum, resulting in a sub-optimally trained network. This problem may be alleviated by the inclusion of a momentum term, α. This involves altering the updating of weights in Step 5, where the following equation is used:

$$\boxed{w_{l,j,i}(k+1) = w_{l,j,i}(k) - \mu \sum_{p=1}^{P} \left.\frac{\partial J_p(\mathbf{w})}{\partial w_{l,j,i}}\right|_{\mathbf{w}(k)} + \alpha\left(w_{l,j,i}(k) - w_{l,j,i}(k-1)\right)} \qquad (4.5)$$

where, $0<\alpha<1$. The momentum term helps the weights move across flat portions and overcome small peaks of the error surface after having descended from the steep portions of the surface.

A further problem is that the performance of the network can be sensitive to the initial condition of the network weights and bias values. A number of networks with identical structure, but initialised with different sets of weights and bias values, can yield different results due to the existence of local minima in the error surface. Therefore, to provide some insensitivity to initial conditions which result in local solutions it is advisable to re-train the same network structure a number of times.

Another important aspect of training is the criterion used to assess network performance. For time series forecasting purposes the aim is to perform a multi-step-ahead forecast, however, the neural network is typically trained using a single-step-ahead prediction criterion. Consequently, this does not always determine the weight set which optimises the multi-step-ahead prediction performance of the network. It is difficult to design backpropagation training with a multi-step-ahead criterion, with the added difficulty that it would be computationally intensive, particularly when the prediction horizon is long. A compromise is to train the network using a single-step-ahead criterion, but also examine the multi-step-ahead performance during training. To do this the data is divided into the following data sets:

- training data set - adapt weights and biases using a single-step-ahead criterion .
- validation data set - examine multi-step-ahead performance for the weights obtained in training.
- prediction data set - perform multi-step-ahead forecast.

Figure 4.5 show an example of data segmentation for a weekly time series application (System A), where a total of ten years of weekly data is available.

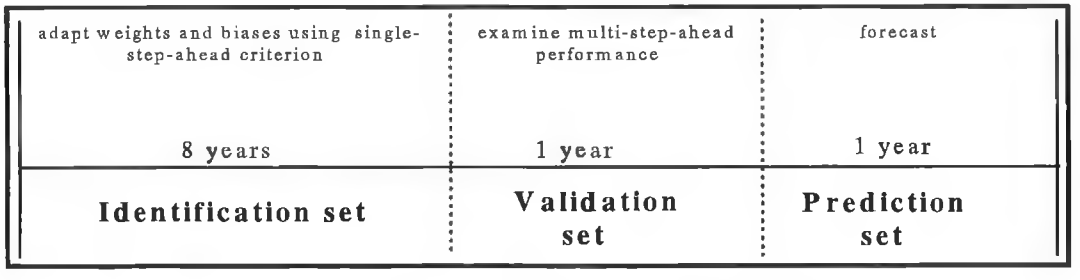

Figure 4.5 Weekly time series data segmentation

The performance criterion used to adapt the weights and biases of the network during training is the traditional single-step-ahead SSE and the performance criterion used to examine the multi-step-ahead performance during validation is the multi-step-ahead MAE (see Section 2.8).

4.2.4 Training Cessation Point

An important decision in neural network training is the selection of the network training cessation point. Consider Figure 4.6 which shows an example of the different network performance criterion at each training epoch for a weekly electricity demand application. The network used in the example has two hidden layers and was trained using the LMS gradient technique with backpropagation, an adaptive learning rate and a momentum term (Section 4.2.3). The total number of training epochs was 5000, however the results are given for the epoch range 1500 epochs - 2000 epochs for visual clarity. The following performance criterion are shown on the graph:

- single-step-ahead SSE (scaled) over the training set.
- multi-step-ahead MAE over the validation set.
- multi-step-ahead MAE over the prediction set.

The graph also shows the adaptive learning rate (scaled).

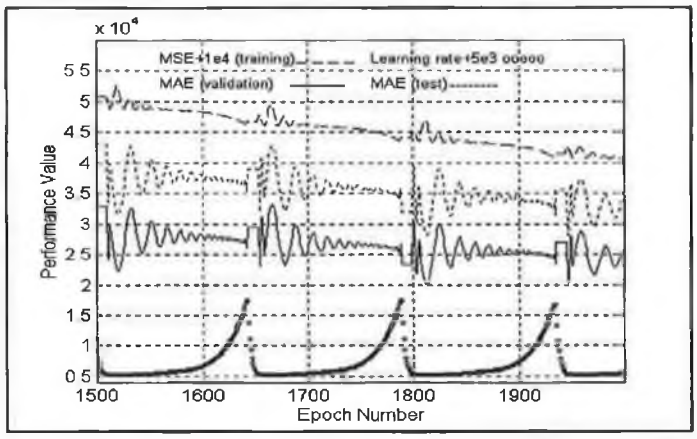

Figure 4.6 Different performance criterion

The profile of the performance criteria vary across the training epochs, where this is due to the effect of the adaptive learning rate during training. As a result of the variation in performance criterion different network training cessation points and thus choice of weight vector result in the achievement of different performance criterion values. It can be seen in Figure 4.6 that the variation in the performance criterion is consistent over the validation and prediction set and it is therefore possible to choose the training cessation point based on the validation set which will give good multi-step-ahead performance on the prediction set. An example of this may be seen by examining Figure 4.6 more closely; that is, in the range 1650 epochs to 1690 epochs, given in Figure 4.7. A choice of weight at epoch 1654 (shown by ● on the graph) yields a

multi-step-ahead performance value of 3.19 x 10^4 on the prediction set while the corresponding value at epoch 1666 (shown by $*$ on the graph) is 4.29 x 10^4 which is approximately 35% worse. Note that the single-step-ahead performance criterion over the training set suggests a cessation point at 1683 (shown by o on the graph) which yields a value of 3.53 x 10^4 on the prediction set, which is approximately 10% worse.

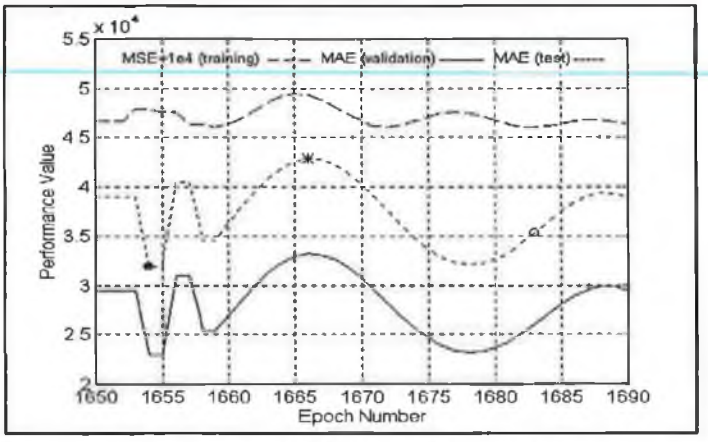

Figure 4.7 Close-up of Figure 4.6

Consequently, the training cessation point is chosen by the examination of the multi-step-ahead performance over the validation set which should ensure good performance over the prediction set. In the thesis the *optimal* multi-step-ahead MAE over the validation set shall be referred to as MMAE_val. In practice the procedure involves training the network for a fixed number of epochs and the calculation of the multi-step-ahead MAE at every epoch.

It is necessary to determine a suitable number of fixed training epochs over which to train the network. Given that the network is being trained to perform a multi-step-ahead prediction and since there is a correlation between the multi-step-ahead performance over the validation and prediction sets, it is acceptable to continue training the network until the multi-step-head MAE over the validation set has stopped decreasing beyond a particular chosen tolerance level. The number of epochs required to reach this point will depend on the size of the network and on the training data set. Consequently, different networks may require to be trained for a different numbers of fixed training epochs. It is desirable to choose as a low a tolerance level as possible, however this may not always be practical since it is necessary to re-train a network a number of times and thus the overall time taken to complete the training of a network can be lengthy. For example, in the System A weekly application typical training times for a 3-5-1 network with 11 inputs is approximately 50 minutes per 10000 epochs trained on a 160 MHz Pentium PC with 16 MB of RAM. It was required to train this particular network for 40000

epochs to reach a tolerance of approximately 0.4% over 10000 epochs. Therefore, training this network for a further 10000 epochs would reduce the MMAE_val by an estimated 0.4%. The overall computation time for training this network over 40000 epochs for 10 training runs takes approximately 33 hours. For the applications dealt with in this chapter a tolerance level of approximately 0.5% over 10000 epochs was chosen for each network. In some cases however it was not always possible to achieve this tolerance level due to unacceptable training times but in this instance a measure of the tolerance level attained is given.

4.2.5 Application Procedure

The objective of this chapter is to address some of the issues involved in the specification of MLP neural networks for time series forecasting purposes, where weekly and yearly electricity demand applications are considered. The basic steps involved in the specification of the MLP network for each application are as follows:

NN-Step 1. Select input and output structure.

NN-Step 2. Perform input and output data normalisation.

NN-Step 3. Construct input/output training, validation and prediction data sets.

NN-Step 4. Select hidden layer architecture: number of hidden layers; number of neurons in hidden layers and neuron activation functions.

There are numerous factors involved in the specification of the MLP network, where at present there is no complete theoretical basis relating these factors to known characteristics of the system being modelled. Consequently, some of the network specifications, such as input/output data normalisation and hidden layer architecture are best determined through experimentation. However, such an experimentation analysis can involve long training times and therefore due to computational time constraints the work dealt with here concentrates on two aspects of the MLP network specification:

- the specification of the input structure.
- the optimal number of neurons in the hidden layers.

where the latter case is only dealt with in the yearly application. The other MLP specifications, such as the input/output data normalisation and the selection of the activation functions are specified based on previous work carried out in this application area (Smith, 1994).

The specification of the MLP networks used in each application was carried out using the procedure, NN-Step 1 to NN-Step 4, outlined above. Each of the MLP networks were trained on a 160 MHz Pentium (586) PC with 16 MB of RAM using the backpropagation LMS training algorithm (with adaptive learning rate and momentum) given in the Neural Network Toolbox For Use with MATALB (1994)

4.3 Application to Weekly Electricity Demand

A neural network analysis was performed on the weekly electricity demand systems, System A and System B, which were previously dealt with using linear analysis in Chapter 3.

4.3.1 System A

Weekly electricity demand (MWh) and weather (AT and HDD_{18}) data is available for System A from 4th April 1982 to 28th December 1991. Figure 3.1 shows a plot of the time series and the characteristics of the system are discussed in Section 3.2. Information already established in the linear analysis developed in Chapter 3 is used in the neural network analysis. The linear models of particular interest are the Box-Jenkins models (2.19) and the structural models (2.41). These models identify specific characteristics within the time series, such as trend and seasonality, and model these elements. The characteristics are handled differently by each model and Table 4.1 compares and contrasts the models and the modelling techniques.

Table 4.1 Comparison of linear models

	Box-Jenkins	**Structural**
Trend component	$(1-a_1q^{-1}-.....-a_pq^{-p})(1-q^{-1})^1 y(k) = \varepsilon(k)$	$\mu(k) = \mu(k-1) + \beta(k-1) + \eta(k)$ $\beta(k) = \beta(k-1) + \zeta(k)$
Seasonal component	$(1-a_{1,s}q^{-s}-.....-a_{P,s}q^{-Ps})(1-q^{-s})^1 y(k) = \varepsilon(k)$	$\gamma(k) = -\sum_{j=1}^{s-1}\gamma(k-j) + \omega(k)$
Identify	*s, d, D, p* and *P*	*s*
Estimate	$a_1,....,a_p$ and $a_{1,s},....,a_{P,s}$	σ_ε^2, σ_η^2, σ_ζ^2, σ_ω^2
Estimation technique	Least Squares /Maximum Likelihood	Maximum Likelihood via PED
Problem type	System parameter identification problem	State estimation problem
Stationarity issues	Differencing applied to induce stationarity in time series	Models non-stationary time series

Two different modelling approaches were adopted. The first approach involves the use of Box-Jenkins univariate (2.19) and bivariate (2.23) linear models to determine the input structure to the MLP used to forecast weekly electricity demand. The second is a hybrid approach which involves using the linear BSM with dummy seasonal component (2.44) to forecast the trend and seasonal components of the weekly electricity demand time series and using an MLP to forecast the irregular component of the series. A single example of a simple black-box neural network modelling approach is also presented to demonstrate such a case, where this involves the use of a purely autoregressive weekly electricity demand input structure to the MLP. Due to computational time constraints only the univariate case in this instance was examined.

4.3.1.1 Box-Jenkins Model Approach

NN-Step 1: Input and output structure

The construction of the MLP input structure based on the Box-Jenkins linear model involves the adaptation of equation (2.19). The specific form of the model identified for this data set is given in Table 3.10, Model No. 1, with the seasonal length $s=52$; the order of nonseasonal differencing $d=1$; the order of seasonal differencing $D=1$. The model is given by:

$$(1-a_1q^{-1}-.....-a_pq^{-p})(1-a_{1,52}q^{-52}-a_{2,52}q^{-104}.....+a_{P,52}q^{-P52})(1-q^{-1})^1(1-q^{-52})^1y(k)=\varepsilon(k) \qquad (4.6)$$

where, p the order of the nonseasonal autoregressive part; P the order of the seasonal autoregressive part. Expansion of the nonseasonal and seasonal regressors in (4.6) yields:

$$\begin{aligned}&(1-a_1q^{-1}-.....-a_pq^{-p}-a_{1,52}q^{-52}+a_1a_{1,52}q^{-53}+.....+a_pa_{1,52}q^{-(52+p)}-.....-a_{P,52}q^{-52P}\\&+a_1a_{P,52}q^{-(52P+1)}+.....+a_pa_{P,52}q^{-(52P+p)})(1-q^{-1})(1-q^{-52})y(k)=\varepsilon(k)\end{aligned} \qquad (4.7)$$

By defining $z(k)$ as:

$$z(k)=(1-q^{-1})(1-q^{-52})y(k) \qquad (4.8)$$

a generalisation of the linear model given by equation (4.7) to the nonlinear case is given by the following:

$$\boxed{\begin{aligned}z(k)=g\big(&z(k-1),.....,z(k-p),z(k-52),.....,z(k-52-p),z(k-52P),\\&z(k-52P-1),.....,z(k-52P-p)\big)+\varepsilon(k)\end{aligned}} \qquad (4.9)$$

where $g(.)$ represents a nonlinear function. An MLP with an input structure described by the nonlinear model given by equation (4.9) is used to forecast $z(k)$. A single neuron at the output

represents the current value of the $z(k)$. The forecast of the original weekly electricity demand time series $y(k)$ is obtained by applying the inverse of the differencing transformation given by equation (4.8). Figure 4.9 describes the total model which is referred to as a Univariate Differenced Box-Jenkins model, denoted as UDBJ_A, where the A denotes the System A data set.

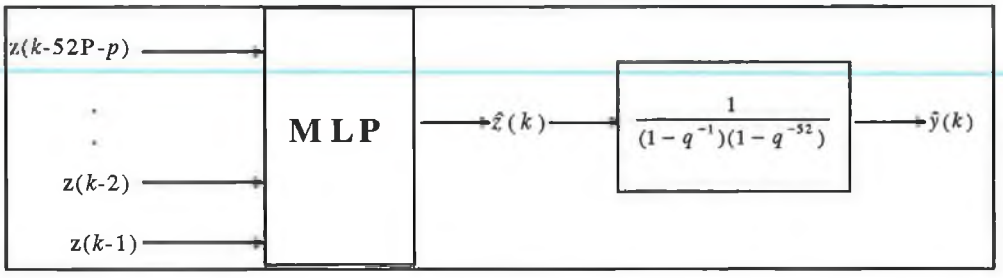

Figure 4.8 UDBJ_A model

A plot of the $z(k)$ time series was shown in Figure 2.10 of Chapter 2.

Further expansion of $z(k)$ in (4.8) yields the following:

$$\begin{aligned}
&\Big(1-a_1q^{-1}-.....-a_pq^{-p}-a_{1,52}q^{-52}+a_1a_{1,52}q^{-53}+.....+a_pa_{1,52}q^{-(52+p)}-.....\\
&-a_{P,52}q^{-52P}+a_1a_{P,52}q^{-(52P+1)}+.....+a_pa_{P,52}q^{-(52P+p)}\\
&-q^{-1}+a_1q^{-2}+....+a_pq^{-(p+1)}+a_{1,52}q^{-53}-a_1a_{1,52}q^{-54}-.....-a_pa_{1,52}q^{-(52+p+1)}+.....\\
&+a_{P,52}q^{-(52P+1)}-a_1a_{P,52}q^{-(52P+2)}-.....-a_pa_{P,52}q^{-(52P+p+1)}\\
&-q^{-52}+a_1q^{-53}+.....+a_pq^{-(52+p)}+a_{1,52}q^{-104}-a_1a_{1,52}q^{-105}-.....-a_pa_{1,52}q^{-(104+p)}+.....\\
&+a_{P,52}q^{-(52P+52)}-a_1a_{P,52}q^{-(52P+53)}-.....-a_pa_{P,52}q^{-(52P+p+52)}\\
&-q^{-53}+a_1q^{-54}+.....+a_pq^{-(p+53)}+a_{1,52}q^{-105}-a_1a_{1,52}q^{106}-.....-a_pa_{1,52}q^{-(105+p)}+.....\\
&+a_{P,52}q^{-(52P+53)}-a_1a_{P,52}q^{-(52P+54)}-.....-a_pa_{P,52}q^{-(52P+p+53)}\Big)y(k)=\varepsilon(k)
\end{aligned} \tag{4.10}$$

A nonlinear model may be adapted from (4.10) and is given by:

$$\begin{aligned}
&y(k)=h\big(y(k-1),.....,y(k-p-1),y(k-52),.....,y(k-52-p-1),y(k-52-p),.....\\
&....,y(k-105-p),.....,y(k-52P-1),y(k-52P),y(k-52P-p-1),y(k-52P-p),....\\
&.....,y(k-52P-54),y(k-52P-p-53)\big)+\varepsilon(k)
\end{aligned} \tag{4.11}$$

where $h(.)$ represents a nonlinear function. An MLP with an input structure described by this nonlinear model (4.11) is used to forecast the time series $y(k)$, where the output of the MLP

represents the current value of $y(k)$. This model is referred to a Univariate Box-Jenkins (UBJ_A) model and is described in Figure 4.9.

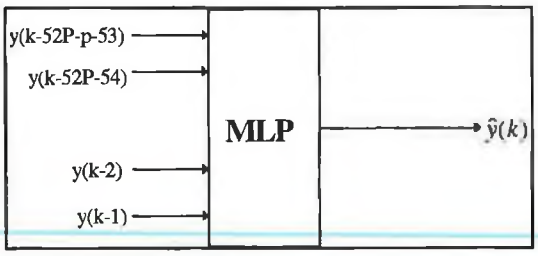

Figure 4.9 Description of UBJ_A

The input structures of the MLP's in the UDBJ_A and UBJ_A models may be extended to include AT and HDD_{18} weather variables. Analysis carried out in Chapter 3 determined the form of the BJTF model (2.22) for the weekly weather and electricity demand variables. This analysis involved the application of differencing transformations to the electricity demand and weather time series and the subsequent use of the SCCF (2.12) to determine the structure of the model. For both AT and HDD_{18} the delay parameter b was determined to be equal to zero, that is the weather in the current week affects the electricity demand in the current week. In addition, the order of the $B(q)$ polynomials was determined to be equal to zero and the order of $F(q)$ equal to 2. Therefore, the temperature in any previous weeks does not effect the electricity demand in the current week. The relationship between electricity demand and weather may be described by the following equation

$$z(k) = \frac{b_0}{(1 - f_1 q^{-1} - f_2 q^{-2})} w(k-b) \tag{4.12}$$

where $z(k)$ represents the differenced weekly electricity demand time series and $w(k)$ represents the differenced weekly weather time series, wAT or $w$$HDD_{18}$. For each of the AT and HDD_{18} time series the differencing transformations were determined in Section 3.4.1.4 to be as follows:

$$wAT(k) = (1 - q^{-1})(1 - q^{-52})\, AT(k) \tag{4.13}$$

$$wHDD_{18}(k) = (1 - q^{-1})(1 - q^{-52})\, HDD_{18}(k) \tag{4.14}$$

The differenced form of the weather variables, wAT and $wHDD_{18}$, given by (4.13) and (4.14) may be included in the nonlinear model UDBJ_A (4.9) to yield a multivariate nonlinear model of the following form:

$$z(k) = G\big(z(k-1),\ldots..,z(k-p),z(k-52),\ldots..,z(k-52-p),z(k-52P),$$
$$z(k-52P-1),\ldots..,z(k-52P-p),wAT(k),wHDD_{18}(k)\big)+\varepsilon(k) \tag{4.15}$$

where G(.) represents a nonlinear function. Expansion of (4.13) and (4.14) yields the following:

$$wAT(k) = AT(k) - AT(k-1) - AT(k-52) + AT(k-53) \tag{4.16}$$

$$wHDD_{18}(k) = HDD_{18}(k) - HDD_{18}(k-1) - HDD_{18}(k-52) + HDD_{18}(k-53) \tag{4.17}$$

Based on (4.16) and (4.17) the nonlinear model UBJ_A (4.11) may be extended to the multivariate case as follows:

$$\begin{aligned} y(k) = H\big(&y(k-1),\ldots..,y(k-p-1),y(k-52),\ldots..,y(k-52-p-1),y(k-52-p),\ldots.. \\ &\ldots..,y(k-105-p),\ldots.....,y(k-52P-1),y(k-52P),y(k-52P-p-1),y(k-52P-p),\ldots.. \\ &\ldots..,y(k-52P-54),y(k-52P-p-53),AT(k),AT(k-1),AT(k-52),AT(k-53) \\ &HDD_{18}(k),HDD_{18}(k-1),HDD_{18}(k-52),HDD_{18}(k-53)\big)+\varepsilon(k) \end{aligned} \tag{4.18}$$

where $H(.)$ represents a nonlinear function. The models which use MLP's with input structure described by equation (4.15) and (4.18) shall be referred to as the a Multivariate Differenced Box-Jenkins (MDBJ_A) and Multivariate Box-Jenkins (MBJ_A) models respectively.

The number of input variables in the input structure of the MLP's in the UDBJ_A, UBJ_A, MDBJ_A and MBJ_A models depend on the value of the p and P parameters in equations (4.9), (4.11), (4.15) and (4.18). The origin of these equations are from the linear Box-Jenkins model (4.7), whose parameters were previously identified in Chapter 3 as p=6 and P=2 (Table 3.10 - Model No. 1). The number of variables in the input structure of the MLP's for p=6 and P=2 are given in Table 4.2

Table 4.2 Number of variables in input structure for p=6 and P=2

Model	No. of Input Variables
UDBJ_A	21
UBJ_A	31
MDBJ_A	23
MBJ_A	39

Since it is desirable that the MLP input structure contains the absolute minimum number of input variables an attempt was made to reduce the order of p and P. This was achieved through the examination of the t-ratio of the estimated parameters of the linear model identified in Chapter 3. The general guideline used in the linear analysis is that if the absolute value of the t-ratio is greater than 2 then the parameter should be included in the model (Bowerman and O'Connell, 1987). Table 4.3 gives the absolute value of the t-ratio for each of the estimated parameters at lags p = 1 to 6 and P = 52, 104.

Table 4.3 *t-ratio* for estimated parameters of linear Box-Jenkins model

Autoregressive lags	t-ratio
1	16.52
2	10.57
3	7.41
4	6.11
5	6.04
6	5.01
52	18.05
104	11.11

Comparing the relative magnitudes of the absolute value of the t-ratio, the larger values are at lags 1 and 2 at the nonseasonal level and at lag 52 at the seasonal level. Consequently, a reduction in the total number of input variables may be obtained by reducing the value p = 6 to p = 2 and P = 2 to P = 1. The UDBJ_A and UBJ_A models with MLP input structures based on p= *6*, P = *2* and p = *2*, P = *1* were trained and the multi-step-ahead MAE results obtained over the validation set were compared. Table 4.4 presents these results.

Table 4.4 Comparison of models with input structures based on different p and P

p and P	Model	No. of Input Variables	Multi-Step-Ahead MAE Validation Set
p=6 P=2	UDBJ_A	21	1.2053×10^4
	UBJ_A	31	1.3000×10^4
p=2 P=1	UDBJ_A	5	0.7149×10^4
	UBJ_A	11	0.6858×10^4

Based on the results given in Table 4.4 the input structure of the MLP's in the UDBJ_A, UBJ_A, MDBJ_A and MBJ_A models are modified to p = 2 and P = 1, where Table 4.5 gives the specific form of the input structure for each MLP.

Table 4.5 MLP input variables based on p=2 and P=1.

Model	MLP Input Variables
UDBJ_A	$z(k\text{-}1)$, $z(k\text{-}2)$, $z(k\text{-}52)$, $z(k\text{-}53)$, $z(k\text{-}54)$
UBJ_A	$y(k\text{-}1)$, $y(k\text{-}2)$, $y(k\text{-}3)$, $y(k\text{-}52)$, $y(k\text{-}53)$, $y(k\text{-}54)$, $y(k\text{-}55)$, $y(k\text{-}104)$, $y(k\text{-}105)$, $y(k\text{-}106)$, $y(k\text{-}107)$
MDBJ_A	$z(k\text{-}1)$, $z(k\text{-}2)$, $z(k\text{-}52)$, $z(k\text{-}53)$, $z(k\text{-}54)$, $w\text{AT}(k)$, $w\text{HDD}_{18}(k)$
MBJ_A	$y(k\text{-}1)$, $y(k\text{-}2)$, $y(k\text{-}3)$, $y(k\text{-}52)$, $y(k\text{-}53)$, $y(k\text{-}54)$, $y(k\text{-}55)$, $y(k\text{-}104)$, ...$y(k\text{-}105)$, $y(k\text{-}106)$, $y(k\text{-}107)$, $w\text{AT}(k)$, $w\text{AT}(k\text{-}1)$, $w\text{AT}(k\text{-}52)$, $w\text{AT}(k\text{-}53)$, $w\text{HDD}_{18}(k)$, $w\text{HDD}_{18}(k\text{-}1)$, $w\text{HDD}_{18}(k\text{-}52)$, $w\text{HDD}_{18}(k\text{-}53)$

NN-Step 2: Input/output data normalisation

The input and output data for each MLP was normalised by dividing each value of the time series making up the input vector by the its corresponding NF value. The NF value are as follows:

- $y(k) = 1 \times 10^6$
- $\text{AT} = 1 \times 10^2$
- $\text{HDD}_{18} = 1 \times 10^3$
- $z(k) = 1 \times 10^6$
- $w\text{AT} = 1 \times 10^1$
- $w\text{HDD}_{18} = 1 \times 10^2$

The normalisation was carried out prior to the construction of the input training vectors. For each MLP the output neuron has a linear activation function and thus has an unrestricted output range, however, to ensure efficient learning the target output data was normalised to the same magnitude as the input data (divide by 1×10^6 for all cases).

NN-Step 3: Construct input/output training, validation and prediction data sets

The time series data was divided into the following training, validation and prediction sets (see Section 4.2.3):

- training data set - weekly time series data from 2/01/1983 - 25/12/1989.
- validation data set - weekly time series data from 31/12/1989 - 23/12/1990.
- prediction data set - weekly time series data from 30/12/1990 - 22/12/1991.

NN-Step 4: Network structure and hidden layer architecture

For the reasons outlined earlier in Section 4.2.2 an MLP with two hidden layers was chosen. It would be desirable to select the optimal hidden layer architecture for each of the different MLP's through a full experimentation analysis, however, for practical reasons this was not possible due to computational time constraints. Work carried out by Smith (1994) was used to aid the determination of a suitable hidden layer architecture, where his work examined different hidden layer architecture for this specific application. Table 4.6 describes MLP structures employed for each model.

Table 4.6 MLP structure for Box-Jenkins models

Model	MLP Structure	Activation Function	
		Hidden Layers	Output Neuron
UDBJ_A	3-5-1	Log-sigmoid	Linear
UBJ_A	3-5-1	Log-sigmoid	Linear
MDBJ_A	10-30-1	Log-sigmoid	Linear
MBJ_A	10-30-1	Log-sigmoid	Linear

4.3.1.2 Structural Model Approach

The approach using the structural model involved modelling the trend and seasonal components of the time series using the linear BSM (2.44) model and using the neural network to model the remaining residual component. The determination of the residual component involves the estimation of the trend and seasonal components of the weekly electricity demand time series through the use of a *fixed interval smoothing* (FIS) algorithm (Appendix A) and the subsequent removal of these components from the data. The residual component is given by:

$$r(k) = y(k) - \hat{\mu}(k) - \hat{\gamma}(k) + \varepsilon(k) \qquad k = 1,\ldots..N \tag{4.19}$$

where $r(k)$ is the residual component; $y(k)$ is the original time series; $\hat{\mu}(k)$ is the estimated trend component; $\hat{\gamma}(k)$ is the estimated seasonal component; $\varepsilon(k)$ is the forecast error and N is the number of observations in the time series. A general autoregressive nonlinear model may then be defined for $r(k)$ as follows:

$$\boxed{r(k) = m\big(r(k-1),\ldots\ldots\ldots,r(k-52),\ldots\ldots\ldots r(k-l)\big) + \varepsilon(k)} \tag{4.20}$$

where $m(.)$ is a nonlinear function and l represents the order of the autoregressive component of $r(k)$.

The modelling approach involved four distinct stages:

Stage 1. Modelling of $\mu(k)$ and $\gamma(k)$ to obtain $r(k)$ data.

Stage 2. Modelling and prediction of $r(k)$ using MLP network

Stage 3. Prediction of $\mu(k)$ and $\gamma(k)$ using BSM model.

Stage 4. Prediction of $y(k)$ is the sum of the predictions of $r(k)$, $\mu(k)$ and $\gamma(k)$.

Figure 4.10 and Figure 4.11 give a description of the stages involved in the modelling procedure.

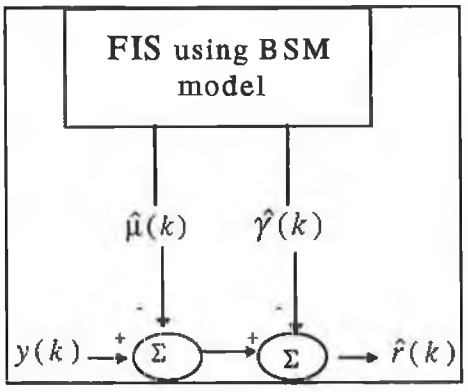

Figure 4.10 Stage 1 of structural modelling procedure

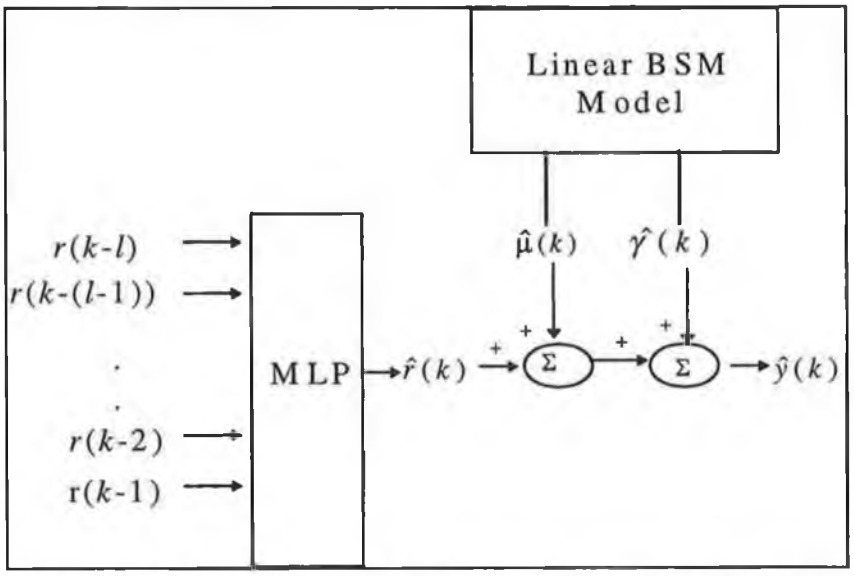

Figure 4.11 Stage 2 to Stage 4 of structural modelling procedure

Stage 1: Modelling of $\mu(k)$ and $\gamma(k)$ to obtain $r(k)$ data

The estimation of the trend, $\mu(k)$, and seasonal component, $\gamma(k)$, of $y(k)$ was carried out using FIS smoothing, with the equations for the smoothing algorithm given in Appendix A ((A6) to (A7)). The FIS algorithm was used to estimate the trend and seasonal components of the time series for every point within the series. For this application the hyperparameters of the basic structural model were estimated previously in Chapter 3 and are described in Table 3.10, Model No. 4. Figure 4.12 and Figure 4.13 show a plot of the Kalman filter estimates of $y(k)$ and the smoothed Kalman filter estimates of $y(k)$ respectively, note that for visual clarity only the first four years of the data are shown.

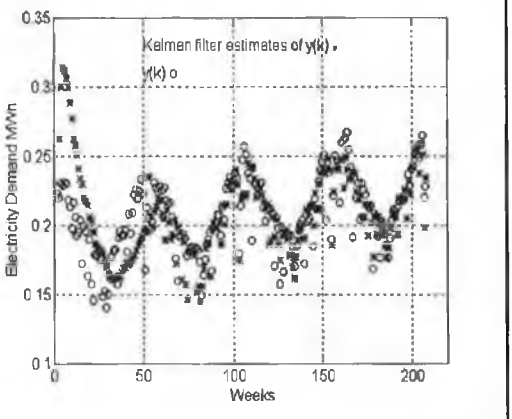

Figure 4.12 Kalman filter estimates of $y(k)$

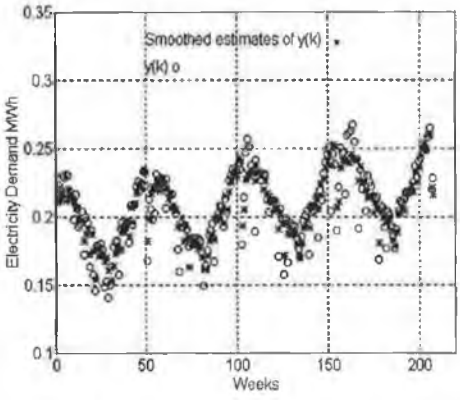

Figure 4.13 Smoothed estimates of $y(k)$

Figure 4.12 shows that the non-smoothed estimates are poor during initial convergence, with the improvement obtained in the smoothed estimates clearly demonstrated in Figure 4.13. The residual component of the time series was determined using equation (4.19) and Figure 4.14 gives a plot of the $r(k)$ time series.

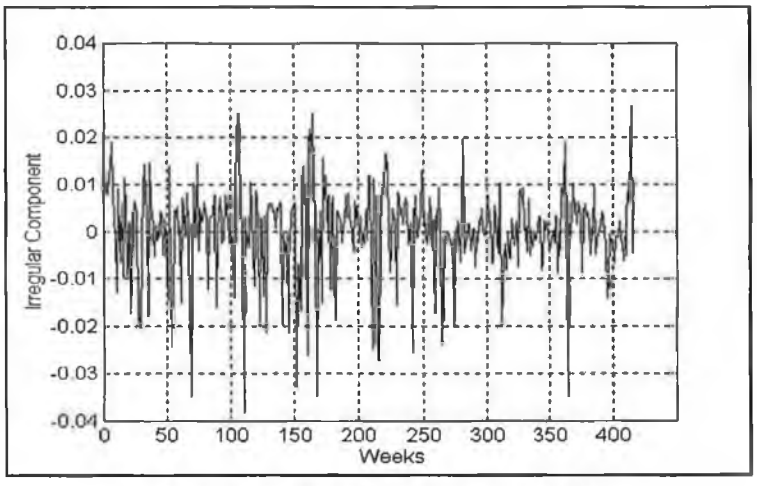

Figure 4.14 Residual component $r(k)$

Stage 2: Modelling and prediction of r(k) using an MLP

The application procedure described in Section 4.2.5 was used to specify the MLP network.

NN-Step 1: Input and output structure

The construction of the input structure for the MLP used to forecast $r(k)$ involved the determination of the lags of $r(k)$ to include in equation (4.20). The SACF (2.8) of $r(k)$ was calculated in an attempt to determine appropriate lags, where the plot of this function is given in Figure 4.15.

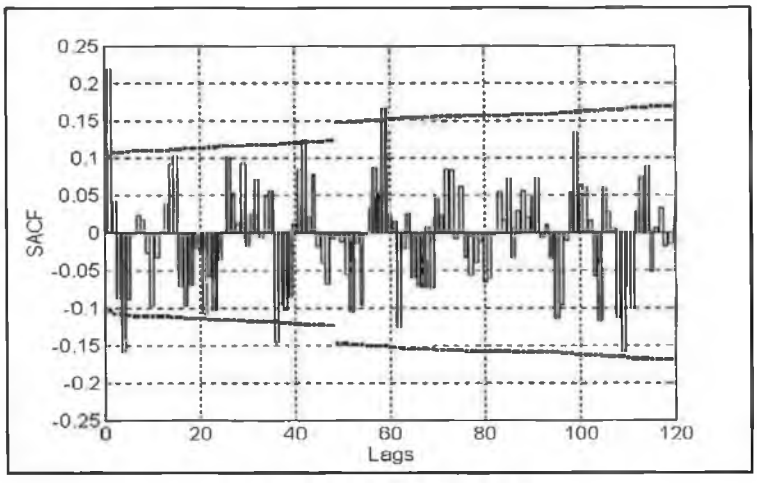

Figure 4.15 SACF of $r(k)$

At the nonseasonal level there are significant lags at 1 and 3 and at the seasonal level at approximately 52 and 104. Thus, the SACF suggests the inclusion of $r(k)$ at lags 1, 2, 3, 52 and 104 in the input structure. These lags were used with the performance of the MLP network further improved when the number of lags at the seasonal level was increased to include 52, 53, 54 and 104, 105, 106. The inclusion of weather variables in the input structure of the MLP was investigated but the MAE multi-step-ahead performance over both validation and prediction was approximately 8% less accurate than the univariate case. Consequently, only past values of the residual component were included in the MLP input structure for $r(k)$. Table 4.7 describes this input structure.

Table 4.7 MLP input structure for $r(k)$

Predict	**MLP Input Variables**
$r(k)$	$r(k\text{-}1)$, $r(k\text{-}2)$, $r(k\text{-}3)$, $r(k\text{-}52)$,$r(k\text{-}53)$,$r(k\text{-}54)$, $r(k\text{-}104)$, $r(k\text{-}105)$, $r(k\text{-}106)$

NN-Step 2: Input/output data normalisation

Since the input structure consists of a single time series, $r(k)$, and given the order of magnitude of this series (1×10^{-1}), shown in Figure 4.14, it was not necessary to normalise the input and output data.

NN-Step 3: Construct input/output training, validation and prediction data set

The data segmentation, used in Section 4.3.1.1 in the Box-Jenkins modelling approach was also used here.

NN-Step 4: Select structure and hidden layer architecture

As in the case of the Box-Jenkins modelling approach it was not possible to carry out a full experimentation analysis into the optimal hidden layer architecture due to computational time constraints. However, using trial and error a network structure which produced good multi-step-ahead performance over the training and validation data set was determined. This structure is described in Table 4.8.

Table 4.8 Network structure for $r(k)$ MLP

MLP Structure	Activation Function	
	Hidden Layers	Output Neuron
20-60-1	Log-sigmoid	Linear

Stage 3: Prediction of $\mu(k)$ and $\gamma(k)$

The prediction of $\mu(k)$ and $\gamma(k)$ was carried out using the BSM with dummy seasonal component already estimated in Chapter 3, where this model is described in Table 3.10, Model No. 4.

Stage 4: Prediction of $y(k)$

The prediction of $r(k)$ obtained in Stage 2 and the prediction of $\mu(k)$ and $\gamma(k)$ obtained in Stage 3 are combined to yield $y(k)$, where the results for $y(k)$ are given in Section 4.3.1.4 in Table 4.18 and Table 4.19.

4.3.1.3 Autoregressive Model

To demonstrate a simple black-box modelling approach an MLP with a purely autoregressive input structure was also used to forecast weekly electricity demand.

NN-Step 1: Input and output structure

A purely autoregressive input structure consisting of 52 past values of the time series, corresponding to a full season of data, was constructed. A nonlinear model for this case may be described by the following:

$$y(k) = M(y(k-1), y(k-2), y(k-3), \ldots\ldots, y(k-51), y(k-52)) + \varepsilon(k) \quad (4.21)$$

where $M(.)$ represents a nonlinear function and $\varepsilon(k)$ represents the forecast error. This is referred to as the AR(52) model.

NN-Step 2: Input/output data normalisation

The weekly electricity demand time series was normalised prior to the construction of the input training vectors by dividing each value of the time series by its NF value (1×10^6) and the output data was normalised to the same order of magnitude as the input data (divide by 1×10^6).

NN-Step 3: Construct input/output training, validation and prediction data sets

The same data segmentation that was used in Box-Jenkins and structural modelling approaches in Section 4.3.1.1 and Section 4.3.1.2 was also used here.

NN- Step 4: Structure and hidden layer architecture

For comparative purposes the MLP network structure of the univariate AR(52) model was chosen to be the same as that used in the univariate UBJ_A and UDBJ_A models. Table 4.9 describes the structure.

Table 4.9 Network structure for AR(52) MLP

MLP Structure	**Activation Function**	
	In the Two Hidden Layers	**Output Neuron**
3-5-1	Log-sigmoid	Linear

4.3.1.4 Network Training and Forecasting Results

The neural networks were trained using the LMS gradient technique with backpropagation (Section 4.2.3), with an adaptive learning rate (initial value of 1×10^{-3}) and a momentum constant equal to 0.95. Each network is re-trained ten times in an attempt to circumvent the sensitivity to initial conditions (Section 4.2.3). The network training cessation point was selected through the examination of the multi-step-ahead MAE performance over the validation

set (MMAE_val) as discussed in Section 4.2.4. The following results are given for each training run (referred to as Result Set 1):

1. MMAE_val.
2. Network training cessation point.
3. Multi-step-ahead MAE over prediction set.

For comparative purposes results are presented where the network training cessation point was selected through the examination of the single-step-ahead SSE performance over the training set. Since the SSE will continue to decrease at each training epoch the optimal SSE occurs at the final training epoch, that is the fixed number of training epochs. For each training run the normalised single-step-ahead SSE (MSE) is given, where this is referred to as SMSE_tr. The following results are given for this case (referred to as Result Set 2):

1. SMSE_tr.
2. Network training cessation point = fixed number of training epochs.
3. Multi-step-ahead MAE over prediction set.

For both result sets the average value over the ten training is presented for each result. Table 4.10 to Table 4.21 give the results for each of the models.

Table 4.10 Results for UDBJ_A model - Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	7.1790×10^3	29881	6.8860×10^3
2	7.2904×10^3	29876	6.7746×10^3
3	7.1118×10^3	29907	6.5865×10^3
4	9.3909×10^3	29806	7.0104×10^3
5	9.0564×10^3	29800	6.8135×10^3
6	1.1197×10^4	29803	8.1852×10^3
7	8.0102×10^3	29807	6.6456×10^3
8	7.1490×10^3	29845	6.5415×10^3
9	8.1541×10^3	29846	6.6951×10^3
10	1.2199×10^4	29807	9.0160×10^3
Average	8.7738×10^4	29838	7.1455×10^3

Table 4.11 Results for UDBJ_A model - Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	293.75	30000	2.2470×10^4
2	283.28	30000	0.8979×10^4
3	273.04	30000	6.3565×10^4
4	289.00	30000	0.7705×10^4
5	278.76	30000	0.7820×10^4
6	274.54	30000	1.5096×10^4
7	292.56	30000	0.7705×10^4
8	285.18	30000	0.7820×10^4
9	298.46	30000	0.7430×10^4
10	294.28	30000	1.0490×10^4
Average	286.28	30000	1.5908×10^4

Table 4.12 Results for UBJ_A model - Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	7.0749×10^3	39086	9.3060×10^3
2	6.8247×10^3	39015	9.0925×10^3
3	7.5083×10^3	39224	9.6032×10^3
4	7.1203×10^3	39359	9.3432×10^3
5	7.5930×10^3	39495	9.8535×10^3
6	6.8589×10^3	38567	9.1489×10^3
7	7.3369×10^3	39129	9.5380×10^3
8	7.3412×10^3	39139	9.4724×10^3
9	6.9347×10^3	39776	9.2060×10^3
10	7.2175×10^3	39565	9.3383×10^3
Average	7.1910×10^3	39236	9.3902×10^3

Table 4.13 Results for UBJ_A model - Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	154.12	40000	1.1929×10^4
2	155.31	40000	1.2094×10^4
3	166.61	40000	1.3385×10^4
4	155.63	40000	1.1433×10^4
5	160.12	40000	1.2601×10^4
6	154.29	40000	1.1755×10^4
7	161.03	40000	1.2337×10^4
8	163.25	40000	1.2062×10^4
9	153.67	40000	1.1080×10^4
10	155.46	40000	1.1399×10^4
Average	157.95	40000	1.2008×10^4

Table 4.14 Results for MDBJ_A model - Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	6.0061 x10^3	15790	6.1044 x10^3
2	6.0773 x10^3	15938	6.2210 x10^3
3	6.1233 x10^3	12141	6.1168 x10^3
4	6.1665 x10^3	15688	6.2274 x10^3
5	6.1099 x10^3	15951	6.5599 x10^3
6	6.4501 x10^3	15963	6.4986 x10^3
7	5.5488 x10^3	7584	6.0582 x10^3
8	6.2693 x10^3	15830	6.3409 x10^3
9	6.3632 x10^3	15986	6.5911 x10^3
10	6.6650 x10^3	15887	7.0494 x10^3
Average	6.1780 x10^3	14676	6.3768 x10^3

Table 4.15 Results for MDBJ_A model - Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	492.57	16000	2.8123 x10^4
2	427.07	16000	3.5201 x10^4
3	419.69	16000	6.5690 x10^3
4	437.27	16000	6.2613 x10^3
5	439.03	16000	3.7607 x10^4
6	419.21	16000	1.0752 x10^4
7	403.13	16000	6.6272 x10^3
8	443.67	16000	6.4607 x10^3
9	441.42	16000	7.0697 x10^3
10	485.92	16000	1.3630 x10^4
Average	440.90	16000	1.5830 x10^4

Table 4.16 Results for MBJ_A model - Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	6.0160 x10^3	49892	8.0084 x10^3
2	6.8663 x10^3	34628	8.2311 x10^3
3	7.6078 x10^3	49857	8.5897 x10^3
4	8.1573 x10^3	49646	9.1075 x10^3
5	7.1181 x10^3	47993	8.7441 x10^3
6	7.1501 x10^3	49927	8.5954 x10^3
7	7.7932 x10^3	49324	8.9950 x10^3
8	7.5659 x10^3	49784	8.1294 x10^3
9	7.6585 x10^3	49922	8.9795 x10^3
10	8.5320 x10^3	46983	9.1812 x10^3
Average	7.5465 x10^3	47796	8.6561 x10^3

Table 4.17 Results for MBJ_A model - Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	154.58	50000	8.2847×10^3
2	147.86	50000	8.2108×10^3
3	151.02	50000	8.9355×10^3
4	160.12	50000	9.6343×10^3
5	155.25	50000	9.2200×10^3
6	151.66	50000	8.7553×10^3
7	164.83	50000	1.0125×10^4
8	157.01	50000	8.1042×10^3
9	149.12	50000	9.4504×10^3
10	163.58	50000	9.8336×10^3
Average	155.50	50000	9.0554×10^3

Table 4.18 Results for structural model - Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	2.8965×10^3	645	7.2739×10^3
2	3.9759×10^3	1453	7.9150×10^3
3	2.7675×10^3	371	7.5447×10^3
4	2.8475×10^3	1389	7.8416×10^3
5	3.8571×10^3	1570	7.5481×10^3
6	2.9405×10^3	854	7.4538×10^3
7	3.9911×10^3	578	7.1716×10^3
8	2.7511×10^3	514	7.1978×10^3
9	3.0468×10^3	632	7.4873×10^3
10	2.8492×10^3	419	7.5554×10^3
Average	3.1923×10^3	842	7.5171×10^3

Table 4.19 Results for structural model-Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	255.59	2000	7.2243×10^3
2	193.60	2000	7.5622×10^3
3	212.34	2000	7.3391×10^3
4	200.14	2000	7.7790×10^3
5	209.03	2000	7.5187×10^3
6	225.33	2000	7.3208×10^3
7	189.36	2000	7.3704×10^3
8	198.20	2000	7.2498×10^3
9	222.48	2000	7.2743×10^3
10	210.65	2000	7.4456×10^3
Average	211.67	2000	7.4084×10^3

Table 4.20 Results for AR(52)_A model -Result Set 1

Run no.	Validation Set		Prediction Set
	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	1.0025×10^4	59677	1.3266×10^4
2	0.9070×10^4	59721	1.1839×10^4
3	1.0465×10^4	59441	1.4157×10^4
4	0.8558×10^4	59970	1.0642×10^4
5	0.9980×10^4	59969	1.2851×10^4
6	0.9508×10^4	59928	1.2089×10^4
7	0.9205×10^4	59911	1.2179×10^4
8	0.9274×10^4	59627	1.4846×10^4
9	1.1063×10^4	59892	1.2761×10^4
10	0.9435×10^4	59250	1.2761×10^4
Average	1.7360×10^4	59739	1.2739×10^4

Table 4.21 Results for AR(52)_A model -Result Set 2

Run no.	Training Set		Prediction Set
	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1	247.00	60000	1.4544×10^4
2	243.35	60000	1.2410×10^4
3	259.24	60000	1.7439×10^4
4	228.45	60000	1.0755×10^4
5	245.35	60000	1.3205×10^4
6	236.01	60000	1.2808×10^4
7	242.59	60000	1.3339×10^4
8	241.07	60000	1.3169×10^4
9	249.17	60000	1.6179×10^4
10	238.64	60000	1.3676×10^4
Average	243.09	60000	1.3752×10^4

Before going on to make some general conclusions some comments are made regarding the training results. In the case of the MDBJ_A, MBJ_A and structural models the networks have been successfully trained within the fixed number of training epochs. However, for the UDBJ_A, UBJ_A and AR(52) this is not the case where the tolerance levels attained were as follows:

- UDBJ_A - 0.56% over 10000 epochs.
- UBJ_A - 0.35% over 10000 epochs.
- AR(52) - 0.36% over 10000 epochs.

It is interesting to compare the number of training epochs required to train the networks to a reasonable degree of accuracy to both network structure and type of training data. This is achieved by comparing the average of the training cessation points (based on the MMAE_val)

calculated over the ten training runs with the different MLP properties. Table 4.22 provides such a comparison, where n/a refers to not applicable.

Table 4.22 Comparison of average number of training epochs to network properties

Model	Average No. of Training Epochs	No. of Inputs	MLP Structure	Training Data	Tolerance Level
UDBJ_A	29838	5	3-5-1	differenced	0.5%
UBJ_A	39236	11	3-5-1	original	0.35%
MDBJ_A	14676	7	10-30-1	differenced	n/a
MBJ_A	47796	19	10-30-1	original	n/a
Structural	842	9	20-60-1	residual	n/a
AR(52)	59739	52	3-5-1	original	0.36%

It can be seen that the required number of training epochs depends on the network size but a more significant factor is the type of training data used. For example, the networks which use pre-processed data in the training vectors take relatively less number of epochs to train successfully. An outstanding example is in the case of the structural model where it only takes on average 842 epochs to train the network. Moreover, comparison of the Box-Jenkins models show that the UDBJ_A and MDBJ_A models took on average less number of iterations to train than the UBJ_A, MBJ_A models which use original data in the training vectors. The reason is that the pre-processing removes a certain amount of information from the data prior to modelling and the system which the network is required to represent has been simplified. This is particularly true in the case of the structural model where the majority of the information has been removed from the training data through pre-processing. In contrast, in the Box-Jenkins case the pre-processing (differencing) removes a proportion of the data but there still remains a reasonable amount of information content which the network is required to model. It is also interesting to note that the multivariate model MDBJ_A took a significantly less amount of time to train than its univariate counterpart UDBJ_A. The reason for this is that the input structure in the MDBJ_A model has the advantage of being pre-processed but also presents a truer representation of the system through the inclusion of weather variables.

The results show that in the majority of cases the selection of the training cessation point based on the MMAE_val produced a more accurate multi-step-ahead MAE performance over the prediction set than the selection based on the SMSE_tr (that is, the final training epoch). The only exception is in the case of the structural model where the average over the 10 training runs of the multi-step-ahead MAE over the prediction set is 7.5171×10^3 using the MMAE_val to choose the training cessation point and 7.4084×10^3 using the SMSE_tr to choose the training cessation point.

The results also demonstrate the sensitivity of the network training results to the stochastic initialisation of the weight and biases during training. A notable example of this is in the UDBJ_A results where there is a difference of approximately 27% in the multi-step-ahead MAE over the prediction set produced by training run 3 (6.5872×10^3) and training run 10 (9.0160×10^3), with a similar variance in the validation set results. Similarly, in the case of the AR(52) model there is a difference of approximately 28% in the multi-step-ahead MAE over the prediction set produced by training run 4 (1.0642×10^4) and training run 8 (1.4864×10^4). In the cases of the other models differences of the order of approximately 13% occurred.

In order to compare the models performance over the prediction set a summary table of the entire set of results is given in Table 4.23 where the multi-step-ahead MAE over the prediction set is given, with the best result out of the ten training runs selected according to the following criterion:

- Minimum MMAE_val out of all training runs.

The following are also included in the table:

- Average of multi-step-ahead MAE over prediction set of the ten training runs.
- Multi-step-ahead MAE over prediction set produced by the equivalent linear model developed in Chapter 3.

Note that, in the case of the MDBJ_A and MBJ_A there is no linear model for which a true direct comparison can be made since a bivariate analysis as opposed to a multivariate was conducted in Chapter 3. However, in this instance the most accurate linear bivariate model result is given, that is BJTF with HDD_{18} as an exogenous variable.

Table 4.23 Summary of forecasting results over the prediction data set

Model	**Minimum MMAE_val Out of Training 10 Runs**		**Average Over 10 Training runs**	**Linear Model**
	Training Run	Multi-Step-Ahead MAE (Prediction Set)	Multi-Step-Ahead MAE (Prediction Set)	Multi-Step-Ahead MAE (Prediction Set)
UDBJ_A	3	0.6586×10^4	0.7146×10^4	1.0691×10^4
UBJ_A	2	0.9093×10^4	0.9390×10^4	
MDBJ_A	7	0.6058×10^4	0.6377×10^4	0.6777×10^4
MBJ_A	1	0.8008×10^4	0.8656×10^4	
Structural	8	0.7198×10^4	0.7517×10^4	0.7891×10^4
AR(52)	4	1.0642×10^4	1.2739×10^4	1.1101×10^4

Comparison of the multi-step-ahead MAE over the prediction set results show that the pre-processing of the training data achieves higher accuracy than the networks which use original data in the training vectors. Comparing the results achieved between UDBJ_A and UBJ_A and between MDBJ_A and MBJ_A this improvement in accuracy is of the order of approximately 27% in the univariate case and 32% in the multivariate case. In addition, as discussed above pre-processing the data has the added advantage that the network take less numbers of iterations to train. It is interesting to note that the univariate structural model does not outperform the univariate UDBJ_A model, where in the linear application this was found to be the case. A likely reason for this is that the multi-step-ahead prediction in the case of the structural model is a combination of forecasts produced by a linear and neural network model, whereas in the UDBJ_A the prediction was produced solely by a neural network.

Although the UBJ_A and MBJ_A models use original data in the input structure the use of linear Box-Jenkins modelling techniques to determine a parsimonious input structure to the MLP may also be considered as a form of pre-processing when compared with the more traditional black-box modelling approach. Comparison of the multi-step-ahead MAE over the prediction set results produced by the UBJ_A and the AR(52) models shows that significant improvement (of the order of approximately 38%) can be achieved through the use of use of Box-Jenkins linear analysis to determine the input structure to the MLP.

Overall comparison of the multi-step-ahead MAE results in Table 4.32 show that all neural network based models outperform their linear counterparts with the exception of the MBJ_A model. The MDBJ_A produces the best multi-step-ahead MAE result over the prediction set (0.6058×10^4). The improvement over its linear counterpart is of the order of 11% and over the second best neural network model (UDBJ_A) of the order of 8%.

Finally, the criterion (minimum MMAE_val) used to select the training run out of the 10 possible training runs is consistent with the multi-step-ahead MAE over the prediction set in all but two cases. The first case is the UDBJ_A model where the most accurate multi-step-ahead prediction is produced by training run 3 over the validation set but by training run 8 over the prediction set. The difference in accuracy is of the order of approximately 0.7%. In the case of the structural model using the minimum MMAE_val criterion training run 8 is selected but the most accurate multi-step-ahead result over the prediction is produced by training run 7. Here the difference in accuracy is of the order of approximately 0.4%. Therefore for this application given the low magnitude in the percentage differences and the overall success of the criterion, the minimum MMAE_val may be considered as a suitable criterion to adopt when the

selecting the best training run out of a set of possible training runs. The full set of forecasting accuracy results for each of the models are given in Table 4.24.

Table 4.24 Forecasting accuracy results for weekly electricity demand

		Multi-Step-Ahead Over Prediction Set			
Model	**Training Run**	**MAE**	**MSE**	**MAPE**	**MPE**
UDBJ_A	3	0.6586×10^4	1.0620×10^8	2.33	0.11
UBJ_A	2	0.9093×10^4	1.9870×10^8	3.21	0.38
MDBJ_A	7	0.6058×10^4	0.9191×10^8	2.16	0.39
MBJ_A	1	0.8008×10^4	1.5884×10^8	2.88	-0.37
Structural	8	0.7198×10^4	1.0104×10^8	2.57	0.12
AR(52)	4	1.0642×10^4	1.9261×10^8	3.74	0.71

The additional forecasting accuracy measures MSE, MAPE and MPE also indicate that the models which use pre-processed training data produce the most accurate forecast over the prediction set. Based on the MAE, MSE and MAPE measures the MDBJ_A produces the most accurate result over the prediction set. The relatively high value of the MPE (0.39) for this model indicates that it tends to make predictions that are lower on average in comparison with the UDBJ_A (0.11) and the structural models (0.12). The UDBJ_A and structural models produce comparatively accurate results over all forecasting accuracy measures.

Figure 4.16 to Figure 4.21 give the actual versus predicted graphical results.

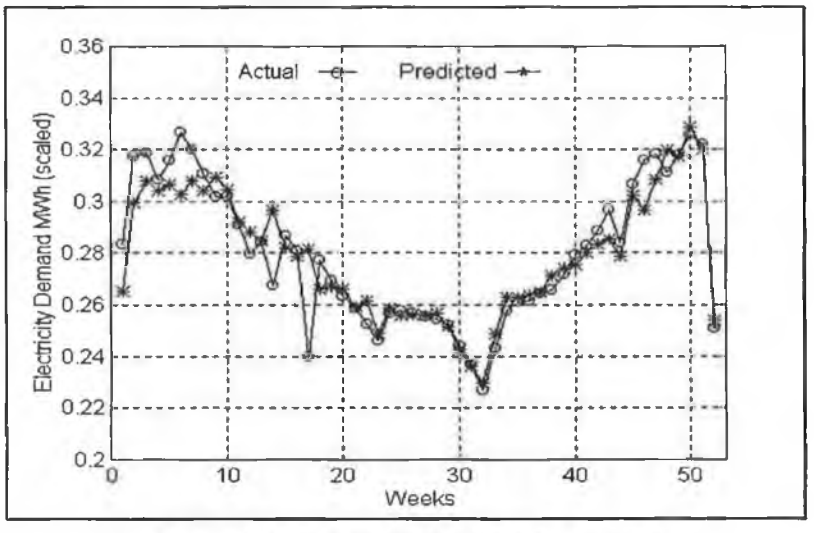

Figure 4.16 Forecast using UDBJ_A model

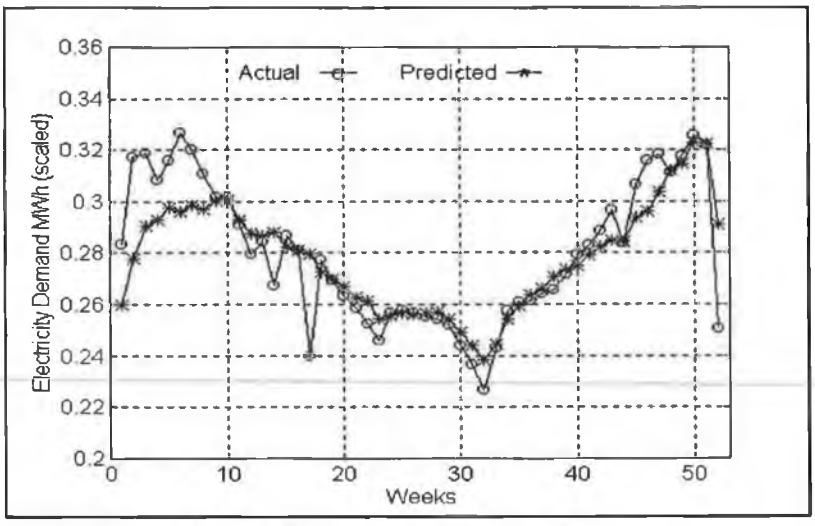

Figure 4.17 Forecast using UBJ_A model

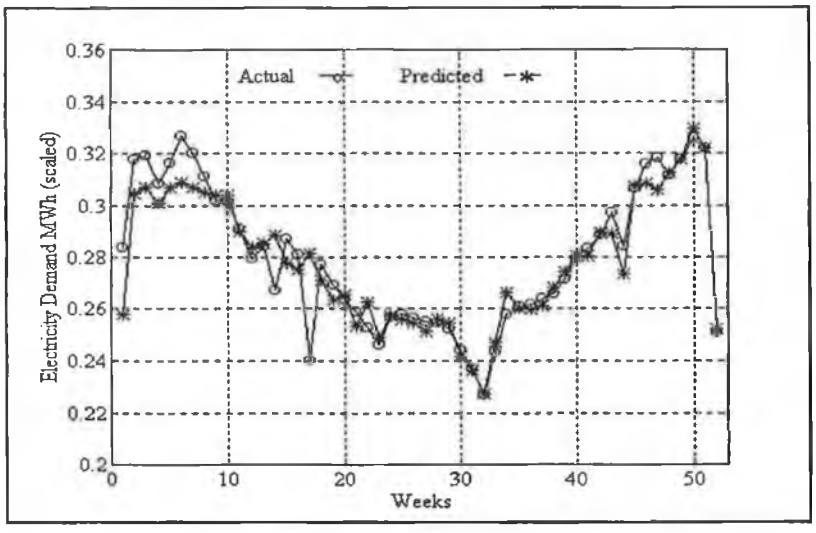

Figure 4.18 Forecast using MDBJ_A model

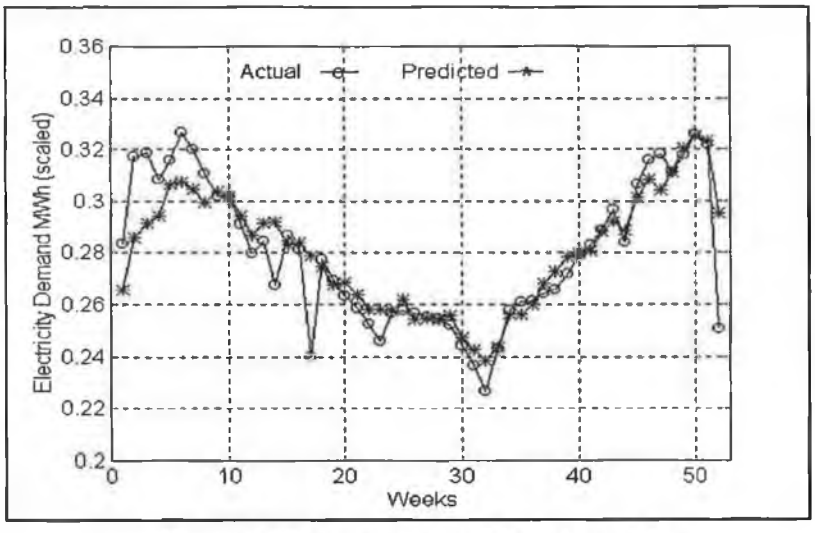

Figure 4.19 Forecast using MBJ_A model

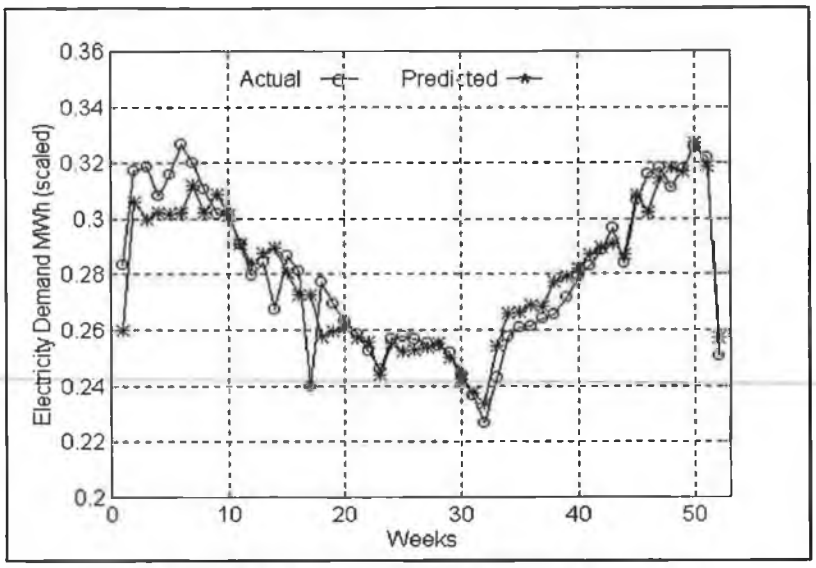

Figure 4.20 Forecast using structural model

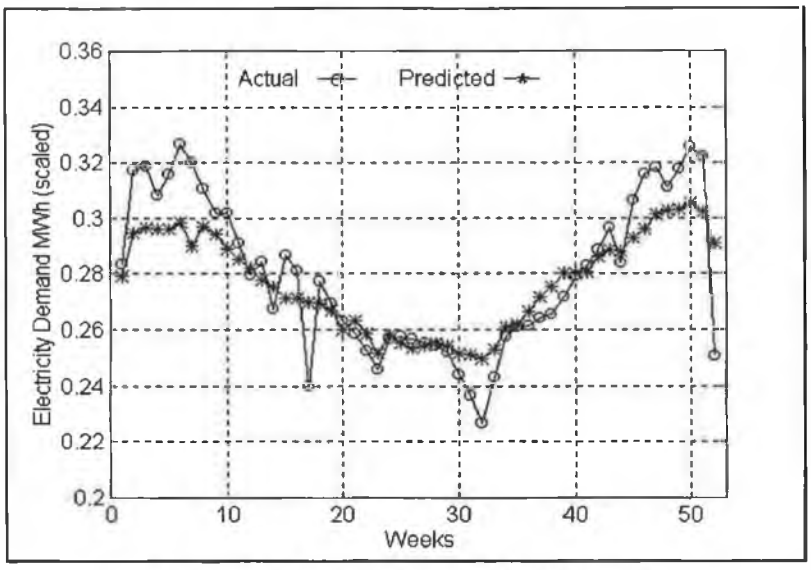

Figure 4.21 Forecast using AR(52) model

Examining the results graphically it can be seen that the greatest variance between the actual and predicted data occurs at Weeks 1 to 8; Week 17; Week 32 and Week 52. The uncharacteristically low demand at Week 17 is not due to weather effects since the AT for this week is $8^{O}C$ which is typical. Therefore, it is assumed that this value is an outlier or that the low demand is due to some other influencing factor. A striking feature of the predictions is that the UDBJ_A, MDBJ_A and the structural models produce a more accurate prediction of Weeks 2 to 8; Week 32 and Week 52 than the UBJ_A, AR(52) and MBJ_A model. In addition, all the models with the exception of the AR(52) produce an inaccurate prediction of Week 1, where the inaccuracy is of the order of approximately 7%. In an attempt to explain these variances the time series data over the prediction and training sets are compared. For

visual clarity only the most recent two years of the training data set are shown. Figure 4.22 shows the weekly electricity demand time series data for the following years:

- 1991- prediction set.
- 1989 - training data.
- 1988 - training data.

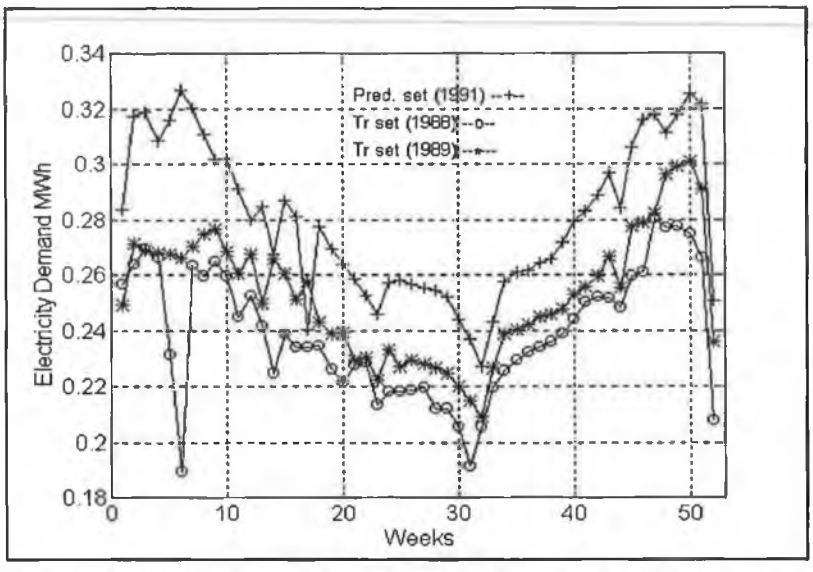

Figure 4.22 Two years of original training data and original prediction data

Examination of Figure 4.22 shows that the three profiles are reasonably consistent during Week 18 to Week 47, with the exception of Week 32 in 1988. However, from Week 1 to Week 18, during the winter (December to February) and spring (March to April) months, the profiles vary mainly due to the dominant effect of weather influences. In addition, there appears to be an outlier at Week 6 in 1988. Comparing the annual increase in demand due to economic and demographic factors, it can be seen that the increase in 1991 is greater relative to the increase in 1989. A more significant difference can be seen during Weeks 1 to 8 where electricity demand is higher in the prediction data than in the training data. One of the reasons for this was an unusually cold spell in 1991 where AT was on average approximately 3°C during this period, whereas in contrast it was on average approximately 7°C in 1988 and 1989. Therefore, the relatively lower demand produced by the univariate UBJ_A and AR(52) models for Weeks 1 to 8 can be matched to the relatively low demand during Weeks 1 to 8 of 1988 and 1989. In the case of the MBJ_A the input weather variables result in the production of a more accurate prediction of demand during Weeks 5 to 8 but this model still under predicts demand for the first 4 weeks of the forecast horizon. Since the UDBJ_A and structural models were also

capable of producing a reasonably accurate prediction of Weeks 2 to 8 but do not contain weather variables a contributing factor to the success of the UDBJ_A, MDBJ_A and structural model must be the fact that these models use pre-processed training data. Figure 4.23 shows a plot of the differenced data and Figure 4.24 shows a plot of the structural model residual data, where for visual clarity the prediction set (1991) and one year (1989) of the training data set is given.

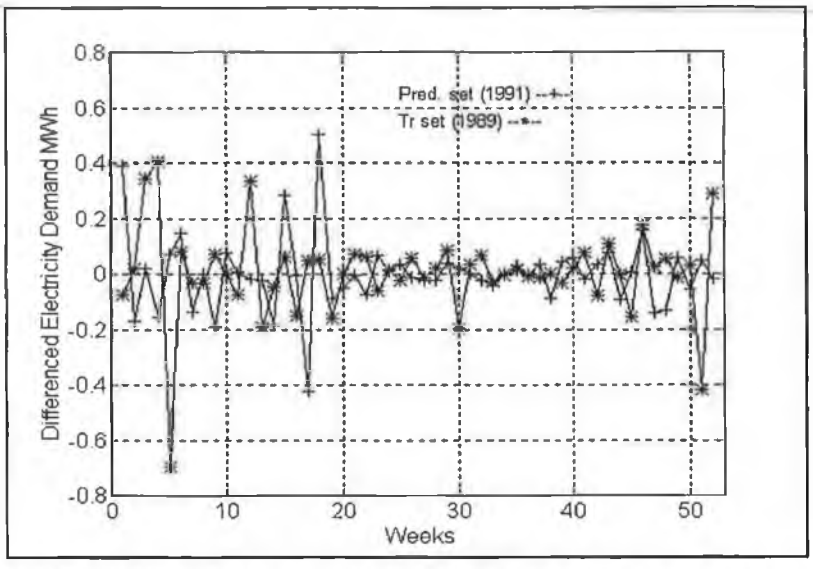

Figure 4.23 One year of differenced training data and differenced prediction data

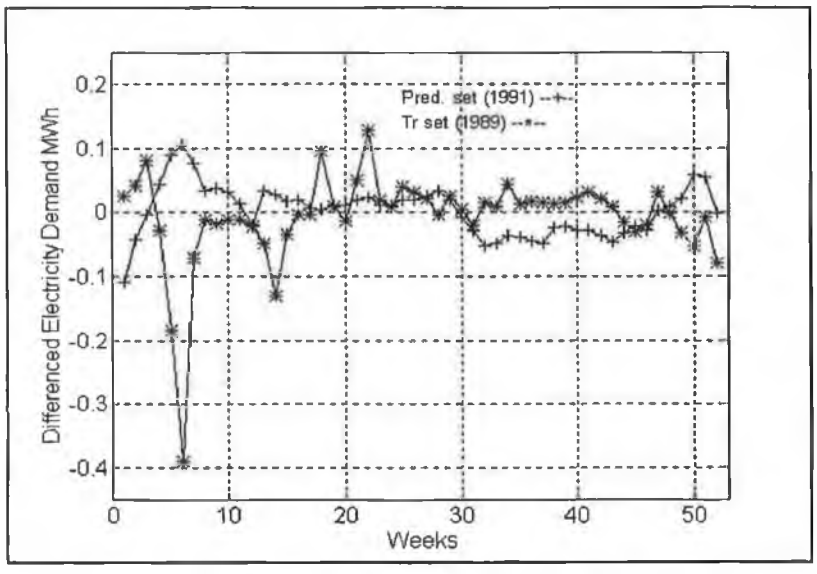

Figure 4.24 One year of residual training data and residual prediction data

The spurious value at Week 5 in the 1989 differenced and residual data is due to the effect of the outlier at Week 6 in the 1988 original data (see Figure 4.22). Similarly, the spurious

values at Weeks 17 and 18 in the 1991 differenced data are due effect of the outlier at Week 17 in the 1991 original data. In each case notwithstanding the effect of the outliers there is still a variance between the training data and prediction data during the winter and spring months (Weeks 1 to 17 and Weeks 48 to 52), however the magnitude of the variance has been reduced through pre-processing. Consequently, the neural networks which are trained on pre-processed data produce have the capability of producing a more accurate prediction over this period.

The important conclusions that were determined from the results of the neural network analysis and that are valid for the System A weekly electricity demand application are summarised as follows:

- the fixed number of training epochs that are required to train the network is highly dependent on the type of training data, that is original or pre-processed data.
- overall the selection of the training cessation point based on the multi-step-ahead performance over the validation set produces good performance over the prediction set.
- it is sufficient to re-train the network 10 times in order to circumvent the sensitivity of the network to initial conditions.
- the minimum MMAE_val is a suitable criterion to use when selecting the best training out of a set of training runs.
- the models that were trained using pre-processed training data produced the most accurate results over the prediction set.

4.3.2 System B

The System B data set that was introduced in Section 3.2 consists of weekly electricity demand (MWh), AT and HDD_{18} data from 5th April 1980 to 25th August 1990. Figure 3.3 gives a plot of this series and Section 3.2 discusses the main attributes of the time series. Neural networks were applied to model the weekly electricity demand system, where a similar analysis to the one performed on the System A data set was carried out. However, preliminary results were not promising. The results showed that in contrast to the System A case the MLP networks did not produce more accurate forecasting results in comparison with their linear counterparts. This was true for both univariate and multivariate neural networks. MLP's with a number of different design configurations were trained but it was not possible to determine a suitable MLP structure which could model the System B data set. Due to the consistent achievement of poor forecasting results it was decided not to continue with a full neural network analysis. A brief summary of the preliminary analysis is given in now given.

4.3.2.1 System B MLP Specification

The Box-Jenkins modelling approach (Section 4.3.1.1) employed in the System A case was also employed here, where the application procedure outlined in Section 4.2.5 was used to specify the MLP networks. An MLP with a purely autoregressive input structure was also used.

NN-Step 1: Input and output structure

Input structures adapted from the Box-Jenkins linear model (2.19) similar to those developed in the System A weekly electricity demand application were constructed. Nonlinear models were adapted based on the linear Box-Jenkins models identified for this data set which are given in Chapter 3, Table 3.13, Model No. 25. Table 4.25 gives a description of the MLP input structures.

Table 4.25 MLP input structures

Model	MLP Input Variables
UDBJ_B	$z(k\text{-}1)$, $z(k\text{-}2)$, $z(k\text{-}3)$, $z(k\text{-}52)$, $z(k\text{-}53)$, $z(k\text{-}54)$, $z(k\text{-}55)$
UBJ_B	$y(k\text{-}1)$, $y(k\text{-}2)$, $y(k\text{-}3)$, $y(k\text{-}4)$, $y(k\text{-}52)$, $y(k\text{-}53)$, $y(k\text{-}54)$, $y(k\text{-}55)$, $y(k\text{-}56)$, $y(k\text{-}104)$, $y(k\text{-}105)$, $y(k\text{-}106)$, $y(k\text{-}107)$, $y(k\text{-}108)$
MDBJ_B	$z(k\text{-}1)$, $z(k\text{-}2)$, $z(k\text{-}3)$, $z(k\text{-}52)$, $z(k\text{-}53)$, $z(k\text{-}54)$,$z(k\text{-}55)$, $w\text{AT}(k)$, $w\text{HDD}_{18}(k)$
MBJ_B	$y(k\text{-}1)$, $y(k\text{-}2)$, $y(k\text{-}3)$, $y(k\text{-}4)$, $y(k\text{-}52)$, $y(k\text{-}53)$, $y(k\text{-}54)$, $y(k\text{-}55)$, $y(k\text{-}56)$, $y(k\text{-}104)$, $y(k\text{-}105)$, $y(k\text{-}106)$, $y(k\text{-}107)$, $y(k\text{-}108)$, $w\text{AT}(k)$, $w\text{AT}(k\text{-}1)$, $w\text{AT}(k\text{-}52)$, $w\text{AT}(k\text{-}53)$, $w\text{HDD}_{18}(k)$, $w\text{HDD}_{18}(k\text{-}1)$, $w\text{HDD}_{18}(k\text{-}52)$, $w\text{HDD}_{18}(k\text{-}53)$
AR(52)_B	$y(k\text{-}1)$,.....$y(k\text{-}52)$

NN-Step 2: Input and output data normalisation

The input data is normalised prior to the construction of the input vectors by dividing each time series by its corresponding NF value which are as follows:

- $y(k) = 1 \times 10^{8}$;
- $\text{AT} = 1 \times 10^{2}$
- $\text{HDD}_{18} = 1 \times 10^{2}$
- $z(k) = 1 \times 10^{7}$;
- $w\text{AT} = 1 \times 10^{1}$
- $w\text{HDD}_{18} = 1 \times 10^{2}$

The output was also normalised to the same order of magnitude of the input data (divide by 1 x 10^8 for UBJ_B; MBJ_B and divide by 1 x 10^7 for MBJ_B; MDBJ_B).

NN-Step 3: Training, validation, prediction data sets

The training, validation and prediction data sets are constructed as follows:

- training data set - weekly time series data from 31/08/1980 - 21/08/1988.
- validation data set - weekly data time series from 28/8/1980 - 20/08/1989.
- prediction data set - weekly data time series from 27/08/1989 - 19/08/1990.

NN-Step 4: Network structure and hidden layer architecture

An MLP with two hidden layers was used, where work carried out by Smith (1994) was used to select the hidden layer architecture. Table 4.26 describes the MLP structures for each model.

Table 4.26 MLP input structure

Model	MLP Structure	Activation Function	
		In the Two Hidden Layers	Output Neuron
UDBJ_B	2-6-1	Log-sigmoid	Linear
UBJ_B	2-6-1	Log-sigmoid	Linear
MDBJ_B	3-9-1	Log-sigmoid	Linear
MBJ_B	3-9-1	Log-sigmoid	Linear
AR(52)_B	2-6-1	Log-sigmoid	Linear

4.3.2.2 Training and Forecasting Results

The networks were trained using the LMS gradient technique with backpropagation, with an adaptive learning rate (initial value of $1x10^{-3}$), a momentum constant equal to 0.95 and each network was re-trained 10 times. For brevity the most accurate result achieved out of the 10 training runs is given. Similar results to those given in the System A analysis are presented in Table 4.27 and Table 4.28, that is Result Set 1 and Result Set 2 respectively.

Table 4.27 Summary of results for System B - Result Set 1

Model	Validation Set		Prediction Set
	Multi-Step-Ahead MAE	Training Cessation @ Epoch	Multi-Step-Ahead MAE
UDBJ_B	2.6710×10^5	22114	3.4817×10^5
UBJ_B	4.1706×10^5	14750	3.8789×10^5
MDBJ_B	4.2970×10^5	22778	3.8560×10^5
MBJ_B	5.1230×10^5	58784	4.2190×10^5
AR(52)_B	2.9940×10^5	58690	3.9686×10^5

Table 4.28 Summary of results for System B - Result Set 2

Model	Training Set		Prediction Set
	Single-Step-Ahead MSE	Training Cessation @ Epoch	Multi-Step-Ahead MAE
UDBJ_B	1.9640×10^{3}	30000	3.4978×10^{5}
UBJ_B	1.5240×10^{3}	40000	4.4126×10^{5}
MDBJ_B	2.1423×10^{3}	30000	4.0060×10^{5}
MBJ_B	2.0019×10^{3}	60000	4.2240×10^{5}
AR(52)_B	2.1327×10^{3}	60000	3.8235×10^{5}

Comparing the results obtained using the MLP's to those obtained using the equivalent linear models in Chapter 3 show that the linear models produce the most accurate results over the prediction set. The variance in the MAE over the prediction set between the UDBJ_B and the equivalent linear model (SARI) is of the order of approximately 7%. In the multivariate case the results are even poorer, with the variance in the MAE over the prediction set of the order of approximately 34% between the MDBJ_B model and the linear BJTF (with HDD_{18}) model. As in the case of System A it can be seen that pre-processing the training data reduces the number of iterations required to train the network. In addition, the selection of the training cessation point using the MMAE_val criterion yields the best result over the predication set compared with the MSE_tr criterion.

It is difficult to analyse the physical significance of the relationship that a neural network constructs between the past and future values of the time series and consequently, it is difficult to ascertain the reasons why the MLP networks did not perform well in this particular application. One possible reason is that System B is a linear as opposed to a nonlinear system and may thus be adequately represented by a linear model. Unfortunately, there is no clear cut strategy available for distinguishing linear time series from nonlinear time series. The common approach is to construct both linear and nonlinear models and then find sufficient evidence for the abandonment of the linear model, where this often involves the use of hypothesis testing (Gooijer and Kumar, 1992). However, the tests available for the comparison of the linear and nonlinear models are model specific. For example the tests presented in Gooijer and Kumar (1992), who provide an overview of existing methods on the subject, deal with tests specifically designed for traditional nonlinear models, such as Autoregressive Conditional Heteroscedastic (ARCH); Bilinear and Threshold Autoregressive (TAR) models.

Another possible reason for the failure of the MLP's in this application could be that the networks are getting trapped in local minimum during training. Since, the error surface is

multi-dimensional it is not possible to examine the shape of this surface in an attempt to verify if this is the case. It is known that many backpropagation error surfaces have extensive flat areas and troughs that have very little slope (Hecht-Neilsen, 1990). Therefore, in these areas it is necessary to move the weight value a substantial distance before a significant drop in error occurs. Consequently, the use of the backpropagation learning rule in this application may have required greater numerical precision and a greater number of training runs in order to make significant progress.

4.4 Application to Yearly Electricity Sales

A neural network analysis was performed on yearly electricity sales (GWh) data which was obtained from the same national power board dealt with in the System A weekly electricity demand application (Table 3.1). Aggregated and disaggregated data is available from 1965 to 1994 and 1971 to 1994 respectively. The disaggregated data is divided into three sectors: industrial; commercial; domestic. The aggregated data is referred to as the *total sales*. Figure 4.25 compares the four electricity sales time series.

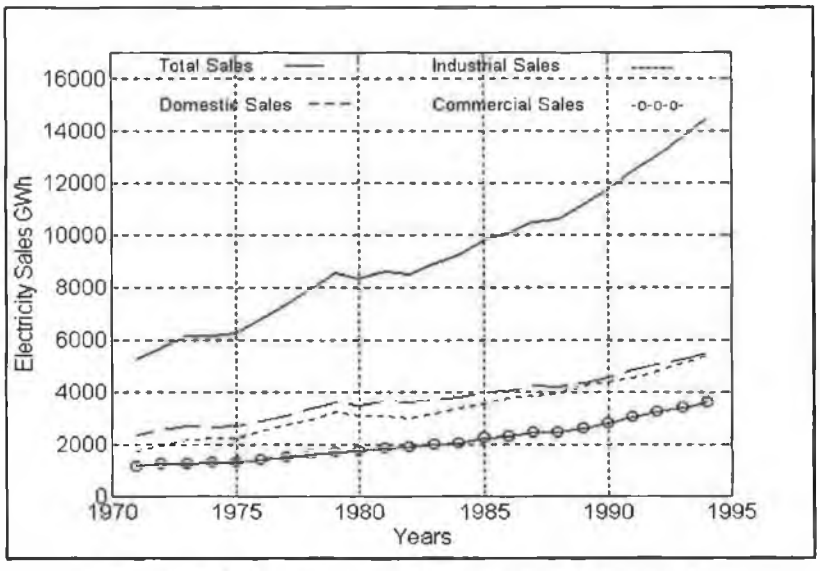

Figure 4.25 Aggregate and disaggregated time series

Each of the time series exhibit a rising trend due to the annual influences of economic and demographic variables. Data is available for these economic (in real terms) and demographic variables, where Table 4.29 gives a description of each variable.

Table 4.29 Economic and demographic variables

Name	Definition
AIW	Average industrial wage (£/week)
GDP	Gross domestic product (£m)
TAUP	Total average unit price (p/KWh)
IAUP	Industrial average unit price (p/KWh)
DAUP	Domestic average unit price (p/KWh)
CAUP	Commercial average unit price (p/KWh)
TNOC	Total number of customers
INOC	Industrial number of customers
DNOC	Domestic number of customers
INOC	Commercial number of customers

A multivariate neural networks analysis was performed on each of the total, industrial, domestic and commercial yearly electricity sales systems, where the application procedure outlined in Section 4.2.5 was followed. The aggregate and disaggregate cases are considered simultaneously.

4.4.1 Specification of MLP Networks

NN-Step 1: Input and output structure

Analysis developed in the weekly electricity demand application showed the advantage to be gained through the incorporation of information obtained through linear time series modelling when constructing the MLP input structure. The construction of the input structure for the yearly application involved the expansion and adaptation of the Box-Jenkins linear model (2.19). The yearly electricity sales time series are nonseasonal series with a rising trend and each have a linear model structure of the following form:

$$(1-a_1q^{-1}-\ldots\ldots-a_pq^{-p})z(k)=\varepsilon(k) \tag{4.22}$$

where p is the order of the autoregressive component and $z(k)=(1-q^{-1})^d y(k)$ is the differencing transformation required to convert the nonstationary electricity sales time series to stationary series. Expansion of (4.22) yields a generalised nonlinear model of the following form:

$$\boxed{z(k) = g\big(z(k-1),\ldots\ldots,z(k-p)\big) + \varepsilon(k)} \tag{4.23}$$

where $g(.)$ is a nonlinear function and $\varepsilon(k)$ is the forecast error. Alternatively a nonlinear model structure may be obtained from the full expansion of the following

$$(1-a_1q^{-1}-\ldots\ldots-a_pq^{-p})(1-q^{-1})^d\, y(k) = \varepsilon(k) \tag{4.24}$$

which yields

$$\boxed{y(k) = h\big(y(k-1),\ldots\ldots,y(k-p)\ldots\ldots y(k-p+d)\big) + \varepsilon(k)} \tag{4.25}$$

where $h(.)$ is a nonlinear function and $\varepsilon(k)$ is the forecast error.

The order of the nonlinear models given by (4.23) and (4.25) depend on the value of the parameters p and d. The SACF (2.8) was used to determine the appropriate order of differencing required to convert the time series to stationary series (see Section 2.4.1.2) and the SPACF (2.9) was used to determine the order of the autoregressive parameter p for each series (see Section 2.4.3.1). Table 4.30 gives the value of the parameters for the electricity sales time series, where *tot*, *ind*, *dom* and *com* represent the total, industrial, domestic and commercial electricity sales time series respectively.

Table 4.30 Value of p and d for electricity sales time series

Time Series	*p*	*d*
tot	2	2
ind	2	2
dom	2	2
com	1	1

Based on the values of p given in Table 4.28 the following nonlinear models may be obtained from equation (4.23):

$$ztot(k) = g_{tot}\big(ztot(k-1), ztot(k-2)\big) \tag{4.26}$$

$$zind(k) = g_{ind}\big(zind(k-1), zind(k-2)\big) \tag{4.27}$$

$$zdom(k) = g_{dom}\big(zdom(k-1), zdom(k-2)\big) \tag{4.28}$$

$$zcom(k) = g_{com}\big(zcom(k-1)\big) \tag{4.29}$$

where, $g_{tot}(.)$, $g_{ind}(.)$, $g_{dom}(.)$ and $g_{com}(.)$ are nonlinear functions and *ztot, zind, zdom* and *zcom* are the differenced *tot, ind, dom* and *com* time series respectively. Based on the values of p and d in Table 4.30 the following nonlinear models may be obtained from equation (4.25):

$$tot(k) = h_{tot}\left(tot(k-1), tot(k-2), tot(k-3), tot(k-4)\right) \tag{4.30}$$

$$ind(k) = h_{ind}\left(ind(k-1), ind(k-2), ind(k-3), ind(k-4)\right) \tag{4.31}$$

$$dom(k) = h_{dom}\left(dom(k-1), dom(k-2), dom(k-3), dom(k-4)\right) \tag{4.32}$$

$$com(k) = h_{com}\left(com(k-1), com(k-2)\right) \tag{4.33}$$

where, $h_{tot}(.)$, $h_{ind}(.)$, $h_{dom}(.)$ and $h_{com}(.)$ are nonlinear functions. The nonlinear models given by equation (4.26) to (4.33) were extended to include the economic and demographic exogenous variables given in Table 4.29. The entire set of available exogenous variables could be included in each model, however, a parsimonious model is desirable and some pre-processing analysis was thus performed on the exogenous variables. The SCCF(2.12) was used to ascertain if a relationship existed between a particular exogenous variable and the electricity sales variable of interest. The SCCF was calculated for each of the exogenous variable/electricity demand pairs. The SCCF analysis involved the conversion of each time series to a stationary one through differencing. Table 4.31 give the exogenous variables which were found to have a relationship with electricity sales for each of the sectors, the corresponding delay parameter b and nb the order of the input variable regressor $B(q)$. Table 4.32 give the degree of differencing applied to each time series to achieve stationarity.

Table 4.31 Relevant exogenous variables

Time Series	**Exogenous variable**	***b***	***nb***
Total	*AIW*	0	0
	GDP	0	0
	TAUP	0	1
Industrial sector	*AIW*	0	0
	GDP	0	0
	IAUP	1	0
Domestic sector	*AIW*	0	1
	GDP	0	0
	DAUP	0	0
Commercial sector	*GDP*	0	0
	TAUP	0	1
	CNOC	0	0

Table 4.32 Value of *d* for exogenous time series

Time Series	*d*
AIW	2
GDP	2
TAUP	2
IAUP	2
DAUP	2
CAUP	1
CNOC	2

Figure 4.26 to Figure 4.29 show the profile of the original electricity sales and the associated exogenous variable time series for each sector, where each time series is scaled to the same order of magnitude.

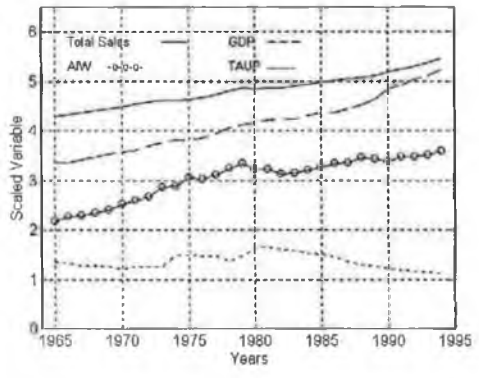

Figure 4.26 Total sales and exogenous variables

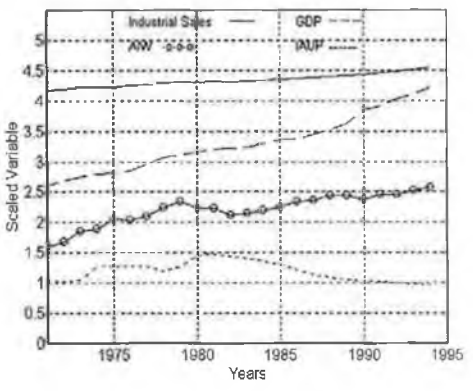

Figure 4.27 Industrial sales and exogenous variables

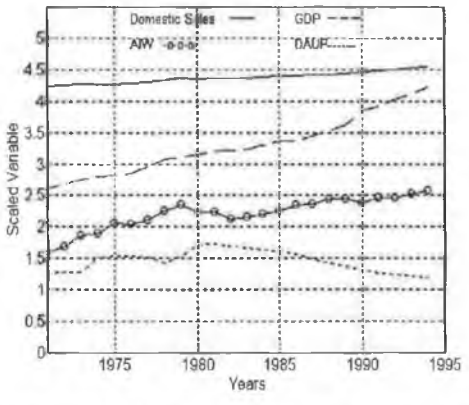

Figure 4.28 Domestic sales and exogenous variables

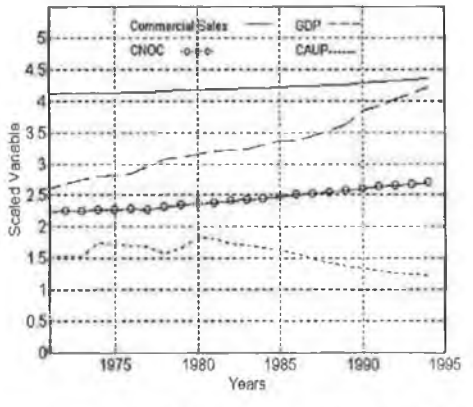

Figure 4.29 Commercial sales and exogenous variables

Based on the results given in Table 4.31 and Table 4.32 the nonlinear models described by equations (4.26) to (4.29) were extended to include the relevant exogenous variables.

$$ztot(k) = G_{tot}\left(wGDP(k), wTAUP(k), wTAUP(k-1), wAIW(k), ztot(k-1), ztot(k-2)\right) \quad (4.34)$$

$$zind(k) = G_{ind}\left(wGDP(k), wIAUP(k-1), wAIW(k), zind(k-1), zind(k-2)\right) \quad (4.35)$$

$$zdom(k) = G_{dom}\left(wGDP(k), wDAUP(k), wAIW(k), wAIW(k-1), zdom(k-1), zdom(k-2)\right) \quad (4.36)$$

$$zcom(k) = G_{com}\left(wGDP(k), wCAUP(k), wCAUP(k-1), wCNOC(k), zcom(k-1)\right) \quad (4.37)$$

where, $G_{tot}(.)$, $G_{ind}(.)$, $G_{dom}(.)$ and $G_{com}(.)$ are nonlinear functions and *wGDP, wAIW, wTAUP, wIUAP, wDAUP* and *wCNOC* are the *differenced GDP, AIW, TAUP, IUAP, DAUP* and *CNOC* time series respectively. Similarly for equations (4.30) to (4.33).

$$\begin{aligned} tot(k) = H_{tot}\big(&GDP(k), GDP(k-1), GDP(k-2), TAUP(k), TAUP(k-1), TAUP(k-2), \\ &TAUP(k-3), AIW(k)AIW(k-1), AIW(k-2), tot(k-1), tot(k-2), tot(k-3), tot(k-4)\big) \end{aligned} \quad (4.38)$$

$$\begin{aligned} ind(k) = H_{ind}\big(&GDP(k), GDP(k-1), GDP(k-2), IAUP(K-1), IAUP(k-2), IAUP(k-3), \\ &AIW(k), AIW(k-1), AIW(k-2), ind(k-1), ind(k-2), ind(k-3), ind(k-4)\big) \end{aligned} \quad (4.39)$$

$$\begin{aligned} dom(k) = H_{dom}\big(&GDP(k), GDP(k-1), GDP(k-2), DAUP(k), DAUP(k-1), DAUP(k-2), \\ &AIW(k), AIW(k-1), AIW(k-2), AIW(k-3), dom(k-1), dom(k-2), dom(k-3), dom(k-4)\big) \end{aligned} \quad (4.40)$$

$$\begin{aligned} com(k) = H_{com}\big(&GDP(k), GDP(k-1), GDP(k-2), CAUP(k-1), CAUP(k-2), \\ &CNOC(k), CNOC(k-1), CNOC(k-2), com(k-1), com(k-2)\big) \end{aligned} \quad (4.41)$$

where, $H_{tot}(.)$, $H_{ind}(.)$, $H_{dom}(.)$ and $H_{com}(.)$ are nonlinear functions. The names of the aggregate and disaggregate models with MLP's input structures based on the nonlinear models given by equations (4.34) to (4.41) are given in Table 4.33.

Table 4.33 Aggregate and disaggregate models

Model	Equation
MDBJ_tot	(4.34)
MDBJ_ind	(4.35)
MDBJ_dom	(4.36)
MDBJ_com	(4.37)
MBJ_tot	(4.38)
MBJ_ind	(4.39)
MBJ_dom	(4.40)
MBJ_com	(4.41)

As in the case of the weekly electricity demand application (Section 4.3.1.1) if the MLP is used to forecast the differenced form of the time series then it is necessary to apply the inverse differencing transformation to obtain the forecast of the original electricity sales time series (see Figure 4.8 for an example of this).

NN-Step 2: Input and output training data normalisation

For each case the input data was scaled using the external normalisation procedure, where each time series was divided by the NF value prior to the construction of the input training vectors. Table 4.34 describes the NF values used for each time series in the MBJ and MDBJ networks.

Table 4.34 NF values for time series in MBJ and MDBJ models

Variable	NF	Variable	NF
tot	1×10^5	*ztot*	1×10^3
ind	1×10^5	*zind*	1×10^3
dom	1×10^5	*zdom*	1×10^3
com	1×10^5	*zcom*	1×10^4
AIW	1×10^3	*wAIW*	1×10^2
GDP	1×10^6	*wGDP*	1×10^1
TAUP	1×10^2	*wTAUP*	1×10^0
IAUP	1×10^2	*wIAUP*	1×10^0
DAUP	1×10^2	*wDAUP*	1×10^0
CAUP	1×10^2	*wCAUP*	1×10^0
INOC	1×10^8	*wINOC*	1×10^4

The output neuron has a linear activation function and thus has an unrestricted output range but to aid the learning process the output data was normalised to the same order of magnitude as the input data (divide by 1×10^5 for MBJ_tot; MBJ_ind; MBJ_dom; MBJ_com and divide by 1×10^3 for MDBJ_tot; MDBJ_ind; MDBJ_dom; MDBJ_com).

NN-Step 3: Construct input/output training, validation and prediction data sets

A typical forecast horizon for yearly electricity sales for this system is five years ahead (Section 1.1) and therefore the prediction data set is made up of five years of data. Due to data limitations only two years data was selected for the validation set the remainder of the data was reserved for the training set. The aggregate electricity sales data is available from 1965 to 1994, whereas the disaggregate electricity sales data is only available from 1972 to 1994 and therefore there was less training data available for the disaggregate case. The data segmentation for each case is as follows:

Aggregate case:

- training data set - yearly time series data from 1965 to 1987.
- validation data set - yearly data time series from 1988 to 1989.
- prediction data set - yearly data time series from 1990 to 1994.

Disaggregate case:

- training data set - yearly time series data from 1972 to 1987.
- validation data set - yearly data time series from 1988 to 1989.
- prediction data set - yearly data time series from 1990 to 1994.

NN-Step 4: Select Structure and Hidden Layer Architecture

Based on the motivation given in Section 4.2.2 an MLP with two hidden layers was chosen as the network structure. An experimentation analysis was carried out to select the optimal number of neurons in the hidden layers of the MLP. Networks with different number of neurons in the hidden layers were trained and their performance over the training, validation and prediction sets were compared. The main objective of the analysis was to ascertain the type of network structures that are suitable for forecasting yearly electricity sales. Moreover, an attempt was made to establish a suitable criterion to employ when selecting the best network structure out of a set of trained network structures. Given a set of trained network structures there are two choices available for comparing their performance. These are

1. SMSE_tr (optimal single-step-ahead MSE over the training set)
2. MMAE_val (optimal multi-step-ahead MAE over the validation set)

Since, a forecast and thus a separate neural network model was required for the aggregate and disaggregate electricity sales, some preliminary analysis was conducted based on the aggregate (total) sales to gain some insight into problem. To achieve the objectives of the analysis ten different MLP networks with different numbers of neurons in the hidden layers were trained and the results analysed. Initial results showed that the MBJ_tot (4.38) model outperformed the MDBJ_tot model (4.34) and therefore the ten MLP structures were trained using the MBJ_tot model. Log-sigmoid activation functions were used in all the neurons in the hidden layers and a single linear neuron was used at the output. The networks were trained using the LMS gradient technique with backpropagation (Section 4.2.3), with an adaptive learning rate (initial value of $1x10^{-3}$) and a momentum constant equal to 0.95. Each network was re-trained 30 times using different random weight initialisations. Since there are 10 different network structures and 30 training runs for each structure, the set of results for

every training run are not given. Instead, for each MLP structure the results are shown for the training run which yielded the minimum MMAE_val. Similar results to those presented in the weekly application are also presented here, that is: (Result Set 1):

1. MMAE_val
2. Network training cessation point.
3. Multi-step-ahead MAE over prediction set.

For the purposes of determining a criteria to select the best network structure out the 10 network structures, each structure is ranked according to the MMAE_val. To aid the analysis the structure which yields the best multi-step-ahead MAE result over the prediction set is highlighted (shaded).

For comparative purposes the results were the network training cessation point was selected through the examination of the single-step-ahead SSE performance over the training set are also given (Result Set 2), where these are:

1. SMSE_tr
2. Network training cessation point.
3. Multi-step-ahead MAE over prediction set.

Similarly, these results are ranked according to the SMSE_tr and the best multi-step-ahead MAE over the prediction set result is highlighted. Table 4.35 and Table 4.36 give the above results for the MBJ_tot model.

Table 4.35 Results for MBJ_tot model - Result Set 1

		Minimum MMAE_val Out of 30 Training Runs		
		Validation Set		**Prediction Set**
Structure	**Rank**	**MMAE_val**	**Training Cessation @ Epoch**	**Multi-Step-Ahead MAE**
1-3-1	10	15.08	30874	205.27
1-9-1	6	1.85	9339	359.14
2-6-1	8	10.85	4652	153.98
2-9-1	1	0.17	8263	375.65
3-5-1	5	1.62	37665	315.69
3-9-1	7	5.04	4636	179.05
3-15-1	3	1.27	19149	369.99
4-10-1	4	1.28	36138	310.37
4-20-1	2	0.96	7687	373.15
10-30-1	9	11.22	32708	309.81

Table 4.36 Results for MBJ_tot model - Result Set 2

		Minimum MMAE_val Out of 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	1	2.204×10^{-6}	40000	203.79
1-9-1	9	4.960×10^{-6}	40000	418.88
2-6-1	4	3.028×10^{-6}	40000	156.77
2-9-1	5	3.260×10^{-6}	40000	344.65
3-5-1	8	3.872×10^{-6}	40000	317.40
3-9-1	3	2.364×10^{-6}	40000	178.70
3-15-1	6	3.468×10^{-6}	40000	464.61
4-10-1	7	3.728×10^{-6}	40000	365.40
4-20-1	2	2.232×10^{-6}	40000	401.14
10-30-1	10	5.480×10^{-6}	40000	402.13

Examination of the results in Table 4.35 show that the MLP structure that yields the best MMAE_val is 2-9-1. This however is inconsistent with the corresponding result over the prediction set, where the structure that yields the best multi-step-ahead forecast is in fact 2-6-1. In Table 4.36 the best multi-step-ahead result over the prediction set is also obtained using a 2-6-1 structure but again this was inconsistent with the result for the SMSE_tr, where if this criterion was used a 1-3-1 structure would be selected. A possible reason for the variance in the case of the MMAE_val criterion is that only two-year-ahead forecast is carried out over the validation set, due to limitations in the total number of data points available for the MBJ_A model (20 points), whereas a five-year-ahead forecast is performed over the prediction set. Consequently, the weight set which produces a good two-year ahead forecast on the validation set may not always produce a comparable five-year-ahead forecast over the prediction set. However, the extension of the validation to five years data would reduce the number of data points in the training set to an unacceptable number of points (15 points) which could result in learning difficulties. In the case of the SMSE_tr criterion a likely reason for the inconsistency is the problem associated with the use of a single-step-ahead criterion to select the training cessation point for a network that is required to perform a multi-step-ahead prediction.

Closer examination of the overall set of results for all training runs shows that some structures are very sensitive to initial conditions. In some cases the MMAE_val and SMSE_tr vary greatly across training runs. Notable example are given in Table 4.37 which shows the minimum MMAE_val achieved out of the 30 training runs and the maximum MMAE_val achieved out of the 30 training runs, with similar results given for the SMSE_tr.

Table 4.37 Network structures that are particularly sensitive to initial conditions

	MMAE_val		SMSE_tr	
Structures	**Minimum**	**Maximum**	**Minimum**	**Maximum**
1-9-1	1.85	1287.01	0.24×10^{-5}	1.36×10^{-5}
2-9-1	0.17	1105.05	0.34×10^{-5}	1.32×10^{-5}
3-15-1	1.27	1194.14	0.35×10^{-5}	1.36×10^{-5}

In contrast, network structures whose performance is relatively stable across training runs are given in Table 4.38.

Table 4.38 Network structures that were not sensitive to initial conditions

	MMAE_val		SMSE_tr	
Structures	**Minimum**	**Maximum**	**Minimum**	**Maximum**
1-3-1	15.08	632.51	0.22×10^{-5}	0.84×10^{-5}
2-6-1	10.85	666.04	0.30×10^{-5}	0.81×10^{-5}
3-5-1	1.62	672.04	0.39×10^{-5}	0.86×10^{-5}
3-9-1	5.04	692.44	0.24×10^{-5}	0.95×10^{-5}

To account for the significant sensitivity of some network structures to initial conditions and in an effort to assess the overall capability of a particular MLP structure the average value over the 30 training runs was calculated for the MMAE_val and SMSE_tr. The results are given in the same format as Result Set 1 and Result Set 2 and are referred to as Result Set 3 and Result Set 4. Table 4.39 and Table 4.40 show the results.

Table 4.39 Comparison of MLP structures for MBJ_tot model - Result Set 3

		Average Value Over 30 Training Runs			
		Validation Set		Prediction Set	
Structure	**Rank**	**MMAE_val**	**@ Epoch**	**Rank**	**Multi-Step-Ahead MAE**
1-3-1	3	241.28	35197	3	205.27
1-9-1	8	349.32	29725		359.14
2-6-1	1	156.85	20907	1	153.98
2-9-1	10	358.31	32863		375.65
3-5-1	2	214.55	28419	4	315.69
3-9-1	4	245.30	30684	2	179.05
3-15-1	6	317.34	29908		369.99
4-10-1	7	346.64	35498	5	310.37
4-20-1	5	288.92	28047		373.15
10-30-1	9	352.24	33523		309.81

Table 4.40 Comparison of MLP structures for MBJ_tot model - Result Set 4

		Average Value Over 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	@ Epoch	Multi-Step-Ahead MAE
1-3-1	3	4.8244×10^{-6}	40000	203.79
1-9-1	8	6.0800×10^{-6}	40000	418.88
2-6-1	1	4.1392×10^{-6}	40000	156.77
2-9-1	10	6.1988×10^{-6}	40000	344.65
3-5-1	2	4.3576×10^{-6}	40000	317.40
3-9-1	4	4.8800×10^{-6}	40000	178.70
3-15-1	6	5.9124×10^{-6}	40000	464.61
4-10-1	7	5.9528×10^{-6}	40000	365.40
4-20-1	5	5.6880×10^{-6}	40000	401.14
10-30-1	9	6.0496×10^{-6}	40000	402.13

The interesting point to note here is that in contrast to the previous results given by Result Set 1 in Table 4.35 and Result Set 2 in Table 4.36 the use of criteria based on average values yield a consistent structure selection over training, validation and prediction set, where this selection is 2-6-1. In addition, the ranking of the network structures based on the MMAE_val MMSE_tr are consistent. However, with the exception of the top ranking structure (2-6-1) the ranking is not consistent with the prediction set, where the most likely reason for this is the large variation in performance across training runs due to the sensitivity to initial conditions. Although the order of ranking is not consistent between the validation and prediction set and also between the training and prediction set the top four ranking structures for all cases are 2-6-1; 3-5-1; 1-3-1 and 3-9-1. Note that, these were the structures that were found to be not as sensitive to initial conditions as the other structures in the set.

Based on the preliminary analysis the following guidelines may be adopted for the remainder of the experimentation analysis:

- to alleviate the problem of sensitivity to initial conditions it was necessary to carry out at least 30 training runs.
- to assess the overall capability of a network structure a criterion based on the average values of the performances over the 30 training runs should be used to select the most suitable network structure out of a set of trained network structures.
- the most suitable network structures for yearly electricity sales forecasting are the 1-3-1; 2-6-1; 3-5-1 and 3-9-1 structures.

One of the objective of the overall experimentation analysis was to determine a suitable criterion to use when selecting the best network structure out of a set of trained network structures. The preliminary analysis suggests that either the average values over 30 training runs of the MMAE_val or the average values over 30 training runs of the SMSE_tr could be used as a criterion. However, in order to make some broader conclusion concerning this objective it was necessary to examine the results for the disaggregate case. However, before proceeding any further the four choices of criterion available to select the network structure are summarised

Criterion 1: Minimum value out of 30 training runs of the MMAE_val - Result Set 1.

Criterion 2: Minimum value out of 30 training runs of the SMSE_tr - Result Set 2.

Criterion 3: Minimum of average over 30 training runs of the MMAE_val - Result Set 3.

Criterion 4: Minimum of average over 30 training runs of the SMSE_tr - Result Set 4.

The results for the MDBJ_tot model are examined before proceeding to the disaggregate cases. Using the guidelines outlined above the MDBJ_tot model was trained with the results, Result Set 1 to Result Set 4 given in Tables 4.41 to Table 4.44 respectively.

Table 4.41 Results for MDBJ_tot model - Result Set 1 (Criterion 1)

		Minimum MMAE_val Out of 30 Training Runs		
		Validation Set		**Prediction Set**
Structure	**Rank**	**MMAE_val**	**Training Cessation @ Epoch**	**Multi-Step-Ahead MAE**
1-3-1	3	199.21	10551	272.41
2-6-1	1	186.02	11471	404.99
3-5-1	2	195.87	6769	346.20
3-9-1	4	202.22	8545	260.11

Table 4.42 Results for MDBJ_tot model - Result Set 2 (Criterion 2)

		Minimum MMAE_val Out of 30 Training Runs		
		Training Set		**Prediction Set**
Structure	**Rank**	**SMSE_tr**	**Training Cessation @ Epoch**	**Multi-Step-Ahead MAE**
1-3-1	4	7.256×10^{-6}	40000	2743.19
2-6-1	2	7.120×10^{-6}	40000	2507.19
3-5-1	3	7.240×10^{-6}	40000	2114.30
3-9-1	1	6.840×10^{-6}	40000	2054.50

Table 4.43 Comparison of MLP structures for MDBJ_tot model - Result Set 3 (Criterion 3)

		Average Value Over 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	@ Epoch	Multi-Step-Ahead MAE
1-3-1	1	300.93	16169	272.41
2-6-1	2	317.44	18899	404.99
3-5-1	3	345.20	10359	346.20
3-9-1	4	349.98	9455	260.11

Table 4.44 Comparison of MLP structures for MDBJ_tot model - Result Set 4 (Criterion 4)

		Average Value Over 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	@ Epoch	Multi-Step-Ahead MAE
1-3-1	3	6.8864×10^{-6}	40000	2743.19
2-6-1	2	6.6396×10^{-6}	40000	2507.19
3-5-1	4	6.9060×10^{-6}	40000	2114.30
3-9-1	1	6.3120×10^{-6}	40000	2054.50

The only criteria that select the optimal network structure for the prediction set, that is 3-9-1, are Criterion 2 and Criterion 4. Moreover, the results produced by the MDBJ_tot networks are approximately 45% worse, over the prediction set, and 40% worse, over the validation set, than the results produced by the MBJ_tot model. Based on the above MDBJ_tot results a final guideline that was ascertained from the total sales analysis is that an MBJ model will outperform an equivalent MDBJ model.

Finally, based on the above preliminary results the final MLP network selected for 1990 to 1994 prediction of total electricity sales is the MBJ_tot model using a 2-6-1 structure. The full set of results for the prediction are given in the Section 4.4.2.

Disaggregate Electricity Sales Analysis:

An experimentation into the most suitable network structure for the disaggregate electricity sales was carried based on the guidelines determined in the preliminary analysis. It was also hoped that some conclusion may be made concerning the criterion to use when selecting the optimal network structure. For the disaggregate networks the fixed number of training epochs was reduced to 20000 as there was no significant improvement in the single-step-ahead MSE over the training set or the multi-step-ahead MAE over the validation set beyond this point. The results for the MBJ_ind model are presented first and are given in Table 4.45 to Table 4.48.

Table 4.45 Results for MBJ_ind model - Result Set 1 (Criterion 1)

		Minimum MMAE_val Out of 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	1	17.41	4189	87.62
2-6-1	3	25.43	5643	96.09
3-5-1	2	21.48	4222	136.96
3-9-1	4	36.96	7039	41.15

Table 4.46 Results for MBJ_ind model - Result Set 2 (Criterion 2)

		Minimum MMAE_val Out of 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	2	1.2352×10^{-4}	20000	201.92
2-6-1	2	1.1986×10^{-4}	20000	212.89
3-5-1	3	1.2495×10^{-4}	20000	224.00
3-9-1	1	1.1117×10^{-4}	20000	107.98

Table 4.47 Results for MBJ_ind model - Result Set 3 (Criterion 3)

		Average Value Over 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	@ Epoch	Multi-Step-Ahead MAE
1-3-1	3	69.49	7729	87.62
2-6-1	1	64.80	8429	96.09
3-5-1	4	71.81	7045	136.96
3-9-1	2	69.15	8941	41.15

Table 4.48 Results for MBJ_ind model - Result Set 4 (Criterion 4)

		Average Value Over 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	@ Epoch	Multi-Step-Ahead MAE
1-3-1	3	1.2765×10^{-4}	20000	201.92
2-6-1	1	1.0245×10^{-4}	20000	212.89
3-5-1	3	1.2765×10^{-4}	20000	224.00
3-9-1	2	1.2353×10^{-4}	20000	107.98

Before examining the performance of the different network structures in the industrial case a comment is made regarding the fixed number of epochs for which the networks were trained. Although Table 4.45 shows that the training cessation points selected using MMAE_val criterion is between approximately 4000 and 7000 examination of the results for all 30

training runs show that the training cessation point varies between 4000 and 18000 epochs. Consequently, it was considered necessary to train the networks for a total of 20000 epochs.

For the MBJ_ind model Criterion 2 is the only criterion that is consistent with the prediction set. The use of Criterion 3, which gave a consistent selection in the MBJ_tot case, results in a 27 % deterioration in the multi-step-ahead MAE over the prediction set and similarly using Criterion 4 results in a 78 % deterioration.

The results for the MDBJ_dom model are given in Table 4.49 to Table 4.52.

Table 4.49 Results for MBJ_dom model - Result Set 1 (Criterion 1)

		Minimum MMAE_val Out of 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	3	12.64	2282	177.89
2-6-1	4	15.71	2401	76.40
3-5-1	1	2.33	2425	52.53
3-9-1	2	6.89	3178	49.46

Table 4.50 Results for MBJ_dom model - Result Set 2 (Criterion 2)

		Minimum MMAE_val Out of 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	1	0.5882×10^{-4}	20000	173.51
2-6-1	1	0.5882×10^{-4}	20000	167.55
3-5-1	2	0.7058×10^{-4}	20000	98.31
3-9-1	1	0.5882×10^{-4}	20000	122.73

Table 4.51 Results for MBJ_dom model - Result Set 3 (Criterion 3)

		Average Value Over 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	@ Epoch	Multi-Step-Ahead MAE
1-3-1	4	50.20	15280	177.89
2-6-1	1	40.81	13377	76.40
3-5-1	2	43.26	14030	52.53
3-9-1	3	50.19	15504	49.46

Table 4.52 Results for MBJ_dom model - Result Set 4 (Criterion 4)

		Average Value Over 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	@ Epoch	Multi-Step-Ahead MAE
1-3-1	1	6.7059×10^{-4}	20000	173.51
2-6-1	1	6.7059×10^{-4}	20000	167.55
3-5-1	2	6.8824×10^{-5}	20000	98.31
3-9-1	3	7.1765×10^{-5}	20000	122.73

For the domestic case there are no consistent results between the validation and prediction set or the training and prediction set. Moreover, using the MMAE_val criterion for the training cessation point the network which performs the best over the prediction set is the 3-9-1 structure, whereas in contrast for the MMSE_tr criterion it is a 3-5-1 structure. The results for the MDBJ_com model are given in Table 4.53 to Table 4.56.

Table 4.53 Results for MBJ_com model - Result Set 1 (Criterion 1)

		Minimum MMAE_val Out of 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	2	0.28	3296	64.11
2-6-1	1	0.18	4488	85.92
3-5-1	1	0.18	5042	74.52
3-9-1	3	0.83	6295	88.10

Table 4.54 Results for MBJ_com model - Result Set 2 (Criterion 2)

		Minimum MMAE_val Out of 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	Training Cessation @ Epoch	Multi-Step-Ahead MAE
1-3-1	3	0.8870×10^{-5}	20000	84.95
2-6-1	1	0.7647×10^{-5}	20000	122.43
3-5-1	2	0.8129×10^{-5}	20000	104.34
3-9-1	4	0.8562×10^{-5}	20000	119.85

Table 4.55 Results for MBJ_com model - Result Set 3 (Criterion 3)

		Average Value Over 30 Training Runs		
		Validation Set		Prediction Set
Structure	Rank	MMAE_val	@ Epoch	Multi-Step-Ahead MAE
1-3-1	3	16.72	13796	64.11
2-6-1	4	17.89	11893	85.92
3-5-1	1	12.45	12630	74.52
3-9-1	2	15.25	13713	88.10

Table 4.56 Results for MBJ_com model - Result Set 4 (Criterion 4)

		Average Value Over 30 Training Runs		
		Training Set		Prediction Set
Structure	Rank	SMSE_tr	@ Epoch	Multi-Step-Ahead MAE
1-3-1	1	0.9688×10^{-5}	20000	84.95
2-6-1	3	1.0188×10^{-5}	20000	122.43
3-5-1	2	1.0125×10^{-5}	20000	104.34
3-9-1	4	1.0813×10^{-5}	20000	119.85

In the case of commercial electricity sales Criterion 4 is the only criterion that produces a consistent result with the prediction set.

Due to the overall inconsistency of the disaggregate results an attempt was made to compare the four possible criterion available over all sectors, total, industrial, domestic and commercial, where the selections made based on each criterion are compared to the actual networks structures which yielded the best multi-step-ahead MAE over the prediction set. The results are split into two groups. The first group gives the structures selected using Criterion 1 and Criterion 2, given in Table 4.57 and the second group gives the structures selected based on Criterion 3 and Criterion 4, given in Table 4.58. Included in each table are the structures that gave the two best (Rank 1 and Rank 2) multi-step-ahead MAE over the prediction set results.

Table 4.57 Comparison of structures selected using Criterion 1 and Criterion 2

Model	Criterion 1	Criterion 2	Multi-Step -Ahead MAE (Prediction Set) Rank 1	Multi-Step -Ahead MAE (Prediction Set) Rank 2
MBJ_tot	1-3-1	2-9-1	2-6-1	3-9-1
MBJ_ind	1-3-1	3-9-1	3-9-1	2-6-1
MBJ_dom	3-5-1	2-6-1	3-9-1	3-5-1
MBJ_com	2-6-1	2-6-1	1-3-1	3-5-1

Table 4.58 Comparison of structures selected using Criterion 3 and Criterion 4

Model	Criterion 3	Criterion 4	Multi-Step -Ahead MAE (Prediction Set) Rank 1	Multi-Step -Ahead MAE (Prediction Set) Rank 2
MBJ_tot	2-6-1	2-6-1	2-6-1	3-9-1
MBJ_ind	2-6-1	2-6-1	3-9-1	2-6-1
MBJ_dom	2-6-1	2-6-1	3-9-1	3-5-1
MBJ_com	3-5-1	1-3-1	1-3-1	3-5-1

Comparison of the results across both tables indicates that Criterion 3 and Criterion 4 appear to produce the most consistent selections over the prediction set with the exception of the

domestic case. In addition, the results suggest that overall a 2-6-1 structure could be a suitable structure for all sectors. In order to compare Criterion 3 and Criterion 4 Table 4.59 gives the multi-step-ahead MAE results over the prediction set obtained using the network structures selected using each of the criterion for each of the models.

Table 4.59 Results over prediction set for selections based on Criterion 3 and Criterion 4

	Criterion 3		Criterion 4	
Model	**Multi-Step-Ahead MAE (Prediction Set)**		**Multi-Step-Ahead MAE (Prediction Set)**	
MBJ_tot	2-6-1	153.98	2-6-1	156.77
MBJ_ind	2-6-1	96.09	2-6-1	212.89
MBJ_dom	2-6-1	76.40	2-6-1	167.55
MBJ_com	3-5-1	74.52	1-3-1	84.95

For the yearly application based on the above results Criterion 3 is the most suitable criterion to use when selecting a network structure out of a set of trained networks structures. The networks structures given in Table 4.59 are the MLP structures that were used to obtain the five-year-ahead prediction of aggregate and disaggregate yearly electricity sales.

4.4.2 Forecasting Results

For comparative purposes BJTF models (2.23) were built for each of the total, industrial, domestic and commercial time series, where the methodology described in Section 2.4.31. of Chapter 2 was employed. A description of the model structure for each time series is given in Table 4.60. The structure of the transfer function (TF), that is the delay parameter b and the order of the $B(q)$ and $F(q)$ polynomials associated with each exogenous variable is given and also the order of the autoregressive polynomial p in the noise model $N(k)$.

Table 4.60 Description of Box-Jenkins transfer function models

	TF				$N(k)$
Time Series	**Variable**	***b***	***nb***	***nf***	***p***
tot	*AIW*	0	0	2	1
	GDP	0	0	0	
	TAUP	0	1	0	
ind	*AIW*	0	0	0	8
	GDP	0	0	0	
	IAUP	1	0	0	
dom	*AIW*	0	1	2	1
	GDP	0	0	0	
	DAUP	0	0	0	
com	*GDP*	0	0	0	4
	CAUP	0	1	0	
	CNOC	0	0	0	

The forecasts produced by the MLP's and linear models are compared graphically in Figure 4.30 to Figure 4.34 with the full set of forecasting accuracy results given in Table 4.61, where Σ_{NN}(ind+dom+com) is the sum of the industrial, domestic and commercial forecasts produced by the neural networks and Σ_{BJTF}(ind+dom+com) is the equivalent result produced by the BJTF models.

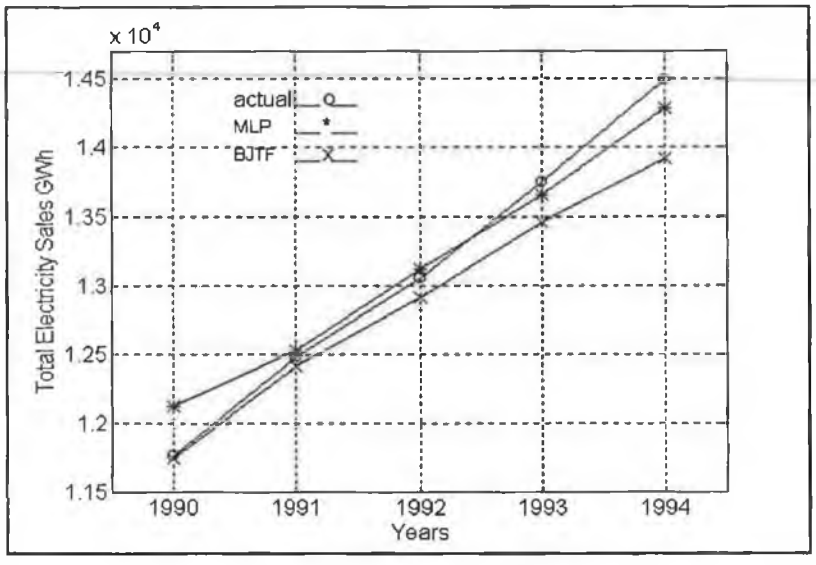

Figure 4.30 Actual vs predicted for total electricity sales

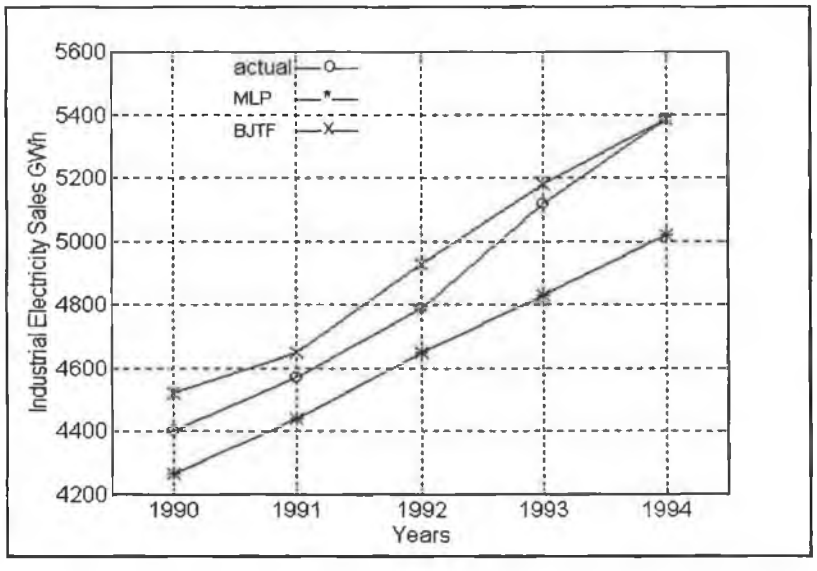

Figure 4.31 Actual vs predicted for industrial electricity sales

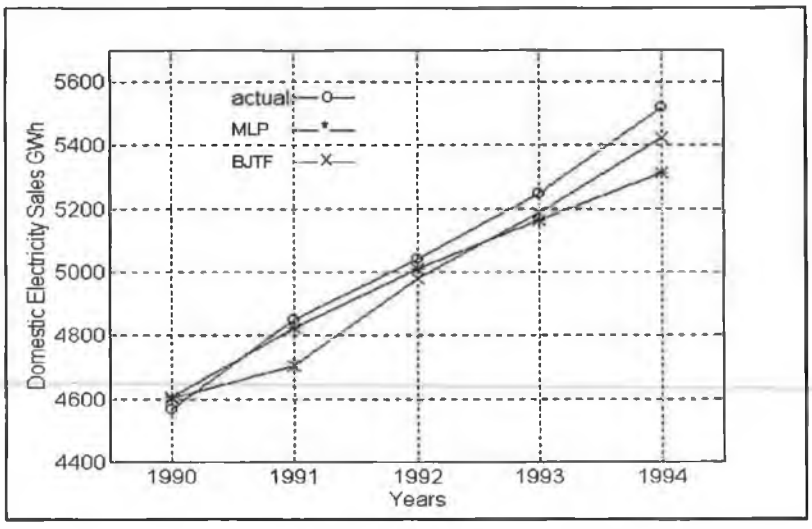

Figure 4.32 Actual vs predicted for domestic electricity sales

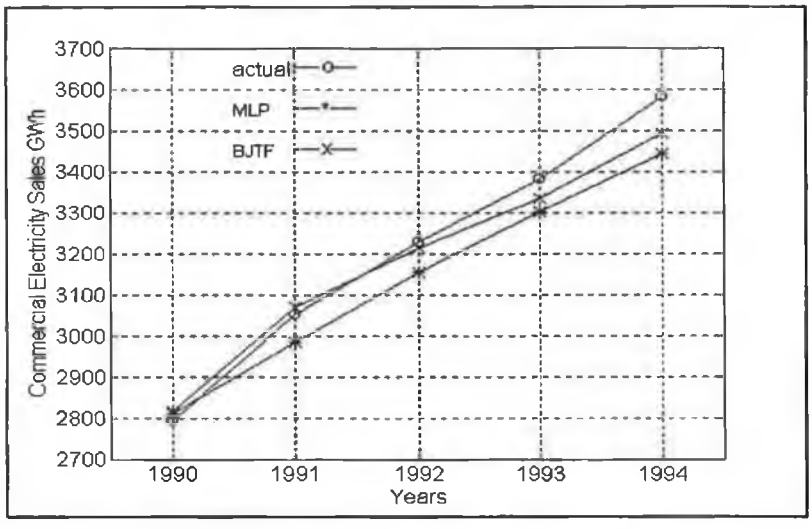

Figure 4.33 Actual vs predicted for commercial electricity sales

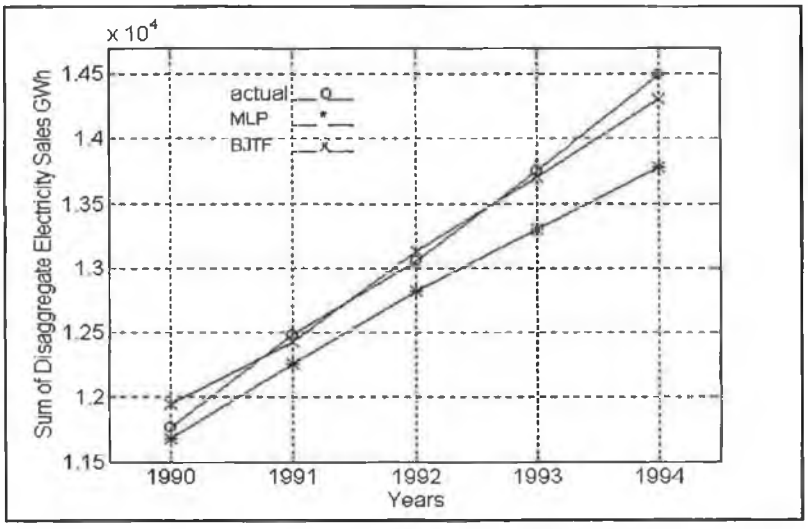

Figure 4.34 Actual vs predicted for sum of disaggregate sales

Table 4.61 Forecasting accuracy results for yearly electricity sales

Model	MAE	MSE	MAPE	MPE
MBJ_tot	153.98	3.653×10^4	1.21	-0.38
MBJ_ind	96.09	1.323×10^4	1.92	1.91
MBJ_dom	76.40	1.034×10^4	1.48	-1.16
MBJ_com	74.52	0.706×10^4	2.24	2.00
Σ_{NN}(ind+dom+com)	342.32	1.644×10^5	2.60	-2.60
BJTF_tot	219.82	8.769×10^4	1.58	1.58
BJTF_ind	81.52	0.917×10^4	1.79	1.79
BJTF_dom	78.39	0.769×10^4	1.59	-1.29
BJTF_com	38.05	0.219×10^4	1.16	-0.59
Σ_{BJTF}(ind+dom+com)	103.80	1.452×10^4	0.79	-0.01

Examination of the total sales graph (Figure 4.30) shows that, in spite of accurate forecasts for the first two years, the BJTF forecast deteriorates considerably for the remaining three years of the forecast horizon. Therefore, overall the neural networks outperforms the BJTF model. This can also be seen in the forecasting accuracy results, with the MAE of the neural networks approximately 40% better than that of the BJTF, the MSE approximately 60% better, with similar improvements in the case of the MAPE and MPE.

In contrast, comparing the BJTF and neural network forecasts in the case of the *sum* of the disaggregate forecasts the BJTF forecast (Σ_{BJTF}(ind+dom+com)) significantly outperforms the neural network forecast (Σ_{NN}(ind+dom+com)). Examination of the profile of the neural network results show although the first step of the forecast is reasonable there is significant deterioration as the forecast progresses through time. In contrast, the sum of the BJTF disaggregate results produce a very accurate result over the entire forecast horizon, where the MAPE(0.79) and MPE (-0.01) forecasting measures also provide an indication of this. A significant contribution to the poor performance in the neural network result for the sum of the disaggregate case is the highly inaccurate commercial electricity sales forecast, where examination of this profile in Figure 4.33 shows that the forecast deteriorates rapidly as it progresses through time. The ability of the linear model to represent the commercial electricity sales time series to a reasonable degree of accuracy suggests that this system is linear and thus does not require a nonlinear function for its representation. It is suggested that since this sector is very diverse further disaggregation may increase the accuracy of the linear result, however, availability of data may pose a problem in such an approach.

Overall, the domestic electricity sales forecast produced by the neural networks and the BJTF are comparable. However, it can seen that in the case of industrial sales neither model seems capable of representing the time series thus suggesting that this system is undergoing a

structural change, where it may only be possible to account for this change as more time series data becomes available.

Therefore, the results indicate that the neural networks are an appropriate modelling tool for the total electricity sales system but does not seem capable of representing the disaggregate sectors. In contrast, in the linear case disaggregation results in the production of a more accurate forecast for the aggregate sector.

4.5 Conclusion

A neural network analysis of the production of medium and long term forecasts of electricity demand was performed. The analysis has shown that the neural networks are highly case dependent, where this was demonstrated in both the weekly and yearly applications. Of particular relevance is the fact that the neural works were not capable of modelling a system that had been disaggregated, whereas they had been able to model the overall aggregate system (total sales). In contrast in the linear modelling case the disaggregation resulted in a more accurate forecast of the overall aggregate system. Therefore, it appears that the neural networks have the ability to decipher the complex dynamics of the aggregate system but a linear model was sufficient to represent this same system when it has been subdivided into its principal sectors.

It has been shown in the applications where neural networks are suitable for representing the electricity demand system that there is a reasonable improvement over the forecasting accuracy results produced by an equivalent linear analysis. For example, in the weekly application the MAPE was improved from 2.47 in the linear case (using a BJTF model with HDD_{18}) to 2.16 in the neural network case (using MDBJ_A) and likewise the MPE was improved from -1.55 to 0.39. Similarly, in the yearly case these measures were improved from 1.58 to 1.12 in the MAPE and from 1.58 to -0.38 in the MPE (total sales). However, the neural network analysis conducted in this chapter demonstrated that the best results were obtained when the analysis was performed in conjunction with a linear analysis. Given that this is the case the resulting overall development effort required to obtain a neural network to represent the electricity demand time series is considerably greater than the development effort required to obtain an equivalent model using a linear analysis.

A substantial bulk of the development effort required in the neural network analysis is spent on the selection of an appropriate configuration for the network. Unfortunately, the lack of specific rules for the design of neural networks can result in the use of a large amount of

computation time due to the requirement of experimentation analysis and also the necessity to alleviate the problem of sensitivity of network training to initial conditions. However, for the applications dealt with in this chapter forecasting is essentially an off-line activity and thus the computation time is not a significant issue.

The analysis carried out in the chapter established that the training times required to train a network can be reduced considerably if the time series training data has been pre-processed using an appropriate transformation. However, depending on the characteristics of the system the transformations can have the effect of improving or deteriorating the forecasts produced by the network in comparison with the forecast produced by a network using original time series training data. The use of pre-processing has been seen to be particularly effective in the case where the time series possesses a seasonal characteristic. In contrast, the pre-processing of the training data based on a time series with a trend component only resulted in a substantial reduction in the forecasting ability of the network. This suggests in this case that either alternative transformation to those presented in the chapter should be investigated, such as a Box-Cox transformation (Janacek and Swift, 1993), or that the original data should be used in cases where the series solely exhibits a trend characteristic.

The work conducted in the chapter also attempted to address the issue of the criteria to use in an experimentation analysis into the optimal configuration of the hidden layers of an MLP. Unfortunately, since the neural networks were not particularly suited to the representation of the disaggregate sectors it was difficult to establish conclusive results. However, in the case where the networks were suitable for representation of the system, that is total electricity sales, reasonably consistent results were obtained. It was established that when assessing the overall performance of a network structure that the average values of the performance criteria calculated over a number of training runs should be compared.

CHAPTER 5

Multi-Time-Scale Integration Technique

5.1 Introduction

This chapter develops a technique which combines forecasts produced on different times scales. The advantage of combining forecasts obtained at different time-scales is that each forecast may be based on information or variables that the other forecast does not consider. In addition, one forecast may make different assumptions about the form of the relationship between the variables. Therefore, a forecast produced by a model based on one time-scale may capture information that the other forecast based on a different time-scale may not.

Specifically, the objective here is to produce a long range prediction on a short sampling interval. The motivation for addressing this problem stems from forecasting problems in the electricity supply industry where forecasts of weekly demand are required for a period of up to five years in advance and the requirement of hourly demand up to a month in advance. The proposed approach involves the improvement of the forecasting accuracy of a relatively long range prediction based on a short-time-scale. This is achieved through the incorporation of information from a long sampling period time series into forecasts produced by a model based on a short sampling period series. Moreover, the framework of the approach is such that the long-time-scale information could in fact come from an independent source, if available.

There are numerous works available on the subject of combining forecasts produced on the same sampling frequency (Bordely, 1982, 1986; Winkler and Makridakis, 1983; Granger and Ramanathan, 1984; Shi and Liu;1993; Holden and Peel; 1986 Diebold and Pauly, 1987; Batchelor and Dua, 1995 and Clemen *et al*, 1995). However, the subject of combining forecasts produced on different time-scales is not dealt with in nearly the same depth. Corrado and Greene (1988) and Howry *et al* (1991) deal with the subject where they use outside information to improve economic forecasts, such as GNP and employment rate. In contrast to the approach presented here each of these methods produce the combined forecast at the lower sampling frequency. However, Fuhrer and Haltmaier (1988) extend the approach put forward by Corrado and Greene to produce the combined forecast at the higher sampling frequency. In all of these cases the methods are demonstrated using quarterly and monthly economic data. Another approach to the problem could be the use of multi-rate analysis. This approach

would involve a discrete time multi-rate model (Berg *et al*, 1988) which would conceptually contain both short and long sampling periods. However, with this approach all models must fit within the same mathematical framework resulting in a structurally unwieldy model.

The current proposed method, which is referred to as multi-time-scale integration, is presented and an example is then provided to demonstrate the approach. The example involves the combination of quarterly and yearly UK primary fuel consumption forecasts (Janecek and Swift, 1993). Chapter 6 presents two different electricity demand applications, where the first of these applications combines weekly and yearly demand forecasts and the second application combines hourly and daily demand forecasts.

5.2 Multi-Time-Scale Model Integration Approach

5.2.1 Overview

The objective of the approach is to combine forecasts produced by models based on different time-scales with the intention of providing a forecast based on a relatively fine sampling period, while retaining the fidelity of longer trends and natural cycles within the data. The underlying principle of the method is demonstrated using an example. Consider Figure 5.1, which shows an electricity demand time series based on two different time-scales.

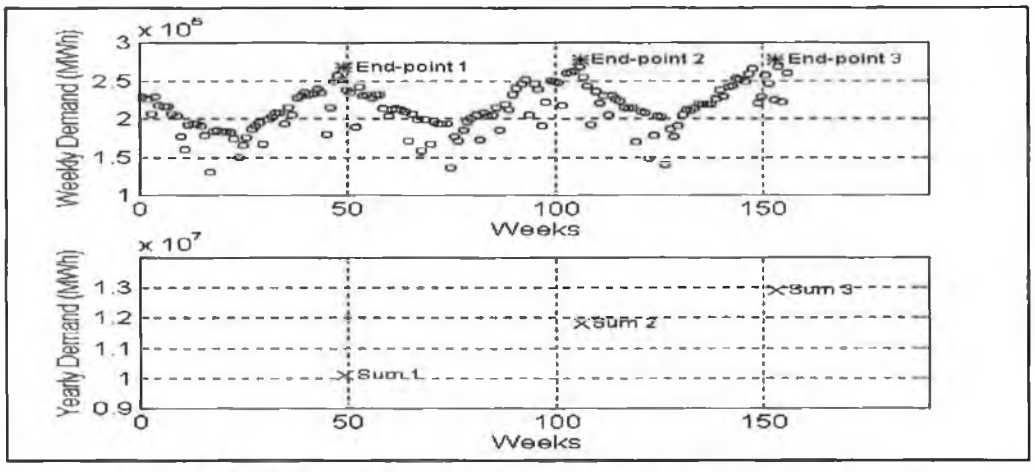

Figure 5.1 Multi-time-scale integration model for electricity demand

Series A: weekly electricity demand (○)

Series B: electricity demand for a single week of the year (∗)

Series C: yearly electricity demand (×)

The two time-scales being dealt with are weekly (Series A) and yearly (Series B and Series C). Therefore, Series A is a short sampling period time series relative to Series B and Series C that are long sampling period time series. The short-time-scale model built to predict Series A will be suitable for predicting weekly electricity demand over a period of one year. That is, it is capable of predicting $y_A(t+l)$ for lead times $l=1, 2, 3, ...N$ weeks ahead, where y_A represents the predictions of Series A and t is the forecasting origin. The longer-time-scale models built to predict Series B and Series C can make yearly predictions over a period of five years. They are capable of predicting $Y_B(T+L)$ and $Y_C(T+L)$ for lead times $L=1, 2,....M$ years ahead, where Y_B and Y_C represent the predictions of Series B and Series C respectively and T is the forecasting origin.

The objective is to integrate the forecasts of the three series A, B and C to produce a long term prediction of Series A where a forecasting horizon of up to five years in advance would be typical for this example. The integration is achieved by using the forecasts of Series B and Series C to impose the longer-time-scale information on the long term prediction of Series A. The forecasts of Series B shall be termed "end-point" information that are intermediate points that Series A should attain. The forecasts from Series C are termed "sum" information that corresponds to values that the accumulation of Series A between the 'sum' points should attain. The forecasts from the long sampling period series therefore provide intermediate 'targets' for the shorter interval forecast to follow. Cholette (1982) also considers this approach of target achievement, where partial prior information in the form of benchmarks are employed as the targets.

The application of the technique involves the generation of appropriate long sampling period sum and end-point time series. The manner in which the series are generated depends on the characteristics of the short sampling period time series. The generation of the sum series involves choosing a suitable time-scale on which to aggregrate the short sampling period series. Consider the example given in Figure 5.1, since Series A is seasonal in nature Series C was generated through the aggregation of Series A at the seasonal level. Factors of consideration when generating the end-point series are the existence of cardinal points (for example a peak or trough) in the short sampling period series. In the example given in Figure 5.1 there are no such cardinal points and thus there was freedom of choice when selecting the end-point value to when generating the end-point series.

The multi-time-scale integration technique developed here requires a state space model for the short sampling period time series. The long-time-scale information may be a forecast produced by any model based on the long sampling period time series or it may come from an outside independent source. The framework of the technique is such that it depends on the quality of the forecasts produced by the long-time-scale models. The production of these forecasts may in turn dependent on the availability of data at the long sampling period. Application of the technique has shown it to be particularly effective in the case where the short sampling period time series exhibits a seasonal component. In this instance the sum and end-point may be chosen such that they are separated by a full season. Therefore, the sum series is generated through the aggregation of the short sampling period time series at the seasonal level and the end-points are chosen to be one season apart. If the time series possesses a trend component then the characteristics of this trend may affect the characteristics of end-point time series. For example, in some cases it may be more straightforward to predict an end-point series that had been generated from a short sampling period series that has a steadily rising trend as opposed to a trend that is shifting up or down over time.

Initially, the technique is developed for the imposition of a single end-point and a single sum value. With reference to the example given in Figure 5.1 this would be the case where predictions of Series A are obtained for lead times $l=1, 2, 3,.....,52$, since the season is of length 52. The required long-time-scale information is given by End-point 1 and Sum 1. The technique is then extended to the case where multi-end-point and multi-sum values are imposed. In relation to the example this would be the case where End-point 1, End-point 2 and End-point 3 and also Sum 1, Sum 2 and Sum 3 are imposed on the short-time-scale prediction obtained for lead times $l=1,2,3,....,156$.

The technique is further extended to the case where additional long-time-scale information, other than the sum and end-point, may be used to provide extra targets for the short-time-scale prediction to follow. This would be applicable if there were a number of cardinal points in the short sampling period series, where one cardinal point is chosen as the end-point and an alternative cardinal point is chosen as the additional point. An example of this would be in the case of the daily electricity demand load profile where the tea-time peak could be chosen as the end-point and the midday peak could be chosen as the additional point. The long sampling period series for the additional point is constructed in the same manner as the end-point, with forecasts of the additional points required as in the case of the end-points.

5.2.2 Short and Long Time-Scale Models

A suitable state space model is required for the short sampling period time series. The choice of model depends on the properties of the time series in question. For example, whether the time series exhibits trend and/or seasonal characteristics or if the series is stationary or non-stationary. In the case of a stationary series an ARMA model (Jones, 1980; Ansley and Kohn, 1985a) could be employed. For some non-stationary series an ARMA model may be used after the time series has been be detrended or alternatively it may be more appropriate to employ an ARMA model after differencing the time series (Harvey and Pierce, 1984; Kohn and Ansley, 1986). However if the time series has a time trend (Harvey, 1989; Kang, 1990) a structural state space model may be more suitable (Harvey, 1984, 1989; Young, 1988, 1994; Young *et al*, 1989, 1991; Harvey and Peters, 1990; Ng and Young, 1990).

The mathematical formulation of the multi-time-scale integration technique is developed for a generalised state space form of the short-time-scale model. However, for the main applications dealt with in the thesis the most suitable model at the short sampling interval is a structural state space model and thus there is an emphasis on issues relating to the application of the technique when such a model is employed. Practical experience has also been gained in the application of the technique using AR (Kohn and Annsley, 1989), ARMA (Jones, 1990) and ARIMA (Annsley and Kohn, 1985a, 1985b) models and some discussion is provided in relation to the use of these models. The work to data has involved the development of the technique for the case where a univariate model is employed at the short sampling period. The extension of the technique to the case where a multivariate model is employed at the short sampling interval is an area of future work and this is discussed further in the conclusions presented in Chapter 7.

There are no restrictions imposed on the long-time-scale models since the focus is on the *forecast* which it produces and not the model itself. Therefore, any appropriate forecasting model may be used to produce the long-time-scale information. Exogenous variables can be used in the long-time-scale models which may lead to the production of more accurate forecasts, with the added advantage that the subsequent application of the multi-time-scale technique could result in the provision of information not already available at the shorter time scale. This flexibility in the choice of the long-time-scale model is demonstrated in the examples given in the thesis, where neural networks are used to forecast the long-time-scale information in two out of the three applications.

5.2.3 Multi-Time-Scale Integration Technique

The multi-time-scale integration procedure involves modifying a long term prediction of a short sampling period time series so that it matches, in a least squares sense, longer-time-scale information. The end-point and sum matching is achieved by adjusting the solution of the state space short-time-scale model through the relaxation of a number of the states in the state vector at the forecasting origin of the prediction. The procedure may be basically described by the following four steps

1. Formulation of the problem in least squares terms, **Ax=b**, where **x** represents the freed (relaxed) states in the state vector at the forecasting origin of the prediction and **A** and **b** are formulated subject to sum and end-point specifications.
2. Solving the least squares formulation for **x** to obtain the new states in the state vector at the forecasting origin of the prediction.
3. Combining the new states **x** with the original states to form the new state vector at the forecasting origin of the prediction.
4. Calculation of the new adjusted short-time-scale prediction using the new state vector.

A weighted least squares problem may be formulated which allows for selective alteration of the original unadjusted solution.

5.2.3.1 Single End-Point and Single Sum Matching

Let the short time-scale model for the observations $y(k)$, $k = 1.....K$ be a general state-space model of the form:

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}(k-1) + \mathbf{G}\boldsymbol{\eta}(k-1) \tag{5.1}$$

$$\mathbf{y}(\mathbf{k}) = \mathbf{H}\mathbf{x}(\mathbf{k}) + \boldsymbol{\varepsilon}(\mathbf{k}) \tag{5.2}$$

where $\mathbf{x}(k) \in \mathrm{R}^{\mathrm{n}}$ is the state vector, $\mathbf{F} \in R^{n\times n}$, $\mathbf{G} \in R^{n\times m}$ and $\mathbf{H}^T \in R^n$ are the system matrices which are assumed to be constant matrices, and $\boldsymbol{\eta}(k)$ and $\boldsymbol{\varepsilon}(k)$ are assumed to be zero mean independent and identically distributed normal random variables. Equation (5.1) is the state equation and (5.2) is the observation equation, together they make up the state space model for a system with n state variables, m system inputs and a single system output. An l-step-ahead forecast of the series is obtained through the following:

$$\hat{\mathbf{x}}(t+l/t) = \mathbf{F}^l\hat{\mathbf{x}}(t) \tag{5.3}$$

$$\hat{y}(t+l/t) = \mathbf{H}\hat{\mathbf{x}}(t+l/t) \tag{5.4}$$

where t represents the forecasting origin, l represents the forecasting lead time and $\hat{\mathbf{x}}(t+l/t)$ is the estimate of the state vector at time $t+l$ given the state estimate $\hat{\mathbf{x}}(t)$at time t. Let the forecast be required for lead times $l = 1, 2, 3,.....N$. It is required to modify the prediction from this state space model so that the forecast obtained at the end-point, N, matches the long-time-scale end-point value. It is also required that the sum of the forecasts, aggregated over the forecast horizon, matches the long-time-scale sum value. The adjustment is achieved by relaxing some of the states in the state vector $\hat{\mathbf{x}}(t)$at forecasting origin t and back-solving for these freed states using the end-point and sum data. Define:

$$\mathbf{\Phi}(l) = \mathbf{F}^l \qquad \in \quad R^{n\times n} \tag{5.5}$$

The forecast from the short-time-scale model at the end-point N of the forecast horizon is given by:

$$\hat{y}(N) = \mathrm{H}\mathbf{\Phi}(N)\mathrm{x}(t) \tag{5.6}$$

Let the value of the end-point predicted by the long-time-scale model be denoted $\hat{Y}_{ep}(L)$. Also, let the number of states which are fixed in the state vector at the forecasting origin $\hat{\mathbf{x}}(t)$ be r, and the number of states which are freed states be $(n\text{-}r)$. The state vector is reconstructed as follows:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1^T & \mathbf{x}_2^T \end{bmatrix}^T \tag{5.7}$$

where

$$\mathbf{x}_1 \in R^r \text{ and } \mathbf{x}_2 \in R^{(n-r)}$$

and with $\mathbf{x}_1$ contains the fixed states and $\mathbf{x}_2$ contains the freed states which it is necessary to solve for using the long-time-scale end-point specification. The matrix $\mathbf{\Phi}$ is partitioned appropriately according to the above construction of $\mathbf{x}(t)$ (5.7) as follows:

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 \end{bmatrix} \tag{5.8}$$

where

$$\mathbf{\Phi}_1 \in R^{n\times r} \text{ and } \mathbf{\Phi}_2 \in \mathrm{R}^{n\times(n-r)}$$

then, using (5.6), (5.7) and (5.8),

$$\hat{Y}_{ep}(L) - \mathbf{H}\Phi_1(N)\mathbf{x}_1(t) = \mathbf{H}\Phi_2(N)\mathbf{x}_2(t) + e_1 \tag{5.9}$$

where $e_1 = \hat{Y}_{ep}(L) - \hat{y}(N)$ represents the error on the end-point specification.

Similarly, if the forecast of the sum from the long-time-scale model is $\hat{Y}_s(L)$ then

$$\hat{Y}_s(L) - \sum_{l=1}^{N} \mathbf{H}\Phi_1(l)\mathbf{x}_1(t) = \sum_{l=1}^{N} \mathbf{H}\Phi_2(l)\mathbf{x}_2(t) + e_2 \tag{5.10}$$

where $e_2 = \hat{Y}_s(L) - \sum_{l=1}^{N} \hat{y}(l)$ represents the error in the sum specification.

Equations (5.9) and (5.10) represents the end-point and sum matching constraints imposed on the short-time-scale prediction.

5.2.3.2 Deviation From the Original Solution

When applying the end-point and sum information it is desirable that the prediction follows, to some degree, the original unaltered solution. Therefore, minimisation of the deviation from the unmodified prediction for *N-1* forecasts, i.e., all the forecasts minus the end-point, is sought. Let $\hat{y}(l)$ and $\hat{y}^*(l)$, l=1,2,.....N, be the unaltered and altered predictions respectively. Also, let $\mathbf{x}(t)$ be the original state vector at the forecasting origin and $\mathrm{x}^*(t)$ be the new altered state vector at the forecasting origin t, then it is required to minimise $\left|\hat{y}(l) - \hat{y}^*(l)\right|^2$ in

$$\mathbf{H}\Phi(l)\mathbf{x}(t) = \mathbf{H}\Phi(l)\mathbf{x}^*(t) + \left(\hat{y}(l) - \hat{y}^*(l)\right) \quad \text{for } l = 1,2,\ldots N-1 \tag{5.11}$$

Partitioning Φ and $\mathrm{x}^*(t)$ according to equations (5.7) and (5.8), equation (5.11) can be written as:

$$\mathbf{H}\Phi(l)\mathbf{x}(t) - \mathbf{H}\Phi_1(l)\mathbf{x}_1^*(t) = \mathbf{H}\Phi_2(l)\mathbf{x}_2^*(t) + e_l \quad \text{for } l = 1,2.....,N-1 \tag{5.12}$$

where $e_l = \left(\hat{y}(l) - \hat{y}^*(l)\right)$ represents the error on the deviation from the original solution for the forecast at lead time l. Therefore, equation (5.12) represents the deviation from the original unadjusted solution constraint. The errors $e_l = \left(\hat{y}(l) - \hat{y}^*(l)\right)$, l=1,2,.....N, are referred to as the *deviation errors*.

5.2.3.3 Weighted Least Squares Solution

A weighted least squares formulation (Franklin *et al*, 1990) of the problem is sought which allows for selective adjustment of the original prediction of the short sampling period time series. The combination of Equations (5.9), (5.10) and (5.12) gives:

$$\begin{bmatrix} \hat{Y}(L) - \mathrm{H}\Phi_1(N)\mathrm{x}_1(t) \\ \hat{Y}_s(L) - \sum_{l=1}^{N} \mathrm{H}\Phi_1(l)\mathrm{x}_1(t) \\ \mathrm{H}\Phi(1)\mathrm{x}(t) - \mathrm{H}\Phi_1(1)\mathrm{x}_1^*(t) \\ \vdots \\ \mathrm{H}\Phi(N-1)\mathrm{x}(t) - \mathrm{H}\Phi_1(N-1)\mathrm{x}_1^*(t) \end{bmatrix} = \begin{bmatrix} [\mathrm{H}\Phi_2(N)] \\ \sum_{l=1}^{N} \mathrm{H}\Phi_2(l) \\ \mathrm{H}\Phi_2(1) \\ \vdots \\ \mathrm{H}\Phi_2(N-1) \end{bmatrix} \mathrm{x}_2(t) + \mathrm{E} \qquad (5.13)$$

where

$$\mathrm{E} = [e_1 \quad e_2 \quad e_3 \quad \ldots\ldots \quad e_{N+1}]^T$$

e_3 ..... e_{N+1} represents the deviation of the modified prediction from the original prediction, where $e_{l+2} = \hat{y}(l) - \hat{y}^*(l)$, l=1,2,.....N, and e_1 and e_2 are obtained from equations (5.9) and (5.10) respectively. Equation (5.13) may be rewritten as:

$$\mathbf{b} = \mathbf{A}\mathbf{x}_2 + \mathbf{E} \qquad (5.14)$$

where

$$\mathbf{b} = \begin{bmatrix} \hat{Y}(L) - \mathbf{H}\Phi_1(N)\mathbf{x}_1(t) \\ \hat{Y}_s(L) - \sum_{l=1}^{N} \mathbf{H}\Phi_1(l)\mathbf{x}_1(t) \\ \mathbf{H}\Phi(1)\mathbf{x}(t) - \mathbf{H}\Phi_1(1)\mathbf{x}_1^*(t) \\ \vdots \\ \mathbf{H}\Phi(N-1)\mathbf{x}(t) - \mathbf{H}\Phi_1(N-1)\mathbf{x}_1^*(t) \end{bmatrix} \qquad A = \begin{bmatrix} [\mathbf{H}\Phi_2(N)] \\ \sum_{l=1}^{N} \mathbf{H}\Phi_2(l) \\ \mathbf{H}\Phi_2(1) \\ \vdots \\ \mathbf{H}\Phi_2(N-1) \end{bmatrix}$$

The matrix **A** is of dimension (N+1) x (n-r), the matrix **b** of dimension (N+1) x 1 and the matrix $\mathbf{x}_2$ is of dimension (n-r) x 1. A weighted least squares solution of the freed states $\mathrm{x}_2(t)$ which minimises $E^T\mathbf{W}E$ is given by:

$$\mathbf{x}_2(t) = \left(\mathbf{A}^{\mathbf{T}}\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{W}\,\mathbf{b} \qquad (5.15)$$

The relative importance of the error minimisation can be specified in the weighting matrix **W** which is chosen to be of the form:

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 & . & . & 0 \\ 0 & w_2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & w_i & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & w_{N+1} \end{bmatrix} \tag{5.16}$$

where the weight w_i corresponds to the error term e_i^2 in $E^T\mathbf{W}E$. The errors may be weighted differently allowing selective control of the modified solution. For example, the deviation errors could be weighted more heavily near the forecasting origin and lighter toward the end. This type of configuration would allow more freedom at the latter end of the forecast in order to achieve the end-point.

Therefore, equation (5.15) is used to solve for $\mathbf{x}_2(t)$ and the new modified state vector given by equation (5.7) is then comprised of the r original states and $(n\text{-}r)$ new modified states. This new state vector at the forecasting origin is now used to obtain the modified short time-scale prediction using equations (5.3) and (5.4).

Before proceeding on to deal with the extension of the technique to deal with more than one sum and end-point value a summary of the technique is given.

1. Use the state space short-time-scale model to calculate the original state vector at the forecasting origin of the short-time-scale prediction.
2. Select r the number of states to fix in the state vector.
3. Use equations (5.9), (5.10) and (5.12) to form the least squares formulation given by (5.13), where (5.9) and (5.10) require forecasts of Y_{ep} and Y_s from the long-time-scale models respectively.
4. Formulate the weighting matrix **W**.
5. Solve the weighted least squares formulation of (5.13) for $\mathbf{x}_2$ using equation (5.15).
6. Combine $\mathbf{x}_1$ and $\mathbf{x}_2$ to form the new state vector at the forecasting origin.
7. Using the new state vector in equation (5.3) obtain the adjusted short-time-scale prediction through (5.4).

5.2.3.4 Extension of Technique to Multi-End-Point and Multi-Sum Matching

The multi-time-scale integration technique has been described for imposing a single end-point and a single sum value on the short-time-scale prediction. It is straightforward to extend the technique to deal with multi-end-points and multi-sum values. In this case j end-points and j sum values are applied to a short-time-scale prediction obtained for lead times $l=1,2,.....jN$, where N represents the number of prediction steps performed between each of the j end-points. There are two choices available for this procedure. The first choice involves augmenting the least squares problem given by equation (5.13) with the new end-points and sum values. The second approach involves reapplying the adjustment after each cycle. For the first choice the least squares formulation (5.13) for the application of j end-points and j sum values becomes:

$$\begin{bmatrix} \hat{Y}_{ep}(L) - \mathbf{H}\Phi_1(N)\mathbf{x}_1(t) \\ \hat{Y}_{ep}(2L) - \mathbf{H}\Phi_1(2N)\mathbf{x}_1(t) \\ \vdots \\ \hat{Y}_{ep}(jL) - \mathbf{H}\Phi_1(jN)\mathbf{x}_1(t) \\ \hat{Y}_{s}(L) - \sum_{l=1}^{N}\mathbf{H}\Phi_1(l)\mathbf{x}_1(t) \\ \hat{Y}_{s}(2L) - \sum_{l=N+1}^{2N}\mathbf{H}\Phi_1(l)\mathbf{x}_1(t) \\ \vdots \\ \hat{Y}_{s}(jL) - \sum_{l=(j-1)N}^{jN}\mathbf{H}\Phi_1(l)\mathbf{x}_1(t) \\ \mathbf{H}\Phi(1)\mathbf{x}(t) - \mathbf{H}\Phi_1(1)\mathbf{x}_1^*(t) \\ \vdots \\ \mathbf{H}\Phi(N-1)\mathbf{x}(t) - \mathbf{H}\Phi_1(N-1)\mathbf{x}_1^*(t) \\ \mathbf{H}\Phi(N+1)\mathbf{x}(t) - \mathbf{H}\Phi_1(N+1)\mathbf{x}_1^*(t) \\ \vdots \\ \mathbf{H}\Phi(2N-1)\mathbf{x}(t) - \mathbf{H}\Phi_1(2N-1)\mathbf{x}_1^*(t) \\ \vdots \\ \mathbf{H}\Phi(((j-1)N)+1)\mathbf{x}(t) - \mathbf{H}\Phi_1(((j-1)N+1)\mathbf{x}_1^*(t) \\ \vdots \\ \mathbf{H}\Phi(jN-1)\mathbf{x}(t) - \mathbf{H}\Phi_1(jN-1)\mathbf{x}_1^*(t) \end{bmatrix} = \begin{bmatrix} [\mathbf{H}\Phi_2(N)] \\ [\mathbf{H}\Phi_2(2N)] \\ \vdots \\ [\mathbf{H}\Phi_2(jN)] \\ \sum_{l=1}^{N}\mathbf{H}\Phi_2(l) \\ \sum_{l=N+1}^{2N}\mathbf{H}\Phi_2(l) \\ \vdots \\ \sum_{l=(j-1)N}^{jN}\mathbf{H}\Phi_2(l) \\ \mathbf{H}\Phi_2(1) \\ \vdots \\ \mathbf{H}\Phi_2(N-1) \\ \mathbf{H}\Phi_2(N+1) \\ \vdots \\ \mathbf{H}\Phi_2(2N-1) \\ \vdots \\ \mathbf{H}\Phi_2(((j-1)N)+1) \\ \vdots \\ \mathbf{H}\Phi_2(jN-1) \end{bmatrix} \mathbf{x}_2(t) + \mathbf{E} \qquad (5.17)$$

where

$$\mathbf{E} = \begin{bmatrix} e_1 & e_2 & . & . & e_j & e_{j+1} & e_{j+2} & . & . & e_{2j} & e_{2j+1} & . & . & . & . & e_{j(N-1)} \end{bmatrix}$$

and where

- $e_1, e_2, \ldots e_j$ represent the errors on the (1 to j) end-point specifications.
- $e_{j+1}, e_{j+2}, \ldots e_{2j}$ represent the errors on the (1 to j) sum specifications.
- $e_{2j+1} \ldots\ldots e_{j(N-1)}$ represents the deviations errors

Equation (5.17) is of the form $\mathbf{b}=\mathbf{A}\mathbf{x}_2$, with $\mathbf{A}$ of dimension $(2j+j(N-1))$ x $(n-r)$, the matrix $\mathbf{b}$ of dimension $(2j+j(N-1))$ x 1 and the matrix $\mathbf{x}_2$ is of dimension $(n-r)$ x 1. As in the single end-point and single sum case a weighted least squares solution of $\mathbf{x}_2(t)$ may be found by minimising $E^T\mathbf{W}E$. In this case $\mathbf{W}$ is the same form as equation (5.16), where it is of dimension $2j+j(N-1)$ x $2j+j(N-1)$.

The second approach involves the application of the technique using the first end-point and first sum value to obtain the new modified state vector at the forecasting origin t, that is $\mathbf{x}^*(t)$. The adjusted short-time-scale prediction is obtained using $\mathbf{x}^*(t)$ for lead times $l=1,\ldots,N$. The second end-point and sum values are now applied to obtain a new state vector at the forecasting origin $t+N$, that is $\mathbf{x}^*(t+N)$ and the short-time-scale prediction is performed using $\mathbf{x}^*(t+N)$ for lead times $l=N+1,..,2N$. The process is repeated j times where the final adjusted short-time-scale prediction is obtained for lead times $l=(j-1)N+1,\ldots,jN$ using the state vector $\mathbf{x}^*(t+jN-N)$ at the forecasting origin $t+jN-N$, where this state vector was obtained through the application of the jth end-point and jth sum specification.

Experience has shown that the approach involving the augmented weighted least squares problem given by equation (5.17) yields the most accurate results. The reason for this is that a solution that simultaneously satisfies the multi-sum and multi-end-point specifications is obtained for this case. Therefore, for the applications dealt with in this thesis this approach is used for all cases involving the use of the technique with more than one-end-point and sum value.

5.2.3.5 Extension of Technique to Include an Additional Long-Time-Scale Point

The technique may be extended so that an additional point in the short-time-scale prediction may also be adjusted. It will be necessary to generate a long sampling period series for the additional point and future long-time-scale predictions of the additional points will also be required. In practice, an additional long-time-scale point is usually chosen as a cardinal point on the short sampling period time series. It is possible to use one or more additional points but the inclusion of additional points will affect the matching on the end-point and sum

specifications, resulting in a trade-off between increased improvement due to the additional point and a reduction in the matching achieved for the end-point and sum values.

If the forecast of the additional point from the long-time-scale model is denoted by $\hat{Y}_a(I)$ then the constraint at this point is given by:

$$\hat{Y}_a(I) - \mathbf{H}\Phi_1(i)\mathbf{x}_1(t) = \left[\mathbf{H}\Phi_2(i)\right]\mathbf{x}_2(t) + e_i \tag{5.18}$$

where, i is the index in the short-time-scale prediction where the additional point occurs. As in the case of the end-point, the constraint which deals with the deviation from the original prediction is relaxed at this point i. Similar constraints may be constructed for each additional point used.

5.2.3.6 Solution of Least Squares Formulation

Equation (5.17) is a least squares problem of the form **Ax=b** which represents $2j+j(N-1)$ equations in $(n-r)$ unknowns. Depending on the value of N the number of predictions carried out between each end-point; j the number of sum and end-point specifications; n the order of the state vector of the short-time-scale model and r the number of fixed states in the state vector, this system is either underdetermined or overdetermined. If $2j+j(N-1) > (n-r)$ then it is overdetermined and if $2j+j(N-1) < (n-r)$ then it is undetermined (Noble and Daniel, 1977). In either case the problem is whether there exists some matrix **Z**, uniquely determined by **A**, such that the unique minimum length of the least squares problem is given by **x** = **Zb**. This matrix **Z** does exist and is called the pseudoinverse of A (Lawson and Hanson, 1974). For a general a x b matrix **A** , the pseudoinverse of **A**, denoted by $\mathbf{A}^+$, is the a x b matrix whose jth column z_j is the unique minimum length solution of the least squares problem

$$\mathbf{A}\mathbf{x}_j = e_j \tag{5.19}$$

where e_j is the jth column of the identity matrix of dimension a x a.

5.2.4 Solution Parameters

The application of the multi-time-scale integration technique involves the selection of the following two parameters:

- the number of states to fix in the state vector at the forecasting origin of the original short-time-scale prediction, that is r.

- the weighting matrix **W** used to obtain the weighted least squares solution to the problem.

The choices available for the parameter r depend on the structure of the state vector of the short-time-scale model. The weighting matrix **W** has been found from experience to be application dependent, thus making it's specification more difficult. The optimal values of r and **W** are determined through the application of the technique using *actual* long-time-scale data for different combinations of the r and **W** parameters. Once determined the technique is then applied using the optimal value of the parameters but where predicted long-time-scale information is used.

5.2.4.1 Selection of r

The value of the parameter r depends on the structure of the state vector of the short-time-scale model. For example, for an AR model (Kohn and Annsley, 1989), ARMA model (Jones, 1990) or ARIMA model (Annsley and Kohn, 1985a, 1985b) the state vector is made up of present and past values of the model output. In this case it is desirable to leave a proportion of the states which represent the most recent values of the model output unchanged since they provide the starting point and higher order derivatives for the solution. Consider an example involving the application of the technique to weekly/yearly System B AT data. It involved the adjustment of a one-year-ahead prediction of weekly AT using yearly sum and end-point long time-scale AT information. The end-point is the last week of the year (Week 52) and the sum is the culmulation of weekly AT from Week 1 in January to Week 52 in December. In Chapter 3 it was determined that an AR(41) model is a suitable model for this data set and this model is thus used as the state space short-time-scale model. The possible values of the parameter r are from 1 to 40. The basic structure of **W** was chosen to a diagonal matrix, with the sum and end-point errors assigned larger values relative to the deviation errors and where the deviation errors were weighted more heavily at the forecasting origin and lighter towards the end. The MAE surface obtained for the adjustment carried out using actual sum and end-point data is shown in Figure 5.2. The MAE is shown on the z-axis, the variation in r, from state 1 to state 40, is shown along the x-axis and the variation in **W** is shown along the y-axis, where **W** was varied by assigning values of varying magnitudes to the deviation errors terms

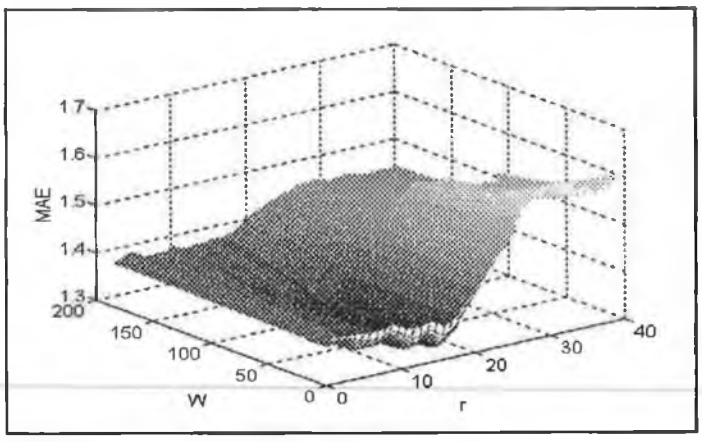

Figure 5.2 MAE surface for AR(41) model

The graph shows that the error surface in the case of the AR model is relatively smooth and it is thus straightforward to determine the optimal combination of the r and **W** parameters. The optimal value for r is 16, therefore the first 16 states in the state vector at the forecasting origin are fixed and the remaining 25 are freed. In the case of the optimal **W** the sum and end-point error weights are equal to 100, the first deviation error weight is equal to 42, the last deviation error is equal to 1, with linear interpolation between these values assigned to the deviation error weights in between. The adjustment of the 52-week-ahead prediction using the optimal value of r and **W** results in a 9% improvement in the MAE and a 20% improvement in the MSE over the original solution.

Further improvement of the adjustment results may be achieved through the overparameterisation of the AR(41) model. If the order of the parsimonious AR(41) model was artificially increased, that is overparameterised, then a greater number of the most recent states could remain fixed while allowing approximately the same number of initial conditions to be freed. For example, when an AR(59) model is used at the short sampling period and adjustment is carried out with the parameter r equal to 34 then approximately the same number of stares are freed as in the AR(41) model case, that is 24. Adjustment carried out in this case results in approximately a 15% improvement in the MAE and a 24% improvement the MSE over the original unadjusted solution.

Experience has shown that relatively smooth error surfaces similar to that obtained in the AR model example given above are also obtained in the case of the ARMA and ARIMA models. Consequently, the search for the optimal value of the parameter r through the examination of the error surface is a straightforward exercise when any of these models are employed at the short sampling interval. However, as a general guideline it has been found that fixing

approximately the first 40% of the states for an AR model, the first 30% of states for an ARMA model and the first 20% of states for an ARIMA model have been found to give good results in practice.

In the case of structural state space models the structure of the state vector is different to the AR, ARMA and ARIMA models in that the states represents characteristics within the data, such as trend and seasonality. For example, in a BSM with a dummy seasonal component (2.44) the first state represents the level of the trend component, the second state the slope of the trend component, the third state represents the current seasonal component and states 4 to n represent past values of the seasonal component, where n is the order of the state vector. Fixing states *1* to *3* states in the state vector yields the original solution and therefore there are only three choices available for r. The choices are as follows:

- fix state 1 and free states 2 to n.
- fix state 2 and free states 1 and states 3 to n.
- fix states 1 and 2 and free states 3 to n.

For a BSM with a trigonometric seasonal component (2.48) the state vector is composed of a trend component and a set of trigonometric terms at seasonal frequencies $\lambda_j=2\pi/s$, for $j=1,.....s/2-1$, for s even and for $j=1,.....s/2$, for s odd, where s is the seasonal length. The state vector is of dimension $(s+2)$ for s even and of dimension $(s+1)$ if s is odd. The first two states make up the level and slope of the trend component and the pairs of states from *3* to $s+2$, if s is even, and *3* to $s+1$, if s is odd, make up the trigonometric terms. For this state space structure the number of choices available for r is when s is even are as follows:

- fix state 1 and free states 2 to $(s+2)$.
- fix state 2 and free states 1 and states 3 to $(s+2)$.
- fix state 3 and free states 1 to 2 and states 4 to $(s+2)$.

- fix state $((s+2)-1)$ and free states 1 to $((s+2)-2)$.
- fix state 1 to 2 and free states 3 to $(s+2)$.
- fix state 1 to 3 and free states 4 to $(s+2)$.

- fix state 1 to $((s+2)-3)$ and free states $((s+2)-2)$ to $(s+2)$

Note that, it is necessary to fix only one of the states in seasonal pairs given by (2.28), that is $\gamma_j(k)$, since $\gamma_j^*(k)$ appears only as a matter of construction and thus the adjustment achieved through fixing state $\gamma_j^*(k)$ is the equal to the adjustment achieved by fixing $\gamma_j(k)$. Based on the above representation the total number of choices available for r is for s even or s odd is $s+2$. Therefore, for applications where s is large the use of a trigonometric seasonal component in the BSM increases the development effort significantly in two ways. Firstly the details of the implementation are more involved. Secondly it results in a multi-time-scale integration method which is work intensive, since the model performance must be examined for each value of r through simulation.

5.2.4.2 Selection of W

The values of the weights in **W** allows the user considerable control over the adjustment of the short-time-scale prediction. For the most part, **W** is application dependent, however some general guidelines and comments are given based on application experience.

The most common form for **W** is given by equation (5.16) where it is only necessary to specify the weights along the diagonal elements. Weights are assigned to the following error terms:

- the sum, end-point and additional point error terms.
- the deviation errors terms.

When a single sum, end-point and additional value is imposed on the short-time-scale prediction larger values, relative to the magnitude of the weights on the deviation errors, are assigned to these error terms reflecting the importance of their minimisation. Experience has shown that good adjustment results may be obtained when the ratio of the magnitude of sum, end-point and additional point weight to the deviation error weights is approximately of the order of 10/1. For cases involving more than one sum, end-point and additional value the weight on the end-point error terms are more heavily penalised than the sum, additional point *and* the deviation error terms, since the end-point of each prediction provides the starting value for the next cycle. In this case the weight on the sum and additional point error is usually reduced to the same order of magnitude as the deviation errors.

There are a number of choices available for the weighting profile on the deviation error terms. For example, the weighting profile may be specified such that the error terms near the forecasting origin are heavily penalised but allowed more freedom near the end-point, in order to allow the solution to achieve the end-point specification. In contrast, the deviation error

terms at the forecasting origin may be given more freedom, to allow the solution to achieve the level of the trend specified by the end-point values, but given heavier weighting toward the end-point. This may be appropriate in cases where the end-point values are difficult to predict and the solution is encouraged to achieve the end-point while attaining the natural course of the original prediction. An alternative is to adopt this profile but to weight a small number of the deviation error terms at the forecasting origin, typically two, the same as the weights near the end-point. Thus, this weighting profile has the combined characteristics of the two previous weighting profiles. In practice, the latter weighting profile has proven to be suitable where adjustment is carried out using more than one end-point and sum value and in addition where the forecast horizon is relatively long.

5.2.5 Multi-Time-Scale Integration Application Procedure

The application of the multi-time-scale integration technique is demonstrated using a standard time series taken from the literature. The technique is applied to the UK quarterly fuel consumption time series dealt with by Janecek and Swift (1993). The purpose of this example is to demonstrate the effectiveness of the technique on an alternative time series and to discuss the issues involved in its practical application. Overall, there are four cardinal steps involved in the application of the technique, where these are as follows:

- Develop the state space short sampling period forecasting model.
- Generate suitable long sampling period time series.
- Develop long sampling period forecasting models.
- Select appropriate r and $\mathbf{W}$.

However, for both clarity of presentation and also for practicality in the application of the technique a comprehensive multi-time-scale integration application procedure was developed. This procedure is used in the example presented here and also in the applications dealt with in Chapter 6. The application procedure is as follows:

MTSI-Step 1. Select sampling period and forecast horizon of short-time-scale prediction.

MTSI-Step 2. Select short sampling period time series which may depend on selection in Step 1

MTSI-Step 3. Generate long sampling period time series which involves the following:

- Generation of the sum long sampling period series which involves selecting the time-scale which to aggregate the short sampling period time

series on. For example, if the time series is seasonal then aggregation could be performed at the seasonal level.

- Generation of the end-point long sampling period series which involves the selection of the end-point, where factors of consideration here may the presence of cardinal points in the series.
- Generation of the additional point(s) series, where again the presence of cardinal points in the series may determine the selection of the additional points.

MTSI-Step 4. Based on MTSI-Step 1 and MTSI-Step 3 determine what long-time-scale information is required to perform the adjustment. The factors involved may be whether a single or multi sum, end-point and/or additional point(s) are being applied.

MTSI-Step 5. Determine a suitable state space short-time-scale model.

MTSI-Step 6. For comparative purposes, perform unadjusted short-time-scale prediction using model found in MTSI-Step 5

MTSI-Step 7. Determine suitable long-time-scale models.

MTSI-Step 8. Using models obtained in MTSI-Step 7 produce predictions of long-time-scale information described in MTSI-Step 4.

MTSI-Step 9. Based on structure of state vector of short time scale model determined in MTSI-Step 5 construct the set of possible values of the parameter r.

MTSI-Step 10. Determine if weighted least squares solution is underdetermined, overdetermined or exact. This depends on the number of equations and the number of unknowns which in turn depend on the following:

- Number of equations depends on the long-time-scale information given in MTSI-Step 4 (single or multi sum, end-point and/or additional long-time-scale information) and on the length of the forecast horizon.
- Number of unknowns depends on the number of free parameters (n-r) in the state vector determined in MTSI-Step 9.

MTSI-Step 11. Construct a suitable set of **W**.

MTSI-Step 12. Determine the optimal combination of r and **W** by examining the forecasting accuracy achieved by carrying out the adjustment using *actual* long-time-scale information. If necessary return to MTSI-11 to construct a new set of **W** until an optimal the combination of r and **W** is determined.

MTSI-Step 13. Using the optimal r and $\mathbf{W}$ determined in MTSI-Step 12 perform the adjustment of short-time-scale prediction using the *predicted* long-time-scale information obtained in MTSI-Step 8.

5.3 Quarterly and Yearly Fuel Consumption Example

5.3.1 Overview

The multi-time-scale integration technique was applied to a quarterly UK primary fuel consumption ($1x10^5$ tons) series (Janecek and Swift, 1993). The fuel data is available from 1965 to 1985. The following predictions of quarterly fuel consumption are performed:

- one-year-ahead - Quarters 1-4 of 1984 and Quarters 1-4 of 1985.
- two-year-ahead - Quarters 1-4 of 1984 to 1985.
- two-quarter-ahead - Quarter 3 to Quarter 4 of 1985.

where the application of the multi-time-scale integration technique for each forecast is dealt with separately.

5.3.2 One-Year-Ahead Forecast

MTSI-Step 1: Short-time-scale forecast

The aim here is to perform two separate one-year-ahead forecasts of quarterly UK primary fuel consumption, where the forecast horizons are Quarters 1- 4 of 1984 and Quarters 1- 4 of 1985

MTSI-Step 2: Short sampling period time series

The short sampling period time series is made up of quarterly UK primary fuel consumption from 1965 to 1985. Figure 5.3 gives a plot of the time series.

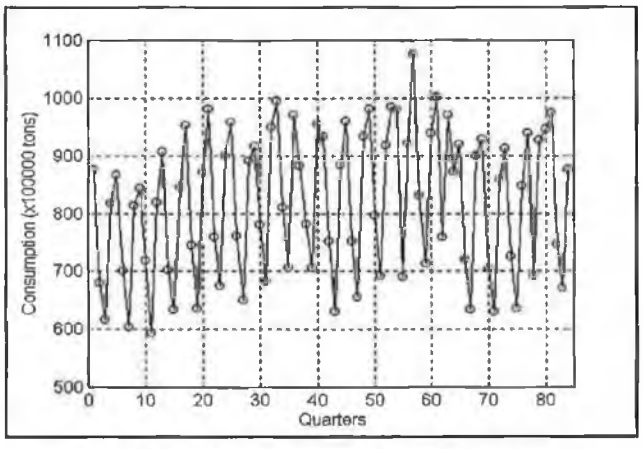

Figure 5.3 UK primary fuel consumption

The time series exhibits an irregular long term trend component and a seasonal component, with a seasonal length of *4*, which is due to the UK seasonal variation in temperature. The seasonal component may be examined further by comparing the quarterly data for each year on the same graph where Figure 5.4 provides such a plot. In this graph quarterly data is over-plotted for two years at the beginning of the series, 1965 and 1966, and two years at the end of the series, 1984 and 1985.

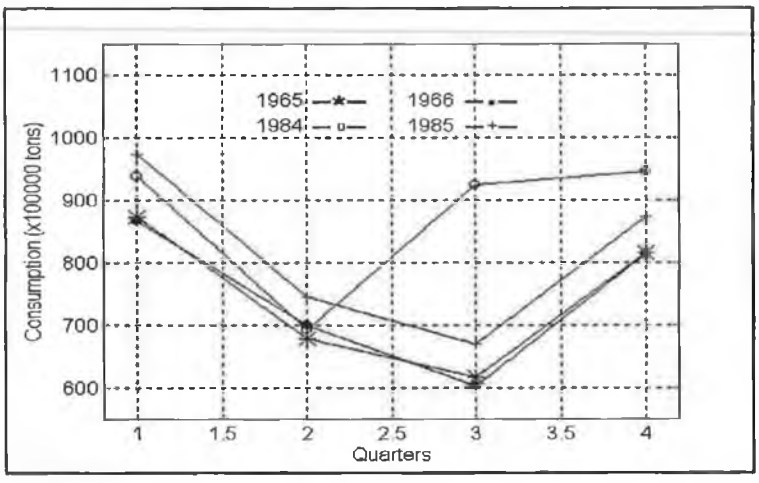

Figure 5.4 Over-plotting of quarterly demand

The 1965, 1966 and 1985 profiles exhibit the same seasonal pattern, however this is not the case for the 1984 profile. Over-plotting of the quarterly data for each year in the entire data set was carried out. The same seasonal pattern is repeated from 1965 up to 1977. However, after 1977 there are variances in the seasonal pattern for three separate years, where these years are 1978, 1980 and 1984. A plot of the data for these years is given on the graph in Figure 5.5 where the 1985 data is also included as a representation of the more regular seasonal pattern for the series.

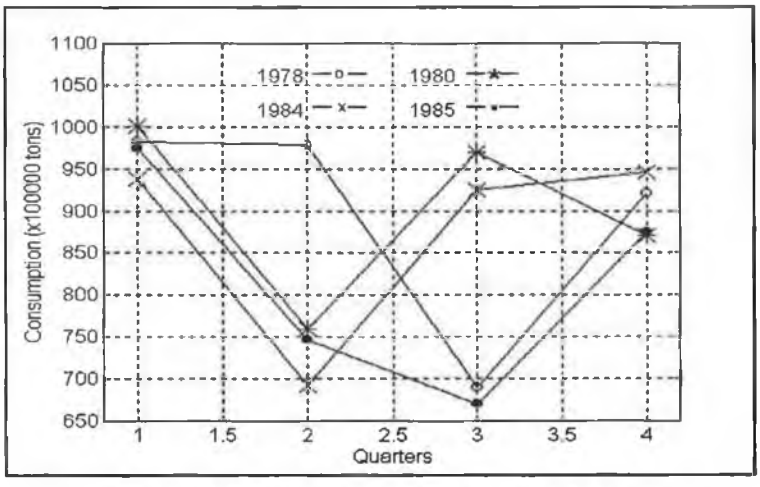

Figure 5.5 Over-plotting of non-typical years

The differences occur at Quarter 2 in 1978, Quarter 3 in 1980 and Quarter 4 in 1984. In all cases the outliers have a higher value of consumption than expected. Unfortunately, at the time of the study access to UK weather data was not possible. However, Irish AT data was examined in an attempt to establish if an unusual cold spell affected the fuel consumption at these particular times, where the assumption was made that the weather conditions in Ireland and the UK would not differ substantially. AT at Quarters 3 and 4 in 1984 were compared to AT at Quarters 3 and 4 in 1985. The AT at Quarter 3 in 1984 is approximately 3 ^{0}C lower than the same quarter in 1985 and is approximately 6 ^{o}C lower in Quarter 4. Irish AT data was not available for the years 1978 and 1980. Based on the analysis for 1984 it is thought that the most likely reason for the outliers in the UK fuel consumption is diverse weather conditions.

MTSI-Step 3: Long sampling period time series

To generate the sum long sampling period series it was necessary to select the time-scale on which to aggregate the short sampling period time series. Since the short sampling period time series is seasonal in nature the series is aggregated at the yearly level. Therefore, the sum series is the aggregate of the fuel consumption from Quarter 1 to Quarter 4 of each year. Because there are no cardinal points in the short sampling period time series there is freedom of choice for the end-point and it was chosen as the fuel consumption in the last quarter of each year. Since the one-year-ahead predictions consists of only 4 points no additional points are chosen for adjustment. Plots of the end-point and sum time series are given in Figure 5.6 and Figure 5.7 respectively.

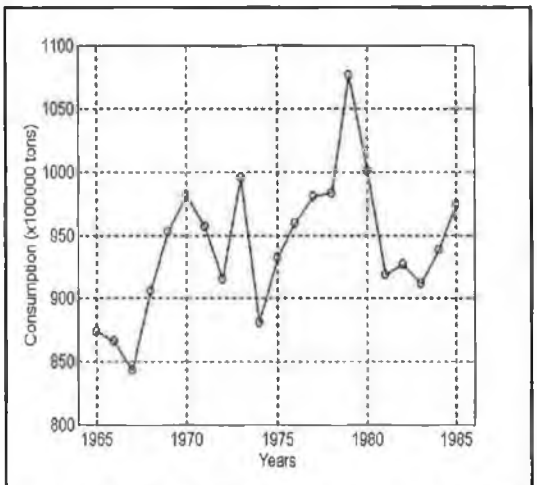

Figure 5.6 End-point long sampling period time series

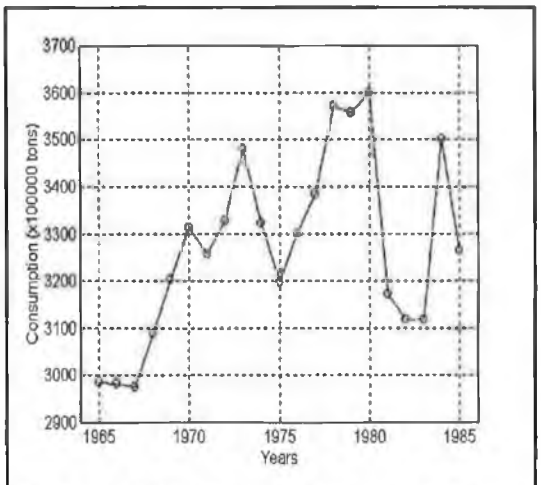

Figure 5.7 Sum long sampling period time series

It can be seen from the graphs that the sum and end-point series follow a rising trend up to the year 1980 mainly due to annual economic and demographic influences. However, a decline in the economic environment starting in 1980 resulted in a decrease in the UK fuel consumption until 1984. By the year 1985 economic conditions were improving resulting in a moderate increase in the fuel consumption in this year. The significantly high sum value in 1984 is due to the high fuel consumption in Quarter 3 and Quarter 4 of that year that were assumed to be due to diverse weather conditions.

MTSI-Step 3: Required long-time-scale information

The long-time-scale information required to carry out the adjustment of the one-year-ahead predictions is described in Table 5.1.

Table 5.1 Required end-point and sum long-time-scale information

Short-Time-Scale Forecast Horizon	End-point	Sum
Quarters 1- 4 of 1984	Quarter 4 1984	Σ(Quarter 1 to Quarter 4 of 1984)
Quarters 1- 4 of 1985	Quarter 4 1984	Σ(Quarter 1 to Quarter 4 of 1985)

MTSI-Step 4: Short time-scale-model

The short sampling period series is a nonstationary time series with trend and seasonal characteristics. Analysis carried out in Chapter 3 determined that a BSM with a dummy seasonal component (2.44) is a suitable model to represent such a series and the use of this model in the current case would have been appropriate. However, given the shortness of the seasonal length, that is $s = 4$, a BSM with a trigonometric seasonal component (2.48) was chosen to model the short sampling period series to demonstrate the multi-time-scale integration approach for such a model. This is in direct contrast to the two applications dealt with in Chapter 6, where the short sampling period time series have seasonal lengths of $s = 52$ (weekly time series) and $s = 24$ (hourly time series) and thus a BSM with dummy seasonal component was employed in each case. Equations (5.20) and (5.21) give the state space form of the quarterly BSM with trigonometric seasonal component:

$$\mathbf{x}(k)=\begin{bmatrix}\mu(k)\\ \beta(k)\\ \gamma_1(k)\\ \gamma_1^*(k)\\ \gamma_2(k)\end{bmatrix}=\begin{bmatrix}1&1&0&0&0\\ 0&1&0&0&0\\ 0&0&0&1&0\\ 0&0&-1&0&0\\ 0&0&0&0&-1\end{bmatrix}\begin{bmatrix}\mu(k-1)\\ \beta(k-1)\\ \gamma_1(k-1)\\ \gamma_1^*(k-1)\\ \gamma_2(k-1)\end{bmatrix}+\begin{bmatrix}\eta(k)\\ \zeta(k)\\ \omega_1(k)\\ \omega_1^*(k)\\ \omega_2(k)\end{bmatrix} \tag{5.20}$$

$$y(k)=\begin{bmatrix}1&0&1&0&1\end{bmatrix}x(k)+\varepsilon(k) \tag{5.21}$$

where

- $\mu(k)$ is the level of the trend component.
- $\beta(k)$ to the slope of the trend component.
- $\gamma_1(k)$ and $\gamma^*_1(k)$ are the pair of states representing the seasonal component at the fundamental frequency.
- $\gamma_2(k)$ represents the seasonal component at the first harmonic.
- $\eta(k)$ is the stochastic disturbance term corresponding to $\mu(k)$
- $\zeta(k)$ is the stochastic disturbance term corresponding to $\beta(k)$
- $\omega_i(k)$ and $\omega^*_i(k)$ are the stochastic disturbance terms corresponding to $\gamma_i(k)$ and $\gamma^*_i(k)$
- $\varepsilon(k)$ is the irregular component.

Note that, the $\gamma_1^*(k)$ appears only as a matter of construction of the state space model and $\gamma_2^*(k)$ is not required as s is even.

The prediction of future values of the time series obtained using equations (5.3) and (5.4) require an estimate of the state vector at the forecasting origin of the forecast horizon. To this end, a different identification data set was used to obtain the model employed to carry out each of the one-year-ahead forecasts described in MTSI-Step 1. The identification data sets and corresponding forecast horizons are given in Table 5.2

Table 5.2 Identification data sets and forecast horizon

Forecast Horizon	Identification Data Set
Quarters 1 to 4 of 1984	Quarter 1 of 1965 - Quarter 4 of 1983
Quarters 1 to 4 of 1985	Quarter 1 of 1965 - Quarter 4 of 1984

A basic description of the techniques used to estimate the hyperparameters of the BSM with trigonometric component was described in Section 2.4.2 of Chapter 2 and full details are given in Appendix A. Tables 5.3 and 5.4 describe the models obtained for each identification data set. The following information is given in each table:

- σ_ε^2, the variance of $\varepsilon(k)$.
- σ_η^2 is the variance of $\eta(k)$.
- σ_ζ^2 is the variance of $\zeta(k)$.
- σ_ω^2 is the variance of $\omega(k)$.

- $\tilde{\sigma}$ is the one step ahead prediction error variance.
- $Q(T)$ is the Box-Ljung statistic (2.35).

Table 5.3 Structural model for 1965/1983 data set

ID data set	Quarter 1 of 1965 - Quarter 4 of 1983					
Forecast Horizon	1-year-ahead: predict Quarters 1 -4 of 1984					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(8)$
	$0.221\text{x}10^{-4}$	$0.013\text{x}\ 10^{-4}$	$0.681\text{x}10^{-8}$	0	0.0054	7.31

Table 5.4 Structural model for 1965/1984 data set

ID data set	Quarter of 1 of 1965 - Quarter 4 of 1984					
Forecast Horizon	1-year-ahead: predict Quarters 1 - 4 of 1985					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(8)$
	$0.280\text{x}10^{-4}$	$0.019\text{x}10^{-4}$	0	$0.007\text{x}10^{-4}$	0.0065	4.56

The chi-squared statistics are $\chi^2_{[0.5]}(8) = 15$ and the models were assumed to be adequate based on the Ljung-Box statistics given by $Q(8)$. For each model the relatively large values of the variance σ_η^2 in comparison with σ_ζ^2 and σ_ω^2 are due to the fact that the trend component of the series is shifting up and down, where the larger the variance σ_η^2 the greater the stochastic movement. The small values of the variance σ_ζ^2 indicate that the slope of the trend component is reasonably constant. The greater the value of σ_ω^2 relative to σ_ε^2 the more past observations are discounted in the prediction of the seasonal component. For the 1965-1983 data set the value of σ_ω^2 is zero, where in contrast it is of the order of 10^{-4} for the 1965-1984 data set. This is possibly due to the effect of the two outliers at Quarters 3 and 4 in 1984, with more discounting in the 1965-1984 data set.

MTSI-Step 6: Unadjusted short time-scale prediction

The unadjusted short-time-scale predictions were obtained using the models determined in MTSI-Step 5. The forecasting accuracy of these results are given in Table 5.5.

Table 5.5 Unadjusted short-time-scale prediction

Forecast Horizon	**MAE**	**MSE x10^{-4}**	**MAPE**	**MPE**
Quarters 1 to 4 of 1984	119.63	2.6830	13.05	11.47
Quarters 1 to 4 of 1985	45.25	0.2712	6.04	-6.04

The relatively high values of the forecasting accuracy measures for the 1984 prediction, particularly in the case of the MSE which penalises large forecasting discrepancies, are due to the outliers at Quarters 3 and 4 in 1984 quarterly data. In the case of the 1985 prediction the high values of the MAPE and MPE show that the presence of the outliers in the 1965-1984 identification data set appear to have affected the performance of the model used to carry out the 1985 prediction.

MTSI-Step 6: Long-time-scale yearly model

MLP neural networks were used to model the yearly sum and end-point long sampling period time series. The neural network analysis developed in Chapter 4 was adopted and the MLP application procedure outlined in Section 4.2.5 was used where both cases are dealt with simultaneously.

NN-Step 1: Input and output structure

The output layer of the MLP consists of a single neuron which represents the current value of the time series.

In Chapter 4 linear forecasting analysis was used to determine the input structure to the MLPs. An attempt was made to employ a similar analysis in the case of the sum and end-point MLPs, however difficulty was encountered in the identification of a linear model structure, particularly in the case of the sum series. The SACF (2.8) for the yearly sum fuel consumption time series is given in Figure 5.8 where it can be seen that there are no cut offs.

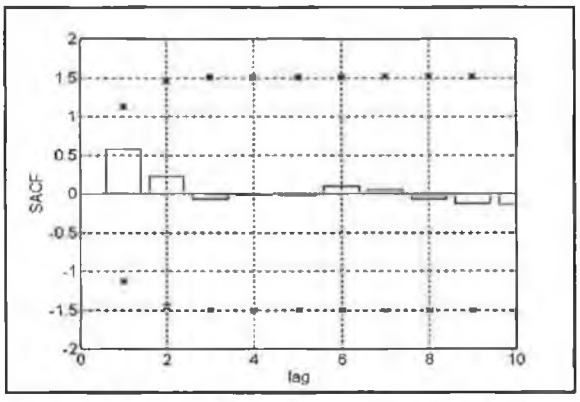

Figure 5.8 SACF of sum long sampling series.

The input structure to the MLP used to predict the sum series was determined on a trial and error basis since it was not possible to use the SACF for its determination. The form of the input structure used may be described by the following nonlinear equation

$$Y_s(k) = g\left(Y_s(k-1), Y_s(k-2), Y_s(k-3)\right) \tag{5.22}$$

where g is the nonlinear function and $Y_s(k)$ is the sum of the fuel consumption for year k.

The SACF for the end-point series is shown in Figure 5.9 , where it can be seen that there is a cut-off at lag 1.

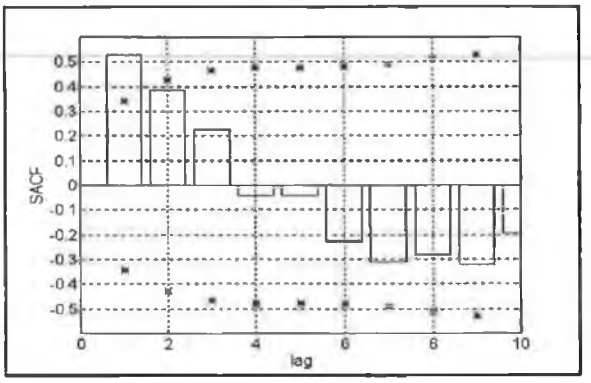

Figure 5.9 SACF of end-point long sampling series

Based on the SACF the end-point at lag one was included in the input structure. It was established that the inclusion of the end-point at lags 2 and 3 further improved the results over the validation set. Thus the MLP input structure for the end-point series may be described by the following nonlinear equation

$$Y_{ep}(k) = f\left(Y_{ep}(k-1), Y_{ep}(k-2), Y_{ep}(k-3)\right) \tag{5.23}$$

where f is the nonlinear function and $Y_{ep}(k)$ is the fuel consumption at the end-point in year k.

NN-Step 2: Input and output data normalisation

The training data for each MLP was normalised by dividing by the NF value of the time series. For the sum series the NF value is 1 x 10^4 and for the end-point series the NF value is 1 x 10^3. The target output data was normalised to the same order of magnitude as the input data as discussed in Section 4.2.2.

NN-Step 3: Construct input/output training data sets

For each case, two years (1982-1983) were used for the validation data set and the remainder of the data was used for the training data set (1965 to 1981).

NN-Step 4: Select structure and hidden layer architecture

Based on the motivation given in Section 4.2.2 of Chapter 4 two hidden layers were used in each MLP. Due to computational time constraints it was not possible to conduct a full

experimentation analysis to select the optimal number of neurons in each of the hidden layers of the networks. Therefore, based on the experimentation analysis performed in Section 4.4.1 of Chapter 4 four network structures were trained and their performance compared. These structures were 1-3-1, 2-6-1, 3-5-1 and 3-9-1. Initial results showed that the 1-3-1 and 2-6-1 network structures did not produce good forecasting results and thus only the 3-5-1 and 3-9-1 structures were trained and the results compared. The MLPs were trained using the LMS gradient technique with backpropagation, with an adaptive learning rate (initial value of 1×10^{-3}) and a momentum constant equal to 0.95. Ten training runs were performed and the training cessation point was determined through the examined of the multi-step-ahead MAE performance over the validation set. The sum and end-point MLPs were trained for 60000 and 50000 epochs respectively. In the experimentation analysis conducted in Section 4.4.1 the average over the total number of training runs of the MMAE_val was used to compare the different network structures and a similar approach is adopted here. A summary of the results of the current analysis is given in Table 5.6 and Table 5.7.

Table 5.6 Comparison of hidden layer architecture for sum MLP

	Average value over 10 training runs	
Structure	**MMAE_Val**	
	@ Epoch	**MAE**
3-5-1	47629	3.51
3-9-1	47387	4.01

Table 5.7 Comparison of hidden layer architecture for end-point MLP

	Average value over 10 training runs	
Structure	**MMAE_Val**	
	@ Epoch	**MAE**
3-5-1	18640	55.19
3-9-1	37022	68.33

Based on the above results a 3-5-1 neural network was used to forecast the sum and end-point long-time-scale information.

MTSI-Step 8: Long-time-scale prediction

Forecasts of the long-time-scale information described in MTSI-Step 4 were obtained using the MLPs obtained in MTSI-Step 7. Two single-step-ahead forecasts were obtained for the

1984 and 1985 sum and end-point sum predictions and the percentage error associated with each of the predictions is given in Table 5.8.

Table 5.8 Percentage accuracy for predicted end-point and sum

Prediction	End-point	Sum
	% Error	% Error
1984	6.55	4.86
1985	0.56	4.22

All the forecasting errors are relatively high with the exception of the end-point forecast in 1985. In the case of the 1984 sum forecast the high forecasting error may be explained to some extent by the presence of outliers in the latter two quarters of this year. However, overall the high forecasting errors in the predictions suggest that a more extension neural network analysis is possibly merited, where a particular area of concern is the input structure of the sum and end-point MLPs. However, due to computational time constraints a more detailed neural network analysis was not performed and the multi-time-scale integration technique was applied using the predictions given in Table 5.8 under the assumption the results may be further improved through the improvement in accuracy of the results of long-time-scale models.

MTSI-Step 9: Selection of possible values of r

The selection of r depends on the structure of the state vector of the short-time-scale model. Given the structure of the state vector (5.20) there are six choices available for r. These choices are as follows:

R1: fix state 1 and free states 2 to 5

R2: fix state 2 and free states 1 and states 3 to 5

R3: fix state 3 and free states 1 to 2 and states 4 to 5

R4: fix state 5 and free states 1 to 4

R5: fix state 1 to 2 and free states 3 to 5

R6: fix state 1 to 3 and free states 4 to 5

MTSI-Step 10: Description of weighted least squares problem

Table 5.9 describes the form of the weighted least squares problem for the adjustment of the one-year-ahead forecast using each of the different choices of r described in MTSI-Step 9, where it can be seen that an overdetermined system of equations is obtained.

Table 5.9 Description of weighted least squares problem

r	(*n-r*) No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R4	4	4	sum and end-point	5	overdetermined
R5	3	4	sum and end-point	5	overdetermined
R6	2	4	sum and end-point	5	overdetermined

MTSI-Step 11: Selection of set of **W**

The weighting matrix **W** was chosen to be a diagonal matrix of the following form:

$$\mathbf{W} = diag\left[w_{ep} \quad w_{s} \quad w_{dev\_qr1} \quad w_{dev\_qr2} \quad w_{dev\_qr3}\right] \tag{5.24}$$

where

- w_{ep} is the weight on the end-point error.
- w_s is the weight on the sum error.
- w_{dev_qri} represents the weight on the error for the deviation from the original forecast at quarter *i*.
- w_{dev_qr1} is referred to as the *starting value* of the deviation error profile.
- w_{dev_qr3} is referred to as the *finishing value* of the deviation error profile.

As discussed in Section 5.2.4.2 the weights on the end-point and sum errors were given large values relative to the weights on the deviation errors, where the order of magnitude is typically a ratio of approximately 10/1.

Three different weighting profiles were considered for the deviation error. An example of the weighting profiles is shown in Figure 5.10:

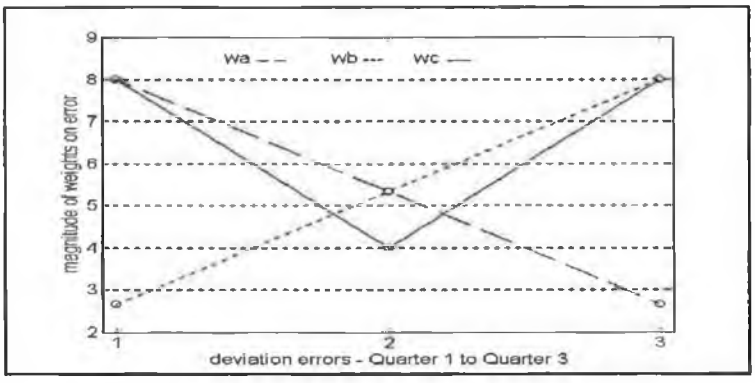

Figure 5.10 Comparison of weighting profiles

where

- *wa* weights the deviation error corresponding to Quarter 1 heavily and weights Quarter 2 and Quarter 3 progressively lighter.
- *wb* weights the deviation error corresponding to Quarter 3 heavily with gradually lighter weighting on Quarter 2 and Quarter 1
- *wc* employs a symmetrical weighting profile on the deviation errors corresponding to Quarter 1 to Quarter 3.

For example, setting up the deviation error weights in the *wa* weighting matrix involved the following steps:

1. select a number for the starting value.
2. decrement this number by a set amount to obtain w_{dev_qr2}, where the set amount = starting value/ the number of deviation errors.
3. decrement w_{dev_qr2} by the set amount to obtain w_{dev_qr3}.

Similarly for *wb*:

1. select a number for the finishing value.
2. decrement this number by a set amount to obtain w_{dev_qr2}, where the set amount = finishing value/ the number of deviation errors.
3. decrement w_{dev_qr2} by the set amount to obtain w_{dev_qr1}

Finally for *wc*:

1. select a number for the starting value.
2. set this number equal to the finishing value.
3. w_{dev_qr2} is equal to half the starting value.

The *wa*, *wb*, and *wc* weighting matrices with starting/finishing values equal to 8 and end-point and sum error weights equal to 100 are as follows:

$$\begin{aligned}
\mathbf{W} &= diag\left[w_{ep} \quad w_{s} \quad w_{dev_qr1} \quad w_{dev_qr2} \quad w_{dev_qr3}\right] \\
wa &= diag\left[100 \quad 100 \quad 8 \quad 5\tfrac{1}{3} \quad 2\tfrac{2}{3}\right] \\
wb &= diag\left[100 \quad 100 \quad 2\tfrac{2}{3} \quad 5\tfrac{1}{3} \quad 8\right] \\
wc &= diag\left[100 \quad 100 \quad 8 \quad 4 \quad 8\right]
\end{aligned} \tag{5.25}$$

where it can be seen that the starting/finishing values are equal to 8. Variations of the *wa*, *wb* and *wc* weighting matrices given in (5.25) may be constructed by choosing different starting and finishing values on the deviation error weighting profiles. For example, changing the starting value of the linear weighting profile on deviation errors in the *wa* matrix from 8 to 5 yields the following:

$$\begin{aligned} \mathbf{W} &= diag\left[w_{ep} \quad w_s \quad w_{dev\_qr1} \quad w_{dev\_qr2} \quad w_{dev\_qr3}\right] \\ wa &= diag\left[100 \quad 100 \quad 5 \quad 3\tfrac{1}{3} \quad 1\tfrac{2}{3}\right] \end{aligned} \tag{5.26}$$

and similarly for *wb*, where the finishing value of the linear weighting profile is changed from 8 to 5:

$$\begin{aligned} \mathbf{W} &= diag\left[w_{ep} \quad w_s \quad w_{dev\_qr1} \quad w_{dev\_qr2} \quad w_{dev\_qr3}\right] \\ wb &= diag\left[100 \quad 100 \quad 1\tfrac{2}{3} \quad 3\tfrac{1}{3} \quad 5\right] \end{aligned} \tag{5.27}$$

Similarly, for the *wc* matrix the starting and finishing values on the symmetrical deviation error profile could be changed as follows:

$$\begin{aligned} \mathbf{W} &= diag\left[w_{ep} \quad w_s \quad w_{dev\_qr1} \quad w_{dev\_qr2} \quad w_{dev\_qr3}\right] \\ wc &= diag\left[100 \quad 100 \quad 5 \quad 2\tfrac{1}{2} \quad 5\right] \end{aligned} \tag{5.28}$$

Different *wa*, *wb* and *wc* weighting matrices were constructed using the steps describes above, where different number were assigned to the starting/finishing values of the deviation error weighting profile. The range of numbers considers for the starting/finishing were from 1 to 20. Therefore, a set of 60 different weighting matrices were considered in total, that is 20 different *wa* matrices with starting values in the range 1 to 20, 20 different *wb* matrices with finishing values in the range 1 to 20 and 20 different *wc* matrices with finishing values in the range 1 to 20.

MTSI-Step 12: Optimal selection of *r* and *W*

This step involved the determination of the optimal *r* and **W** out of the set of parameter values chosen in MTSI-Step 10 and MTSI-Step 11 respectively. This was achieved through the comparison of the forecasting accuracy results obtained for the adjustment of the one-year-

ahead quarterly predictions using *actual* end-point and sum data. The objective here was to determine the optimal combination of the r and **W** parameters. Graphical comparison of the MAE and MSE error surfaces for the different combinations of the parameter values were used to aid their determination. For each of the different weighting matrices, *wa*, *wb* and *wc*, the MAE, r and the range of starting/finishing values of the deviation error profile were plotted against each other. A similar graph was plotted for the MSE. The graphs are shown in Figure 5.11 to Figure 5.16 for 1984 and Figure 5.17 to Figure 5.22 for 1985, where in each of the graphs the axis represent the following:

x-aixs: the different choices for the parameter r - R1 to R6.

y-axis: the variation in the starting/finishing values on the deviation error profile:

- the different starting values for the deviation error for *wa*.
- the different finishing values for the deviation error for *wb*.
- the different starting and finishing values for *wc*.

z-axis: MAE or MSE.

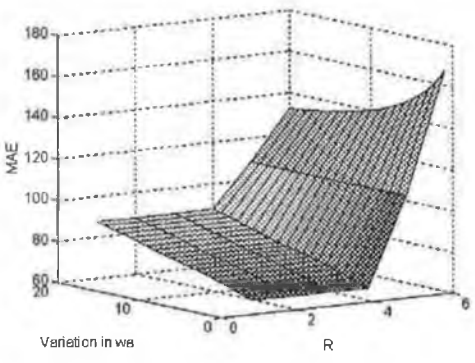

Figure 5.11 MAE for 1984 using *wa*

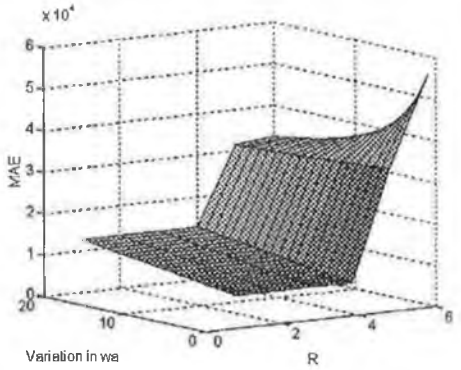

Figure 5.12 MSE for 1984 using *wa*

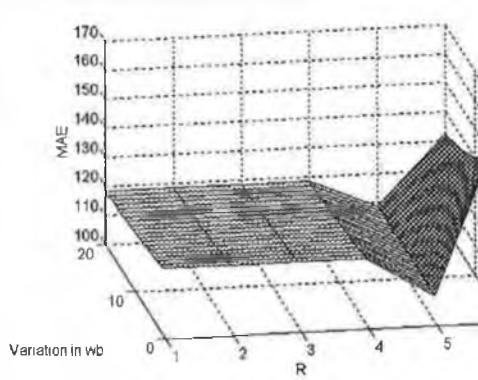

Figure 5.13 MAE for 1984 using *wb*

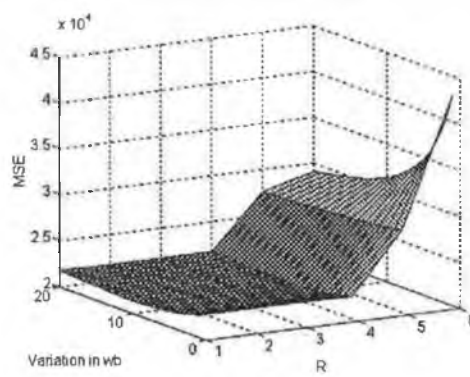

Figure 5.14 MSE for 1984 using *wb*

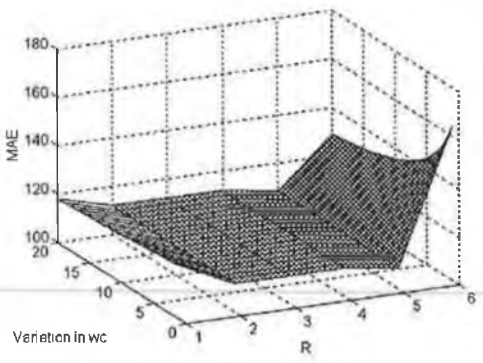

Figure 5.15 MAE for 1984 using *wc*

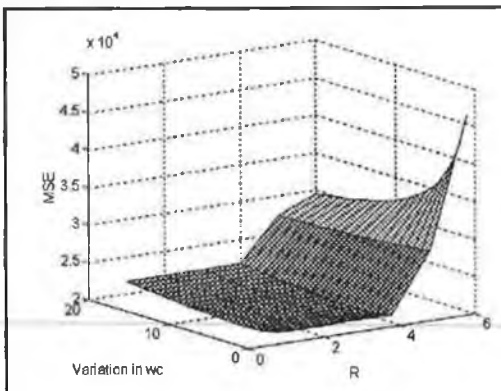

Figure 5.16 MSE for 1984 using *wc*

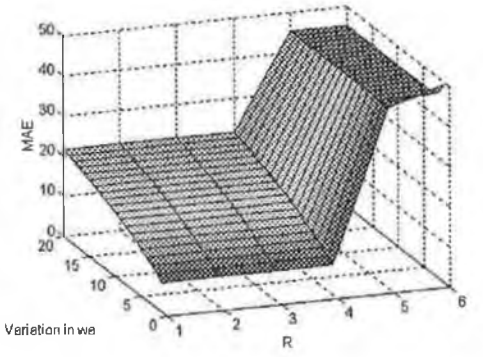

Figure 5.17 MAE 1985 using *wa*

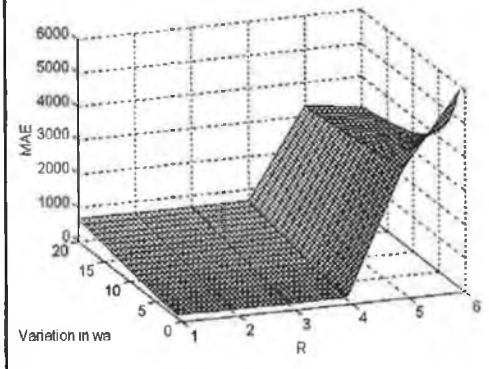

Figure 5.18 MSE 1985 using *wa*

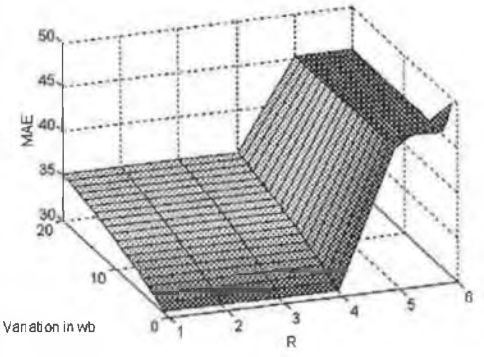

Figure 5.19 MAE for 1985 using *wb*

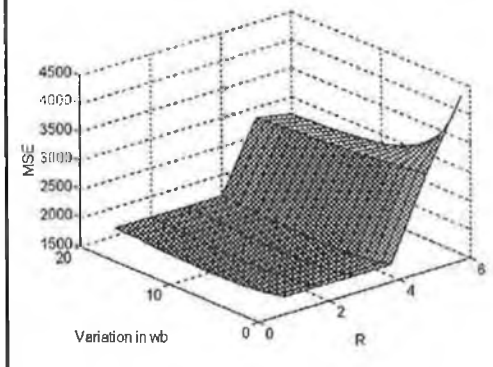

Figure 5.20 MSE for 1985 using *wb*

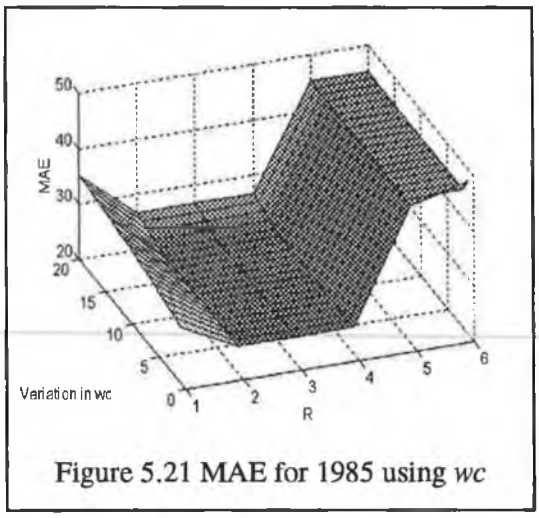

Figure 5.21 MAE for 1985 using *wc*

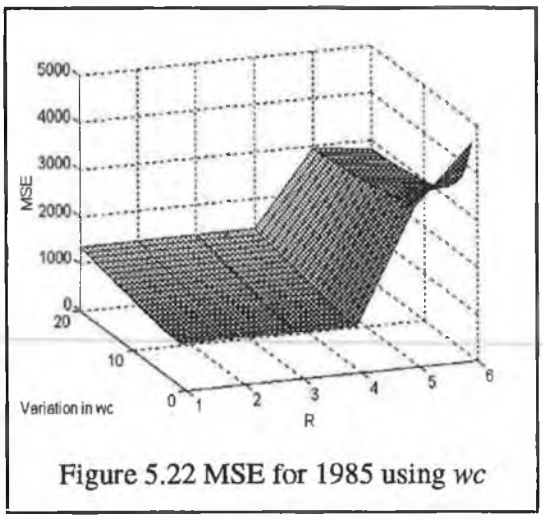

Figure 5.22 MSE for 1985 using *wc*

Examination of the graphs allowed the optimal combination of the r and **W** parameters to be determined. Comparing the MAE and the MSE achieved for each of the different weighting matrices, *wa*, *wb* and *wc* shows that the optimal weighting matrix structure out of three matrices is *wa*, with a starting value equal to 1. The graphs obtained using the *wa* matrix were examination further to determine the optimal choice for the parameter r. It is clear from these graphs that R5 and R6 do not yield good adjustment results but it also appears as though the results obtained for R1 to R4 are practically the same. Closer examination of the MAE and MSE showed that the most accurate results were obtained for R4, however the differences in the results between this choice and R1 to R3 are negligible. The reason for this is that the same number of states in the original state vector were freed (4 states) for the adjustment carried out for these choices of r, in order to achieve the single sum and end-point specifications over the relatively short forecast horizon (4 points). In contrast, the choices that produced inferior adjustment results, that is R5 and R6, only freed 3 and 2 states respectively. Therefore, the results indicate that the parameter r should be chosen such that the maximum number of states are freed in order to achieve the sum and end-point specifications.

Table 5.10 summarises the optimal r and **W** for the adjustment of one-year-ahead 1984 and 1985 predictions.

Table 5.10 Optimal r and W for one-year-ahead predictions

Year	r	W
1984	R4	$wa = diag\left[10 \quad 10 \quad 1 \quad \frac{2}{3} \quad \frac{1}{3}\right]$
1985	R4	$wa = diag\left[10 \quad 10 \quad 1 \quad \frac{2}{3} \quad \frac{1}{3}\right]$

Note that, the weight on the end-point and sum error terms for the *wa* matrix with starting value equal to 1 is such that the order of magnitude of the end-point weight to deviation error weight and sum weight to deviation error weight is in the ratio of 1/10.

MTSI-Step 13: Adjustment of short-time-scale prediction

Adjustment of the one-year-ahead predictions was carried out, with the optimal values of the *r* and **W** parameters determined in MTSI-Step 12, but where the predicted end-point and sum data values determined in MTSI-Step 8 were used.

Figure 5.23 and Figure 5.24 give the actual vs. predicted graphs for the adjusted one-year-ahead short-time-scale prediction, where the unadjusted solution is included for comparative purposes.

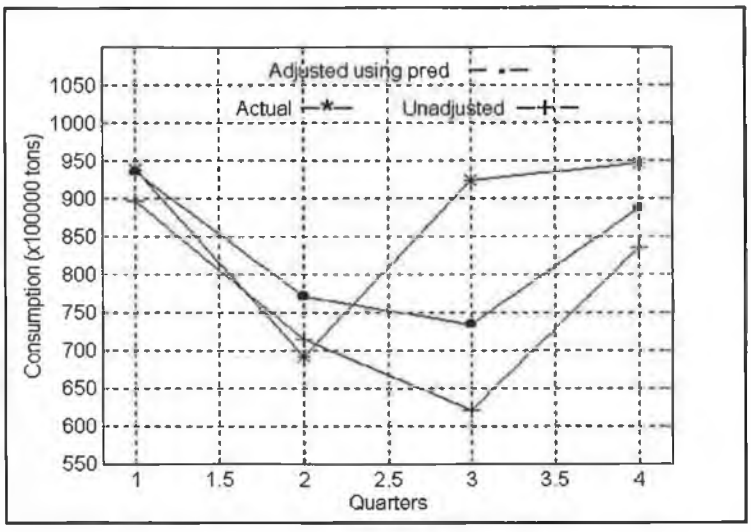

Figure 5.23 Actual vs. predicted for 1984

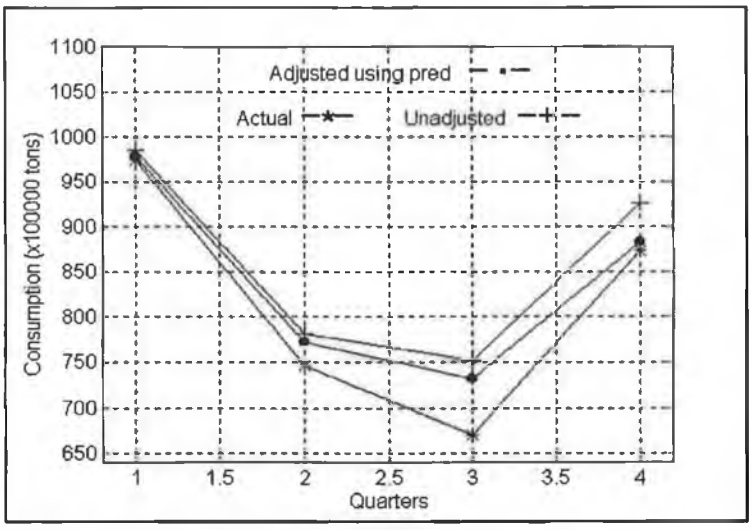

Figure 5.24 Actual vs. predicted for 1985

The graph of the 1984 prediction shows that although there are significant errors at Quarter 3 and Quarter 4 of the adjusted solution there is considerable improvement over the original unadjusted solution at these points. Over and above the improvement at Quarters 3 and 4 the achievement of the sum and end-point specifications yielded a minor error at Quarter 1 but resulted in a higher error in the adjusted solution over the original solution at Quarter 2. In the case of the 1985 prediction the effect of the adjustment is to move the prediction closer to the actual points at all quarters of the forecast horizon, with significant improvement at Quarter 3.

A measure of the improvement achieved may be seen in forecasting accuracy results given Table 5.11 and Table 5.12. The table includes the unadjusted short-time-scale prediction and the adjusted short-time-scale prediction using actual end-point and sum data. The results for the adjustment carried out using the actual long-time-scale information provides a measure of the maximum improvement that may be obtained.

Table 5.11 Adjustment results for 1984 prediction

Forecast	MAE	MSE x 10^{-4}	MAPE	MPE
Unadjusted	119.63	2.6830	13.05	11.47
Adjusted using actual	65.75	0.7451	8.97	-0.81
Adjusted using predicted	82.79	1.1535	9.63	3.94

Table 5.12 Adjustment results for 1985 prediction

Forecast	MAE	MSE x 10^{-4}	MAPE	MPE
Unadjusted	45.25	0.2712	6.04	-6.04
Adjusted using actual	7.49	0.0086	0.90	0.23
Adjusted using predicted	25.17	0.1137	3.52	-3.52

The forecasting accuracy results clearly demonstrate the overall improvement achieved through the adjustment of the original solution in each of the 1984 and 1985 cases. Although reasonable improvement was obtained through the adjustment of the 1984 prediction the presence of the outliers result in a relatively high MAPE measure of 9.63, where the MAPE provides an overall measure of accuracy over time. However, comparison of the MSE measures shows that there is considerable improvement (57%) at the outliers over the original case since the MSE penalises large errors. It can be seen that the presence of the outliers in the data set have an effect on the original 1985 prediction, where the MAPE of the solution is as high as 6.04. For the adjusted solution of the 1985 prediction this value is approximately halved to 3.52. In addition, in the 1985 case, as in the case of the 1984 prediction, there is significant improvement (58%) in the MSE measure. An outstanding factor in the overall set

of results is the maximum achievable improvement that could be obtained (adjustment using actual), where there is a considerable difference between the unadjusted and adjusted results for this case. The reason for this substantial improvement is again the presence of the outliers. However, although it is possible that the forecasting accuracy results of the long-time-scale models could be improved to some extent, the very presence of the outliers are likely to hinder the production of these results.

5.3.3 Two-Year-Ahead Forecast

MTSI- Step 1 to MTSI-Step 3

In this case a two-year-ahead forecast of quarterly UK primary fuel consumption, Quarters 1-4 of 1984 to 1985, was performed. As in the one-year-ahead case the short sampling period time series is made up of quarterly UK primary fuel consumption from 1965 to 1985, where Figure 5.2 gives a plot this series. The long sampling period time series are the yearly sum and end-point series described in the one-year-ahead case and plots of the time series are given in Figure 5.6 and Figure 5.7.

MTSI-Step 4: Long-time-scale information required

The long-time-scale information used to carry out the adjustment of the two-year-ahead prediction is described in Table 5.13.

Table 5.13 Required yearly end-point and sum long-time-scale information

Short-Time-Scale Forecast Horizon	**End-point**	**Sum**
Quarters 1- 4 1984 - 1985	1. Quarter 4 1984 2. Quarter 4 1985	1. Σ(Quarter 1 to Quarter 4 of 1984) 2. Σ(Quarter 1 to Quarter 4 of 1985)

MTSI-Step 5: Short-time-scale-model

As in the one-year-ahead case a BSM with a trigonometric seasonal component was used to obtain the two-year-ahead forecast of quarterly fuel consumption. The identification data set and corresponding forecast horizon is given in Table 5.14.

Table 5.14 Identification data sets and forecast horizon

Forecast Horizon	**Identification Data Set**
Quarters 1- 4 1984 - 1985	Quarter 1 of 1965 - Quarter 4 of 1984

The model for this identification data set was already determined in MTSI-Step 5 of the one-year-ahead case in Section 5.3.2 and the model was described in Table 5.4.

MTSI-Step 6: Unadjusted short time-scale prediction

The forecasting accuracy results of the unadjusted two-year-ahead prediction obtained using the short-time-scale model are given in Table 5.15.

Table 5.15 Unadjusted two-year-ahead prediction

Forecast	MAE	MSE x10-^{4}	MAPE	MPE
1984/1985	91.07	1.5514	10.35	9.56

So that a comparison may be made between the results obtained in the one-year-ahead case and the results obtained here; Table 5.16 gives the MAE and MSE for each of the individual years, 1984 and 1985, of the two-year-ahead prediction.

Table 5.16 Single years of unadjusted two-year-ahead prediction

Forecast	MAE	MSE x10-^{4}	MAPE	MPE
1984	119.63	2.6830	13.05	11.47
1985	62.50	0.4198	7.66	7.66

As expected the accuracy in 1985 of the multi-step-ahead prediction is lower than that obtained in the single-step-ahead prediction for that year obtained in Section MTSI-Step 6 of Section 5.3.2.

MTSI-Step 7 - MTSI-Step 8: Long-time-scale yearly model and prediction

The long-time-scale information required to carry out the adjustment of the quarterly two-year-ahead prediction was described in MTSI-Step 4. A two-step-ahead prediction of the required end-point and sum data were obtained using the MLPs that were developed in the one-year-ahead case in Section 5.3.2. The percentage accuracy of the forecasts for each step of the two-year-ahead prediction are given in Table 5.17.

Table 5.17 Percentage accuracy for predicted end-point and sum

Prediction	End-point	Sum
	% Error	% Error
1984	6.55	4.86
1985	1.14	4.99

The multi-step-ahead prediction has resulted in a further loss in accuracy in the 1985 prediction, particularly in the case of the end-point forecast.

MTSI-Step 9: Selection of possible set of r

The set of six possible choices available for r were already discussed in the one-year-ahead case in Section 5.3.2.

MTSI-Step 10: Form of weighted least squares solution

The form of the weighted least squares solution is given in Table 5.18.

Table 5.18 Description of weighted least squares solution

r	$(n-r)$ No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R4	4	4	sum and end-point	10	overdetermined
R5	3	4	sum and end-point	10	overdetermined
R6	2	4	sum and end-point	10	overdetermined

As in the one-year-ahead case for all choices of the parameter r the WLS problem is overdetermined but where the number of equations has now been increased to 10.

MTSI-Step 11: Selection of possible set of **W**

In this step a set of appropriate weighting matrices were constructed, where the weighting vector for the adjustment of the two-year-ahead prediction is of the following form:

$$\mathbf{W} = diag\left[w_{ep\_yr1} \quad w_{ep\_yr2} \quad w_{s\_yr1} \quad w_{s\_yr2} \quad w_{dev\_qr1\_yr1} \quad w_{dev\_qr2\_yr1} \quad w_{dev\_qr3\_yr1} \quad w_{dev\_qr1\_yr2} \quad w_{dev\_qr2\_yr3} \quad w_{dev\_qr3\_yr2}\right] \tag{5.29}$$

where

- w_{ep_yrj} is the weight on the end-point error in year j.
- w_{s_yrj} is the weight on the sum error in year j.
- $w_{dev\_qri\_yrj}$ represents the weight on the error for the deviation from the original forecast at quarter i in year j.
- $w_{dev\_qr1\_yrj}$ is the *starting value* of the deviation error profile for year j.
- $w_{dev\_qr3\_yrj}$ is the *finishing value* of the deviation error profile for year j.

Similar weighting matrices to those used in the one-year-ahead case were also considered here, that is the *wa*, *wb* and *wc* matrices. The weight on the end-point error is more heavily

penalised than the weight on the sum error since the end-point of each year provides the starting point for the next year's prediction as discussed in Section 5.2.4.1. Thus the sum weight is reduced to the same order of magnitude as the weights on the deviation errors. It would have been possible to assign different deviation error weighting profiles to each year of the two-year-ahead prediction. However, this would increase the computational time involved in the selection of the optimal value of **W** significantly because the adjustment of the short-time-scale prediction must be performed for each **W**. Consequently, the same deviation error profile was adopted for each year. Examples of these weighting matrices are as follows:

$$\mathbf{W} = diag[w_{ep\_yr1} \quad w_{ep\_yr2} \quad w_{s\_yr1} \quad w_{s\_yr2} \quad w_{dev\_qr1\_yr1} \quad w_{dev\_qr2\_yr1} \quad w_{dev\_qr3\_yr1} \quad w_{dev\_qr1\_yr2} \quad w_{dev\_qr1\_yr2} \quad w_{dev\_qr3\_yr2}]$$

$$wa = diag[100 \quad 100 \quad 10 \quad 10 \quad 8 \quad 5\tfrac{1}{3} \quad 2\tfrac{2}{3} \quad 8 \quad 5\tfrac{1}{3} \quad 2\tfrac{2}{3}]$$

$$wb = diag[100 \quad 100 \quad 10 \quad 10 \quad 2\tfrac{2}{3} \quad 5\tfrac{1}{3} \quad 8 \quad 2\tfrac{2}{3} \quad 5\tfrac{1}{3} \quad 8] \qquad (5.30)$$

$$wc = diag[100 \quad 100 \quad 10 \quad 10 \quad 8 \quad 4 \quad 8 \quad 8 \quad 4 \quad 8]$$

Different weighting matrices were constructed with different values assigned to the starting/finishing points of the deviation error weighting profile, where values in the range from 1 to 20 were considered.

MTSI-Step 12: Selection of optimal r and **W**

Adjustment was carried out on the two-year-ahead prediction with actual end-point and sum data and using the set of r and **W** parameter outlined in MTSI-Step 10 and MTSI-Step 11. The optimal values of the parameters were selected through the examination of the MAE and MSE results obtained. Similar graphs to those given in the one-year-ahead case in Section 5.3.2 are presented. Figure 5.25 to Figure 5.30 gives the graphs.

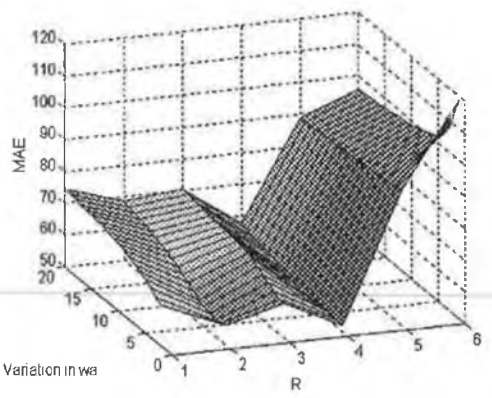

Figure 5.25 MAE for adjustment of 1984/1985 prediction using *wa*

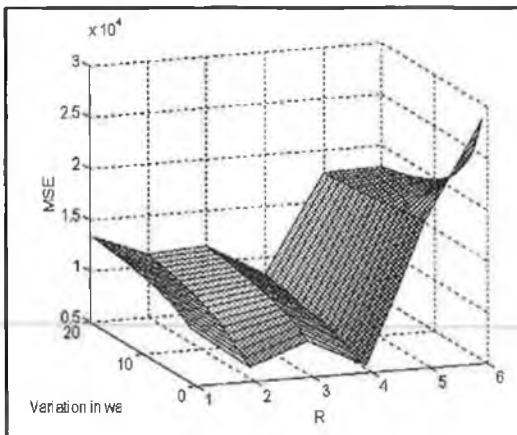

Figure 5.26 MSE for adjustment of 1984/1985 prediction using *wa*

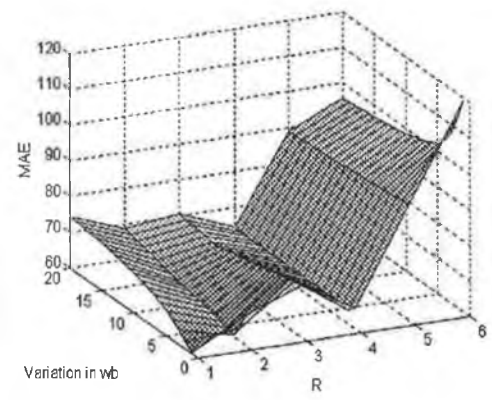

Figure 5.27 MAE for adjustment of 1984/1985 prediction using *wb*

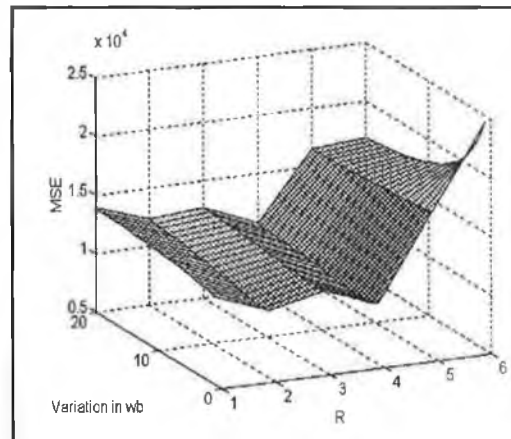

Figure 5.28 MSE for adjustment of 1984/1985 prediction using *wb*

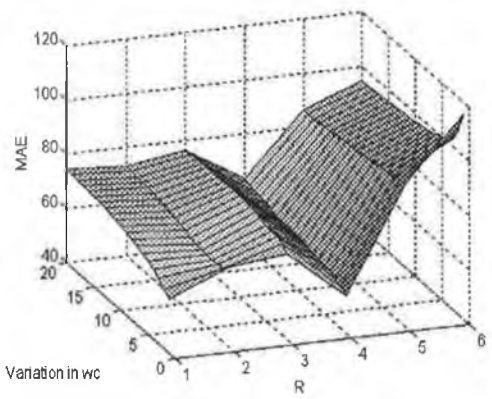

Figure 5.29 MAE for adjustment of 1984/1985 prediction using *wc*

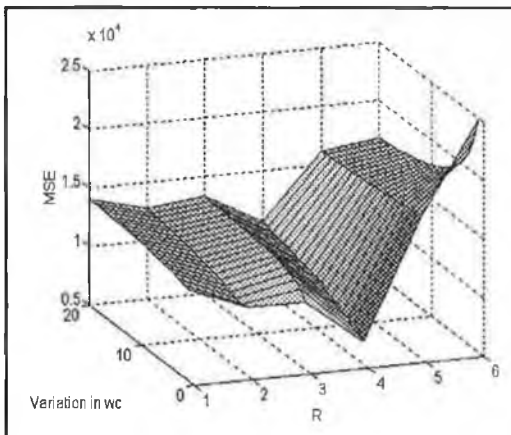

Figure 5.30 MSE for adjustment of 1984/1985 prediction using *wc*

Examination of the graphs show that using both the MAE and MSE as a criterion that the optimal value for the parameter r may clearly be chosen as R4. In contrast to the one-year-ahead case the results obtained for each choice of r are different since it is necessary to satisfy multi-sum and multi-end-point-specifications over a relatively long forecast horizon. The graphs also indicate that either *wa* or *wc* is the optimal weighting matrix structure. In order to determine the most suitable weighting matrix the MAE and MSE values were examined more closely for the following for R4 with the matrix structure *wa* and R4 with the matrix structure *wc*.

Using both the MAE and MSE as a criterion the most accurate results were obtained using the *wa* matrix, with a starting value of 1. The optimal r and **W** parameters used to carry out the adjustment of the two-year prediction are summarised in Table 5.19.

Table 5.19 Optimal r and **W** for two-year-ahead prediction

r	**W**
R4	$wa = diag[100 \quad 100 \quad 10 \quad 10 \quad 1 \quad 0.6 \quad 0.3 \quad 1 \quad 0.6 \quad 0.3]$
R4	$wa = diag[100 \quad 100 \quad 10 \quad 10 \quad 1 \quad 0.6 \quad 0.3 \quad 1 \quad 0.6 \quad 0.3]$

MTSI-Step 13: Adjustment of short-time-scale prediction

Adjustment of the two-year-ahead prediction was carried out using the optimal r and **W** parameters determined in MTSI-Step 12 but where the predicted sum and end-point data determined in MTSI-Step 8 were used. Graphs of actual vs. predicted are given in Figure 5.31 and Table 5.20 give the forecasting accuracy results obtained. Table 5.21 and Table 5.22 give the forecasting accuracy results for each of the individual years, 1984 and 1985, of the two-year-ahead prediction.

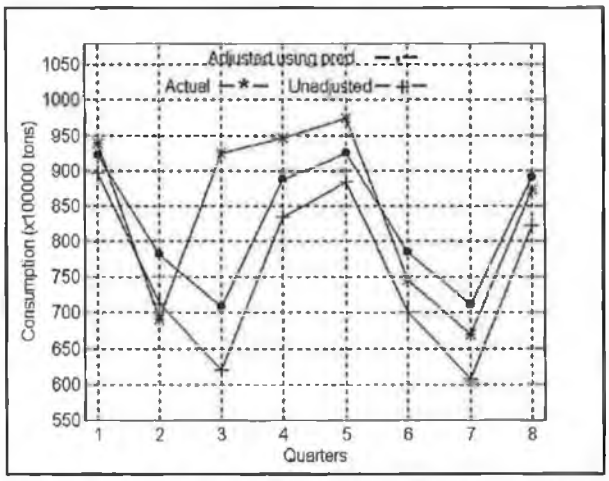

Figure 5.31 Actual vs. predicted for 1984 to 1985

Table 5.20 Adjustment results for two-year-ahead case

Forecast	MAE	MSE x10^{-4}	MAPE	MPE
Unadjusted	91.07	1.5514	10.35	9.56
Adjusted using actual	52.13	0.5327	6.28	0.40
Adjusted using predicted	66.15	0.8077	7.86	1.25

Table 5.21 1984 of two-year-ahead prediction

Forecast	MAE	MSE x10^{-4}	MAPE	MPE
Unadjusted	119.63	2.6830	13.05	11.47
Adjusted using actual	71.13	0.8930	8.59	0.08
Adjusted using predicted	95.41	1.4641	11.09	4.58

Table 5.22 1985 of two-year-ahead prediction

Forecast	MAE	MSE x10-4	MAPE	MPE
Unadjusted	62.50	0.4198	7.66	7.66
Adjusted using actual	33.12	0.1723	3.90	0.71
Adjusted using predicted	36.90	0.1513	4.63	-2.08

The graph shows that with the exception of Quarter 2 in each year of the prediction the result of the adjustment is to move the solution closer to the actual values at each quarter. As in the one-year-ahead case considerable improvement is seen at the outliers in the 1984 prediction. Overall, the original forecast of 1985 is poor due to the outliers in the preceding two quarters at the start of this prediction. Notable improvement through adjustment of the prediction can be observed at Quarter 1 and Quarter 4 (the end-point) of the 1985 prediction. The degree of improvement obtained may be seen in the forecasting accuracy measures. As in the one-year-ahead case, in Section 5.3.2, significant improvement over the original solution can be seen in the MSE measure. Comparison of the results of the forecasting accuracy measures of the single years of the two-step-ahead prediction to the results obtained in the one-year-ahead case shows that overall moderately greater improvement was achieved in the latter case. The reason for this is the relatively poorer forecasting of the long-time-scale information in the multi-step-ahead case.

5.3.4 Two-Quarter-Ahead Forecast

The purpose of the two-quarter-ahead example is to illustrate the case where the least squares problem is an underdetermined system of equations.

MTSI- Step 1 to MTSI-Step 3

A two quarter-ahead forecast of quarterly UK primary fuel consumption, Quarter 3 to Quarter 4 of 1985 is performed. The short and long sampling period time series are as in the one-year-ahead and two-year-ahead cases, where a plot of the short sampling period series is given in Figure 5.3 and plots of the sum and end-point series are given in Figure 5.6 and Figure 5.7 respectively.

MTSI-Step 4: Required long-time-scale information

The end-point and sum data required to carry out the adjustment of the two-quarter-ahead prediction is described in Table 5.23.

Table 5.23 Required end-point and sum information

Short-Time-Scale Forecast Horizon	**End-point**	**Sum**
Quarters 3- 4 1985	Quarter 4 1985	Σ(Quarter 3 to Quarter 4 of 1985)

MTSI-Step 5 to MTSI-Step 6: Short-time-scale-model and prediction

As in the one-year-ahead and two-year-ahead cases a BSM with a trigonometric seasonal component was used to obtain the two-quarter-ahead forecast of quarterly fuel consumption. The identification data set is described in Table 5.24 and the parameters of the estimated model are given in Table 5.25.

Table 5.24 Identification data set and forecast horizon

Forecast Horizon	**Identification Data Set**
Quarters 3- 4 of 1985	Quarter 1 of 1965 - Quarter 2 of 1985

Table 5.25 Structural model for Quarter 1 of 1965 to Quarter 2 of 1985 identification data set

ID data set	Quarter 1 of 1965 - Quarter 2 of 1985					
Forecast Horizon	2-quarters-ahead: predict Quarters 3 - 4 of 1985					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(8)$
	$0.274\text{x}10^{-4}$	$0.017\text{x}10^{-4}$	0	$0.002\text{x}10^{-4}$	0.0032	6.68

As expected a similar model to that obtained for the Quarter 1 of 1965 - Quarter 4 of 1984 identification data set used in the one-year-ahead case was also obtained here, since there are only two extra data points in the identification data set for this case. Table 5.26 gives the forecasting accuracy measures of the unadjusted solution.

Table 5.26 Unadjusted two-quarter-ahead prediction

Forecast Horizon	MAE	MSE	MAPE	MPE
Quarters 3 and 4 1985	58.42	3697.04	8.00	-8.00

MTSI-Step 7 - MTSI-Step 8: Long-time-scale yearly model and prediction

The predicted end-point value, Quarter 4 of 1985, was previously obtained in Section 5.3.2 in the adjustment of Quarter 1 to Quarter 4 of 1985 of the one-year-ahead case, where Table 5.8 gives the percentage accuracy associated with this forecast. The predicted sum value, Σ(Quarter 3 to Quarter 4 of 1985), is obtained using the predicted sum value determined in the 1985 one-year-ahead case, where this is as follows:

$$\hat{Y}_S = \hat{Y}_{S\_1985} - \sum_{k=QR1}^{QR2} y(k) \tag{5.31}$$

where $y(k)$ is the actual quarterly fuel consumption series, $\hat{Y}_{S\_1985}$= Σ(Quarter 1 to Quarter 4 of 1985) is the predicted sum value obtained for the 1985 one-year-ahead case and $\hat{Y}_S$ denotes the sum value required for the current case.

MTSI-Step 9: Selection of possible set of *r* and description of weighted least squares problem

The possible values of the parameter *r* were dealt with in Section 5.3.2. However, in order to demonstrate the underdetermined case the possible choices of the parameter are R1 to R4. Table 5.27 describes the characteristics of the weighted least squares formulation for this case.

Table 5.27 Description of weighted least squares problem

r	(*n*-*r*) No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R4	4	2	sum and end-point	3	underdetermined

MTSI-Step 11: Selection of possible set of **W**

The form of the weighting matrix for the two-quarter-ahead case is given in equation (5.32)

$$\mathbf{W} = diag\left[w_{ep} \quad w_{s} \quad w_{dev\_qr1}\right] \tag{5.32}$$

The only decision to be made here is the relative value of the sum and end-point error weights to the deviation error weight w_{dev_qr1}. Different weighting matrices were constructed where the end-point and sum weights set equal to 10 and the deviation error weight was varied from 5 to 15.

MTSI-Step 12: Selection of optimal r and $\mathbf{W}$

Adjustment was carried out using actual sum and end-point data using the choices of the r and $\mathbf{W}$ parameters described in MTSI-Step 11 and MTSI-Step 12. The difference between the adjustment obtained for the different choices of the r parameters was minor (approximately of the order 10^{-12}), with the minimum MAE and MSE occurring for R3. Therefore, as in the one-year-ahead and two-year-ahead cases the optimal choice of r involves freeing the states representing the level and slope of the trend component. The optimal weighting matrix was determined to be $\mathbf{W} = diag[10 \quad 10 \quad 5]$.

MTSI-Step 13: Adjustment of short-time-scale prediction

The adjustment was carried out using the predicted long-time-scale data described in MTSI-Step 8. Table 5.28 gives the forecasting accuracy results and a plot of the actual versus predicted is given in Figure 5.32.

Table 5.28 Adjustment results for two-quarter-ahead

	MAE	MSE	MAPE	MPE
Unadjusted	58.42	3697.04	8.00	-8.00
Adjusted using actual	28.22	885	3.88	-1.73
Adjusted using predicted	48.15	3507.88	6.95	-6.95

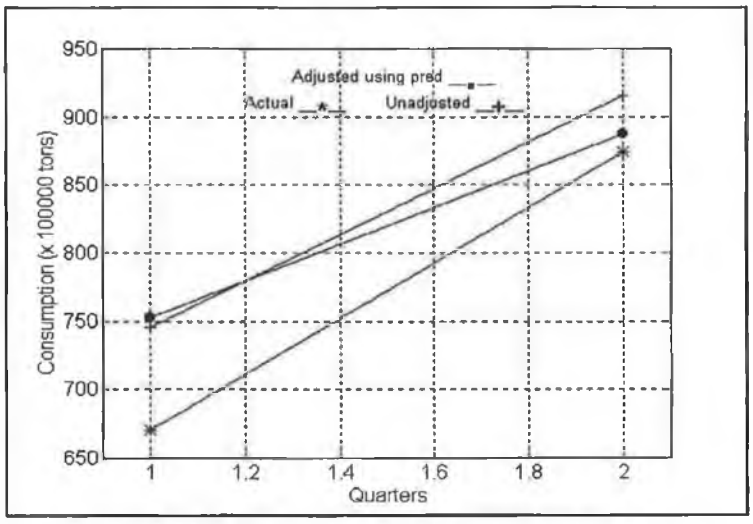

Figure 5.32 Actual vs predicted 2-quarter-ahead case

The graph shows the effect of the adjustment is to move the end-point at Quarter 4 closer to the actual value. The prediction of Quarter 3 in both the unadjusted and adjusted solution are equally inaccurate. The forecasting accuracy results also indicate the achievement obtained at the end-point in the adjusted solution.

5.4 Conclusions

The multi-time-scale integration approach was described and the mathematical formulation of the technique was presented. For clarity a multi-time-scale integration application procedure was developed.

The application of the technique to the quarterly fuel consumption example demonstrated its effectiveness on a variety of prediction horizons. Improvement over the original prediction was achieved in all scenarios. In the case of the 1984 prediction the presence of outliers resulted in a high degree of inaccuracy in the original prediction at these points. The adjustment of the prediction using the multi-time-scale technique significantly improved the overall prediction especially at the outliers. The quarterly fuel consumption time series is influenced by weather and the inclusion of such variables in the short-time-scale model for this example may provide an even more accurate forecast at the outliers. However, at present the technique has only been developed for a univariate model at the short sampling period, where the extension to the multivariate case is an area for future work.

The comparison between the maximum achievable improvement and the improvement obtained using the predicted long-time-scale data provides a measure of the techniques dependency on the accuracy of the predicted long-time-scale information. However, in the case of the 1984 one-year-ahead prediction the variance between the adjustment achieved using actual and that achieved using predicted data is inflated due to the presence of the outliers.

The flexibility of the technique was illustrated where a neural network was employed to model the long-time-scale information. Socio-economic exogenous variables are likely to have an influence on the fuel consumption and their inclusion in the long-time-scale models would have been of benefit. Unfortunately, access to this data was not readily available. However, the application of the technique in this example illustrates that considerable improvement may be attained even when exogenous influencing variables are not included in the models at either time-scale.

The selection of the parameter r was demonstrated in the application. In the case of the structural model the selection of this parameter is more straightforward than in the case of the classical models since the states in the state vector have a physical representation. The optimal choice for r involved freeing the maximum number of states possible in order to attain the sum and end-point specifications. Both states representing the trend component were freed so that the level of end-point could be achieved. In addition, the state representing the fundamental harmonic was also freed so as to allow enough freedom to attain the sum specification.

The application also demonstrated the use a variety of weighting matrices. The selection of this matrix allows the user selective control over the adjustment; particularly in the assignment of the weights to the deviation error terms; in the case of the application various configurations were considered.

CHAPTER 6

Multi-Time-Scale Integration Technique Applications

6.1 Introduction

The multi-time-scale integration technique developed in Chapter 5 was applied to two different electricity demand (MWh) examples. The electricity demand data was obtained from the System A power board which was previously dealt with in Chapter 3, where weather, economic and demographic variables were also available for this data set. The first application involved the combination of weekly and yearly electricity demand forecasts, with the overall objective of producing a one to three-year-ahead forecast of weekly electricity demand. The second application involved the combination of hourly and daily electricity demand forecasts, with the objective of producing a 'next day' (24-hour-ahead) and a three-day-ahead prediction of hourly electricity demand. In each case the multi-time-scale integration application procedure developed in Section 5.2.5 was applied.

6.2 Weekly and Yearly Electricity Demand Integration

For this case, the short and long sampling period time series are made up of weekly and yearly electricity demand respectively. Forecasts produced by the long-time-scale models were used to adjust the weekly forecast. In Chapter 4 weather, economic and demographic variables were employed to forecast yearly electricity sales. In this application these exogenous variables were used as inputs to the models employed to predict the long sampling period yearly electricity demand. Therefore, the adjustment carried out using the long-time-scale forecasts may provide supplemental information not already available at the shorter sampling period. Exogenous weather variables are available at the short sampling period but they were not employed in the short-time-scale model used in this particular application, however this is an area of future work and will be discussed in Chapter 7.

6.2.1 One-Year-Ahead Forecast

MTSI-Step 1: Short-time-scale forecast

Three separate one-year-ahead forecasts of weekly electricity demand were performed over the following forecast horizons:

- Week 30 1987 to Week 29 1988

- Week 30 1988 to Week 29 1989
- Week 30 1989 to Week 29 1990

MTSI-Step 2: Short sampling period time series

The short-time-scale forecast of electricity demand was performed on a weekly sampling period, where Figure 6.1 gives a plot of the electricity demand short sampling period series from July 1982-July 1990, note that the data is scaled for confidentiality reasons.

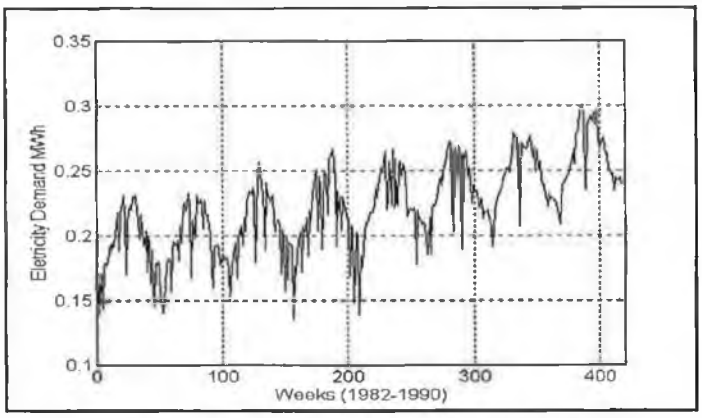

Figure 6.1 Weekly electricity demand

From the plot of time series it may be seen that the salient features are a slowly rising trend due to annual influences and seasonality of reasonably sustained amplitude mainly due to seasonal weather factors, where the seasonal length is 52 weeks. Closer graphical examination of the profile shows that there are changes in the seasonal pattern as the time series progresses through time, where the profile appears to be smoother at the latter end of the time-series. To examine this more closely weekly electricity demand is over-plotted on the same graph for two seasons at the beginning of the time series and for two seasons at the end of the time series, given in Figure 6.2.

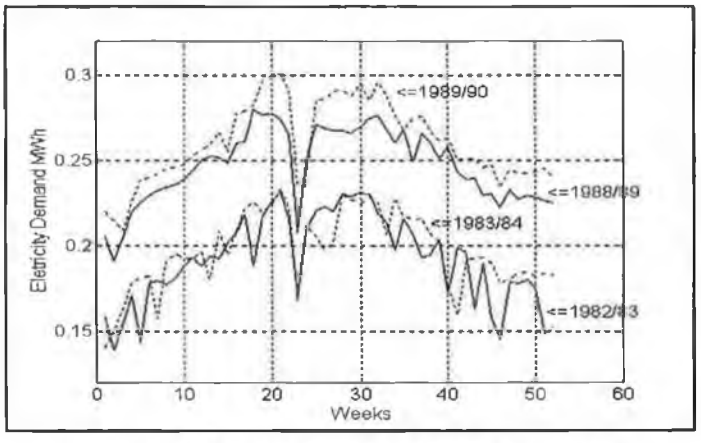

Figure 6.2 Over-plotting of weekly demand

The profile is less erratic in the late 1980's than in the early 1980's, particularly during the months of August to December. The corresponding plots of AT are given in Figure 6.3, where only 1982/1983 and 1989/1990 are presented for visual clarity. There does not appear to be any major differences in AT during these two years and the changes in the seasonal pattern are thus not attributed to this particular weather variable. There may, however, be different weather variables or other influencing factors affecting the series. The plot of AT for each year over the full data set was examined but no distinguishable variance between each year could be determined.

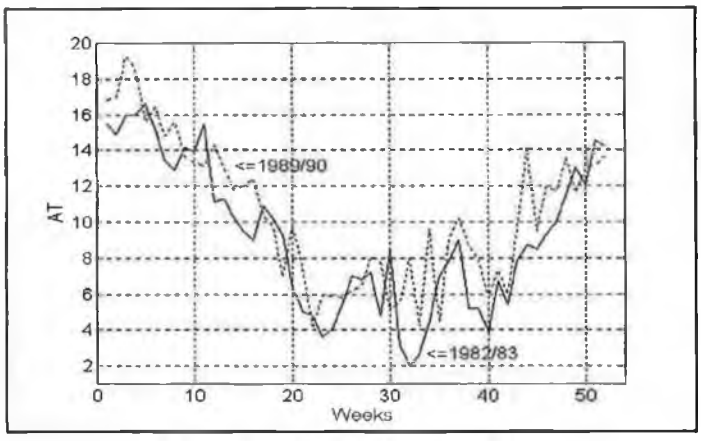

Figure 6.3 Over-plotting of AT

Graphical techniques have shown that the time series has two main components, trend and seasonal and that the smoothness of the seasonal pattern changes over time, where this change is assumed to be attributable to some influencing factor other than the weather.

MTSI-Step 3: Long sampling period time series

Generation of the long sampling period time series which involved making the following selections:

1. The time-scale on which to aggregate the short sampling period time series - yields the sum long sampling period time series.
2. The end-point - yields the end-point long sampling period time series.

The seasonal nature of the weekly short sampling period time series suggests an aggregation of the series up to an annual level. Consequently, the sum time series is made up of the cumulation of a full season of weekly electricity demand data for each year from 1982 to 1990. Based on this selection of aggregation level the end-point was chosen to be one season apart. Since there are no cardinal points in the weekly seasonal profile the end-point may be

chosen arbitrarily and was chosen as the electricity demand in the last week of July. This occurs approximately at Week 29 of the year, where Week 1 is the first week in January and Week 52 is the last week in December.

Therefore, the electricity demand long sampling period sum and end-point time series are based on a yearly sampling period, where Table 6.1 describes each time series.

Table 6.1 Description of sum and end-point series

Sum	Σ(Week 30 year(k-1) to Week 29 year(k)), k=1982 to 1990
End-point	Week 29 of year(k), k=1982 to 1990

Since there are a limited number of data points available at the long sampling period, other variables which are known to have an effect on electricity demand were employed when forecasting the end-point and sum data. In Chapter 3 it was shown that AT and HDD_{18} affects weekly electricity demand and thus these variables were considered as inputs to the models employed to predict the end-point data. In Chapter 4 a number of economic variables were utilised when forecasting yearly 'total' electricity sales and these variables were also considered for use when forecasting the required yearly sum data. The economic variables are GDP, AIW and Average Unit Price (AUP). In addition, a demographic variable, number of customers (NOC), was also considered in the exogenous variable set in order to further increase the size of the data set.

Figure 6.4 gives a plot of the end-point and corresponding weather data, where the data is scaled to the same order of magnitude. It can be seen from the end-point profile that the annual influences result in a moderate increase in demand from one year to the next. In addition, the graph shows the correlation between electricity demand and the weather variables. An exception to this is in 1989 where it is expected that the relatively high value of HDD_{18} would have resulted in a higher value of electricity demand in that week than actually occurred. Figure 6.5 to Figure 6.8 give plots of the sum data with each of the economic and demographic influencing variables. In each case the data is scaled to the same order of magnitude. The annual influences are more significant in the sum case, where it can be seen that the sum series coupled with GDP, AIW and NOC following a rising trend, with AUP following a declining trend.

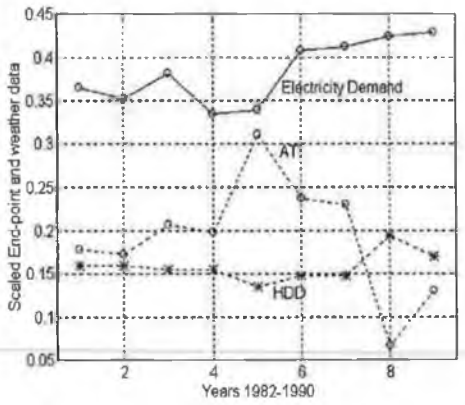

Figure 6.4 End-point and AT series

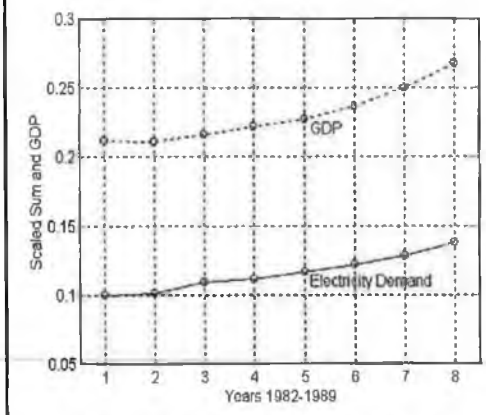

Figure 6.5 Sum and GDP series

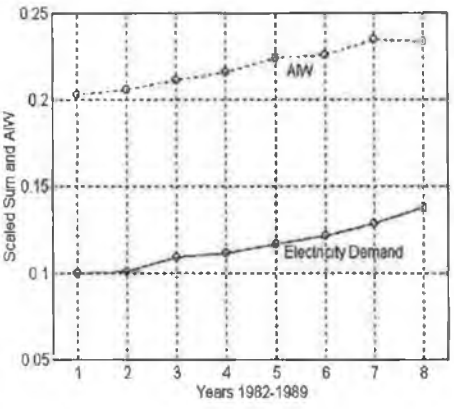

Figure 6.6 Sum and AIW series

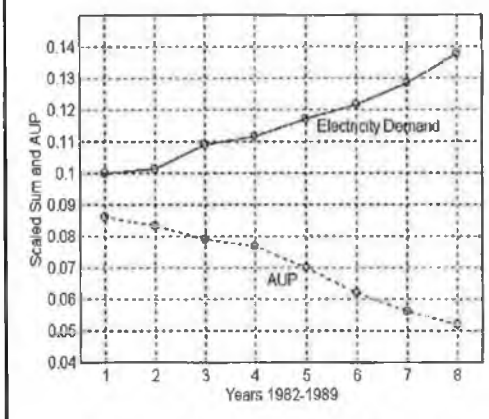

Figure 6.7 Sum and AUP series

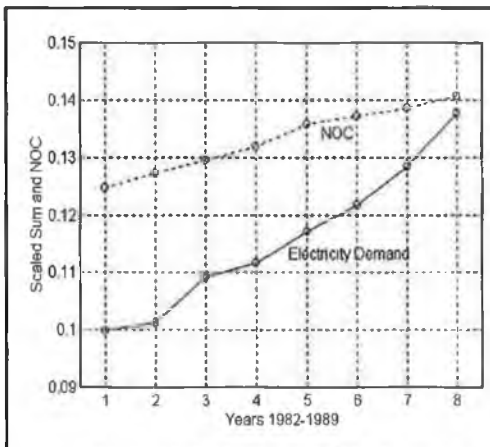

Figure 6.8 Sum and NOC series

MTSI-Step 4: Description of required long-time-scale scale information

The specific form of the long-time-scale information used to carry out the adjustment of the one-year-ahead predictions is described in Table 6.2

Table 6.2 Yearly end-point and sum long-time-scale information for one-year-ahead case

Short-Time-Scale Forecast Horizon	**End-point**	**Sum**
Week 30 1987 - Week 29 1988	Week 29 1988	Σ(Week 30 1987 - Week 29 1988)
Week 30 1988 - Week 29 1989	Week 29 1989	Σ(Week 30 1988 - Week 29 1989)
Week 30 1989 - Week 29 1990	Week 29 1990	Σ(Week 30 1989 - Week 29 1990)

MTSI-Step 5: Short-time-scale model

The short sampling period time series exhibits both trend and seasonal characteristics. A BSM with a dummy seasonal component was employed to model the time series. This type of model was used in the weekly electricity demand application dealt with in Chapter 3 where it outperformed the classical linear statistical models. The state space structure of the model is given by (2.44) and the techniques used to estimate the model are described in Section 2.5.3 and also in Section 3.4.2.2. Note that, the seasonal length is 52 and thus the state vector is of dimension $n = 53$.

An estimate of the state vector at the forecasting origin of the one-year-ahead predictions was required (5.3 and 5.4) to carry out a forecast. The identification data set used to obtain the estimate of the state vector at the forecasting origin of the prediction for each of the one-year-ahead cases is given in Table 6.3.

Table 6.3 Short-time-scale identification data sets and forecast horizon

Forecast Horizon	**Identification Data Set**
Week 30 1987 - Week 29 1988	Week 30 1982 - Week 29 1987
Week 30 1988 - Week 29 1989	Week 30 1982 - Week 29 1988
Week 30 1989 - Week 29 1990	Week 30 1982 - Week 29 1989

Table 6.4 to Table 6.6 describe the different models obtained for each identification data set where the following information is given:

- σ_ε^2 is the estimated variance of $\varepsilon(k)$.
- σ_η^2 is the estimated variance of $\eta(k)$.

- σ_ξ^2 is the estimated variance of $\zeta(k)$.
- σ_ω^2 is the estimated variance of $\omega(k)$.
- $\tilde{\sigma}$ is the one step ahead prediction error variance.
- $Q(T)$ is the Box-Ljung statistic (2.35)

Table 6.4 Structural model for 1982/1987 data set

ID data set	Week 30 1982 - Week 29 1987					
Forecast Horizon	1-year-ahead: predict Week 30 1987 - Week 29 1988					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	*Q(16)*
	$0.170\text{x}10^{-3}$	0	$0.517\text{x}10^{-10}$	0	0.0161	21.65

Table 6.5 Structural model for 1982/1988 data set

ID data set	Week 30 1982 - Week 29 1988					
Forecast Horizon	1-year-ahead : predict Week 30 1988 - Week 29 1989					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	*Q(17)*
	$0.179\text{x}10^{-3}$	0	$0.134\text{x}10^{-10}$	0	0.0161	23.48

Table 6.6 Structural model for 1982/1989 data set

ID data set	Week 30 1982 - Week 29 1989					
Forecast Horizon	1-year-ahead: predict Week 30 1989 - Week 29 1990					
	σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	*Q(19)*
	$0.164\text{x}10^{-3}$	0	$0.132\text{x}10^{-10}$	$0.129\text{x}10^{-5}$	0.0154	22.31

The chi-squared statistics are $\chi^2_{[0.5]}(16) = 26$; $\chi^2_{[0.5]}(17) = 28$; $\chi^2_{[0.5]}(19) = 30$ and thus each of the model were assumed to be adequate based on the Ljung-Box statistics given by Q. For each of the three models σ_η is equal to zero. This is to be expected because the trend of the weekly time series is a steadily rising trend and the effect of $\eta(k)$ is to allow the trend to shift up and down, where the larger the variance the greater the stochastic movement. Also, as expected the variance σ_ζ^2 which determines the change in the slope of the trend component is of approximately the same order of magnitude for each model. In a BSM with dummy seasonal component the greater the value of $\sigma_\omega 2$ relative to σ_ε^2 the more past observations are discounted in the prediction of the seasonal component. The value of σ_ω^2 in the case of the 1982/1987 and 1982/1988 models is zero but is $0.129\text{x}10^{-5}$ in the 1982/1989 model. The reason for the change in σ_ω^2 in the 1982/1989 model is due to the change in the smoothness of

the seasonal pattern toward the end of the 1982/1989 data set, with the model discounting more observations nearer the start of the series where the seasonal pattern is more erratic.

MTSI-Step 6: Unadjusted short-time-scale prediction

For comparative purposes unadjusted short-time-scale predictions were obtained using the BSM models determined in MTSI-Step 5. The forecasting accuracy results are given in Table 6.7

Table 6.7 Forecasting accuracy of unadjusted one-year-ahead forecasts

Forecast Horizon	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
1987/1988	7.669	1.4023	3.46	-0.92
1988/1989	7.788	1.3470	3.22	0.89
1989/1990	6.228	0.6275	2.36	1.49

Comparison of the MSE and MAPE forecasting accuracy results indicate that as the number of data points in the identification data set increases the forecasting accuracy of the predictions produced by the model improves. The MPE, which provides an estimate of accuracy over time but also includes the effect of bias, indicates that as the identification data set increases the model changes from producing forecasts that are that on average higher than actual demand to forecasts that on average lower than actual demand.

MTSI-Step 7: Long-time-scale yearly models

In Chapter 4 it was established that MLP neural networks were capable of modelling 'total' yearly electricity sales. Under the assumption that there is a close association between yearly electricity sales and yearly electricity demand neural networks were employed to model the yearly sum and end-point long sampling period time series. The neural network analysis developed in Chapter 4 was adopted here and the MLP application procedure outlined in Section 4.2.5 was used. Both the sum and end-point cases are dealt with simultaneously.

NN-Step 1: Input and output structure

The output layer of the sum and end-point MLPs consists of a single neuron which represents the current value of the time series. Linear forecasting analysis was used to determine the input structure of the yearly electricity sales MLPs developed in Section 4.4.1 of Chapter 4. In that application it was possible to include a number of autoregressive electricity demand variables in the input structure. However, in the current application due to data limitations the input structure was confined to only a single autoregressive electricity demand variable at lag 1, for both the end-point and sum, and the current value of the weather, economic and demographic exogenous variables.

Therefore, the input structure for the MLP used to forecast the sum long sampling period time series may be described by the following nonlinear equation

$$Y_s(k) = g\left(GDP(k), AUP(k), AIW(k), NOC(k), Y_s(k-1)\right) \tag{6.1}$$

where

- g is a nonlinear function
- $Y_s(k)$ is the sum of weekly electricity demand for year k.
- $GDP(k)$ is the GDP (£m) for year k.
- $AUP(k)$ is the AUP (pence(KWh) for year k.
- $AIW(k)$ is the AIW (pence(KWh) for year k.
- $NOC(k)$ is the number of customers in year k.

In the case of the end-point series results showed that the inclusion of HDD_{18} in the input structure did not produce accurate forecasting results and therefore AT is the only exogenous variable used in this case. Consequently, the end-point MLP input structure may be described by the following nonlinear equation:

$$Y_{ep}(k) = f\left(AT(k), Y_{ep}(k-1)\right) \tag{6.2}$$

where

- f is a nonlinear function
- $Y_{ep}(k)$ is the electricity demand at the end-point in year k.
- $AT(k)$ is the average temperature over the end-point week in year k.

NN-Step 2: Input and output data normalisation

The training data for each MLP was normalised, prior to the construction of the input training vectors, by dividing each value of each time series making up the input vector by the its corresponding NF value. The target output data is normalised to the same order of magnitude as the input data, see Section 4.2.2. The NF values for each of the time series are as follows.

- sum - 1×10^8
- AIW - 1×10^3
- GDP - 1×10^6
- AUP- 1×10^2
- NOC - 1×10^8
- end-point - 1×10^6
- AT - 1×10^2

NN-Step 3: Construct input/output training data sets

A separate MLP was trained for each of the one-year-ahead cases, 1987/1988, 1988/1989, 1989/1990. For each case, a single year was used for the validation data set and the remainder of the data was used for the training data set.

NN-Step 4: Select structure and hidden layer architecture

Based on the motivation given in Section 4.2.2 of Chapter 4 two hidden layers were used in each MLP. An appropriate hidden layer architecture was determined using a similar experimentation analysis to that carried out in the yearly electricity sales application dealt with in Section 4.4.1. However, due to computational time constraints it was not possible to conduct as extensive an analysis and based on the results obtained in Section 4.4.1 the performance of four MLP structures were compared. These structure were 1-3-1, 2-6-1, 3-5-1 and 3-9-1. Initial results showed that the 3-9-1 structure did not produce good forecasting accuracy results and this structure was subsequently dropped from the set. The MLPs were trained using the LMS gradient technique with backpropagation, with an adaptive learning rate (initial value of 1×10^{-3}) and a momentum constant equal to 0.95. It was established that it was necessary to train the sum MLPs for approximately 100000 fixed training epochs and the end-point MLPs for approximately 120000 epochs. Each training run took approximately 15 hours in the case of the sum network and 18 hours in the case of the end-point network, run on a 33 MHz IBM Compatible 486 PC with 16 MB of RAM, and thus it was not practical to re-train the networks more than four times. Determination of the training cessation point was performed through the examination of the single-step-ahead MAE over the validation set. The results of the analysis for each of the three one-year-ahead cases are summarised in Table 6.8 for the sum MLPs and Table 6.9 for the end-point MLPs, where based on the analysis developed in Section 4.4.1 of Chapter 4 the average values of the results produced over the four training runs are given.

Table 6.8 Comparison of different sum MLPs

	1987/1988 Case		1987/1988 Case		1987/1988 Case	
	Average Over 4 Training Runs		**Average Over 4 Training Runs**		**Average Over 4 Training Runs**	
Structure	**Single-Step-Ahead MAE (Validation Set)**		**Single-Step-Ahead MAE (Validation Set)**		**Single-Step-Ahead MAE (Validation Set)**	
	MAE	@ Epoch	MAE	@ Epoch	MAE	@ Epoch
1-3-1	0.04×10^2	65927	0.26×10^2	79840	2.89×10^2	79472
2-6-1	7.76×10^3	75438	1.68×10^3	79754	0.41×10^3	72053
3-5-1	4.77×10^3	79346	0.03×10^3	59809	1.13×10^3	59316

Table 6.9 Comparison of different end-point MLPs

	1987/1988 Case		1987/1988 Case		1987/1988 Case	
	Average Over 4 Training Runs		**Average Over 4 Training Runs**		**Average Over 4 Training Runs**	
Structure	**Single-Step-Ahead MAE (Validation Set)**		**Single-Step-Ahead MAE (Validation Set)**		**Single-Step-Ahead MAE (Validation Set)**	
	MAE	@ Epoch	MAE	@ Epoch	MAE	@ Epoch
1-3-1	1544	36322	18493	59662	9932	21883
2-6-1	415	26710	11925	47119	363	96795
3-5-1	1094	68488	17421	88879	1551	31808

Based on the above analysis a 1-3-1 network structure was used to predict the future values of the sum long sampling period series and 2-6-1 network was used to predict the required future values of the end-point series.

In order to predict the sum and end-point time series future values of the exogenous variables, AT, GDP, AIW, AUP and NOC were required. Data was available for each of the economic and demographic variables from 1965 to 1990 and for AT from 1982 to 1991. Box-Jenkins linear autoregressive forecasting models (2.19) were used to forecast the exogenous variables. The structure of such models has been described in Chapter 2 Section 2.4.2.1 and the identification and estimation techniques were described in Section 2.4.3. It was found, however, that the linear forecasting results for the GDP time series were poor and thus an MLP was used to model this series which resulted in an improvement in the forecasting accuracy of the predictions. Table 6.10 gives details of the linear Box-Jenkins models used to forecast the exogenous long sampling period time series.

Table 6.10 Exogenous variable autoregressive model structures

Exogenous Variable	**Box-Jenkins Model Structure**
AT	SARI(4,1,0)(1,52,0)
AIW	ARI(6,2,0)
AUP	ARI(1,2,0)
NOC	ARI(3,2,0)

Details of the neural network analysis used to forecast GDP are as follows:

NN-Step 1: Input and output structure

The output layer of the MLP consists of a single neuron which represents the current value of the GDP time series. The input structure to the MLP used to forecast GDP was adapted from the linear analysis and can be described by the following nonlinear equation

$$GDP(k) = h\big(GDP(k-1), GDP(k-2), GDP(k-2)\big) \tag{6.3}$$

where is a nonlinear function and $GDP(k)$ is the GDP in year k.

NN-Step 2: Input and output data normalisation

The training data for the MLP was normalised by dividing by the NF factor which was 1 x 10^6 and the target output data was also normalised using this value.

NN-Step 3: Construct input/output training data sets

The training data set for the MLP consisted of GDP data from 1965 to 1984 and the validation set of GDP from 1985 to 1986.

NN-Step 4: Select structure and hidden layer architecture

A two layer MLP was used where an appropriate hidden layer architecture was determined through experimentation, where the 1-3-1, 2-6-1, 3-5-1 and 3-9-1 network structures were compared. As before the MLPs were trained using the LMS gradient technique with backpropagation (Section 4.2.3), with an adaptive learning rate (initial value of 1 x 10^{-3}) and a momentum constant equal to 0.95. The number fixed training epochs was 15000 epochs and the networks were re-trained four times. Table 6.11 gives a the results of the analysis where the average values of the results produced over the four training runs are given.

Table 6.11 Comparison of different sum MLPs

	1987/1988 Case	
	Average Over 4 Training Runs	
Structure	**Multi-Step-Ahead MAE (Validation Set)**	
	MAE	@ Epoch
1-3-1	263.79	11942
2-6-1	314.46	8034
3-5-1	117.93	4936
3-9-1	305.41	6607

Based on the above analysis a 3-5-1 network structure was used to predict the future values of the GDP time series.

MTSI-Step 8: Long-Time-Scale Prediction

The one-year-ahead forecasts of the yearly long-time-scale information described in MTSI-Step 4 were obtained using the long-time-scale models developed in MTSI-Step 7. Table 6.12 gives the percentage accuracy of the forecasts of the yearly sum predictions, with the accuracy of the predictions of the exogenous variables required to obtain the sum forecasts given in Table 6.13. Similar results are presented in Table 6.14 for the predicted end-point values, with the percentage accuracy of the AT values given in Table 6.15.

Table 6.12 Accuracy (%) for predicted yearly sum

Short-Time-Scale Forecast Horizon	Sum	% Error
Week 30 1987 - Week 29 1988	Σ(Week 30 1987 - Week 29 1988)	0.89
Week 30 1988 - Week 29 1989	Σ(Week 30 1988 - Week 29 1989)	0.20
Week 30 1989 - Week 29 1990	Σ(Week 30 1989 - Week 29 1990)	0.06

Table 6.13 Accuracy (%) of predicted economic and demographic exogenous variables

Sum Forecast Required For	GDP	AIW	AUP	NOC
	% Error	% Error	% Error	% Error
Σ(Week 30 1987 - Week 29 1988)	1.92	0.51	2.69	0.90
Σ(Week 30 1988 - Week 29 1989)	0.47	0.95	1.04	1.62
Σ(Week 30 1989 - Week 29 1990)	0.01	0.29	0.12	0.25

Table 6.14 Accuracy (%) for predicted yearly end-point

Short-Time-Scale Forecast Horizon	End-point	% Error
Week 30 1987 - Week 29 1988	Week 29 1988	0.19
Week 30 1988 - Week 29 1989	Week 29 1989	5.17
Week 30 1989 - Week 29 1990	Week 29 1990	0.15

Table 6.15 Accuracy (%) of predicted AT

End-point Forecast Required For	AT
	% Error
Week 29 1988	1.89
Week 29 1989	22.84
Week 29 1990	3.51

Examination of the results show that there is a high degree of accuracy in the yearly sum results for all of the one-year-ahead cases in spite of the poor accuracy of some of the economic and demographic exogenous variable forecasts, where examples of this are in the case of AUP and NOC. In contrast, there is a strong relationship between the accuracy of the predicted AT values and the accuracy of the predicted end-point values. For example, the

accuracy of the prediction of the end-point in Week 29 of 1989 is 5.31% with the accuracy of the AT in that week at 22.84%. This is to be expected since AT is highly correlated with the end-point electricity demand and in addition it is the only exogenous variable used to obtain the end-point prediction. However, in the case of the exogenous variables used to obtain the sum predictions the highest degree of inaccuracy occurs in the NOC variable which is known to be the least correlated with yearly electricity demand. Furthermore, the accuracy of GDP which is a significant influencing variable on yearly electricity demand is relatively good.

MTSI-Step 9: Selection of possible values of *r*

The selection of *r* depends on the structure of the state vector of the state space model. For a BSM with a dummy seasonal component the state vector consists of the following states:

- State 1 = the level of the trend component.
- State 2 = the slope of the trend component.
- State 3 = the current seasonal component.
- States 4 to 53 = past values of the seasonal component.

Fixing states 1 to 3 in the state vector yields the original unmodified solution and therefore there are only three choices available when choosing *r*. The choices are as follows:

R1: fix state 1 and free states 2 to 53.

R2: fix state 2 and free states 1 and states 3 to 53.

R3: fix states 1 and 2 and free states 3 to 53.

MTSI-Step 10: Form of weighted least squares solution

The form of the weighted least squares solution depends on the selection of the parameter *r* described in MTSI-Step 9 above, the length of the forecast horizon and also on the number of end-point and sum values used in the adjustment. Table 6.16 describes the form of the least squares solution for each of the different cases R1 to R3, where it can be seen that there is an overdetermined system of equations in all cases.

Table 6.16 Description of weighted least squares solution

r	(*n*-*r*) No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R2	52	156	sum and end-point	53	overdetermined
R3	51	156	sum and end-point	53	overdetermined

MTSI-Step 11: Construction of W

For the adjustment of a one-year-ahead prediction the weighting matrix was chosen to be of the following form:

$$W = diag\left[w_{ep} \quad w_s \quad w_{dev\_wk1} \quad w_{dev\_wk2} \quad . \quad . \quad . \quad . \quad w_{dev\_wk51}\right] \tag{6.4}$$

where

- w_{ep} is the weight on the end-point error
- w_s is the weight on the sum error.
- w_{dev_wki} represents the weight on the error for the deviation from the original forecast at week i.
- w_{dev_wk1} is the starting value of the deviation error profile.
- w_{dev_wk51} is the finishing value of the deviation error profile.

The *wa*, *wb* and *wc* weighting matrices (see Figure 5.10) considered in the fuel consumption example presented in Section 5.3 of Chapter 5 were also considered in this application. Results showed that *wc* matrix, which has a quadratic deviation error weighting profile, did not yield good adjustment results and was therefore not used. An alternative to the *wb* weighting matrix was considered. This involved adapting the *wb* matrix so that the first two deviation errors at the forecasting origin (the starting value and the second deviation error) were assigned the same weight as the weight on the last deviation error (the finishing value). This weighting matrix, which is referred to as *wd*, has a combination of the properties of the *wa* and *wb* matrices.

The setting up the deviation error weighting profile in the *wa* weighting matrix involved the following steps:

1. select a number for the starting value.
2. assign the starting value to w_{dev_wk1} and assign the value 1 to w_{dev_wk51}.
3. assign the values corresponding to the linear interpolation between the starting value and 1 to the deviation errors w_{dev_wk2} to w_{dev_wk50}.

Similarly for *wb*:

1. select a number for the finishing value.
2. assign the finishing value to w_{dev_wk51} and assign the value 1 to w_{dev_wk1}.

3. assign the values corresponding to the linear interpolation between the finishing value and 1 to the deviation errors w_{dev_wk50} to w_{dev_wk2}.

Finally for *wd*:

1. select a number for the finishing value.
2. assign the finishing value to w_{dev_wk51}, w_{dev_wk1} and w_{dev_wk2} and assign the value 1 to w_{dev_wk3}.
3. assign the values corresponding to the linear interpolation between the finishing value and 1 to the deviation errors w_{dev_wk50} to w_{dev_wk4}.

Examples of *wa*, *wb* and *wd* matrices with starting/finishing values equal to 10 and the end-point and sum error weight equal to 100 are shown as follows:

$$
\begin{array}{lllllllllllll}
W = diag[& w_{ep} & w_s & w_{dev\_wk1} & w_{dev\_wk2} & w_{dev\_wk3} & . & . & w_{dev\_wk49} & w_{dev\_wk50} & w_{dev\_wk51} &] & \\
wa = diag[& 100 & 100 & 10 & 9.82 & 9.64 & . & . & 1.36 & 1.18 & 1 &] & \\
 & & & & & & & & & & &] & (6.5) \\
wb = diag[& 100 & 100 & 1 & 1.18 & 1.36 & . & . & 9.64 & 9.82 & 10 &] & \\
wd = diag[& 100 & 100 & 10 & 10 & 1 & . & . & 9.67 & 9.86 & 10 &] &
\end{array}
$$

The *wa*, *wb* and *wd* weighting matrices with different starting/finishing values in the range 1 to 20, were constructed. Note that, in each of the matrices a stating/finishing value equal to 1 corresponds to a constant deviation error weighting profile, that is each deviation error is assigned a value of 1. The weights on the end-point and sum errors were given large values in relation to the weights on the deviation error. The ratio of the order of magnitude of the end-point and sum error weights to the deviation error weights was approximately equal to 10/1, as discussed in Section 5.2.4.2. Therefore, the values assigned to the end-point and sum error weights in all the *wa*, *wb* and *wd* weighting matrices were $w_{ep} = w_s = 100$.

MTSI-Step 12: Determination of optimal *r* and **W**

The optimal combination of *r* and **W** was determined using similar methods to those described in the fuel consumption example described in Section 5.3.2 and Section 5.3.3 of Chapter 5. The MAE and MSE results obtained through the adjustment of the one-year-ahead predictions using actual long-time-scale information were compared for different combinations of the *r* and **W** parameters determined in MTSI-Step 10 and MTSI-Step 11. Table 6.17 show the optimal values of the parameters for the adjustment of each of the one-year-ahead predictions.

In the case of the optimal weighting matrices the corresponding starting/finishing value is given.

Table 6.17 Optimal *r* and *W* for adjustment of one-year-ahead predictions

Forecast Horizon	*r*	W	Starting/Finishing Value on Deviation Error Weighting Profile
1987/1988	R2	*wd*	$W_{dev\_wk51} = 17$
1988/1989	R2	*wa*	$W_{dev\_wk1} = 5$
1989/1990	R2	*wb*	$W_{dev\_wk51} = 2$

For each prediction the optimal choice for the parameter *r* was determined to be R2. This corresponds to fixing the slope of the trend component in the short-time-scale model and freeing the remaining states. A possible reason for this choice may be due the fact that it is necessary to free the level of the trend component in order to achieve the end-point specification. It can seen that the optimal weighting matrix varies across the predictions, thus illustrating that this parameter depends on the specific characteristics of the forecast which is being adjusted. The weighting matrices given in Table 6.17 are the *optimal* weighting matrices, that is the matrices that give the maximum improvement achieved through the adjustment of the original prediction. However, it is important to note that improvement over the original prediction was achieved for all weighting matrices in all cases for the choice of parameter *r* equal to R2. For example, consider Figure 6.9 and Figure 6.10 which show the MAE and MSE performance achieved through the adjustment of the original 1987/1988 prediction using the different weighting matrices, *wa*, *wb* and *wd*. The *y*-axis gives the variation in the starting/finishing values for each of the matrices and the solid horizontal line in each graph shows the MAE or MSE of the original unadjusted prediction.

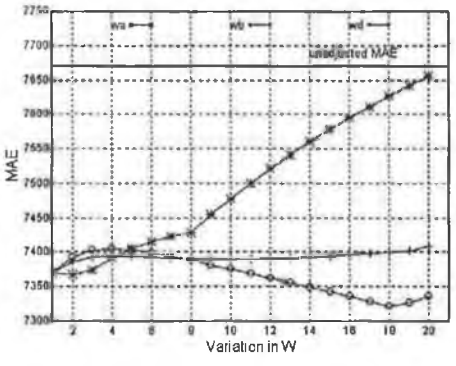

Figure 6.9 Comparison of MAE performance for different W 1987/1988

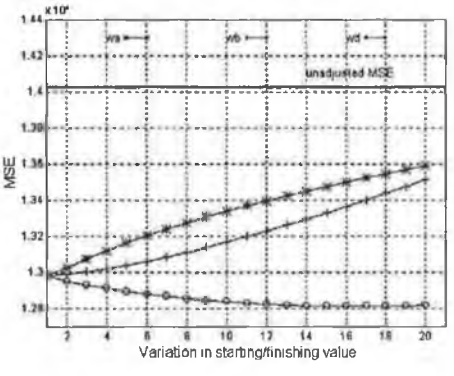

Figure 6.10 Comparison of MSE performance for different W - 1987/1988

The graphs show that for both the MAE and MSE an improvement over the original prediction is achieved through adjustment using any one of the weighting matrices. The maximum improvement achieved using the *wd* matrix is clear from the graphs. Figure 6.11 to Figure 6.14 show similar plots for the 1988/1989 and 1989/1990 predictions.

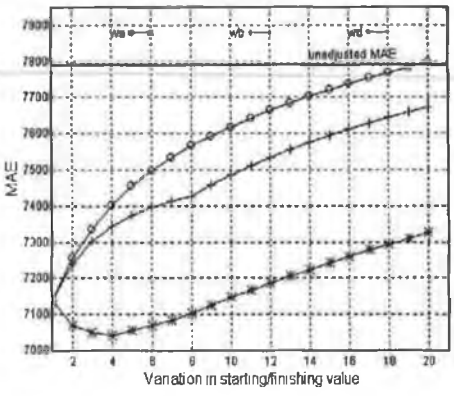

Figure 6.11 Comparison of MAE performance for different **W** -1988/1989

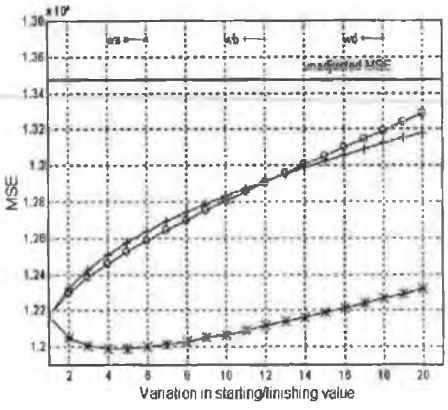

Figure 6.12 Comparison of MSE performance for different **W** 1988/1989

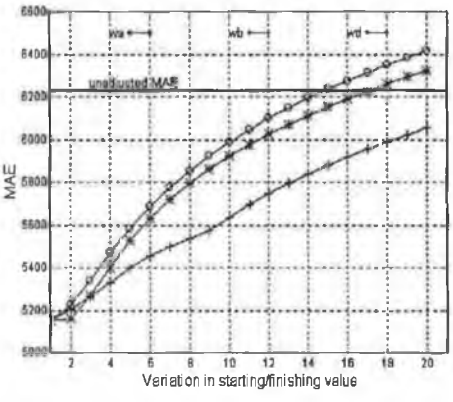

Figure 6.13 Comparison of MAE performance for different **W** -1989/1990

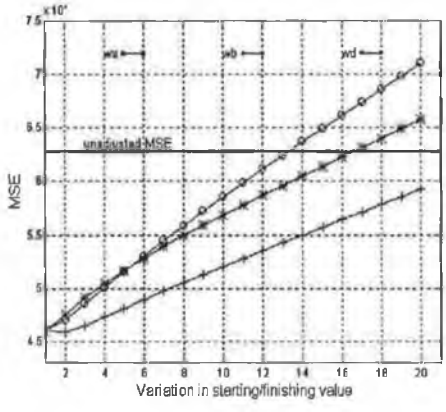

Figure 6.14 Comparison of MSE performance for different **W** - 1989/1990

Figure 6.11 to Figure 6.14 shows that in the case of the *sub optimal* matrices, that is *wb* and *wd* in the 1988/1989 prediction and *wa* and *wd* in the 1989/1990 prediction, that as the starting/finishing values increase from 1 to 20 that the forecasting accuracy of the adjustment decreases. In fact the use of a constant deviation error weighting profile (starting/finishing value equal to 1) yields a more accurate result than the use of a sub optimal matrix that has variation on the deviation error weighting profile.

MTSI- Step 13: Adjustment of short-time-scale prediction

The one-year-ahead short time scale predictions were adjusted using the optimal values of r and **W**, described in Table 6.17 of MTSI-Step 12, using the predicted long-time-scale information obtained in MTSI-Step 8.

Figure 6.15 to Figure 6.17 give the actual vs. predicted graphs for each of the adjusted one-year-ahead predictions. The unadjusted solution is also included for comparative purposes.

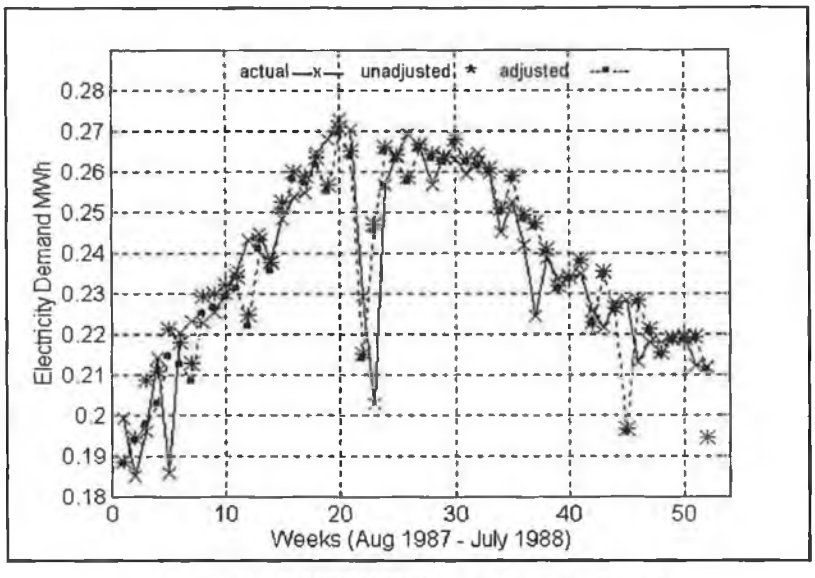

Figure 6.15 1987/1988 Actual vs predicted

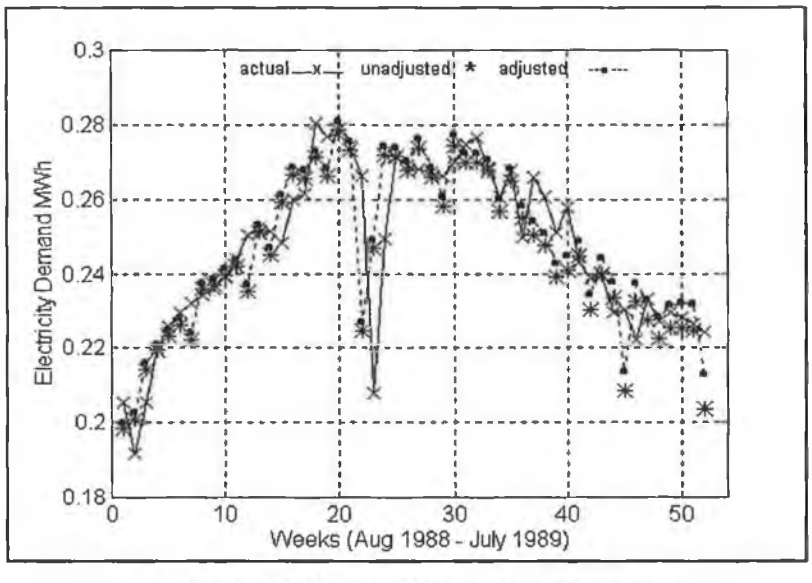

Figure 6.16 1988/1989 Actual vs predicted

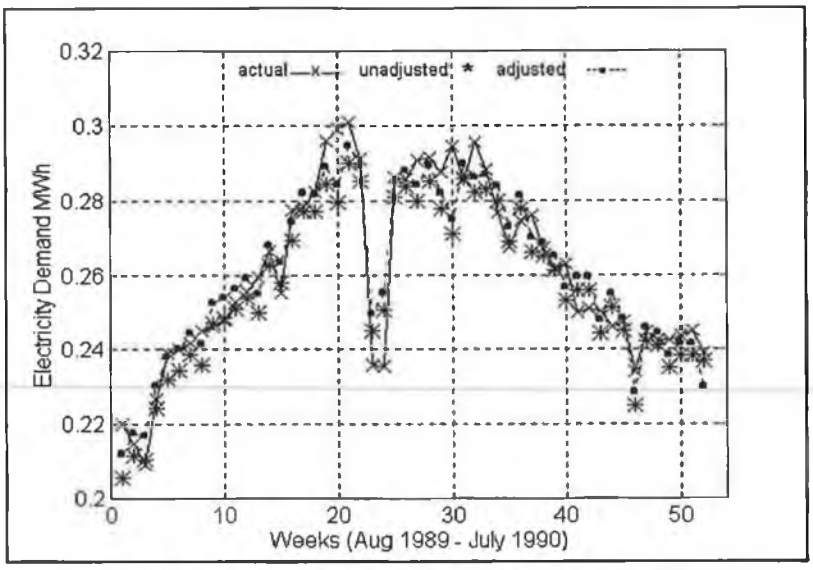

Figure 6.17 1989/1990 Actual vs predicted

Comparison of the profile of the adjusted and unadjusted predictions illustrate the effect of the choice of **W** on the solution. For example, in the 1987/1988 and 1989/1990 predictions the majority of the adjustment occurred during the first 10 weeks of the predictions due to the deviation error weighting profile of the *wd* and *wb* weighting matrices respectively. In contrast, in the 1988/1989 prediction the heavier weighting of deviation errors near the forecasting origin in the *wa* weighting matrix results in the majority of the adjustment occurring during the last 16 weeks of the forecast horizon. Thus, it is can seen that the weighting matrix **W** allows selective control over the adjustment of the original solution. Consequently, it would be possible to choose a particular weighting matrix such that a particular type of adjustment would be performed depending on the forecast requirement.

Further examination of the graphs show that the low error (0.2%) on predicted end-point of the 1987/1988 prediction resulted in approximately an 8% improvement over the original solution at this point. In the case of the 1988/1989 prediction in spite of the high error (5.31%) on the predicted end-point there is still approximately a 4% improvement over the original solution. This demonstrates that although the forecasting accuracy of the predicted end-point value is an important factor in the adjustment, it is still possible to achieve a reasonable degree of improvement even when the end-point error is relatively high. In contrast, the original 1989/1990 prediction produced a more accurate prediction (0.1%) at the end-point than the adjusted solution (3.8%). However, the adjustment of this prediction produced a greater improvement than that achieved in the 1987/1988 and 1989/1988 cases. The improvement over the unadjusted solution is of the order of approximately 14% in the MAE and 24% in the MSE. The reason for this is the combination of the very low errors on

both the sum and (0.06%) and end-point (0.15%) predictions for this case It is expected that the degree of accuracy achieved through the adjustment of the original prediction would be more sensitive to the error on the predicted sum value, since this value affects the overall prediction, where this has been found to be true form practical experience.

Table 6.18 to Table 6.20 gives these results where the results for the unadjusted prediction and adjusted prediction using actual long-time-scale data are also included for comparative purposes. The adjustment of the prediction using actual long-time-scale information provides a measure of the maximum achievable improvement.

Table 6.18 Adjusted and unadjusted results for 1987/1988 one-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	7.669	1.4023	3.46	-0.92
Adjusted using actual	7.328	1.2813	3.26	-0.14
Adjusted using predicted	7.326	1.2795	3.27	-0.22

Table 6.19 Adjusted and unadjusted results for 1988/1989 one-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	7.788	1.3470	3.22	0.89
Adjusted using actual	7.055	1.1991	2.92	-0.14
Adjusted using predicted	7.399	1.2305	3.07	-0.33

Table 6.20 Adjusted and unadjusted results for 1989/1990 one-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	6.228	0.6275	2.36	1.49
Adjusted using actual	5.149	0.4587	2.00	-0.10
Adjusted using predicted	5.392	0.4765	2.08	-0.14

Examination of the forecasting accuracy measures shows that improvement over the original prediction is achieved in all cases for all forecasting accuracy measures. In addition, for all cases the improvement in the MSE is approximately twice the order of the improvement achieved in the MAE. The reason for this is that the MSE penalises large errors, where examples of these are at Weeks 23, 45 and the end-point in the 1987/1988 and 1988/1989 predictions and at weeks 19 to 21 and at Weeks 30 to 34 in the 1989/1990 prediction. Examination of the MPE values shows that the effect of the adjustment is to produce a solution that tends to forecast on average higher than actual values. With the exception of the 1987/1988 prediction the error on the sum and end-point information resulted in approximately a 5% loss in the maximum achievable improvement. In the case of 1987/1988 prediction the maximum improvement possible was attained to within 0.02 %.

6.2.2 Three-Year-Ahead Forecast

MTSI-Step 1 to MTSI-Step 3

A three year multi-step-ahead forecast of weekly electricity demand from Week 30 1987 to Week 29 1990 was performed. The short and long sampling period time series are the same as in the one-year-ahead case dealt with in Section 6.2.1.

MTSI-Step 4: Description of long-time-scale scale information

The long-time-scale information used to carry out the adjustment of the three-year-ahead prediction is described in Table 6.21.

Table 6.21 Long-time-scale information used for adjustment of three-year-ahead prediction

Short-Time-Scale Forecast Horizon	End-point	Sum
Week 30 1987-Week 29 1990	1. Week 29 1988	1. Σ(Week 30 1987 - Week 29 1988)
	2. Week 29 1989	2. Σ(Week 30 1988 - Week 29 1989)
	3. Week 29 1990	3. Σ(Week 30 1989 - Week 29 1990)

MTSI-Step 5: Short-time-scale model

The BSM with dummy seasonal component that was used to carry out the one-year-ahead 1987/1988 prediction was used to obtain the three-year-ahead prediction from 1987/1990. The identification data set and forecast horizon are given in Table 6.22

Table 6.22 Short-time-scale identification data sets and forecast horizon

Forecast Horizon	Identification Data Set
Week 30 1987 - Week 29 1990	Week 30 1982 - Week 29 1987

The model for this identification data set was given earlier in Table 6.4.

MTSI-Step 6: Unadjusted short-time-scale prediction

The forecasting accuracy results of the unadjusted three-year-ahead forecast obtained using the BSM with dummy seasonal component model are given in Table 6.23.

Table 6.23 Unadjusted three-year-ahead prediction

Forecast Horizon	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
1987/1990	8.621	1.7358	3.61	-0.82

For comparative purposes Table 6.24 gives the forecasting accuracy results for each of the individual years of the three-year-ahead forecast.

Table 6.24 One-year-ahead results for unadjusted three-year-ahead prediction

Single Year of 3-Year-Ahead Forecast	MAE x 10^{-3}	MSE x10^{-8}	MAPE	MPE
1st-step-ahead - 1987/1988	7.669	1.4023	3.46	-0.92
2nd-step-ahead - 1988/1989	8.882	1.6621	3.71	1.17
3rd-step-ahead - 1989/1990	9.314	2.1133	3.67	-0.36

As expected all forecasting accuracy measures are greater than the results obtained in each of the one-year-ahead cases.

MTSI-Step 7 and MTSI-Step 8: Long-time-scale models and predictions

A multi-step-ahead forecast of the end-point and sum long-time-scale information described above in MTSI-Step 4 was obtained using the neural networks developed in Section 6.2.1 of the one-year-ahead case. The percentage accuracy associated with each step in the multi-step-ahead prediction is given, where Table 6.25 shows the results obtained for the three-step-ahead forecast of the sum series and Table 6.26 gives the results for the associated exogenous variables. Similar results are shown in Table 6.27 and Table 6.28 for the end-point and AT series.

Table 6.25 Accuracy (%) for predicted yearly sum

Step-Ahead of Forecast	Sum	% Error
1st-step-ahead	1. Σ(Week 30 1987 - Week 29 1988)	0.89
2nd-step-ahead	2. Σ(Week 30 1988 - Week 29 1989)	0.19
3rd-step-ahead	3. Σ(Week 30 1989 - Week 29 1990)	2.90

Table 6.26 Accuracy (%) of predicted exogenous variables used to obtain forecast of sum

Step-Ahead of Forecast	GDP	AIW	AUP	NOC
	% Error	% Error	% Error	% Error
1st-step-ahead	1.92	0.51	3.50	0.90
2nd-step-ahead	5.31	0.53	3.54	2.45
3rd-step-ahead	6.95	0.35	0.12	2.58

Table 6.27 Accuracy (%) for three-year-ahead prediction of end-point

Step-Ahead of Forecast	End-point	% Error
1st-step-ahead	1. Week 29 1988	0.19
2nd-step-ahead	2. Week 29 1989	4.16
3rd-step-ahead	3. Week 29 1990	4.50

Table 6.28 Accuracy (%) of predicted AT used to obtain forecast of end-point

Step-Ahead of Forecast	AT
	% Error
1st-step-ahead	1.89
2nd-step-ahead	21.7
3rd-step-ahead	11.87

For the most part the error on the predictions obtained from the three-step-ahead forecast are greater than the errors on the single step-ahead predictions obtained in the one-year-ahead case in Section 6.2.1, one exception is the 2nd step of the 1987/1990 sum prediction.

MTSI-Step 9: Selection of possible values of *r*

The structure of the state vector in the BSM with dummy seasonal component short-time-scale model yields three possible choices for *r* and these choices were described previously in MTSI-Step 9 of Section 6.2.1.

MTSI-Step 10 Description of weighted least squares problem

Table 6.29 describes the form of the weighted least squares solution of the three-year-ahead case.

Table 6.29 Form of weighted least squares solution

r	(*n-r*) No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R2	52	156	sum and end-point	159	overdetermined
R3	51	156	sum and end-point	159	overdetermined

The number of equations in the weighted least squares solution has increased relative to the one-year-ahead case with the increase in the forecast horizon, with an overdetermined system of equations for all choices of the parameter *r*.

MTSI-Step 11: Construction of set of *W*

For the adjustment of the three-year-ahead prediction the weighting matrix was of the following form:

$$W = diag\left[w_{ep\_yr1} \quad w_{ep\_yr2} \quad w_{ep\_yr3} \quad w_{s\_yr1} \quad w_{s\_yr2} \quad w_{s\_yr3} \quad w_{dev\_wk1\_yr1} \quad \cdots \quad w_{dev\_wk51\_yr1} \quad w_{dev\_wk1\_yr2} \quad \cdots \quad w_{dev\_wk51\_yr2} \quad w_{dev\_wk1\_yr3} \quad \cdots \quad w_{dev\_wk51\_yr3}\right] \tag{6.6}$$

where

- w_{ep_yrj} is the weight on the end-point error in year j.
  w_{s_yrj} is the weight on the sum error in year j.
- $w_{dev\_wki\_yrj}$ represents the weight on the error for the deviation from the original forecast at week i in year j.
- $w_{dev\_wk1\_yrj}$ is the starting value of the deviation error profile of year j.
- $w_{dev\_wk51\_yrj}$ is the finishing value of the deviation error profile of year j.

The weighting matrix structures considered in the one-year-head case, *wa*, *wb* and *wd* were also considered here. The weight on the end-point error was more heavily penalised than the weight on the sum error, since the end-point of each year provides the starting point for the next year's prediction, as discussed in Section 5.2.4.2. Therefore, the sum weight was reduced to the same order of magnitude as the deviation error weights and thus the end-point and sum weights were as follows: as w_{ep} = 100 and w_s = 10. For the reasons outlined in the two-year-ahead fuel consumption example in Section 5.3.3 the same weighting values were assigned to the deviation error profile of each year ($w_{dev\_wk1\_yrj}$.....$w_{dev\_wk51\_yrj}$,, j=1:3).

Weighting matrices *wa*, *wb* and *wd* were constructed with different starting/finishing values assigned to the deviation error weighting profile, with the range of values considered from 1 to 20. Results showed that the error was still continuing to decrease for starting/finishing values equal to 20 and the range of values for the starting/finishing values was thus increased from 20 to 50. Therefore, the total set of weighting matrices considered was 150, that is 50 *wa* matrices, 50 *wb* matrices and 50 *wd* matrices.

MTSI-Step 12: Determination of optimal *r* and **W**

The MAE and MSE results obtained through the adjustment of the three-year-ahead prediction using actual long-time-scale information for different combinations of the *r* and **W** parameters compared using the methods described in Section 5.3 of Chapter 5. Table 6.30 given the optimal values of the parameters for three-year-ahead case.

Table 6.30 Optimal *r* and *W* for adjustment of three-year-ahead predictions

Forecast Horizon	***r***	**W**	**Finishing Value on Deviation Error Profile**
1987/1990	R2	*wd*	W_{dev_wk51} = 38

It can be seen that the optimal choice for the parameter *r* is again R2, as in the case of the adjustment of the one-year-ahead predictions, and that *wd* is the optimal weighting matrix. The MAE and MSE achieved using the different weighting matrices are shown in Figure 6.18 and Figure 6.19 respectively. A similar effect to that observed in the one-year-ahead case can be seen, with the improvement achieved using the sub-optimal matrices decreasing as the starting/finishing value increases.

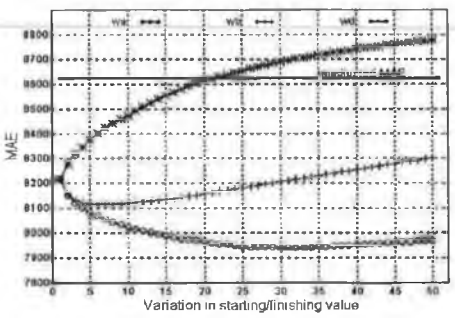

Figure 6.18 Comparison of MAE performance for different **W** -1987/1990

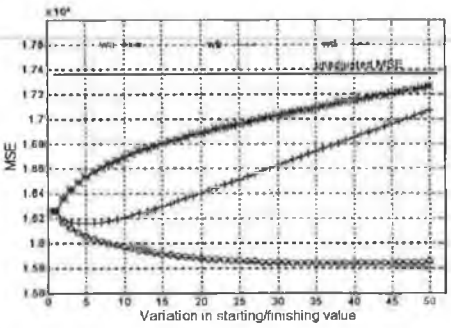

Figure 6.19 Comparison of MSE performance for different **W** - 1987/1988

MTSI-Step 13: Adjustment of short-time-scale prediction

The three-year-ahead forecast was adjusted, using the optimal values of *r* and **W** determined in MTSI-Step 12 above, with the predicted long-time-scale information obtained in MTSI-Step 8. Figure 6.20 gives the actual vs. predicted graphs for the three-year-ahead case. For visual clarity a graph of each year of the three-year-ahead forecast is also given in Figure 6.21 to Figure 6.23.

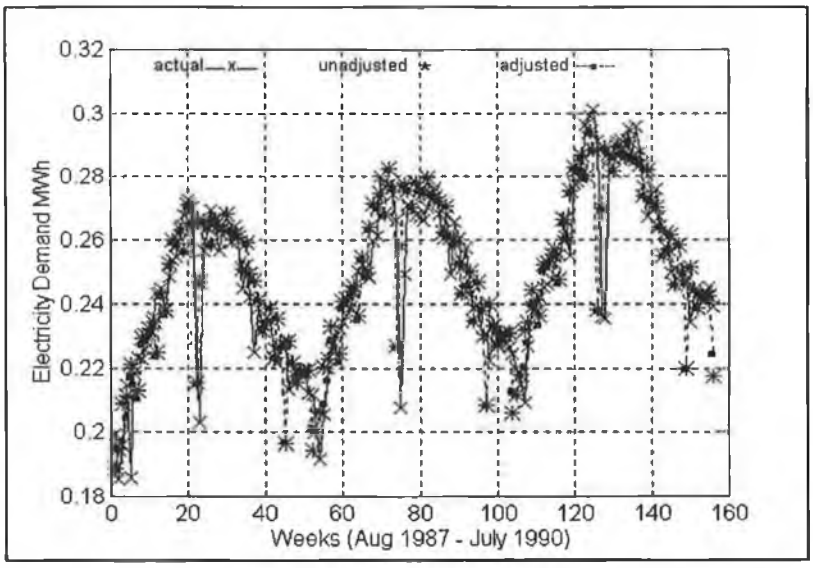

Figure 6.20 1987/1990 - actual vs predicted

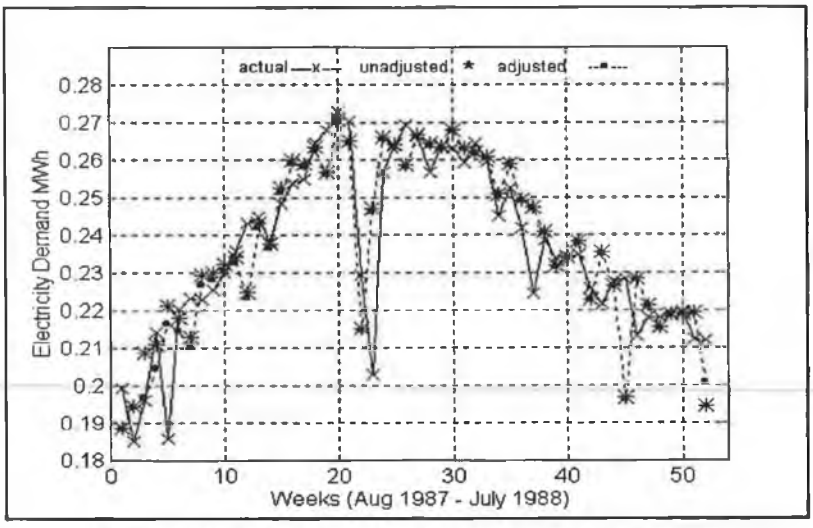

Figure 6.21 Year 1 of 1987/1990 - actual vs predicted

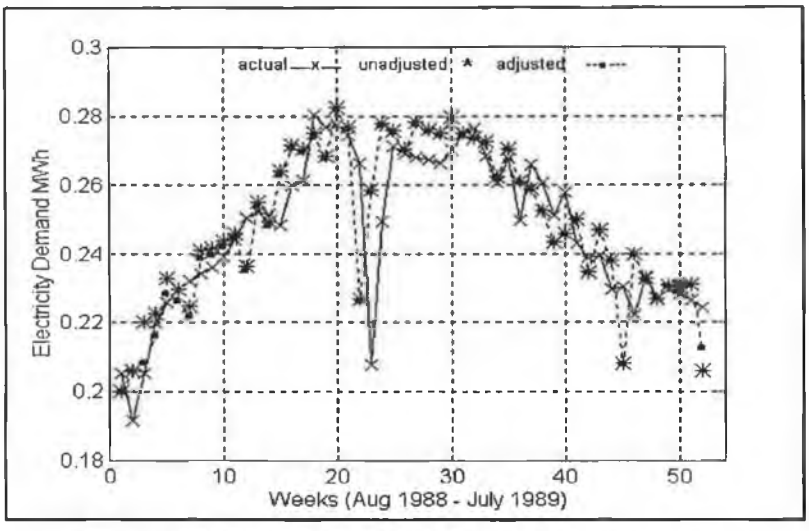

Figure 6.22 Year 2 of 1987/1990 - actual vs predicted

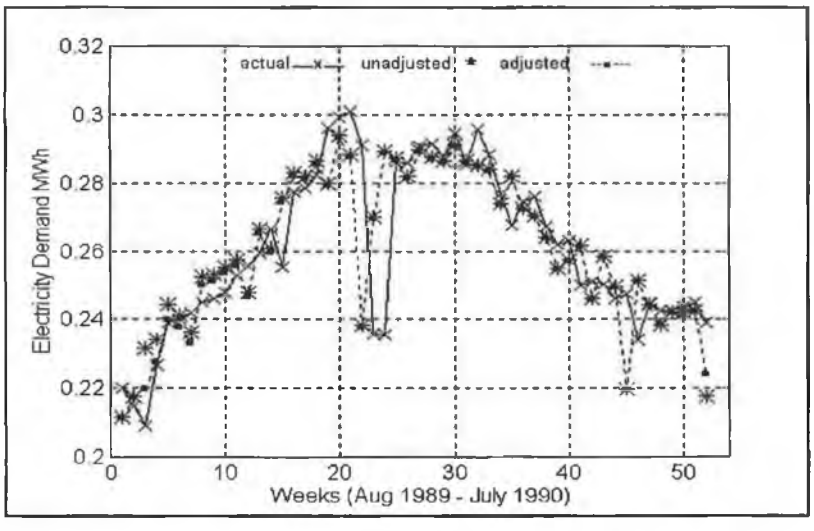

Figure 6.23 Year 3 of 1987/1990 - actual vs predicted

The forecasting accuracy measures are given in Table 6.31 and Table 6.32 to Table 6.34 give the results for each of the individual years of the three-year-ahead adjusted forecast.

Table 6.31 Adjusted and unadjusted results for 1987/1990 three-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	8.621	1.7358	3.61	-0.82
Adjusted using actual	7.9478	1.5838	3.29	-0.27
Adjusted using predicted	8.1595	1.6186	3.39	-0.34

Table 6.32 Results for 1987/1988 of three-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	7.669	1.4023	3.46	-0.92
Adjusted using actual	7.3437	1.2813	3.28	-0.36
Adjusted using predicted	7.4193	1.3028	3.32	-0.43

Table 6.33 Results for 1988/1989 of three-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	8.882	1.6621	3.71	1.17
Adjusted using actual	8.0161	1.5240	3.32	-0.63
Adjusted using predicted	8.2981	1.5533	3.45	-0.69

Table 6.34 Results for 1989/1990 of three-year-ahead prediction

Forecast	MAE $x10^{-3}$	MSE $x10^{-8}$	MAPE	MPE
Unadjusted	9.314	2.1133	3.67	-0.36
Adjusted using actual	8.4807	1.9460	3.29	0.16
Adjusted using predicted	8.7621	1.9997	3.41	0.09

Examination of the graphs of the predictions for each the individual years of the three-year-ahead forecast shows that the effect of the adjustment is to moderately improve the forecast at each step of the prediction over the whole forecast horizon. An exception to this is at the end-point, where relatively speaking there is significant improvement over the original prediction. This smooth improvement over the whole of the forecast horizon is in contrast to the one-year-ahead cases, where there was a variation in the magnitude of adjustment achieved throughout the forecast. The more even adjustment in the three-year-ahead case is required in order to simultaneously satisfy the multi-sum and multi-end-point specifications over the three-year-ahead forecast horizon.

Examination of the results in Tables 6.31 to Tables 6.34 shows that improvement over the original prediction is achieved over all forecasting accuracy measures. Comparing the results for each of the individual years of the three-year-ahead prediction shows that in contrast to the

one-year-ahead predictions comparable improvement was achieved between each year. In addition, the significant improvement in the MSE experienced in the one-year-ahead cases is not seen here. The reason for the above is the even adjustment achieved over the entire forecast horizon.

Comparing the MAPE, which provides a measure of accuracy over time, between the adjusted and unadjusted results for the one-year-ahead and individual years of the three-year-ahead cases shows that on average a greater improvement was achieved in the adjustment of the three year-ahead prediction. This is to be expected since the three-year-ahead prediction will deteriorate as the forecast horizon increases beyond the seasonal length and the effect of the adjustment in this case will have a greater improvement relative to the one-year-ahead cases.

6.3 Hourly and Daily Electricity Demand Integration

In this application the multi-time-scale integration technique was applied to hourly and daily electricity demand short and long period time series respectively. This application is being considered principally for the illustration of the technique as an interpolation technique associated with the short-time-scale load profile. It is employed to interpolate between cardinal points on the daily load profile; such as the tea-time peak, the midday peak and the overnight minimum. The cardinal points at the daily sampling period make up the long-time-scale data. For brevity, future predictions of the cardinal points were obtained using a simple univariate structural model, however a comprehensive model for the production of such forecasts would normally include exogenous weather variables.

Both twenty-four-hour-ahead and three-day-ahead predictions of weekday forecasts during the month the of November were produced. For confidentiality reasons the specific dates associated with each of the forecasts performed are not provided and only the day(s) of the week associated with the forecast is given.

6.3.1 Twenty-Four-Hour-Ahead Forecast

MTSI-Step 1 Required short-time-scale forecast

A 24-hour-ahead prediction of hourly electricity demand was performed for the following forecast horizon:

- 1900 hrs Tuesday to 1800 hrs Wednesday during the month of November.

MTSI-Step 2: Short sampling period time series

The short sampling period time series is hourly electricity demand data. Before proceeding any further the daily load profile is examined. The shape of the daily load profile depends mainly on the month of the year and on the day of the week. Figure 6.24 shows a plot of a typical week-day daily load profile during the month of November

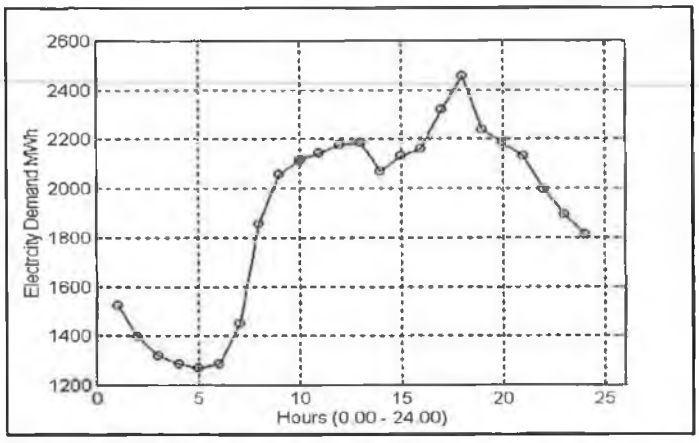

Figure 6.24 Daily load profile for a typical day in November

For this particular day the peak occurs at 1800 hrs and the minimum at 0500 hrs. Other significant turning points occur at lunch-time (1300 hrs) followed by a sharp dip at 1400 hrs. For contrast, Figure 6.25 shows the daily load profile for a Wednesday during the second week of the month for October, November and December.

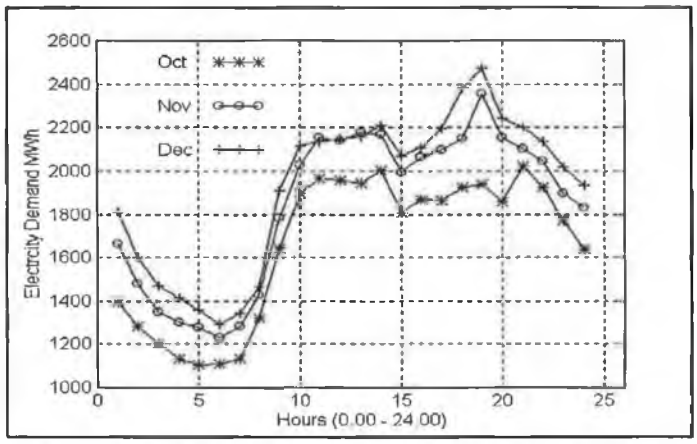

Figure 6.25 Wed. load profile for Oct., Nov. and Dec.

The difference in the October load profile can be seen clearly in the graph, where in particular there is an absence of a tea-time peak load. The load profiles for November and December are reasonably consistent. This is to be expected because it is only the second week of the month and the Christmas factor in December has not yet begun to take effect. Similar

graphical comparisons were carried out for typical week-days for each month of the year. The common characteristics in the daily load profile are as follows:

- An overnight minimum between the hours of 0400 hrs and 0600 hrs.
- A sharp increase between the hours of 0600 hrs and 1000 hrs.
- A plateau between 1100 hrs and 1200 hrs.
- A lunchtime peak at 1300 hrs.
- A sharp decrease after lunch-time at 1400 hrs.
- In winter, the peak load of the day is at 1800 hrs (attributed to higher heating and lighting requirements during the winter months).
- In summer, there is a late high at 2000 hrs or 2100 hrs.

The shape of the daily load profile also varies according to the day of the week, with the main difference occurring between the week-days and the weekend. Figure 6.26 gives a plot of hourly demand for a typical week in November from 000 hrs Monday to 2400 hrs Sunday.

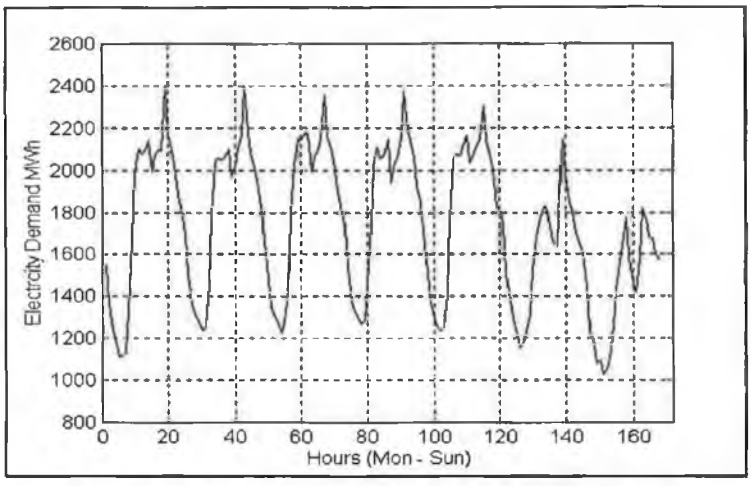

Figure 6.26 One week of hourly data in November

It is easy to distinguish the Saturday and Sunday load profiles due to the pronounced reduction in electricity demand on these days. The graph shows that the load profiles for each week-day are similar with a slight reduction in the load as the week progresses. Furthermore, the peak occurs at 1800 hrs for each day of the week.

Since the 24-hour-ahead prediction falls on a week-day which occurs during the month of November the short sampling period time series is made up of week-day hourly electricity demand data recorded during the month of November. Figure 6.27 shows a plot of the time series for the month of November for two consecutive years. It is seen that the time series

exhibits a trend component which shifts up and down and has a small slope. In addition it has a seasonal component of seasonal length equal to 24.

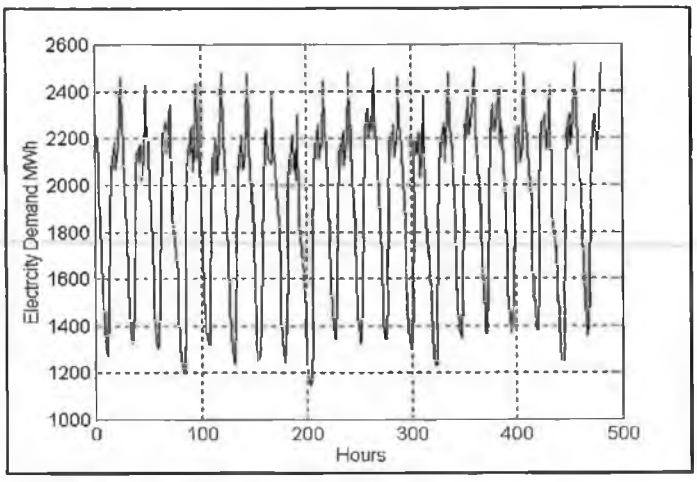

Figure 6.27 Hourly data

MTSI-Step 3: Generate long sampling period series

Given the seasonal nature of the short sampling period time series a daily aggregation level was chosen to generate the sum series. Consequently, the sum time series is made up of the cumulation of hourly electricity demand over a 24 hour period. A number of cardinal points on the daily load profile were chosen as intermediate target values for the adjusted short-time-scale prediction to follow. These points are as follows:

- the peak load of the day at 1800 hrs.
- the 0500 hrs overnight minimum.
- the lunch-time peak at 1300 hrs.
- the 1400 hrs decrease after lunch-time.

One of the above cardinal points was chosen as the end-point and the remaining points as the additional intermediate points (see Section 5.2.3.5). The end-point value was chosen as the peak load of the day at 1800 hrs. The reason for choosing this point as the end-point is for demonstrative purposes since, although it is not the focus of the work carried out in this thesis, generally there will be a reasonably accurate forecast of the peak load of the day available to the utility.

The electricity demand long sampling period sum, end-point and additional point time series are based on a daily sampling period. Table 6.35 describes each time series.

Table 6.35 Description of long sampling period time series

Time Series	Description
Sum	Σ(1900 hrs one day to 1800 hrs of the next day)
End-point	1800 hrs
Additional point 1	1400 hrs
Additional point 2	1300 hrs
Additional point 3	0500 hrs

Each time series is made up of data recorded during the month of November recorded over 9 years. The omission of data for Saturday and Sunday reduces the number of points in the time series from 245 points to only 175 and it was therefore decided to include both week-day and weekend data points in each series. The number of points in each time series is therefore 245 days or 35 weeks. Figure 6.28 to Figure 6.32 show plots of the long sampling period time series.

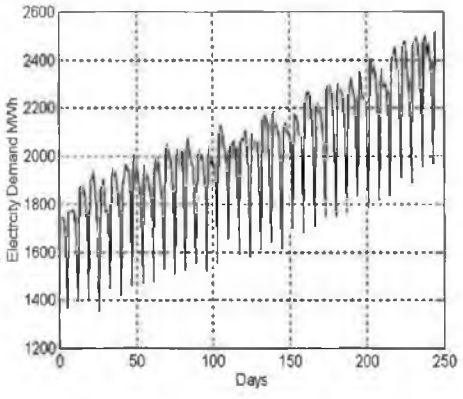

Figure 6.28 End-point time series

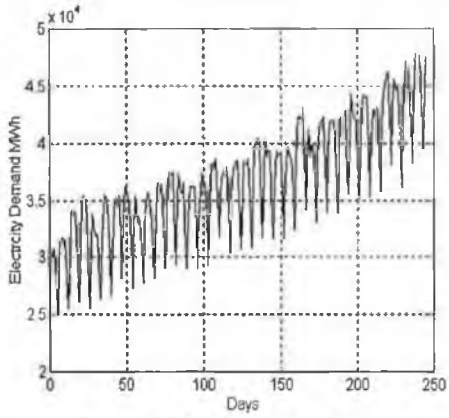

Figure 6.29 Sum time series

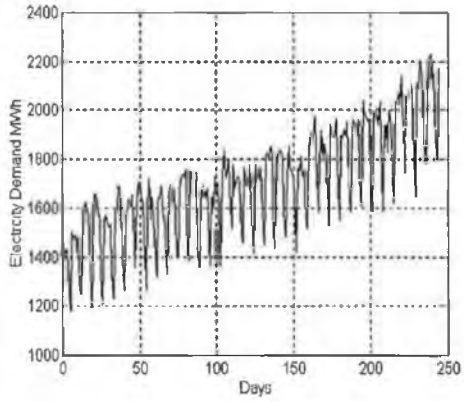

Figure 6.30 1300 hrs time series

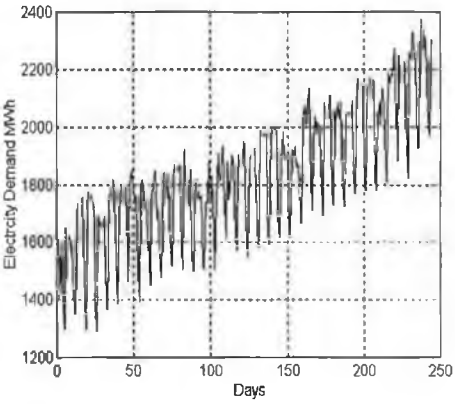

Figure 6.31 1400 hrs time series

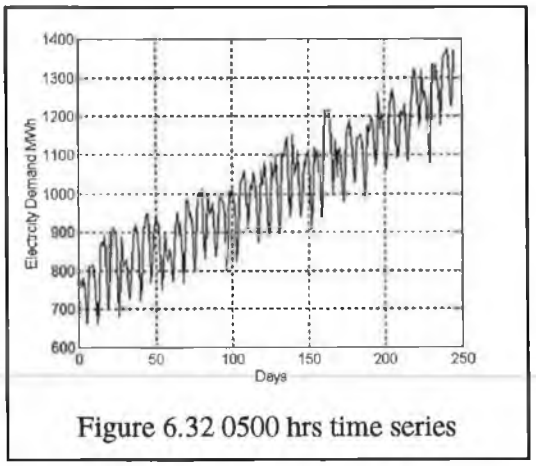

Figure 6.32 0500 hrs time series

Each of the time series exhibit a rising trend due to annual influences and seasonality with a seasonal length of 7, corresponding to one weeks data, as might be expected. The seasonal component may be seen more clearly by over-plotting various weeks on the same graph. Weeks 1, 11, 23 and 35 of the 35 week data set are over-plotted, with Figure 6.33 to Figure 6.37 showing such graphs, where it can be seen that the seven day profile for each week is very similar with the possible exception of Week 23 for the 1400 hrs time series.

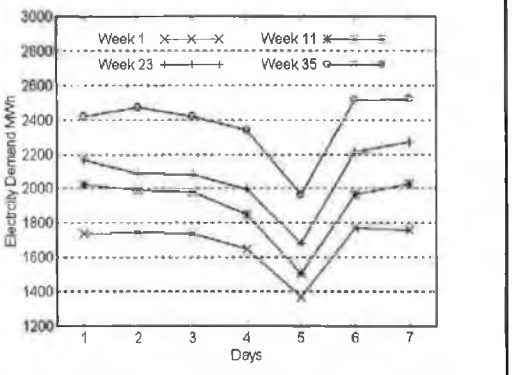

Figure 6.33 Over-plotting for end-point

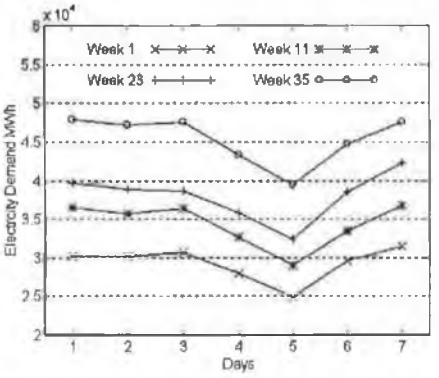

Figure 6.34 Over-plotting for sum

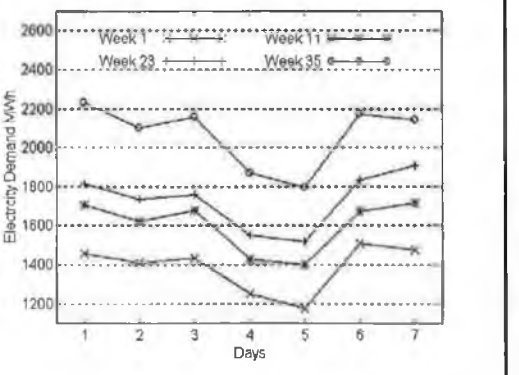

Figure 6.35 Over-plotting for 1300 hrs

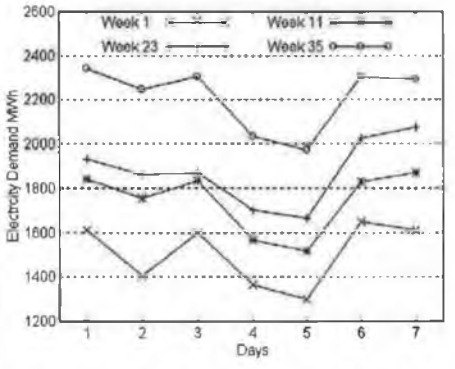

Figure 6.36 Over-plotting of 1400 hrs

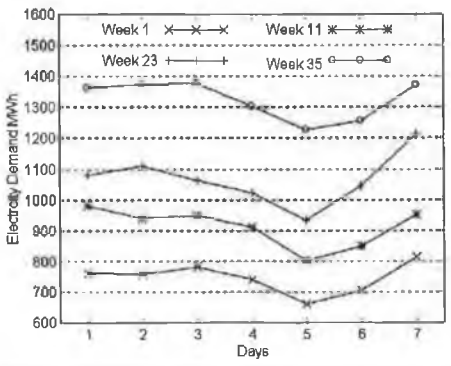

Figure 6.37 Over-plotting of 500 hrs

For this application at the daily sampling period the only exogenous variables available are AT or HDD_{18}. However, given the difficulty associated with predicting these variables at the daily sampling period and since it would be necessary to predict future values of the variables for each of the five models it was decided not to include the exogenous weather variables in the long-time-scale models.

MTSI-Step 4: Description of required long-time-scale scale information

The long-time-scale information that was used in the adjustment of the 24-hour-ahead prediction is described in Table 6.36.

Table 6.36 Long-time-scale information required to carry out adjustment

Short-Time-Scale Forecast Horizon	**End-point and additional points**	**Sum**
1900 hrs Tues to 1800 hrs Wed	1800 hrs, 1400 hrs, 1300 hrs and 0500 hrs on Wed	Σ(1900 hrs Tues to 1800 hrs Wed)

MTSI-Step 5 Short-time-scale model

The hourly short sampling period time series is a seasonal series with a long term trend. An appropriate model structure for a time series exhibiting these characteristics is the BSM with a dummy seasonal component (2.44). Note that, the seasonal length is 24 and thus the dimension of the state vector is $n = 25$.

An estimate of the state vector at the forecasting origin was required to perform the 24-hour-ahead prediction (5.3 and 5.4). The identification data set used to obtain this estimate is made up of 17 days of hourly electricity demand data taken from the month of November. The estimated state space model for this identification data set is described in Table 6.37.

Table 6.37 Short-time-scale model for 24-hour-ahead prediction

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(18)$
0	$0.155\text{x}10^{-4}$	0	$0.004\text{x}10^{-4}$	0.0047	23.67

Since $\chi^2_{[0.5]}(18) = 38$ the model was assumed to be adequate based on the Ljung-Box statistic $Q(18)$. The variance of the stochastic disturbance term corresponding to the slope of the time series is zero as expected and the trend component is allowed to shift up and down due to the relatively large value of σ_η^2. The value of $\sigma_\omega^2 = 0.004\text{x}10^{-4}$ relative to $\sigma_\varepsilon^2 = 0$ indicates that there is a small amount of discounting of past value when predicting the seasonal component.

MTSI-Step 6: Unadjusted short-time-scale prediction

The forecasting accuracy results for the unadjusted 24-hour-ahead prediction obtained using the short-time-scale model determined in MTSI-Step 5 are given in Table 6.30

Table 6.38 Unadjusted 24-hour-ahead forecast

Forecast Horizon	MAE	MSE x 10^{-3}	MAPE	MPE
1900 hrs Tues to 1800 hrs Wed	45.54	2.8074	2.19	1.89

MTSI-Step 7: Long-time-scale models

The long sampling period time series described in MTSI-Step 3 are all seasonal time series with seasonal length equal to 7 and with a rising trend characteristic. Consequently, a BSM with a dummy seasonal component was used to forecast the time series. The identification data set is made up of data taken for the month of November over a period of 9 years. Details of the estimated models are given in Tables 6.39 Table 6.43.

Table 6.39 Daily sum structural model

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(15)$
$0.385\text{x}10^{-10}$	$0.385\text{x}10^{-4}$	0	$0.005\text{x}10^{-5}$	0.0337	17.39

Table 6.40 Daily end-point structural model

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(15)$
$0.188\text{x}10^{-5}$	$0.654\text{x}10^{-5}$	0	$0.006\text{x}10^{-5}$	0.0195	21.98

Table 6.41 Daily 1400 hrs structural model

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(15)$
$0.142\text{x}10^{-4}$	$0.0435\text{x}10^{-4}$	0	$0.004\text{x}10^{-5}$	0.0053	22.34

Table 6.42 Daily 1300 hrs structural model

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(15)$
$0.139\text{x}10^{-4}$	$0.038\ \text{x}10^{-4}$	0	$0.002\text{x}10^{-5}$	0.0051	17.87

Table 6.43 Daily 0500 hrs structural model

σ_ε^2	σ_η^2	σ_ζ^2	σ_ω^2	$\tilde{\sigma}$	$Q(15)$
$0.177\text{x}10^{-5}$	$0.397\text{x}10^{-5}$	0	$0.001\text{x}10^{-5}$	0.0028	20.66

Since $\chi^2_{[0.5]}(15) = 25$ all the models are assumed to be adequate based on the Ljung-Box statistic $Q(15)$. In all cases, the variance corresponding to the slope of the trend component, σ_ζ^2, is zero indicating that each series has a constant slope which does not vary stochastically with time. The relatively large values of the variance σ_η^2 account for the fact the trend component is shifting up and down as the time series progresses through time. In each model the values of σ_ω^2 are less than the values of σ_ε^2 and thus the model does not discount a large number of past observations when estimating the seasonal component.

MTSI-Step 8 Long-time-scale prediction

The forecasting accuracy of the long-time-scale forecasts obtained using the models given in MTSI-Step 7 are given in Table 6.44.

Table 6.44 Accuracy (%) of long-time-scale information for 24-hour-ahead prediction

Long-Time-Scale Forecast	**% Error**
Σ(1900 hrs Tues 12/11/1991 to 1800 hrs Wed 3/11/1991)	1.56
1800 hrs Wed 13/11/1991	2.90
1400 hrs Wed 13/11/1991	1.11
1300 hrs Wed 13/11/1991	3.67
0500 hrs Wed 13/11/1991	2.13

The relatively high errors on the end-point and 1300 hrs forecast suggest that although an adequate model was identified for both time series that either these models could be improved upon or that an alternative modelling approach should be adopted. The use of neural networks is a possible option since as discussed in Chapter 2 they have been widely applied to the area of short term and peak load forecasting. However, the results obtained using the BSM with dummy seasonal component long-time-scale models were used to carry out the adjustment of the short-time-scale prediction under the assumption that any improvement

obtained may possibly be further improved if an alternative modelling approach was adopted for the long sampling period series.

MTSI-Step 9: Selection of possible values of r

The structure of the state vector of the short-time-scale model determines the possible choices available for the parameter r. For the current case the dimension of the state vector is 25 and the choices available for r are as follows:

R1: fix state 1 and free states 2 to 25.

R2: fix state 2 and free states 1 and states 3 to 25.

R3: fix states 1 and 2 and free states 3 to 25.

MTSI-Step 10: Form of weighted least squares solution

The form of the weighted least squares solution, for each of the different cases of R1 to R3, is described in Table 6.45, where for each choice of the parameter r an overdetermined system of equations is obtained.

Table 6.45 Form of weighted least squares solution for 24-hour-ahead prediction

r	$(n-r)$ **No. of Unknowns**	**Forecast Horizon**	**Long-Time-Scale Information**	**No. of Equations in WLS Solution**	**Type of System**
R1 - R2	24	24	sum, end-point and additional points	25	overdetermined
R3	23	24	sum, end-point and additional points	25	overdetermined

MTSI-Step 11: Construction of set of **W**

For the adjustment of the 24-hour-ahead prediction the weighting matrix is of the following form:

$$W = diag\left[w_{ep} \quad w_s \quad w_{addpt} \quad w_{dev\_1900\ hrs} \quad w_{dev\_2000\ hrs} \quad \cdots \quad w_{dev\_1700\ hrs}\right] \qquad (6.7)$$

where

- w_{ep} is the weight on the end-point error.
- w_s is the weight on the sum error
- w_{addpt} is the weight on the additional point error

- $w_{dev\_i00\ hrs}$ represents the weight on the error for the deviation from the original forecast at *i00* hrs.
- $w_{dev\_1900\ hrs}$ is the starting value on the deviation error weighting profile.
- $w_{dev\_1700\ hrs}$ is the finishing value on the deviation error weighting profile.

The weighting matrices *wa*, *wb* and *wd* considered in the weekly/yearly application dealt with in Section 6.2 were also considered here. Different matrices were constructed with starting/finishing values in the range 1 to 20. The weight on the end-point and sum was chosen relative to the deviation errors and was thus equal to $w_{ep} = w_s = w_{addpt}$=100.

MTSI-Step 12: Determination of optimal *r* and **W**

In order to determine the optimal values of the *r* and **W** parameters adjustment of the 24-hour-ahead prediction was carried out with the set of *r* and **W** parameters determined in MTSI-Step 10 and MTSI-Step 11 using actual long-time-scale information. Adjustment was carried out for a number of different combinations of the long-time-scale information, where Table 6.46 describes each of the different cases.

Table 6.46 Different adjustments performed

Long-Time-Scale Data Used for Adjustment	Reference
sum and end-point	PA
sum, end-point and 1300 hrs	PB
sum, end-point and 1400 hrs	PC
sum, end-point and 0500 hrs	PD
sum, end-point, 1300 hrs and 1400 hrs	PE
sum, end-point, 1300 hrs and 0500 hrs	PF
sum, end-point, 1400 hrs and 0500 hrs	PG

The adjustment carried out using PF and PG were not found to give good results and were thus discarded from the set. Using the graphical techniques described in the fuel consumption example in Section 5.3 of Chapter 5 the optimal value of *r* and **W** for each of the cases PA to PE were determined. Table 6.47 summarises the results of this analysis.

Table 6.47 Optimal *r* and **W** for adjustment of 24-hour-ahead prediction

Forecast Horizon	Case	*r*	Weighting Matrix	
			Type	Starting Value on Deviation Error Profile
1900 hrs Tues to 1800 hrs Wed	PA to PE	R2	*wa*	$w_{dev\_1900\ hrs} = 2$

As in the case of the weekly/yearly application the optimal choice of the parameter *r* is R2. The optimal weighting matrix out of the set of weighting matrices was determined to be *wa* but improvement over the original prediction could also be achieved using the *wb* and *wd* weighting matrices. Figure 6.38 and Figure 6.39 show an example of this where the MAE and MSE performances are compared for the different **W** for the adjustment case PA.

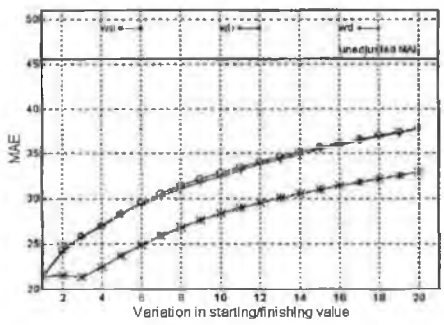

Figure 6.38 Comparison of MAE performance for different **W** -PA

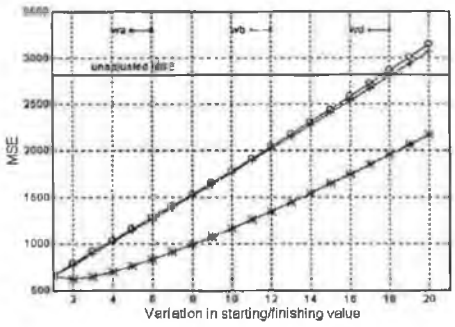

Figure 6.39 Comparison of MSE performance for different **W** -PA

Examination of the profiles show that in contrast to the weekly/yearly application the similarity between the *wb* matrix and the *wd* matrix can be seen in the adjustment results. A possible reason for the difference between the two applications is that the 24-hour-ahead profile is smoother and more consistent than the weekly one-year-ahead profile and thus the effect of the heavier weighting on the first two deviation errors is not as significant as in the weekly/yearly case. In addition, the performance of the *wb* and *wd* sub optimal matrices deteriorates as the starting/finishing values increase from 1 to 20 and as in the case of the weekly/yearly application a finishing value equal to 1 yields the more accurate result.

MTSI-Step 13: Adjustment of short-time-scale prediction

In this step adjustment of the 24-hour-ahead prediction was carried out using the predicted long-time-scale information obtained in MTSI-Step 8. Figure 6.40 to Figure 6.44 give the actual vs. predicted graphs which include the adjusted and unadjusted 24-hour-ahead predictions and Table 6.48 gives the forecasting accuracy results.

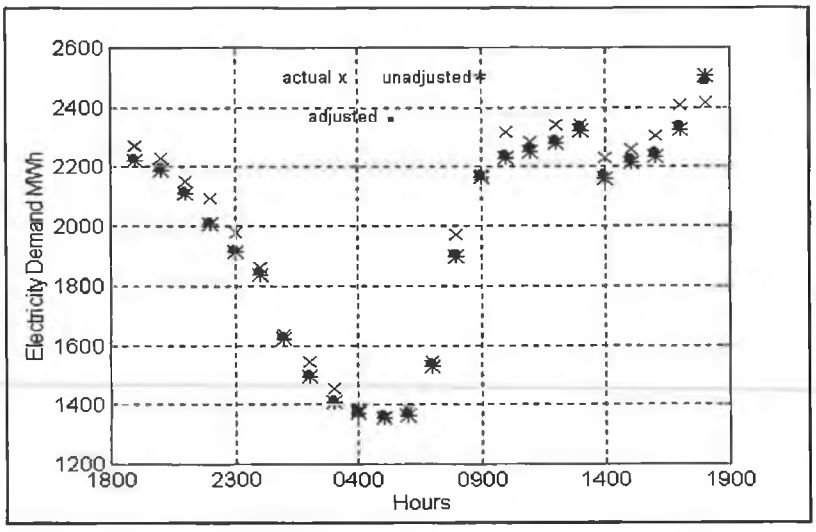

Figure 6.40 Actual vs. predicted for PA

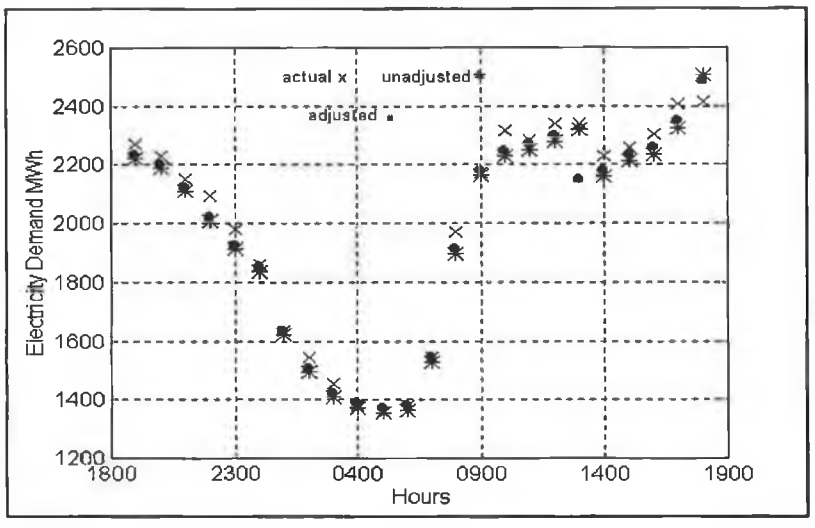

Figure 6.41 Actual vs. predicted for PB

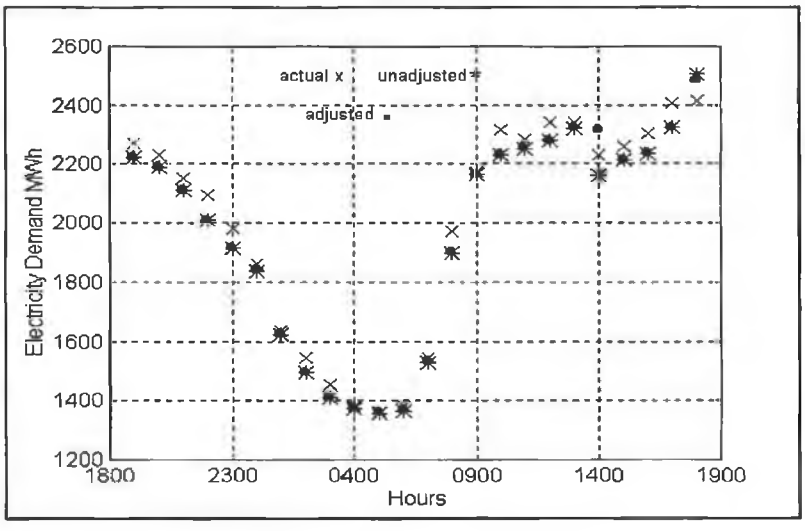

Figure 6.42 Actual vs. predicted for PC

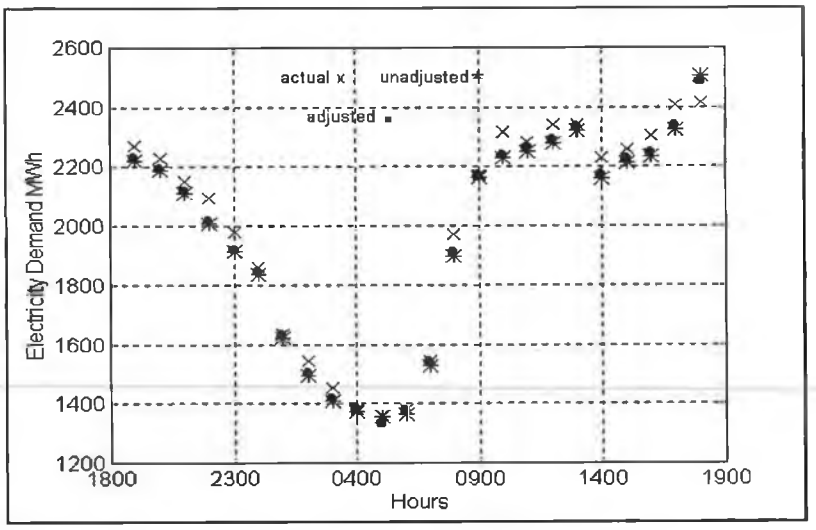

Figure 6.43 Actual vs. predicted for PD

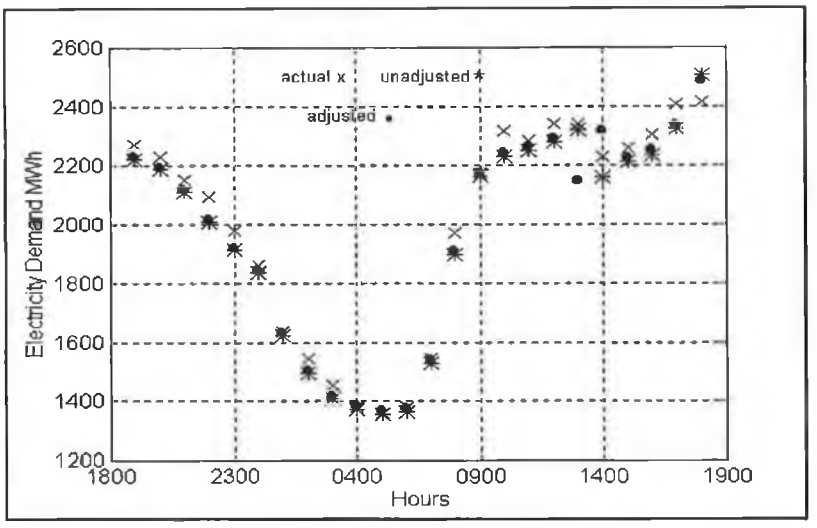

Figure 6.44 Actual vs. predicted for PE

Table 6.48 Results for adjustment of 24-hour-ahead prediction

Forecast	MAE	MSE x 10^3	MAPE	MPE
Unadjusted	45.54	2.8074	2.19	1.89
PA using actual	21.34	652.10	1.14	-0.09
PA using predicted	37.22	2.0407	1.78	1.52
PB using actual	23.85	1.1328	1.27	-0.18
PB using predicted	39.65	3.1053	1.86	1.44
PC using actual	26.16	71.166	1.34	-0.09
PC using predicted	44.12	2.6909	2.11	1.56
PD using actual	19.69	566.30	1.05	-0.09
PD using predicted	37.21	1.9880	1.80	1.54
PE using actual	29.04	1.6280	1.49	-0.14
PE using predicted	44.92	3.6723	2.10	1.48

Before going on to discuss the adjustment results some comments are made regarding the forecast produced by the univariate structural model. Figure 6.45 gives a plot of the actual and predicted load (unadjusted and adjusted PA), with the AT profile for the twenty-four-hour-ahead forecast also shown. For comparative purposes the AT for the previous day is also given.

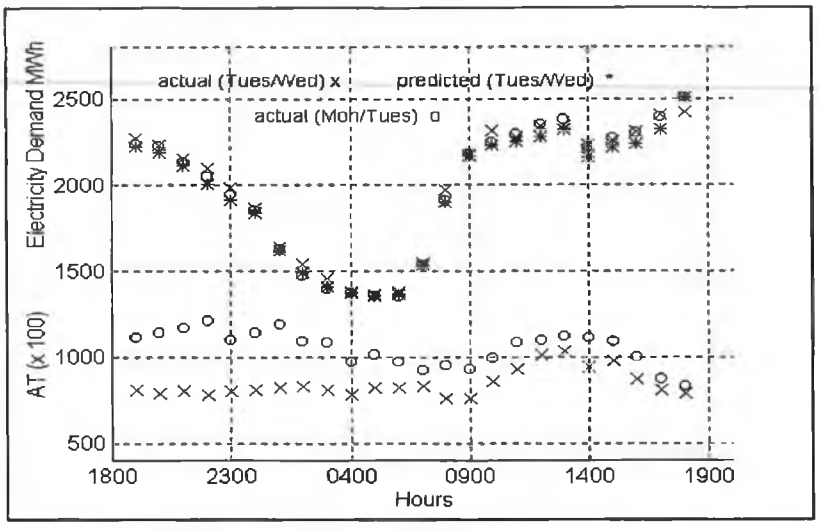

Figure 6.45 Actual vs predicted Tues/Wed including AT data

Examination of the graph shows that the main differences in AT occurs overnight, when weather is not expected to have a significant influence on the load. During the hours of 0900 hrs and 1800 hrs the difference in the AT is small, approximately 2°C and therefore, for this particular day, the univariate structural model was capable of producing a reasonably accurate forecast of the load.

Examining the adjusted forecasts the most accurate results were obtained for the PA and PD adjustments. Examination of the PA and PD graphs show that the effect of the adjustment is to moderately change each hour of the prediction so that it is closer to the actual load. For the PA and PD cases the most significant variance between the actual and predicted profiles occur at 0800 hrs, 1000 hrs, 1200 hrs, 1700 hrs and 1800 hrs. On average an 18 % improvement in MAE and 28% improvement in MSE was achieved, with the MAPE and MPE equal to 1.79 and 1.53 respectively. However, the average maximum achievable improvement that could be obtained in PA and PD cases is of the order of 54% in the MAE, 78% in the MSE, with an MAPE of 1.04 and MPE of -0.09. This indicates that the case for the improvement of the predicted long-time-scale information is strong. Consider Figure 6.46 and Figure 6.47 which show the actual load versus the adjusted load using actual long-time-scale information for PA

and PD respectively, where for comparative purposes the adjustment using predicted long-time-scale data is also included.

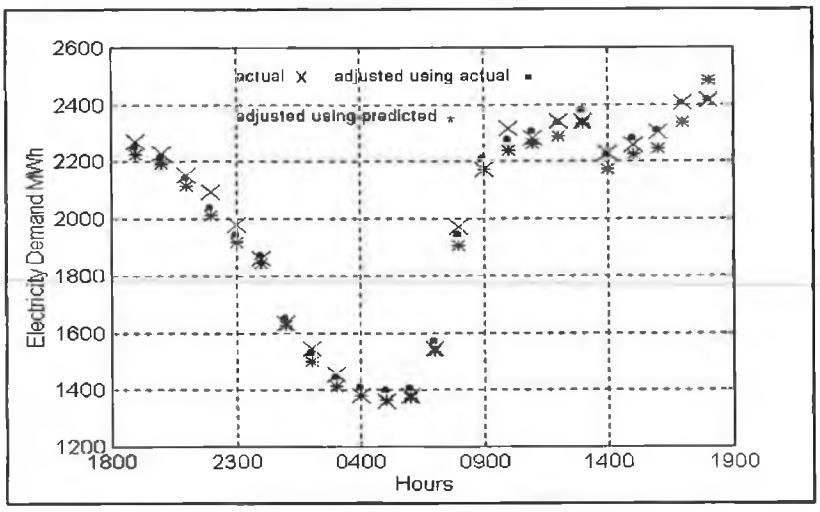

Figure 6.46 Actual vs. adjusted using actual long-time-scale data for PA

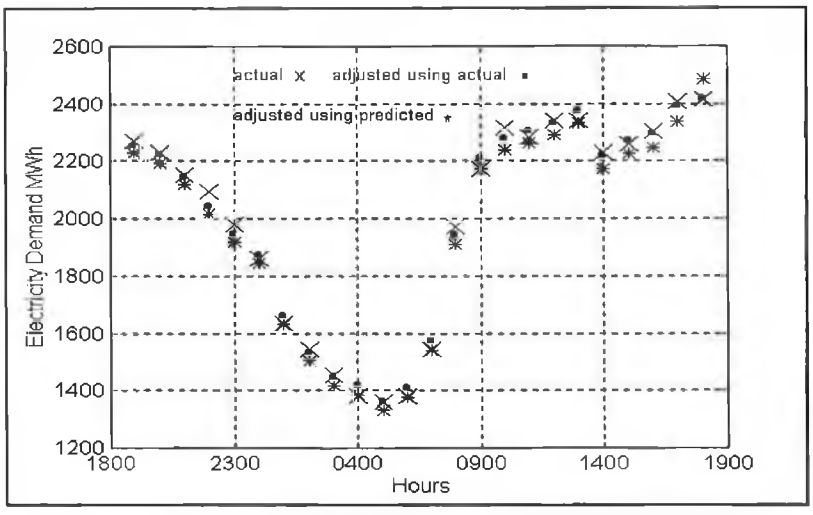

Figure 6.47 Actual vs. adjusted using actual long-time-scale data for PD

Comparison of the adjustment using actual long-time-scale data to the adjustment using predicted long-time-scale data shows that overall, in addition to the obvious improvement at the end-point, the effect of an improvement in the sum information is to push the prediction even closer to the actual load at each hour of the forecast. Of particular significance are the hours at which the greatest error in the adjustment using predicted long-time-scale data occurred, that is 0800 hrs, 1000 hrs, 1200 hrs and 1700 hrs. Exceptions to the overall improvement using the actual data are at 0400 hrs, 0600 hrs, 0700 hrs, 0900 hrs and 1200 hrs. Consequently, since the improvement that could be obtained using more accurate long-time-scale data is significant it is suggested that further improvement of the long-time-scale models would be desirable. The inclusion of a daily weather variable in the structural models

or the use of neural networks are possible directions for future work in this particular application.

In the case of the PB, PC and PE adjustments significant loss of accuracy occurred due to the inaccurate predictions of the additional long-time-scale data, where the difference between the original solution and the adjusted solution at this points is clearly shown on the graph. As expected the forecasting accuracy measures for these solutions are comparable to the original solution. However, examination of the maximum achievable improvement demonstrates again the case for the improvement of the forecasting accuracy of the long-time-scale models. The maximum achievable improvement over the MAE of the original prediction is on average of the order of 42%, of the order of 53% in the MSE, with the MAPE and MPE equal to on average 1.36 and -0.13.

For an arbitrarily chosen day the results have shown that an improvement over the original solution may be obtained through the application of the technique. However, no comment may be made regarding the quality of the results produced through the application of the technique to a set of different days, for example, a weekday occurring during the summer months.

6.3.2 Three-Day-Ahead Forecast

MTSI-Step 1 - MTSI-Step 3

The 24-hour-ahead forecast was extended to a three-day-ahead forecast of hourly electricity demand, that is:

- 1900 hrs Tuesday to 1800 hrs Friday during the month of November.

The short and long sampling period time series used in the 24-four-hour-ahead case were also used in this case.

MTSI-Step 4: Description of required long-time-scale scale information

The long-time-scale information which was used in the adjustment of the three-day-ahead prediction is described in Table 6.49

Table 6.49 Required daily end-point, sum and additional information forecasts

Short-Time-Scale Forecast Horizon	End-point and additional points	Sum
1900 hrs Tues to 1800 hrs Fri	0500 hrs, 1300 hrs, 1400 hrs and 1800 hrs on Wed	Σ(1900 hrs Tues to 1800 hrs Wed)
	0500 hrs, 1300 hrs, 1400 hrs and 1800 hrs on Thur	Σ(1900 hrs Wed to 1800 hrs Thur)
	0500 hrs, 1300 hrs, 1400 hrs and 1800 hrs on Fri	Σ(1900 hrs Thur to 1800 hrs Fri

MTSI-Step 5: Short-time-scale models

The BSM with dummy seasonal component used for the 24-hour-ahead case was used to obtain the three-day-ahead prediction, where a description of the estimated model was given in Table 6.37 in Section 6.3.1.

MTSI-Step 6: Unadjusted short-time-scale prediction

An unadjusted three-day-ahead prediction was obtained using the short-time-scale model described in MTSI-Step 5 above, where the forecasting accuracy result are given in Table 6.50.

Table 6.50 Unadjusted three-day-ahead forecast

Forecast	MAE	MSE x 10^{-3}	MAPE	MPE
1900 hrs Tues to 1800 hrs Fri	47.79	3.0308	2.42	1.34

Since the main interest in the area of short-term load forecasting is in the daily load profile the forecasting accuracy results are calculated for each of the 24 hour periods in the three-day-ahead prediction, where these results are given in Table 6.51.

Table 6.51 Individuals days of unadjusted three-day-ahead forecast

Forecast	MAE	MSE x 10^{-3}	MAPE	MPE
1900 hrs Tues to 1800 hrs Wed	45.54	2.8074	2.19	1.89
1900 hrs Wed to 1800 hrs Thur	54.45	3.6420	2.83	0.79
1900 hrs Thur to 1800 hrs Fri	43.40	2.6429	2.26	1.34

MTSI-Step 7 and MTSI-Step 8: Long-time-scale model and prediction

The long-time-scale models determined in the 24-hour-ahead case, in Section 6.3.1, for each of the long sampling period time series were used to obtain a three-step-ahead forecast of the long-time-scale information. These models were described in Table 6.39 to Table 6.43. The forecasting accuracy results of the three-step-ahead long-time-scale predictions are given in Table 6.52 to Table 6.56. The percentage accuracy associated with each step of the forecast is given.

Table 6.52 Percentage accuracy for sum

Step-Ahead of Forecast	% Error
1st-step-ahead	1.56
2nd-step-ahead	0.22
3rd-step-ahead	0.47

Table 6.53 Percentage accuracy for end-point

Step-Ahead of Forecast	% Error
1st-step-ahead	2.90
2nd-step-ahead	0.28
3rd-step-ahead	0.20

Table 6.54 Percentage accuracy for 1400 hrs

Step-Ahead of Forecast	% Error
1st-step-ahead	1.11
2nd-step-ahead	2.58
3rd-step-ahead	0.65

Table 6.55 Percentage accuracy for 1300 hrs

Step-Ahead of Forecast	% Error
1st-step-ahead	3.67
2nd-step-ahead	1.85
3rd-step-ahead	0.33

Table 6.56 Percentage accuracy for 0500 hrs

Step-Ahead of Forecast	% Error
1st-step-ahead	2.13
2nd-step-ahead	1.17
3rd-step-ahead	2.03

The prediction of the long-time-scale information on the Wednesday are relatively inaccurate in comparison with the long-time-scale predictions of Thursday and Friday, where an exception to this is in the 1400 hrs prediction.

MTSI-Step 9: Selection of possible values of r

The three possible choices available for r have been described previously in Section 6.3.1 in MTSI-Step 9 of the 24-hour-ahead case.

MTSI-Step 10: Description of weighted least squares problem

Table 6.57 describes the form of the weighted least squares solution of the three-day-ahead prediction, where the number of equations relative to the 24-hour-ahead prediction has been increased from 25 to 75 and for each choice of r the system of equations is overdetermined.

Table 6.57 Description of weighted least squares solution for three-day-ahead prediction

r	$(n-r)$ No. of Unknowns	Forecast Horizon	Long-Time-Scale Information	No. of Equations in WLS Solution	Type of System
R1 - R2	24	72	sum, end-point and additional points	75	overdetermined
R3	23	72	sum, end-point and additional points	75	overdetermined

MTSI-Step 11: Construction of set of **W**

For the adjustment of the 24-hour-ahead prediction the weighting matrix is of the form:

$$\mathbf{W} = diag\Big[w_{ep\_wed} \quad w_{ep\_thurs} \quad w_{ep\_fri} \quad w_{s\_tues/wed} \quad w_{s\_wed/thurs} \quad w_{s\_thurs/fri} \quad w_{dev\_1900\ hrs\_tues} \quad \cdots\cdots\cdots\cdots \quad w_{dev\_1700\ hrs\_wed} \quad w_{dev\_1900\ hrs\_wed} \quad \cdots\cdots\cdots\cdots \quad w_{dev\_1700\ hrs\_thurs} \quad w_{dev\_1900\ hrs\_thurs} \quad \cdots\cdots\cdots\cdots \quad w_{dev\_1700\ hrs\_fri}\Big] \tag{6.8}$$

where

- w_{ep_j} is the weight on the end-point error on day j.
- w_{s_j} is the weight on the sum error on day j.
- $w_{dev\_i00\ hrs\_j}$ represents the weight on the error for the deviation from the original forecast at hour $i00$ on day j.
- $w_{dev\_1900\ hrs\_j}$ is the starting value on the deviation error weighting profile on day j.
- $w_{dev\_1700\ hrs\_j}$ is the finishing value on the deviation error weighting profile on day j.

The weighting matrices *wa*, *wb* and *wd* were considered with starting/finishing values in the range 1 to 20. Since the adjustment was being carried out on a three-ahead-ahead prediction using multiple long-time-scale points, the weight on the sum error was reduced to the same order of magnitude as the weight on the deviation error, as discussed in Section 5.2.4.2. Therefore, the weights on the end-point and sum were as follows: $w_{ep} = 100$ and $w_s = 10$.

MTSI-Step 12: Determination of optimal *r* and **W**

Using the methods outlined in Chapter 5 Section 5.3 the optimal *r* and **W** parameters for the cases PA to PE were determined, where a description of the parameters is given in Table 6.58.

Table 6.58 Optimal *r* and **W** for adjustment of 24-hour-ahead prediction

Forecast Horizon	**Case**	***r***	**Weighting Matrix**	
			Type	Starting Value on Deviation Error Profile
1900 hrs Tues - 1800 hrs Wed	PA	R2	*wd*	$w_{dev\_1700\ hrs} = 7$
	PB	R2	*wd*	$w_{dev\_1700\ hrs} = 5$
	PC	R2	*wd*	$w_{dev\_1700\ hrs} = 8$
	PD	R2	*wd*	$w_{dev\_1700\ hrs} = 6$
	PE	R2	*wd*	$w_{dev\_1700\ hrs} = 5$

The optimal choice for the parameter *r* is again R2, which is consistent with all other applications using the BSM with dummy seasonal component and the optimal weighting matrix *wd* is consistent with the multi sum and multi end-point adjustment case in the weekly/yearly application. As in the case of the other applications improvement over the original solution was also achieved using the other weighting matrices *wa* and *wb* but the optimal weighting matrix is *wd*. Figure 6.48 and Figure 6.49 the MAE and MSE performances for the different weighting matrices for the adjustment case PA.

As in the case of the 24-hour-ahead prediction the performance of the *wb* and *wd* weighting matrix is similar due to the similarity of their structure. In the case of the MSE performance for the *wa* matrix improvement over the original solution is limited to a starting values in the range 1 to 6. In addition, the performance of the *wa* sub optimal weighting matrix deteriorates as the starting value increases from 1 to 20, where a similar effect was also observed in the sub optimal matrices of the other applications.

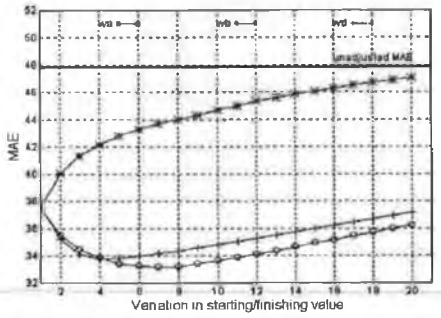

Figure 6.48 Comparison of MAE performance for different W-PA

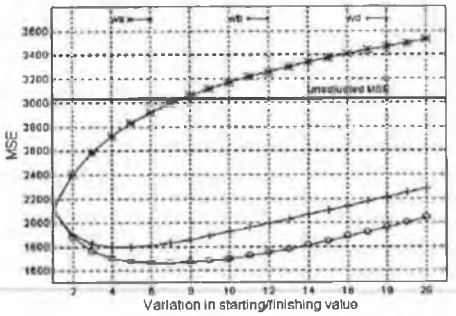

Figure 6.49 Comparison of MSE performance for different W-PA

MTSI-Step 13: Adjustment of short-time-scale prediction

Adjustment was performed using the optimal values of the r and **W**, where Figure 6.53 to Figure 6.57 (see pages) give the actual vs. predicted graphs which include the adjusted and unadjusted three-day-ahead predictions. Table 6.59 gives the forecasting accuracy results for the 3-day-ahead prediction and Table 6.60 to 6.62 give the results for each 24 hour period making up the three-day-ahead prediction.

Table 6.59 Results for adjustment of 3-day-ahead prediction

Forecast	1900 hrs Tues to 1800 hrs Fri			
	MAE	**MSE x 10^{-3}**	**MAPE**	**MPE**
Unadjusted	47.79	3.0308	2.42	1.34
PA using actual	33.20	1.6593	1.63	0.04
PA using predicted	36.96	1.9879	1.84	0.69
PB using actual	36.31	2.3886	1.76	-0.04
PB using predicted	39.45	2.6943	1.92	0.62
PC using actual	37.72	2.4875	1.86	0.08
PC using predicted	38.9	2.2664	1.95	0.72
PD using actual	32.97	1.6575	1.62	0.05
PD using predicted	37.43	1.9886	1.88	0.72
PE using actual	40.45	3.1311	1.95	0.01
PE using predicted	40.74	2.9240	1.98	0.63

Table 6.60 Results for Wednesday of 3-day-ahead prediction

Forecast	1900 hrs Tues to 1800 hrs Wed			
	MAE	MSE x 10^{-3}	MAPE	MPE
Unadjusted	45.54	2.8074	2.19	1.89
PA using actual	29.20	1.2121	1.42	0.59
PA using predicted	31.92	1.6282	1.51	1.25
PB using actual	35.00	2.4397	1.69	0.50
PB using predicted	37.77	2.9147	1.78	1.16
PC using actual	30.42	1.3859	1.46	0.64
PC using predicted	31.01	1.6015	1.48	1.27
PD using actual	28.13	1.1322	1.36	0.60
PD using predicted	31.44	1.5711	1.50	1.27
PE using actual	36.22	2.5431	1.73	0.55
PE using predicted	35.81	2.8599	1.68	1.18

Table 6.61 Results for Thursday of 3-day-ahead prediction

Forecast	1900 hrs Wed to 1800 hrs Thur			
	MAE	MSE x 10^{-3}	MAPE	MPE
Unadjusted	54.45	3.6420	2.83	0.79
PA using actual	42.29	2.5994	2.09	-0.51
PA using predicted	46.68	2.8081	2.36	0.15
PB using actual	41.94	2.7358	2.05	-0.58
PB using predicted	45.32	2.8336	2.25	0.07
PC using actual	48.71	4.0044	2.42	-0.47
PC using predicted	50.13	3.3551	2.55	0.16
PD using actual	42.99	2.6725	2.12	-0.50
PD using predicted	47.71	2.8573	2.44	0.17
PE using actual	48.05	4.0467	2.35	-0.55
PE using predicted	48.22	3.3155	2.40	0.09

Table 6.62 Results for Friday of 3-day-ahead prediction

Forecast	1900 hrs Thur to 1800 hrs Fri			
	MAE	MSE x 10^{-3}	MAPE	MPE
Unadjusted	43.40	2.6429	2.26	1.34
PA using actual	28.10	1.1663	1.39	0.04
PA using predicted	32.28	1.5274	1.65	0.70
PB using actual	31.98	1.9903	1.53	-0.04
PB using predicted	35.27	2.3345	1.72	0.61
PC using actual	34.01	2.0772	1.69	0.08
PC using predicted	35.81	1.8428	1.83	0.72
PD using actual	27.80	1.1677	1.37	0.05
PD using predicted	33.15	1.5374	1.71	0.72
PE using actual	37.07	2.8033	1.78	-0.001
PE using predicted	38.19	2.5996	1.87	0.63

Examining the unadjusted results produced by the univariate structural model for each of the twenty-four-ahead periods making up the three-day-ahead forecast (1900 hrs Tuesday to 1800 hrs Wednesday; 1900 hrs Wednesday to 1800 hrs Thursday; 1900 hrs Thursday to 1800 hrs Friday) it can be seen that the error associated with the Wednesday/Thursday prediction is higher than the other two days of the forecast. In an attempt to find a reason for this AT and actual and predicted load are plotted on the same graph in Figure 6.50. For comparative purposes AT and load for the day prior to the three-day-ahead forecast, that is Monday/Tuesday, is also given.

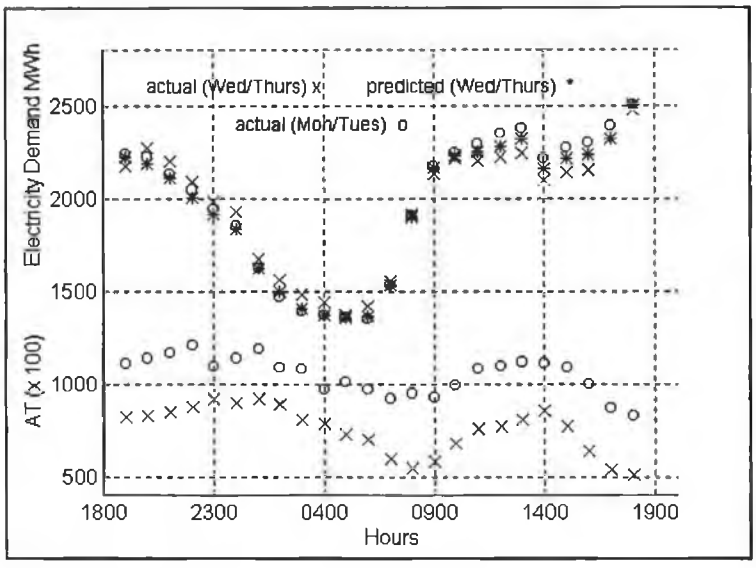

Figure 6.50 Actual vs predicted Wed/Thurs including AT data

The effect of weather on the daily load profile is clearly demonstrated in the graph, with the lower AT on Wednesday/Thursday resulting for the most part in a relatively higher load on this day in comparison with Monday/Tuesday. Examining the actual versus predicted profiles for Wednesday/Thursday shows that the largest discrepancy occurs between the hours of 1100 hrs and 1600 hrs on Thursday. Comparing the actual load of 1900 hrs on Wednesday to 0600 hrs on Thursday with 1900 hrs on Monday to 0600 on Tuesday shows that the lower AT on Wednesday/Thursday has the effect of increasing the load relative to Monday/Tuesday. However, a similar comparison of the load between 1100 hrs and 1600 hrs shows that although the AT on Thursday is still low relative to Tuesday the load is lower than expected. It is assumed that other influencing factors, possibly weather related, resulted in this reduced load. Figure 6.51 and Figure 6.52 show that this discrepancy is not experienced in the other two days of the three-day-ahead forecast. Consequently the forecast error for these two days is lower than on the error on the Wednesday/Thursday prediction.

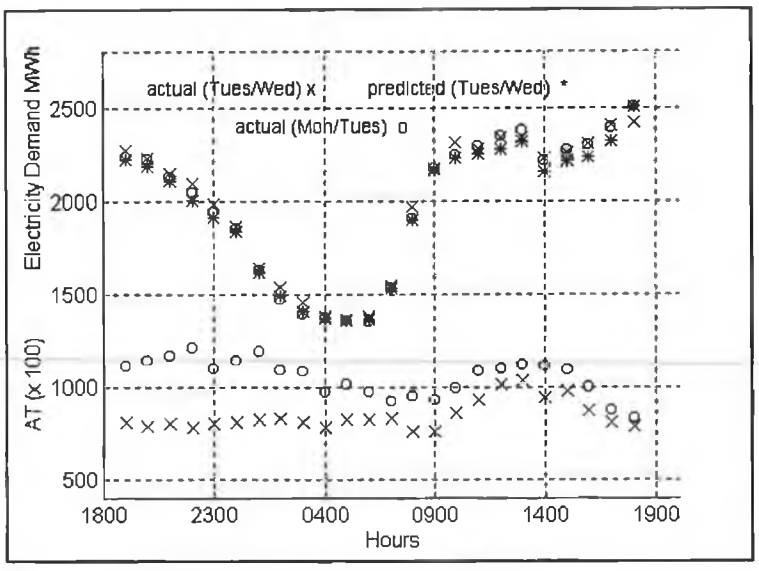

Figure 6.51 Actual vs predicted Tues/Wed including AT data

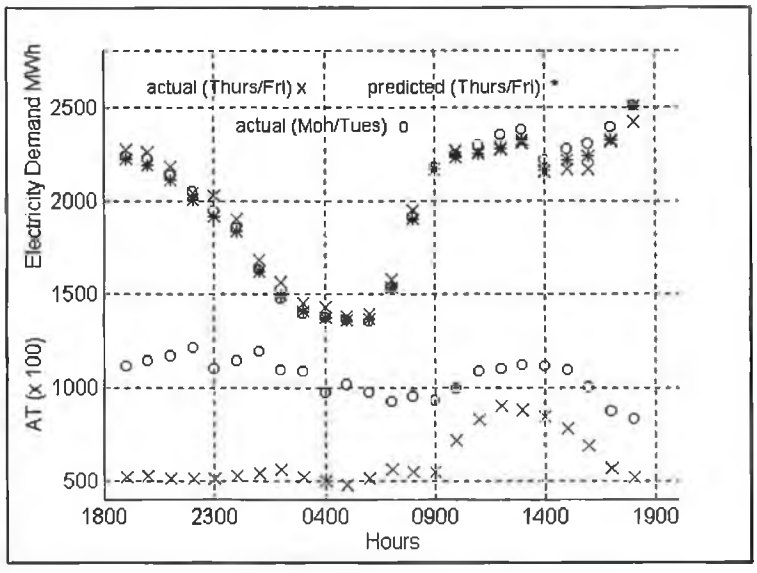

Figure 6.52 Actual vs predicted Thurs/Fri including AT data

Examining the forecasting accuracy of the adjusted prediction it can be seen that the results obtained for Tuesday/Wednesday of the 3-day-ahead forecast are more accurate than those obtained for the adjustment of the 24-hour-ahead forecast over this time period. This is the case for all adjustments PA to PE. For example, in the case of the PC adjustment the predicted 1400 hrs specification is 2315 MWh. In the 24-hour-ahead adjustment a value of 2323 MWh was attained but in the 3-day-ahead case a value of 2327 MWh was attained, which is in fact closer to the actual value of 2341 MWh. The difference in the adjustment of

the 3-day-ahead prediction is that it has the additional effect of the specifications on the other two days of the forecast, which in this particular case has had a beneficial effect.

As in the 24-hour-ahead case the adjustment carried using PA and PD produced the most accurate results. Examination of the improvement over the whole forecast horizon shows that the improvement in MAE is on average of the order of approximately 22%, the improvement in MSE is of the order of approximately 34%, with a MAPE and MPE of on average 1.86 and 0.70 respectively. Comparing the adjustments, PA to PE, for each day of the forecast horizon shows that the MAE measures are comparable but the MSE results indicate the effect of the inaccurate predictions of the long-time-scale information on the PC to PE adjusted predictions, where this inaccuracy can also be seen clearly in the graphs. Examination of the maximum achievable improvement in all cases demonstrates the need for more accurate predictions of the long-time-scale data.

Examination of the graphs show that for the most part the adjusted solution has been moved closer to the actual load for each point of the 3-day-ahead forecast. The relatively accurate predictions of the sum long-time-scale data in the case of Wednesday/Thursday, 0.22%, and Thursday/Friday, 0.47%, combined with the relatively accurate predictions at the end-point, that is 0.28% and 0.20% respectively, on these days has resulted overall in good adjustment of this forecast for the PA and PD cases. This is in spite of the relatively inaccurate predictions of the sum and end-point on the Tuesday/Wednesday, where the error on the sum is equal to 1.56% and the error on the end-point is equal to 2.90%. In addition, in the PD case the relatively inaccurate prediction of the 0500 hrs point over all days does not appear to have had a significant effect on the adjustment, where the errors on this point are 2.13%, 1.17% and 2.03% for Wednesday, Thursday and Friday respectively.

Finally, comparing the maximum achievable improvement of the PA to PE adjustment results indicates that the inclusion of these additional points in the application of the technique does not result in an increased improvement over the use of the sum and end-point only (PA). An exception to this is in the case of PD (additional equal to 0500 hrs) where comparable results to the adjustment attained using sum and end-point only was obtained.

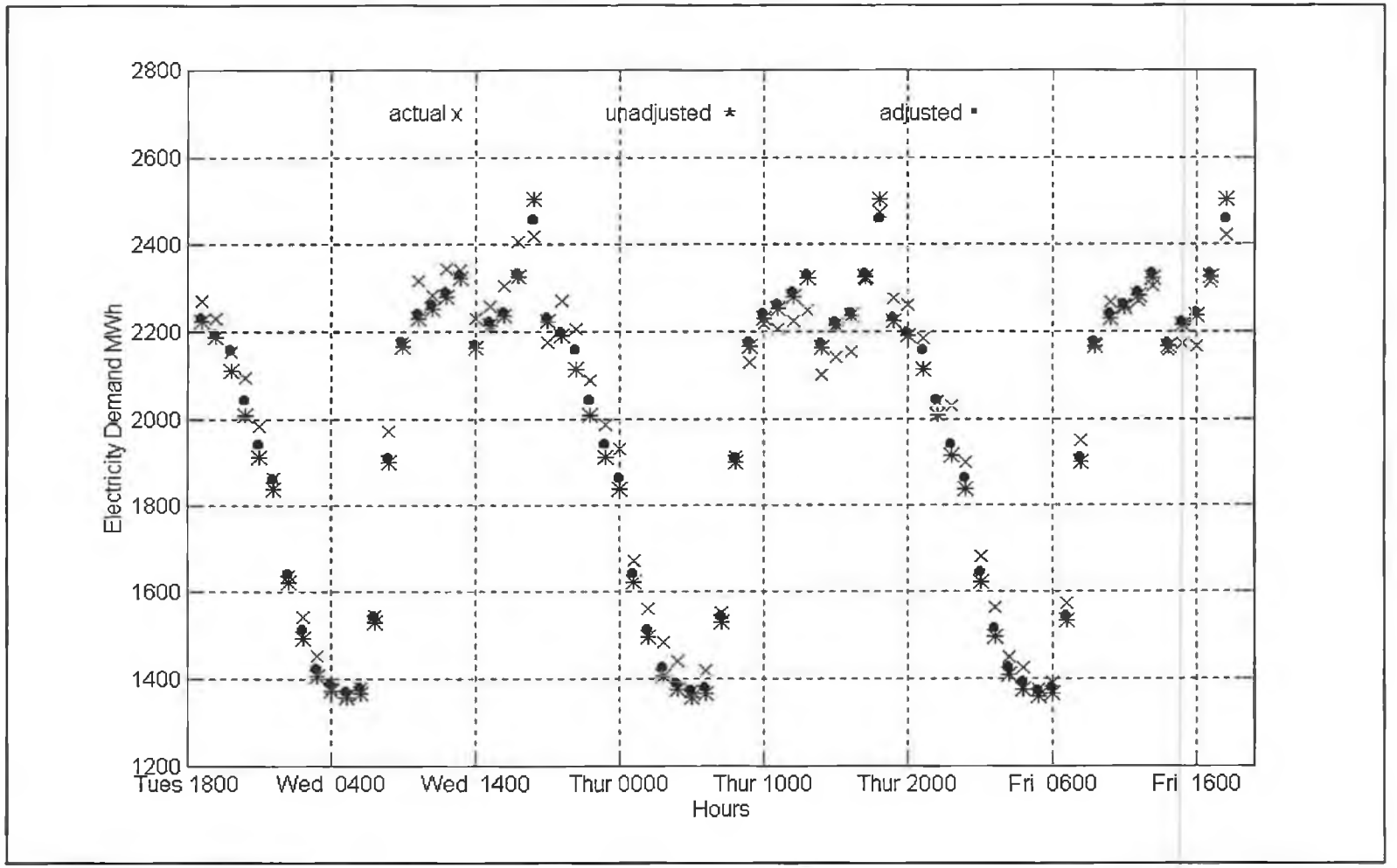

Figure 6.53 Three-day-ahead actual vs predicted for PA

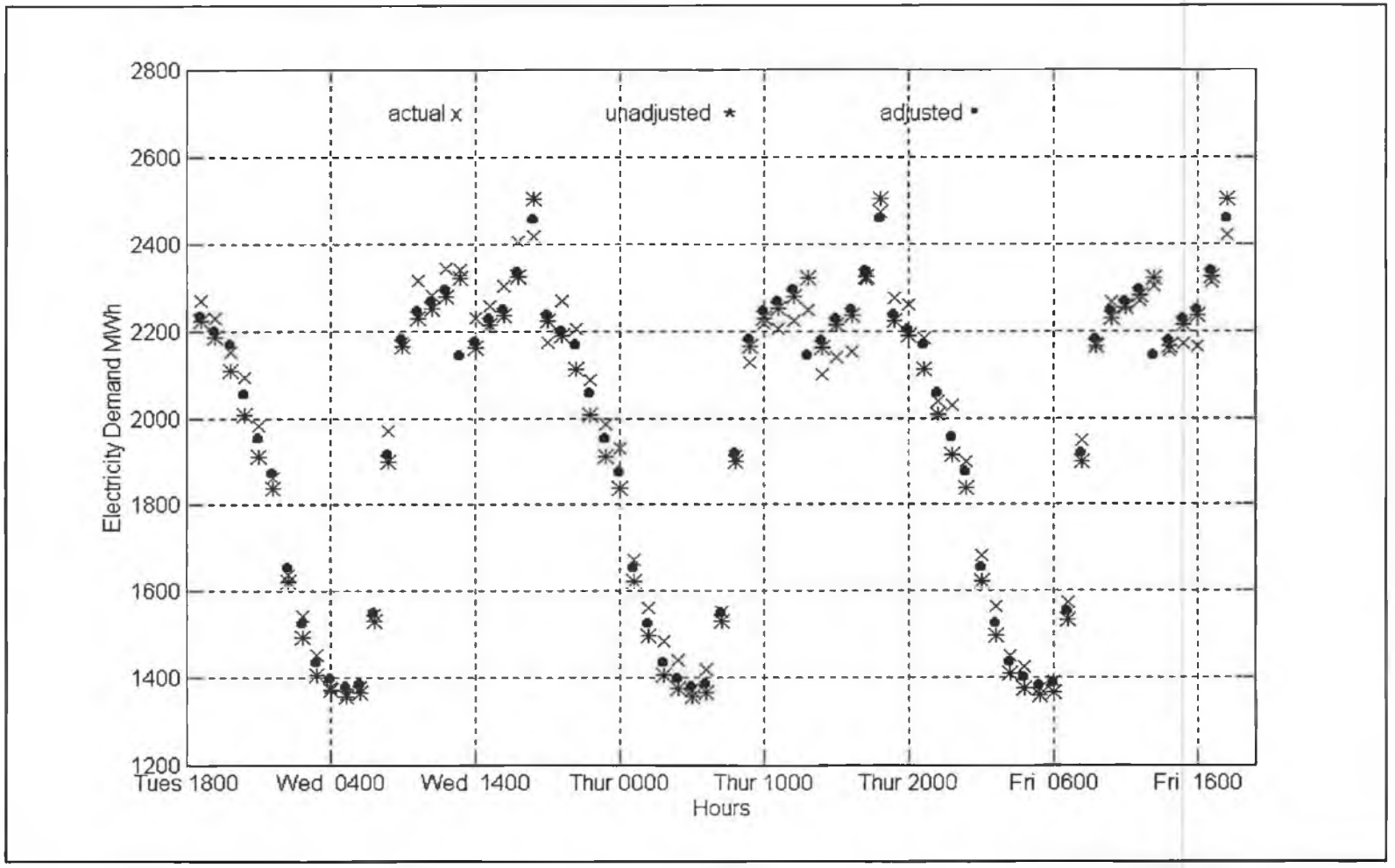

Figure 6.54 Three-day-ahead actual vs predicted for PB

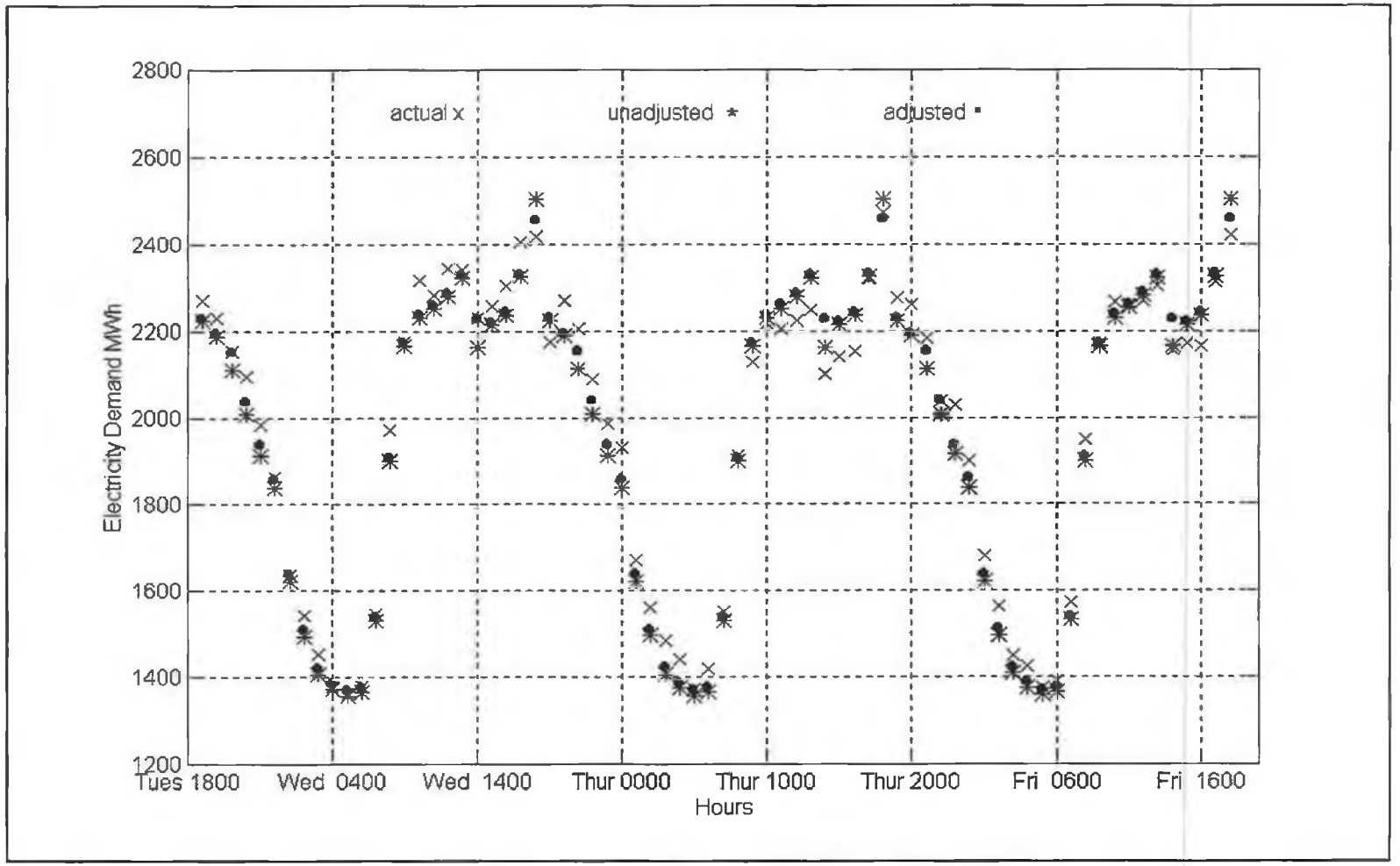

Figure 6.55 Three-day-ahead actual vs predicted for PC

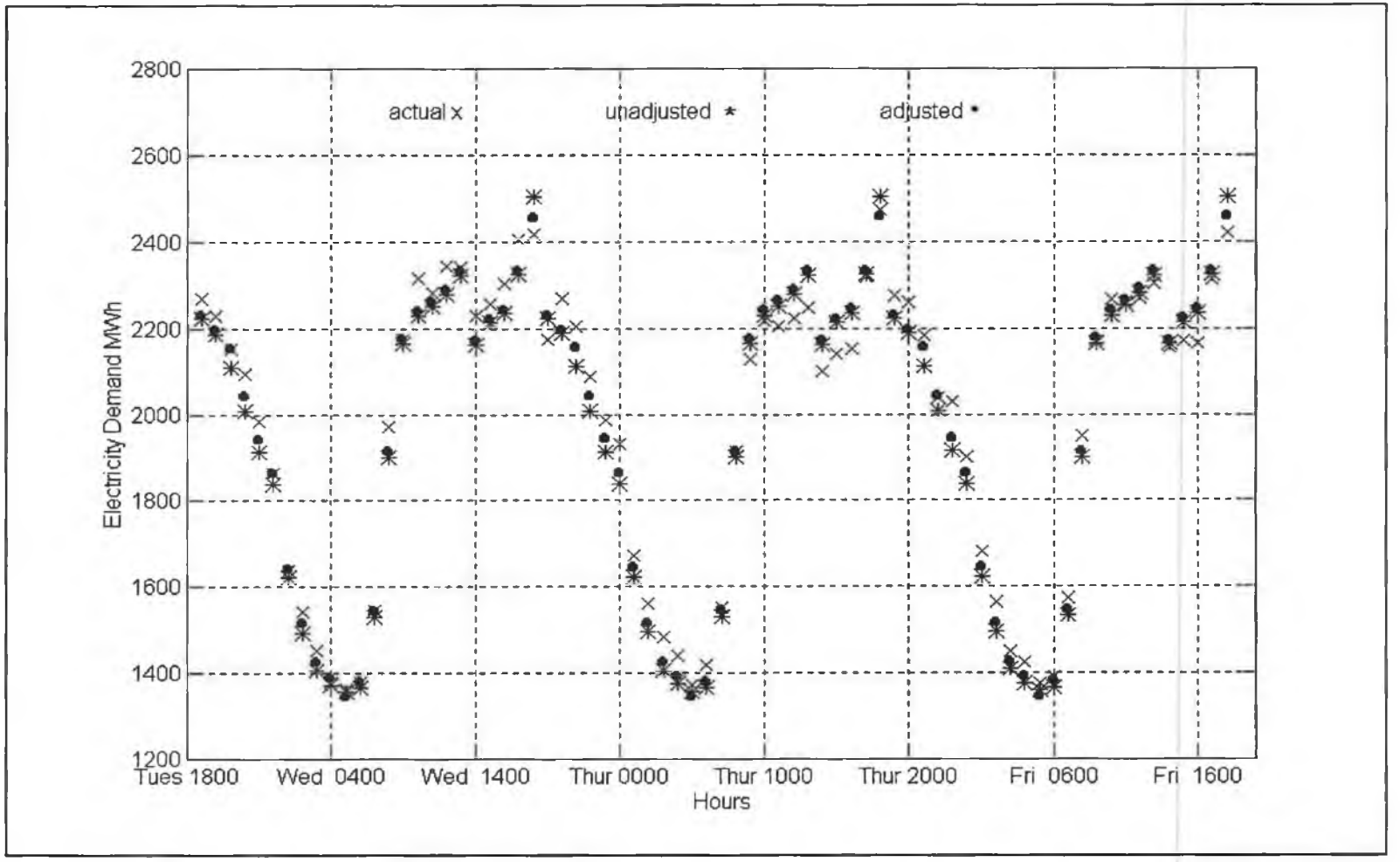

Figure 6.56 Three-day-ahead actual vs predicted for PD

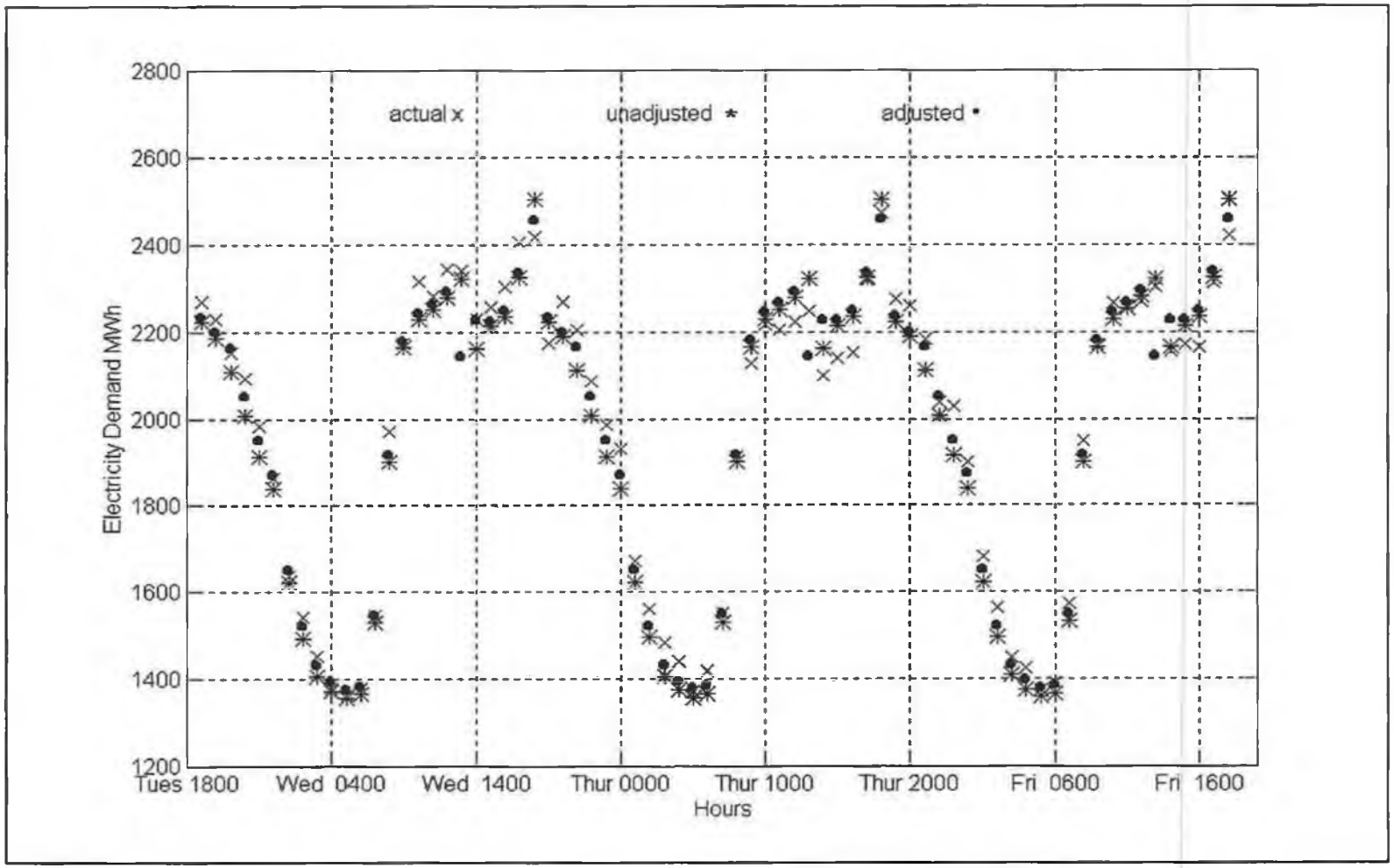

Figure 6.57 Three-day-ahead actual vs predicted for PE

6.4 Conclusions

The multi-time-scale integration technique was applied to two electricity demand applications. In all cases where adjustment was applied subject to sum and end-point specifications there was a good improvement over the original unadjusted solution. The improvement achieved in the hourly/daily application was considerable; an example of this is in the 3-day-ahead hourly/daily application where the improvement obtained was from 2.42 to 1.89 in the MAPE and from 1.34 to 0.69 in the MPE.

The effect of the accuracy of the predicted long-time-scale data was clearly established in both applications. This was significant in the hourly/daily application where inaccurate predictions of the additional long-time-scale data hindered the achievement of good adjustment results. Of particular importance is the accuracy of the sum data, where since this effects the overall forecast the accuracy of this specification is important.

The scope for the application of the technique in the hourly/daily application is wide. The characteristics of the daily load profile are such it is straightforward to select intermediate target points for the hourly prediction to follow. Examination of the maximum achievable improvement for this application shows that if more accurate predictions of the cardinal points were available then greater improvement than was achieved could be obtained. The use of an alternative long-time-scale model such as neural networks would be appropriate. The reason being that neural networks have been widely applied to the problem of predicting the cardinal points on the daily load profile (Bacha and Meyer 1992; Caire *et al*, 1992; Tamura *et al*, 1993; Papalexopoulos *et al* 1994) and they offer a real alternative to the structural models.

The results also demonstrated the effect of the form of the weighting matrices on the solution. In each application in the case where multiple long-time-scale specifications were applied the *wd* weighting matrix was the optimal matrix structure. In this matrix the weights on the deviation errors on either side of the end-point are heavy penalised. Therefore, the short-time-scale prediction follows the long term path determined by the end-points but does not deviate form the original solution significantly at the start of each prediction cycle. Thus ensuring sensible adjustment over the entire long term prediction.

The optimal choice for the parameter r in all cases was shown to be R2, that is fix the state representing the slope of the trend and free the states representing the level of the trend and the seasonal component. The freedom in the level of the trend allows for the solution to achieve

the level of the long-time-scale end-points and the freedom in the seasonal component allows for the achievement of the sum specification.

Ultimately, the aim is to combine the hourly/daily and weekly/yearly applications so that a prediction of hourly load could be produced over a period of three years. The implementation of this application requires the resolution of a number of issues which are the areas of future work; a discussion of these issues is given in the Conclusions in Chapter 7.

CHAPTER 7

Conclusions

7.1 Evaluation of Short, Medium and Long Term Forecasting Techniques

This thesis has dealt with a number of forecasting methods for the production of short, medium and long term electricity forecasts based on a single-time-scale and also addresses the problem of multi-time-scale forecasting.

The multi-time-scale integration technique was shown to be a valuable forecasting tool for the production of hourly short term load forecasts. The characteristic of the daily load profile is such that the application of the technique is well suited to this problem area. However, in the work carried out in the thesis weather variables were not included at the short sampling interval, where it is known that they have a significant effect but this is an area of future work (see Section 7.3.2). In addition, the structural long-time-scale models employed in the application did not produce accurate predictions and this substantially affected the attainment of good adjustment results. Therefore, there is a requirement for the use of alternative long-time-scale models in this case; this is also an area for future work (see Section 7.3.2).

In the case of medium term forecasting it was established that the employment of exogenous weather variables can improve the accuracy of the results produced by linear and nonlinear forecasting models. The use of such variables are appropriate even in the case where weather is not a dominant influencing factor in the system. However, it is important to examine the optimal form of the available weather variable. For example, in the applications dealt with in the thesis the use of HDD as the exogenous weather variable in a linear forecasting model produced more accurate results than the use of AT.

The results also determined that neural networks in some cases are suitable for representing a medium term electricity demand time series; these are discussed further in the Section 7.2. In cases where a linear model is appropriate, comparative results showed that structural time series models produced more accurate forecasting results than classical models. In addition, the development of structural models is more straightforward and is not as highly dependent on user interaction. However, comparing the univariate (2.41) and bivariate (2.49) structural

model results show that the multivariate structural model is inadequate at representing a system with exogenous variables.

The application of the multi-time-scale integration technique to the problem of medium term forecasting was shown to be of considerable benefit. The production of the forecasts over a prediction of a year or more allowed the effect of economic and demographic influences to be accounted for at the medium term time-scale.

In the case of long term forecasting, the results indicated that neural networks are a suitable modelling tool for the representation of an aggregate system but when the system is disaggregated linear models may be adequate to represent the simpler systems. However, this would require an examination of case by case. Therefore, depending on the modelling tool employed the analysis of the system may benefit from disaggregation.

7.2 Assessment of Neural Networks as a Forecasting Tool

The work conducted in the thesis established that neural networks can be a powerful modelling tool for medium and long term electricity forecasting. However, it was also established that they are highly case dependent. Furthermore, a successful neural network analysis relies on the use of prior knowledge of the system, where this may be obtained from a linear time series analysis.

A number of guidelines that may be employed when conducting a medium and long term neural network analysis were determined and these are described as follows:

1. If the time series has trend and seasonal characteristics then pre-processing the training data using a linear transformation reduces the time required to train the network considerably.
2. The use of linear *differencing* transformations for seasonal time series results in a substantial improvement in the network performance.
3. Linear analysis should be employed to select a parsimonious input structure to the neural network.
4. If the network is required to perform a multi-step head prediction but is being trained only on a single-step-ahead criterion, then the performance of the network should be assessed by examining the multi-step-ahead performance over a validation data set.
5. It is necessary to re-train a network a number of times to circumvent the problem of sensitivity to initial conditions.

6. When performing an experimentation analysis to determine an appropriate network structure the overall capability of a structure should be assessed based on the average values of the performance criterion and not on the best result, calculated over a suitable number of training runs.

Three areas were identified for further investigation, as follows:

1. Determination of an appropriate test for nonlinearity.
2. Use of alternative neural network structures.
3. Use of genetic algorithms in the neural network analysis.

(1) In spite of the fact that neural networks can be very successful in the area of medium and long term forecasting a significant problem is that they are especially case dependent. The problem can arise that the system that the neural network is required to represent is a linear and not a nonlinear system. The determination of a neural networks suitability to represent a time series usually involves examining the forecasting accuracy of the results produced on a test (out of sample) set. However, this can entail a considerable amount of effort, since the design of the network often requires extensive experimentation analysis. In classical analysis there are a number of hypothesis based tests available to test if a nonlinear linear model is suitable to represent the system but unfortunately the majority of them are model specific (Gooijer and Kumar, 1992). The employment of a nonlinear test for the time series *data* is desirable. The use of third and fourth order cumulants to detect the existence of system nonlinearities (Emara-Shabaik and Moustafa, 1994) are possible directions for future work in this area.

(2) The MLP's employed in the medium and long term applications dealt with in the thesis are feedforward MLP networks. However, it would be of value to investigate the possible improvement of these networks through the extension of the network structure. One such extension is to that of the Dynamical Neural Network (Hush and Horne, 1993). This involves feeding the output of the network back to the input during training through a tapped delay line (see Figure 2.14) and the use of a gradient descent algorithm for learning. Alternatively, the MLP could be extended to the case where in addition to the nonlinear connections between the hidden units there is also a direct linear connection from each input to the output. This type of architecture can extract the linearly predictable part early in the learning process and free the nonlinear resources to be employed were they are really needed (Weigend and Gersenfield,

1992). Other possibilities are the use of Radial Basis Function (Hush and Horne, 1993), Feature Maps (Hertz *et al*, 1991) or the Recurrent networks (Hertz *et al*, 1991)

(3) The optimal design of a neural network for a particular application can be difficult to determine since there are a large number of factors involved. Genetic algorithms are general purpose adaptive optimisation search algorithms that could be employed in the design of the neural network. Specifically they could be used to determine the following:

(i) networks weights.
(ii) input structure.
(iii) number of neurons in each layer.
(iv) number of layers.

In the case of (i), the genetic algorithms search the weight space without use of any gradient information. The advantage is that they perform a global search and do not have the local minima problem associated with backpropagation learning. However, there is a high computational penalty associated where the gradient information is not used. A possible compromise would be the use of genetic algorithms to conduct an initial search followed by a gradient method. In the case of (ii), (iii) and (iv), the genetic algorithms could be employed to search the space of all possible input structures and network architectures. However, for applications requiring large networks this would require an inordinate amount of computation time.

7.3 Multi-Time-Scale Forecasting Technique

7.3.1 Evaluation of Multi-Time-Scale Integration Technique

The multi-time-scale integration technique proposed in the thesis has been applied to a number of applications with the achievement of consistently good results. In addition to the development of the mathematical formulation of the technique many issues regarding its practical implementation were resolved. However, there still remains some issues that require further attention; this is an intended area of future work and is discussed in Section 7.3.2.

It has been shown that in addition to its use for the production of a short-time-scale prediction over a relatively *long* forecast horizon, the technique is also ideal for the improvement of a short-time-scale prediction over a *short* forecast horizon. For example, in the case of the hourly/daily application, where methods for the prediction of the cardinal points on the daily

load profile, such as the midday peak, the tea-time peak and the overnight minimum may already be in place.

The versatility of the technique in comparison with other multi-time-scale techniques (Corrado and Greene, 1994) was demonstrated in the applications, where a variety of models were employed at the long sampling period. However, the dependency of the technique on the accuracy of the predicted long-time-scale information was also established, where the accuracy of the long-time-scale sum prediction is particularly important.

The technique has been shown to be particularly well suited to applications involving the prediction of a short sampling period time series that has a seasonal component. In the majority of systems the electricity demand time series will have seasonal characteristics due to weather and temporal factors, with the exception of yearly electricity demand. In such a case the long sampling period time series are generated at the seasonal level.

The analysis conducted in the thesis identified a number of guidelines that may be employed for the determination of the parameters involved in the application of the technique. The guidelines are summarised as follows

Selection of *r* when a structural model is employed at short sampling interval

1. Free the maximum number of states possible.
2. If possible free both states representing the trend component, that is the level and slope.
3. If it is necessary to fix a state in the trend component then the state corresponding to the slope of the trend component should be fixed.
4. In a BSM with trigonometric seasonal component model free the state representing the fundamental harmonic.

Selection of *r* when AR, ARMA or ARIMA model is employed at short sampling interval

1. Fix approximately the first 40% of state and free the remaining states.
2. Fix approximately the first 30% of state and free the remaining states.
3. Fix approximately the first 20% of state and free the remaining states.

Form of W when single sum, end-point and additional point are applied

1. the ratio of the end-point, sum and additional point error weights to deviation error weights is approximately of the order of 10/1.
2. use a deviation error weighting profile that is heavy at the forecasting origin and lighter towards the end-point (*wa* see Figure 5.10)

Form of **W** when multi sum, end-point and additional point are applied

1. the ratio of the end-point error weight to the sum, additional point and deviation error weights is approximately of the order of 10/1.
2. use a deviation error weighting profile that is heavy on either side of the end-point and lighter for the remaining deviation errors (*wd* see Figure 5.10)

7.3.2 Areas for Further Research

The following issues require further attention and are the subject of ongoing research:

1. Numerical optimisation of r and **W**.
2. Extension of technique so that a multivariate short-time-scale model could be employed at the short sampling period.
3. Extension of the technique to deal with multiple levels.
4. Use of other long-time-scale models.

(1) This involves the optimisation of the error $E(r, \mathbf{W})$ with respect to the parameters using an appropriate numerical optimisation routine. The use of such an approach in the case where an AR, ARMA or ARIMA model is employed at the short sampling period would in practical terms be reasonably straightforward since the value of r is simply a proportion of the states in the state vector. In addition, analysis has shown (see Figure 5.2) that the error surface, as a function of r and **W**, associated with these model is relatively smooth. In contrast, in the case of the structural model the different choices of r do not result in such a smooth error surface. However, since the states in this model have a physical significance the selection of the r is simpler and the requirement for numerical optimisation in this case not as important.

(2) The use of a multivariate model at the short sampling period requires the mathematical formulation of the technique for a multivariate model in state space form.

As it stands, the technique may be applied when an ARX model is employed at the short sampling interval if this model is rewritten as a multivariable AR model with multiple outputs (Isaksson, A, 1993). This involves writing $z(k) = [y(k)\ \mathbf{u}(k)]^T$, where $y(k)$ is the univariate time series and $\mathbf{u}(k)$ is a vector of exogenous variables and the state vector is formulated as past values of $z(k)$. In the same manner, the technique could also be employed in its present form where a BSMX model (2.49) is used at the short sampling interval. However, analysis has shown that this model is not a particularly effective model for representing a system with exogenous variables and an alternative structural model approach would be more appropriate.

One possibility is the use of the structural model presented by Ng and Young (1994) which involves the use of a Deterministic Transfer Function (DTF) model to represent the relationship between the $y(k)$ and $u(k)$ time series, and the use of similar models to those employed in the BSM with dummy seasonal component (IRW and DPRW) to represent the trend and seasonal components. In either of the above cases investigation of the use of such models on the r and $\mathbf{W}$ parameters would be necessary.

(3) The extension of the technique to the case where more than two time-scales can be dealt with is required. In relation to the applications dealt with in thesis this would involve the combination of the hourly/daily and weekly/yearly applications, namely the application of the technique on four time-scales. This would result in the production of hourly forecasts over a yearly forecast horizon. The practical implementation of the technique to such an application would require the resolution of a number of issues, as follows:

(i) Mathematical formulation of the technique on multiple time-scales. It is anticipated that this will be a natural extension of the two-time-scale case.
(ii) On the hourly time-scale issues concerning the STLF problem in itself will require consideration. The use of the standard data concept (Longergan, 1995) may be applicable, where multiple models across different days and different seasons could employed.
(iii) The optimisation of r and $\mathbf{W}$ obviously explodes exponentially for the multi-level case.
(iv) The inclusion of exogenous variables on all time-scales, where this is related the discussion in (2) above.

(4) An investigation into the use of alternative models at the long- time-scale is desirable. For the applications dealt with in the thesis this is particularly important in the hourly/daily application. In this case improved methods of producing the prediction of the cardinal points on the daily load profile are required. Longergan (1995) deals specifically with this problem using a linguistic model, where this deals with qualitative rather than quantitative information.

7.4 General Areas for Future Research

This thesis provides a comprehensive set of forecasting tools that can be applied to the area of short, medium and long term electricity demand forecasting. Of considerable importance is the provision of a solution to the relatively new problem in the area that has arisen due to changing market structures in the power industry world-wide, that is the production of short term forecasts over a long forecast horizon. The area of electricity demand forecasting is a

vast and well researched subject; however there still remains a number of areas that have not received much attention. Two such important areas are as follows:

1. Measurement of the stochastic component in the load.
2. Investigation of the relationship between yearly electricity demand and the socio-economic and demographic variables.

(1) On any time-scale the electricity time series may be divided into a *predictable* component and an irregular *stochastic* component. This irregular component can be very difficult to predict and the majority of existing forecasting methodologies focus on the modelling of the predictable part as accurately as possibly. Obviously, an accurate prediction of the irregular component would be desirable, however in some cases it is sufficient to know what proportion of the load this component makes up. For example, an estimate of the value of this component would be of great benefit at power system operational level (hourly) for the purposes of improving the unit commitment and economic dispatch functions.

(2) In the short and medium term the relationship between weather, the significant influencing factor, and electricity demand is relatively easy to understand. However, the relationship between yearly electricity demand and the socio-economic and demographic variables, such as GDP, price of electricity and unemployment and changes in the population growth rate are more complex and less well understood. Changes in living patterns and social attitudes, for example more women in the workforce and rapid technological growth all have an impact on these socio-economic and demographic factors and ultimately on the behaviour of electricity demand. To fully understand the effect of such changes it would be necessary to conduct an investigation at both the aggregate and disaggregate levels.

8. REFERENCES

Abraham, B and Ledolter, J., *Statistical Methods for Forecasting*, John Wiley & Sons, 1983.

Ackerman, G., "Short-term Load Prediction for Electric-utility Control of Generating Units", in: Bunn, D. W. and Farmer, E, D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985, 33-42.

Adegbulugbe, A. O. and Dayo, F. B., "A Synthesis Of Monte Carlo And Noise-In-Variable Model Techniques For Energy Demand Modelling", in: *IFAC Energy Systems, Management and Economics,* Toyko, Japan, (1989) 83-86.

Akaike, H. "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models", Biometrika, **60**, (1973), 255-265.

Aleksander, I., "Introduction to Neural Nets", in Warick, K. *Applied Artificial Intelligence*, Peregrinus, 1991, 13-30.

Al-Zayer, J. and Al-Ibrahim, A., "Modelling the Impact of Temperature on Electricity Consumption in the Eastern Province of Saudi Arabia", *Journal of Forecasting*, **15** (1996), 97-106.

Anderson, B. D. O. and Moore, J. B., *Optimal Filtering*, Englewood Cliffs, NJ:Prentice-Hall, 1979.

Ansley, C. F. and Kohn, R., "A Structured State Space Approach to Computing the Likelihood of an ARIMA Process and its Derivatives", *Journal of Statistical Computer Simulations*, **21** (1985a), 135-169.

Ansley, C. F. and Kohn, R., "Estimation, Filtering, And Smoothing In State Space Models With Incompletely Specified Initial Conditions", *The Annals of Statistics*, **13** (1985b), 1286-1316.

Arellano, C. and Pantula, S. G., "Testing For Trend Stationarity Versus Difference Stationarity", *Journal of Time Series Analysis*, **16** (1995), 147-164.

Azoff, M. E. "Reducing Error in Neural Network Time Series Forecasting", *Neural Computing & Applications*, **1** (1993),240-247.

Azoff, M.E., *Neural Network Time Series Forecasting of Financial Markets*, John Wiley and Son, 1994.

Azzam-ul-Asar and McDonald, J. R., "A Specification of Neural Network Applications in the Load Forecasting Problem", *IEEE Transactions on Control Systems Technology*, **2** (1994), 135-141.

Bacha, H. and Meyer, W., "A neural Network Architecture for Load Forecasting", in: *Proceedings of the International Joint Conference on Neural Networks*, Baltimore, 1992, 442-447.

Baker, A., "Load Forecasting for Scheduling Generation on a Large Interconnected System", in: Bunn, D. W. and Farmer, E, D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985, 57-68.

Bakirtzis, A. G., Petridis, V., Klartzis, S. J., Alexladis, M. C. and Malssis, A. H., "A Neural Network Short Term Load Forecasting Model For The Greek Poser System", *IEEE Transactions on Power Systems*, **11** (1996), 858-863.

Banim, J. and Hodnett, P.F., "Use of Pattern Recognition in Electricity Load Forecasting" in: *Proceedings of the 5th European Conference on Mathematics in Industry*, (1991) 77-81.

Barakat, E. H., Qayyum, M. A., Hamed, M. N. and Rashed, S. A., "Short-term Peak Demand Forecasting in Fast Developing Utility with Inherit Dynamic Load Characteristics - Part I-Application of Classical Time Series Methods", *IEEE Transactions on Power Systems*, **5** (1990), 813-819.

Batchelor, R. and Dua, P., "Forecaster Diversity and the Benefits of Combining Forecasts", *Management Science*, **41** (1995), 68-75.

Beltratti, A, Margarita, S. and Pieto, T. *Neural Networks for Economic and Financial Modelling*, International Thomson Computer Press, 1996.

Berg, M. C., Amit, N. and Powell, D. J., "Multirate Digital Control System Design", *IEEE Transactions on Automatic Control*, **33** (1988), 1139-1150.

Billings, S. A., Jamaluddin, H. B. and Chen, S., "Properties of Neural Networks with Applications to Modelling Non-Linear Dynamic Systems", *International Journal of Forecasting*, 55 (1992), 193-224.

Blake, J., Francino, P., Catot, J. M. and Sole, I "A Comparative Study for Forecasting using Neural Networks vs Genetically Identified Box & Jenkins Models", *Neural Computing & Applications*, **3** (1995), 139-148.

BMDP Statistical Software Manual, BMDP, 1990.

Bordley, R. F., "Linear Combination of Forecasts with an Intercept: A Bayesian Approach", *Journal of Forecasting*, **5** (1986), 243-249.

Bordley, R. F., "The Combination of Forecasts: a Bayesian Approach", *Journal of the Operational Research Society*, **33** (1982), 171-174.

Bornholdt, S. and Graudenz, D, "Genetic Asymmetric Neural Networks and Structure Design by Genetic Algorithms", *Neural Networks*, **5** (1992), 327-334.

Bowerman, B. and O"Connell, R. T., T*ime Series Forecasting: Unified Concepts and Computer Implementations*, Duxbury Press, 1987.

Box, G. and Jenkins, G., *Time Series Analysis Forecasting and Control*, Holden-Day, 1976.

Box, G. E. P., Pierce, D. A. and Newbold, P., "Estimating Trend and Growth Rates in Seasonal Time Series", *Journal of American Statistical Association*, **82** (1987), 276-282.

Brockwell, P. J. and Davis, R. A., *Time Series: Theory and Methods*, Springer-Verlag, 1987.

Bunn, D. W. and Farmer, E. D. "Review of Short-term Forecasting Methods in the Electric Power Industry", in: Bunn, D. W. and Farmer, E, D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985.

Caire, P., Hatabin, P. and Muller, G., "Progress in Forecasting by Neural Networks", in: *Proceedings of the International Joint Conference on Neural Networks*, Baltimore, 1992, 540-545.

Carbone, R., Bilongo, R., Piat-Corson, P. and Nadeau, S., "AEP Filtering", in: Makridakis, A., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., *The Forecasting Accuracy of Major Time Series Methods*, John Wiley and Sons, 1984, 201-220.

Chakraborty, K., Mehrotra, K., Mohan, C.K. and Ranka, S. "Forecasting the Behaviour of Multivariate Time Series Using Neural Networks", *Neural Networks*, **5**, (1992) pp 961-970.

Chandrasekharan, R., Moriarty, M. M. and Wright, G. P., "Testing for Unreliable Estimators and Insignificant Forecasts in Combined Forecasts", *Journal of Forecasting*, **13** (1994), 611-624.

Chang, M. C. and Dickey, D. A., "Recognising Overdifferenced Time Series", *Journal of Time Series Analysis*, **15** (1994), 1-18.

Chen, S. and Billings, S. A., "Modelling and Analysis of Non-linear Time Series", *International Journal of Control*, **50** (1989), 2151-2171.

Chen, S. and Jarrett, J., "On improving Time Series Forecasting", *Omega*, **19** (1991), 505-505.

Chen, S-T, Yu, D. C. and Moghaddanjo, A. R., "Weather Sensitive Short-Term Load Forecasting Using Nonfully Connected Artificial Neural Network", *IEEE Transactions on Power Systems*, **7** (1992), 1098-1105.

Chester, D., "Why Two Hidden Layers are Better than One", in: *Proceedings of the International Joint Conference on Neural Networks*, 1990, 265-268.

Cholette, P., "Prior Information and ARIMA Forecasting", *Journal of Forecasting*, **1** (1982), 375-383.

Clemen, R. T., "Linear Constraints and Efficiency of Combined Forecasts", *Journal of Forecasting*, **5** (1986), 31-38.

Clemen, R. T., Murphy, A. H. and Winkler, R. L., "Screening Probability Forecasts: Contrasts Between Choosing and Combining", *International Journal of Forecasting*, **11** (1995), 133-146.

Connor J. T., Martin, R. D. and Atlas, L. E., "Recurrent Neural Networks and Robust Time Series Prediction" Elimination" *IEEE Transactions on Neural Networks*, **6** (1994), 240-254.

Corrado, C. and Greene, M., "Reducing Uncertainty in Short-Term Projections: Linkage of Monthly and Quarterly Models", *Journal of Forecasting*, **7** (1988), 77-102.

Cottrell, M., Girad, B., Girad, Y., Mangeas, M, and Muller, C., "Neural Modelling for Time Series: A Statistical Stepwise Method for Weight Elimination" *IEEE Transactions on Neural Networks*, **6** (1995) 1355-1363.

Demirovic, E. A., "Short Term Forecasting Algorithm", in: *IASTED International Symposium on Modelling, Identification and Control*, Switzerland (1988) 233-236.

Dharan, B. L., "A Priori Sample Size Evaluation and Information Matrix Computation for Time Series Models", *Journal of Statistical Computer Simulations*, **21** (1985), 171-177.

Di Caprio, U., Genesio, R., Pozzi, S. and Vicino, A., "Comparison of ARMA and Extended Wiener Filtering for Load Prediction at ENEL", in: Bunn, D. W. and Farmer, E, D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985, 109-130.

Dickey, D. A. and Fuller, W. A. "Distributions of the Estimators for Autoregressive Time Series with a Unit Root", *Journal of American Statistical Association*, **74**,(1979), 427-431.

Diebold, F. X. and Pauly, P., "Structural Change and the Combination of Forecasts", *Journal of Forecasting*, **6** (1987), 21-40.

Dodds, G. I. and Beattie, W.C., "Electrical Demand Forecasting Through Exponential Smoothing And Additional External Parameters", in*: Proceedings of the 25th Universities Power Engineering Conference*, Aberdeen, UK, (1990) 473-476.

Dodds, G. I., Irwin, G. W. and Beattie, W. C., "Electricity Demand Modelling from Disaggregate Data", *Electricity Power & Energy Systems*, **12** (1990), 50-60.

Donaldson, R. G. and Kamstra, M., "Forecast Combining with Neural Networks", *Journal of Forecasting*, **15** (1996), 49-61.

El-Sharkawi, M. A., "Short Term Load Forecasting Using Adaptively Trained Peceptrons", in: *Proceedings of the 1st International Forum on Applications of Neural Networks to Power Systems*, Seattle, (1989) 23-26.

Emara-Shabaik, H. and Moustafa, K. "Characterisation of Dynamic System Nonlinearities via Probabilistic Approach", *International Journal of System Science*, **3** (1994), 603-611.

Engle R. F., Chowdhury M. and Rice R., "Modelling peak electricity demand", *Journal of Forecasting*, 11 (1992), 241-251.

Engle, R. F., Brown, S. J. and Stern, G., "A Comparison of Adaptive Structural Forecasting Methods for Electricity Sales" *Journal of Forecasting*, **7** (1988), 149-172.

Engle, R. F., Mustafa, C. and Rice, J., "Modelling Peak Electricity Demand", *Journal of Forecasting*, **11** (1992) 241-251.

Espada, A. and Pena, D., "The Decomposition of Forecast in Seasonal ARIMA Models", *Journal of Forecasting*, **14** (1995), 565-583.

Fildes, R. and Makridakis, S. "The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting", *International Journal of Forecasting*, (1991).

Fildes, R. "Bayesian Forecasting", in: Makridakis, A., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., *The Forecasting Accuracy of Major Time Series Methods*, John Wiley and Sons, 1984, 221-238.

Franklin, G.F., Powell, J.D. and Workman, M. L., *Digital Control Systems*, Addison and Wesley, 1990.

Franses, P. H., "Seasonality, Non-stationarity and the Forecasting of Monthly Time Series", *International Journal of Forecasting*, **7** (1991), 199-208.

Fuller, W. A., *Introduction to Statistical Time Series*, John Wiley & Sons, 1996.

Furher, J. and Haltmaier, J., "Minimum Variance Pooling of Forecasts at Different Levels of Aggregation", *Journal of Forecasting*, **7** (1988), 63-73.

Gardner, E.S. and McKenzie, E., "Seasonal Exponential Smoothing with Damped Trends", *Management Science*, **35** (1989), 372-376.

Gelb, A., *Applied Optimal Estimation*, MIT Press, 1974.

Giles C. L., Kuhn, G. M. and Williams R. J., "Dynamic Recurrent Neural Networks: Theory and Applications", *IEEE Transactions on Neural Networks*, **5** (1994), 153-155.

Gooijer, J. J. and Kumar, K., "Some Recent Developments in Non-linear Time Series Modelling, Testing, and Forecasting", *International Journal of Forecasting*, **8** (1992), 135-156.

Granger, C. W. J. and Andersen, A. P., *Introduction to Bilinear Time Series Models*, Vandenhoeck and Ruprecht, Gottingen, 1978.

Granger, C. W. J. and Ramanathan, R., "Improved Methods of Combining Forecasts", *Journal of Forecasting*, **3** (1984), 197-204.

Gupta, P., "Adaptive Short-term Forecasting of Hourly Loads Using Weather Information", in: Bunn, D. W. and Farmer, E. D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985.

Haido, T., and Muto, S., "Regression Based Peak Load Forecasting Using a Transformation Technique", *IEEE Transactions on Power Systems*, **9** (1994), 1788-1794.

Harris, J. L and Liu, M., "Dynamic Structural Analysis and Forecasting of Residential Electricity Consumption", *International Journal of Forecasting*, **9** (1993), 437-455.

Harrison , P. J. and Stevens, C. F., "A Bayesian Approach to Short-term Forecasting", *Operations Research Quarterly*, **22** (1971), 341-362.

Harvey, A. C. , *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, 1989.

Harvey, A. C. and Durbin, J., "The Effects of Seat Belt Legislation on British Road Causalities: A Case Study in Structural Time Series Modelling", *Journal of the Royal Statistical Society Association*, **149** (1986), 187-227.

Harvey, A. C. and Pierse, R. G., "Estimating Missing Observations in Economic Time Series", *Journal of the American Statistical Association*, **79** (1984), 125-131.

Harvey, A. C., "A Unified View of Statistical Forecasting Procedures", *Journal of Forecasting*, **3** (1984), 245-275.

Harvey, A. C., and Peters, S, "Estimation Procedures for Structural Time Series Models", *Journal of Forecasting*, **9** (1990), 89-108.

Harvey, A. C., *The Econometric Analysis of Time Series*, Oxford Press, 1981.

Hect-Neilsen, R. *Neurocomputing*, Addison Welsley, 1990.

Hegazy, Y. G. and Salama, M. A., "An Efficient Algorithm for Identifying the Structure of Artificial Neural Networks for Forecasting Problems", in: *Proceedings of the 37th Midwest Symposium on Circuits and Systems*, (1994) 618-621.

Hertz, J., Krogh, A. and Palmer, R. G., *Introduction to the Theory of Neural Computing*, Addison-Wesley,1991.

Hibon, M. "Naive, Moving Average, Exponential Smoothing and Regression Methods", in: Makridakis, A., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., *The Forecasting Accuracy of Major Time Series Methods*, John Wiley and Sons, 1984, 239-244.

Holden, K. and Peel, D. A., "An Empirical Investigation of Combinations of Economic Forecasts", *Journal of Forecasting*, **5** (1986), 229-242.

Howrey, E. P., Hymans, S. H. and Donihue, M. R., "Merging Monthly and Quarterly Forecasts: Experience with MQME", *Journal of Forecasting*, **10** (1991), 255-268.

Hudson, J. L., Kube, M., Adomaitis, R. A., Kevrekidid, I.G., Lapedes, A. S. and Farber, R. M., "Nonlinear Signal Processing and System Identification Applications to Time Series from Electrochemical Reactions", *Chemical Engineering Science*, **45** (1990) 2075-2081.

Hunt, K. A. and Sbarboro, D., "Studies in Neural Network Control", in Warick, K. *Applied Artificial Intelligence*, Peregrinus, 1991, 94-120.

Hush, D. R. and Horne, B. G., "Progress in Supervised Networks", *IEEE Signal Processing Magazine*, (January 1993), 8-39.

Hyde, O. and Hodnett, P.F., "An Electricity Load Forecasting System" in: *Proceedings of the 5th European Conference on Mathematics in Industry*, (1991) 99-103.

Isakasson, A. "Identification of ARX Subject to Missing Data", *IEEE Transactions on Automatic Control*, **5**, 813-819.

Janecek, G. and Swift, L., *Time Series Forecasting, Simulation, Applications*, Ellis Horwood Ltd., 1993.

Jin, L., Nikiforuk, P. N. and Gupta, M., "Adaptive Control of Discrete-Time Nonlinear Systems using Recurrent Neural Networks" *IEEE Proceedings of Control Theory and Applications*, **141** (1994), 169-176.

Jones, A. J., "Genetic Algorithms and Their Applications to the Design of Neural Networks", *Neural Computing & Applications*, **1** (1993) 32-45.

Jones, R. H., "Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations", *Technometrics*, **22** (1980), 389-395.

Kalman, R. E. "On the General Theory f Control Systems", in: *Proceedings of the 1st IFAC Congress*, Moscow, (1960), 481-492.

Kang, H., "A composite Model for Deterministic and Stochastic Trends", *International Journal of Forecasting*, **6** (1990), 175-186.

Kermanshahi, B. S., Poskar, C. H., Swift, G., McLaren, P., Pedryez, W., Bhur, W. and Silk, A., "Artificial Neural Network for Forecasting Daily Loads of a Canadian Electric Utility", in: *Proceedings of the 2nd International Forum on Applications of Neural Networks to Power Systems*, Yokohama, Japan, (1993) 302-307.

Kiernan, L. A., Hannan, J.M., Bishop, M., Mitchell, R.J. and Kambhampati, C. "Neural Networks for Load Reshaping", in: *Proceedings of the 29th Universities Power Engineering Conference*, Galway, Ireland (1994) 358-361.

Kitagawa, G. and Gerch, W. "A Smoothness Priors-State Space Modelling of Time Series With Trend and Seasonality", *Journal of American Statistical Association*, **79** (1984), 378-389.

Kitagawa, G., "A Nonstationary Time Series Model And Its Fitting By A Recursive Filter", *Journal of Time Series Analysis*, **2** (1981),103-116.

Kohn, R. and Ansley, C. F., "Estimation, Prediction and Interpolation for ARIMA Models With Missing Data", *Journal of American Statistical Association*, **81** (1986), 751-761.

Kohn, R. and Ansley, C. F., "Filtering And Smoothing Algorithms For State Space Models", *Computers and Mathematics with Applications*, **18** (1989), 515-528.

Komprej, I. and Zunko, P., "Short Term Load Forecasting", in: *6th Mediterranean Electrotechnical Conference Proceedings,* Ljublijana, Yugoslavia (1991) 1470-3.

Koreisha, S. G. and Pukkila, T., "The Identification Of Seasonal Autoregressive Models", *Journal of Time Series Analysis*, **16** (1995), 267-290.

Kuschewski, J. G., Hui, S. and Zak, S. *Application of Feedforward Neural Networks to Dynamical System Identification and Control*, IEEE Trans. on Control System Tech., **1** (1993).

Lachtermacher, G. and Fuller, J. D., "Backpropagation in Time-series Forecasting", *Journal of Forecasting*, **14** (1995), 381-393.

Lawson, C. L. and Hanson, R. J., *Solving Least Squares Problems*, Prentice Hall, 1974.

Lewandowski, R., "Sales Forecasting by FORSYS", *Journal of Forecasting*, **1** (1982), 205-214.

Lin, W. T., "Modelling and Forecasting Hospital Patient Movements: Univariate and Multiple Time Series Approaches", *International Journal of Forecasting*, **5** (1989), 195-208.

Lippmann, R. P., "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, (April 1987), 4-22.

Ljung, L., *System Identification - Theory for the User*, Prentice Hall, 1987.

Lonergan, T and Ringwood, J.V. "Linguistic Modelling of Short-Time-Scale Electricity Consumption Using Fuzzy Modelling Techniques", in *Proceedings of the 6th Irish DSP and Control Colloquium*, Belfast, (1995), 171-178.

Lu, C. N., Wu, H. T. and Vemuri, S. "Neural Network Based Short Term Load Forecasting", *IEEE Transactions on Power Systems*, **8** (1993), 336-342.

Mahmoud, E. and Pegels, C., "An Approach for Selecting Time Series Forecasting Models", *International Journal of Quality Management*, **10** (1989), 50 61.

Makridakis, A., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition", *Journal of Forecasting*, **1**, (1982), 111-153.

Makridakis, A., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R., *The Forecasting Accuracy of Major Time Series Methods*, John Wiley and Sons, 1984.

MATLAB Identification Toolbox For Use with MATLAB User's Guide, The MATHWORKS Inc., 1992.

MATLAB NAG Foundation Toolbox For Use with MATLAB User's Guide, The MATHWORKS Inc., 1995.

MATLAB Neural Network Toolbox For Use with MATLAB User's Guide, The MATHWORKS Inc., 1994.

MATLAB User's Guide, The MATHWORKS Inc., 1993.

Matthew"s, B. P. and Diamantopoulos, A., "Towards a Taxonomy of Forecast Error Measures", *Journal of Forecasting*, **13** (1994), 409-416.

Mbamalu, G. A. N. and El-Hawary, M. E., "Load Forecasting Via Seasonal Autoregressive Models And Iteratively Reweighted Least Squares Estimation", *IEEE Transactions on Power Systems*, **8** (1993), 343-348.

McCafferty, P., Beattie, W. and Dodds, G., "Time Series Modelling of Industrial Electricity Demand", in: *Proceedings of the 25th Universities Power Engineering Conference*, Aberdeen, UK, (1990) 595-598.

Mills, T. C. and Stephenson, M. J., "Forecasting Contemporaneous Aggregates and the Combination of Forecasts: the Case of the U.K. Monetary Aggregates" *Journal of Forecasting*, **4** (1985), 273-281.

Mitchell J., "Comparing Feedforward Neural Network Models for Time Series Prediction", in: *Proceedings Neural Computing Research and Applications*, Belfast, Northern Ireland, (1992) 75-182.

Moghrqm, I. and Rahman, S., "Analysis and Evaluation of Five Short-term Load Forecasting Techniques", *IEEE Transactions on Power Systems* **4** (1989) 1484-91.

Morioka, Y., Sakurai, K., Yokoyama, A. and Sekine, Y., "Next Day Peak Load Forecasting Using Multi Layer Neural Network with an Additional Learning", in: *Proceedings of the 2nd International Forum on Applications of Neural Networks to Power Systems*, Yokohama, Japan, (1993) 60-5.

Morrison, D., *Multivariate Statistical Methods*, McGraw-Hill, New York, 1976.

Natarajan, B. and Tandan, G., "Methodology for On-Line Forecasting of Loads For Electric Utilities", in: *IFAC Energy Systems, Management and Economics*, Toyko, Japan, (1989) 247-250.

Ng C. N. and Young P. C., "Recursive Estimation and Forecasting of Non-stationary Time Series" *Journal of Forecasting*, **9** (1990), 173-203.

Ng, C. N., Young, P. C. and Wang, C., "Recursive Identification, Estimation And Forecasting Of Multivariate Time-Series", in: *IFAC Identification and System Parameter* Estimation, Beijing, PRC, (1988) 593-598.

O'Sullivan, C. "Time Series Modelling Using Fuzzy Logic", Final Year Dissertation, School of Electronic Engineering, Dublin City Univ., Dublin , 1994.

Onoda, T., "Next day"s Peak Load Forecasting Using an Artificial Neural Network", in: *Proceedings of the 2nd International Forum on Applications of Neural Networks to Power Systems*, Yokohama, Japan, (1993) 284-289.

Pankratz, A., *Forecasting with Univariate Box-Jenkins Models Concepts and Cases*, John Wiley & Sons, 1983.

Papalexopoulos, A. D., Hao, S. and Peng, T-M., "An Implementation Of A Neural Network Based Load Forecasting Model For The EMS", *IEEE Transactions on Power Systems*, **9** (1994), 1956-1962.

Park, D.C., El-Sharkawi, R. J., Marks, I. I. R. J., Atlas, L. E. and Damborg, M. J., "Electric Load Forecasting Using an Artificial Neural Network", *IEEE Transactions on Power Systems*, **6** (1991), 442-449.

Parzan, E., "ARARMA Models for Time Series Analysis and Forecasting" *Journal of Forecasting*, **1** (1982), 67-82.

Peng, T. M., Hubele, N. F. and Karady, G. C., "An Adaptive Network Approach to One-Week Ahead Load Forecasting" *IEEE Transactions on Power Systems*, **8** (1993), 1195-1201.

Piggott, J. L. "Short-term Forecasting at British", in: Bunn, D. W. and Farmer, E, D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985, 173-212.

Priestly, M. B., *Spectral Analysis and Time Series. Volume 1: Univariate Series*, Academic Press, 1981.

Priestly, M. B., *Spectral Analysis and Time Series. Volume 2: Multivariate Series, Prediction and Control*, Academic Press, 1981.

Qayyum, M. A. and Hamed, M. N., "Short Term Peak Demand Forecasting in Fast Developing Utility with Inherent Dynamic Load Characteristics: Part -1. Application of Classical Time Series Methods", *IEEE Transactions on Power Systems*, **5** (1990), 813-819.

Rahman, R. and Hazim, O., "A generalised Knowledge-Based Short-Term Load-Forecasting Technique", *IEEE Transactions on Power Systems*, **8** (1993), 508-514.

Rissanen, J. "Modelling by Shortest Data Description", *Automatica,* **14**, 465-471.

Rosenblat, F., *Principles of Neurodynamics*, Spartan, 192

Rummelhart, D.E., Hinton, G.E. and Williams, R.J., " Learning Internal Representations by Error Propagation", , in: Rummelhart and McClelland, J.L., eds., *Parallel Distributed Processing*, MIT Press, 1986.

Schneider, A., Takenawa, T. and Schiffman, D., "24-hour Electric Load Forecasting", in: Bunn, D. W. and Farmer, E. D., eds., *Comparative Models for Electrical Load Forecasting*, John Wiley & Sons, 1985.

Sharda, R. and Patil, R. "*Neural Networks as Forecasting Experts: An Empirical Test*", in; *Proceedings of the International Joint Conference on Neural Networks*, 1990, 491-494.

Shi, S. and Lui, B., "Nonlinear Combination Of Forecasts With Neural Networks", in: *Proceedings of 1993 International Joint Conference on Neural Networks*, Nagoya, Japan (1993) 959-962.

Smyth, I. , "The Prediction of Weekly Electrical Energy Consumption Using Input Driven Neural Networks", MEng. Dissertation, School of Electronic Engineering, Dublin City Univ., Dublin , 1994.

Snedecor, G. and Cochran, W., *Statistical Methods* Iowa State University Press, Iowa, 1980.

Sobajic, D.J. and Pao, Y.H., "Artificial Neural Networks Based Dynamics Security Assessment for Electric Power Systems", *IEEE Transactions on Power Systems*, **4**, (1989), 220-228.

Soderstrom, T. and Stoica, P., *System Identification*, Prentice Hall International, 1989.

Srinivasan, D., Chang, C. S. and Liew, A. C., "Demand Forecasting Using Fuzzy Neural Computation, With Special Emphasis On Weekend And Public Holiday Forecasting", *IEEE Transactions on Power Systems*, **10** (1995), 1897-1903.

Stetson, L. E. and Stark, G. L., "Peak Electricity Demands of Individuals and Groups of Rural Residential Customers", *IEEE Transactions on Industry Applications*, **24** (1988), 772-776.

Taha, H. A., *Operations Research: An Introduction*, Collier MacMillan, 1987.

Tamura, Y., Suszuki, H. and Mori, H. "Another Look at Forecasting with Neural Networks", in *Proceedings of the 2nd International Forum on Applications if Neural Networks to Power Systems*, Japan, 1993.

Tang, Z., Almeida, Fishwick, P.A. "Time Series Forecasting Using Neural Networks vs Box-Jenkins Methodology", *Simulation*, **57** (1991), 101-310.

Tang, Z., *Nonlinear Time Series a Dynamical Systems Approach*, Oxford University Press, Oxford, 1990.

Train, T., Ignelz, P., Engle, R., Granger, C. and Ramanathan, R., "The Billing Cycle And Weather Variables In Models Of Electricity Sales", *Energy*, **9** (1984) 1041-1047.

Vemuri, V. R. and Rogers, R. D., *Artificial Neural Networks: Forecasting Time Series*, IEEE Press, 1994.

Vermuri, V. and Rogers, R. D., *Artificial Neural Networks: Forecasting Time Series*, IEEE Computer Society Press, 1994.

Villiers, J de., and Barnard, E., "Backpropagation Neural Nets with One and Two Hidden Layers", *IEEE Transactions on Neural Networks*, **4** (1992), 136-141.

Warwick, K., Irwin, G. W. and Hunt, K. J., *Neural Networks for Control and Systems*, Peter Peregrinus Ltd., 1992.

Weber, K., "Forecasting Software: Evaluation and Usage", *Methods of Operational Research*, **62** (1990), 591-600.

Weigend, A. S. and Gershenfeld, N. A., *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley, 1992.

Winkler, R. L. and Makridakis, S., "The Combination of Forecasts", *Journal of the Royal Statistical Society Association*, **146** (1983), 150-157.

Wong, F. S., "Time Series Forecasting Using Backpropagation Neural Networks", *Neurocomputing*, **2** (1990/91), 147-159.

Wu, L. S. -Y., Ravishanker, N. and Hosking, J. R. M., "Forecasting for Business Planning: A Case Study of IBM Product Sales", *Journal of Forecasting*, **10** (1991), 579-595.

Yang, H-T, Huang, C-M. and Huang, C. L, "Identification of ARMAX Model for Short-Term Load Forecasting: An Evolutionary Programme Approach", *IEEE Transactions on Power Systems*, **11** (1996), 403-408.

Young, P. C. , *Recursive Estimation and Time-Series Analysis*, Berlin: Springer-Verlag, 1984.

Young, P. C., "Recursive Extrapolation, Interpolation And Smoothing Of Nonstationary Time-Series", in: *IFAC Symposium on Identification and System Parameter Estimation*, Beijing, PRC, (1988) 35-46.

Young, P. C., "Time-variable Parameter and Trend Estimation in Non-stationery Economic Time Series", *Journal of Forecasting*, **13** (1994), 179-210.

Young, P. C., Ng, C. N., and Armitage, P., "A Systems Approach To Recursive Economic Forecasting And Seasonal Adjustment", *International Journal on Computers and Mathematics with Applications*, **18** (1989), 481-501.

Young, P. C., Ng, C. N., Lane, K., and Parker, D., "Recursive Forecasting, Smoothing and Seasonal Adjustment of Non-stationary Environmental Data", *Journal of Forecasting*, **10** (1991), 57-89.

Zaknich, A, deSilvia, C.J.S. and Attikiouzel, Y., "A Modified Probabilistic Neural Network for Nonlinear Time Series Analysis", in: *Proceedings International Joint Conference on Neural Networks*, (1991) 1530-1535.

APPENDIX A

The Kalman Filter and Maximum Likelihood Estimation via Prediction Error Decomposition

A.1 The Kalman Filter

Considering the state space form given in (2.37) and (2.38) let the optimal estimator of $\mathbf{x}(k\text{-}1)$, based on observations up to and including $y(k)$, be given by $\hat{\mathbf{x}}(k-1)$. Let $\mathbf{P}(k\text{-}1)$ denote the n x n covariance matrix of the estimation error:

$$\mathbf{P}(k-1) = E\left[(\mathbf{x}(k-1) - \hat{\mathbf{x}}(k-1))(\mathbf{x}(k-1) - \hat{\mathbf{x}}(k-1)^T\right] \tag{A1}$$

Given $\hat{\mathbf{x}}(k-1)$ and $P(k\text{-}1)$, the optimal estimator of $x(k\text{-}1)$ is given by

$$\hat{\mathbf{x}}(k|k-1) = \mathbf{F}\hat{\mathbf{x}}(k-1) \tag{A2}$$

with the covariance matrix of the estimation error is

$$\mathbf{P}(k|k-1) = \mathbf{F}\mathbf{P}(k)\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \tag{A3}$$

Equations (A2) and (A3) are known as the prediction equations. Once the latest observation becomes available the estimator of $\mathbf{x}(k)$ can be updated using the following updating equations:

$$\mathrm{x}(k) = \mathrm{x}(k|k-1) + \mathrm{P}(k|k-1)\mathrm{H}^T\Sigma(k)^{-1}\left(y(k) - \mathrm{H}\mathrm{x}(k|k-1)\right) \tag{A4}$$

and

$$\mathrm{P}(k) = \mathrm{P}(k|k-1) - \mathrm{P}(k|k-1)\mathrm{H}^{\mathrm{T}}\Sigma(k)^{-1}\,\mathrm{H}\mathrm{P}(k|k-1) \tag{A5}$$

where $\Sigma(k) = \mathrm{H}^T\mathrm{P}(k|k-1)\mathrm{H} + \sigma_\varepsilon^2$.

If the recursions are started at k=1 the estimator of the state at k=0, $\mathbf{x}_0$ and $\mathbf{P}_0$ are required where the prediction equations are used to claculate $\hat{\mathbf{x}}(1|0)$ and $\mathbf{P}(1|0)$. Given the initial conditions, $\mathbf{x}_0$ and $\mathbf{P}_0$ the Kalman filter computes the optimal estimator of the state vector as each new observation becomes available. When all N observations have been processed, the filter gives the optimal estimator of the current state vector based on the full set of observations. This estimator of the current state contains all the information required to make optimal future predictions of the state and thus also future predictions of the observations.

The Kalman filter may also be used to carry out fixed interval smoothing (FIS), where this consists of a set of recursions which start with the final quantities $\mathbf{x}(N)$ and $\mathbf{P}(N)$ given by the Kalman filter and working backwards. The equations are: (Harvey, 1989)

$$\hat{\mathbf{x}}(k|N) = \hat{\mathbf{x}}(k) + \mathrm{P}(k)\mathrm{F}\mathrm{P}^{T}(k+1|k)\big(x(k+1|N) - \mathrm{F}x(k)\big) \tag{A6}$$

$$\mathrm{P}(k|N) = \mathrm{P}(k) + \mathrm{P}(k)\mathrm{F}\mathrm{P}^{T}(k+1|k)\big(\mathrm{P}(k+1|N) - \mathrm{P}(k+1|k)\big)\big(\mathrm{P}(k)\mathrm{F}\mathrm{P}^{T}\big)^{T} \tag{A7}$$

with $x(N|N) = x(N)$ and $\mathbf{P}(N|N)=\mathbf{P}(N)$.

A.2 Maximum Likelihood Estimation for Time Series

Maximum likelihood is first considered for the non-time series case. Given a sample of independent and identically distributed observations $y(k)$, $k=1,....,N$, each with p.d.f $f(y(k))$, the joint density function is

$$f\big(y(1), y(2), ..., y(N)\big) = \prod_{k=1}^{N} f\big(y(k)\big) \tag{A8}$$

The likelihood function is this joint probability but is a function of the parameter vector ψ and is written

$$L(\psi) = \prod_{k=1}^{N} f\,(y(k);\psi) \tag{A9}$$

Therefore $\mathrm{L}(\psi)$ is interpreted as the likelihood function and the maximum likelihood estimates (MLEs) are found by maximising this function with respect to ψ.

The problem with time series is that it is assumed that the observations are not independent and thus (A7) is not applicable. Denote all the observations up to and including $y(k)$ by $\mathbf{Y}(k)$, that is $\mathbf{Y}(k)=\{y(k),y(k\text{-}1), y(k\text{-}2),.....,y(1)\}$.

$$\begin{aligned} f(y(1), y(2),.., y(N)) &= f(y(N)|Y(k))f(y(1), y(2),..., y(N-1)) \\ &= f(y(N)|Y(N-1))f(y(N-1)|Y(N-2))f(y(1), y(2),..., y(N-2)) \end{aligned} \tag{A10}$$

and so on until

$$f(y(1), y(2),....., y(N)) = \prod_{k=1}^{N} f(y(k)|\mathbf{Y(k-1)}) \tag{A11}$$

where $f(y(k)|Y(k-1))$ is the distribution of $y(k)$ conditional on the information set $\{y(k\text{-}1), y(k\text{-}2),.....,y(1)\}$. The likelihood of the time series $y(k)$ can be written as the product of these conditional distributions given by:

$$L(\psi) = \prod_{k=1}^{N} f(y(k)|y(1), y(2),...., y(k-1)) \tag{A12}$$

Equivalently the log-likelihood can be calculated:

$$\log L(\psi) = \prod_{k=1}^{N} \log f(y(k)|y(1), y(2),...., y(k-1)) \tag{A13}$$

A.3 Maximum Likelihood Estimation via Prediction Error Decomposition

For the general state space model (2.37 and 2.38) it was assumed that $\varepsilon(k)$ is a white noise distrubance term and $\eta(k)$ is a vector white noise distrubance term, therefore only the second order moments are known and the form of their probabilty distributions has not been specified. To use maximum likelihood estimation it is necessary to determine the probabilty distribution of these disturbances and thus of the time series $y(k)$. The usual assumption is that $\varepsilon(k)$ is normally distributed and that $\eta(k)$ has a multivariate normal distribution. Using the state space model each observation $y(k)$ is a linear function of normal distrubance terms and therefore the time series $y(k)$ is also multivariate normal. Consequently the distribution of any observation conditional on any other observation must also be normal, by the properties of multivariate normal distribution. Specifically the conditional distribution of

$f(y(k)|y(1), y(2), \ldots, y(k-1))$ which is required for the likelihood is normal. Furthermore the mean and covariance matrix of this conditional distribution are given directly by the Kalman filter (Janecek and Swift 1993). Conditional on $\{y(k\text{-}1),\ y(k\text{-}2), \ldots, y(1)\}$ $\mathbf{x}(k)$ is normally distributed with a mean $\hat{\mathbf{x}}(k|k-1)$ and a covariance $\mathbf{P}(k|k\text{-}1)$. If the measurement equation (2.37) is written

$$y(k) = \mathbf{H}\hat{\mathbf{x}}(k|k-1) + \mathbf{H}\left(\mathbf{x}(k) - \hat{\mathbf{x}}(k|k-1)\right) + \boldsymbol{\varepsilon}(k) \tag{A14}$$

then the conditional distribution of $y(k)$ is normal with mean

$$E\left(y(k)\right) = \mathbf{H}\hat{\mathbf{x}}(k|k-1) = \mathbf{H}\hat{\mathbf{x}}(k|k) \tag{A15}$$

and a covariance matrix given by $\Sigma(k) = \mathrm{H}^T \mathrm{P}(k|k-1)\mathrm{H} + \sigma_\varepsilon^2$ (Janecek and Swift, 1993)

For this normal model the likelihood of (A13) can be written (Harvey, 1981):

$$\log f\left(y(k)|y(1), y(2), \ldots, y(k-1)\right) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \Sigma(k) - \frac{1}{2\Sigma(k)}\left(v(k)\right)^2 \tag{A16}$$

where $v(k) = \left(y(k) - \hat{y}(k|k-1)\right) = \left(y(k) - \mathbf{H}\hat{\mathbf{x}}(k|k-1)\right)$, with the log-likelihood given by:

$$\log L(\boldsymbol{\psi}) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{N}\log \Sigma(k) - \frac{1}{2}\sum_{k=1}^{N}\frac{v(k)^2}{\Sigma(k)} \tag{A17}$$

The vector $v(k)$ is interpreted as a vector of prediction errors since $\hat{y}(k|k-1)$ is the optimal one step predictor of $y(k)$ at time $k\text{-}1$ as well as the conditional expectation of $y(k)$ at $k\text{-}1$. For this reason (A17) is often referred to as the prediction error decomposition of the likelihood. The values of $\Sigma(k)$ and $v(k)$ can be computed using the Kalman filter and the likelihood function (A17) may be found. The likelihood function is then maximised with respect to the unknown parameters $\boldsymbol{\psi}$, where this is usually carried out using some kind of numerical optimisation.