UNIVERSITY OF AGDER

Short-term Forecasting of Electricity Consumption using Gaussian Processes

Milindanath Samarasinghe and Waseem Al-Hawani

Supervisor
Ole-Christoffer Granmo

This Master's Thesis is carried out as a part of the education at the University of Agder and is therefore approved as a part of this education. However, this does not imply that the University answers for the methods that are used or the conclusions that are drawn.

University of Agder, 2012
Faculty of Engineering and Science
Department of Information and Communication Technology

Abstract

Forecasting of electricity consumption is considered as one of the most significant aspect of effective management of power systems. On a long term basis, it allows decision makers of a power supplying company to decide when to build new power plants, transmission and distribution networks. On a short term basis, it can be used to allocate resources in a power grid to supply the demand continuously.

Forecasting is basically divided into three categories : *short-term*, *medium-term*, and *long-term*. Short-term refers to an hour to a week forecast, while medium-term refers to a week to a year, and predictions that run more than a year refers to long-term.

In this thesis, we forecast electricity consumption on a short-term basis for a particular region in Norway using a relatively novel approach: Gaussian process. We design the best feature vector suitable for forecasting electricity consumption using various factors such as *previous consumptions*, *temperature*, *days of the week* and *hour of the day*. Moreover, feature space is scaled and reduced using reduction and normalization methods, and different target variables are analysed to obtain better accuracy.

Furthermore, GP is compared with two traditional forecasting techniques : Multiple Back-Propagation Neural Networks (MBPNN), and Multiple Linear Regression (MLR). Finally we show that GP is as better as MBPNN and far better than MLR using empirical results.

Preface

This thesis is submitted in partial fulfilment of the degree of Master of Science in Information and Communication Technology (ICT) at University of Agder, Grimstad, Norway. The thesis was carried out under the supervision of professor Ole-Christoffer Granmo. We are delighted to take the opportunity to thank all the people who have made valuable support in completing this work.

Professor Ole-Christoffer Granmo was the strong guiding force in our work, who led us in the correct path towards achieving our final goal. His knowledge and experience in the field and the willingness to share his insights with us in this research was remarkable. Therefore we are grateful to him for all that support. We should also thankful to Eidsiva Energy for providing the dataset that was the basic foundation of our thesis. Our acknowledgement also goes to Noel Lopes for providing his research paper for the neural network testing. Finally, we would like to pay our gratitude to our fellow students who helped us in numerous ways to make this thesis a success.

Grimstad, May 2012
Milindanath Samarasinghe
Waseem Al-Hawani

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In this chapter we direct the reader to the basic introduction of the thesis. In section 1.1, we outline the basic background and motivation of the thesis. We define the thesis definition in section 1.2 and present the basic research questions need to be solved in section 1.3. The remaining three sections: section 1.5, 1.6 and 1.7 detail the contributions supplemented to the research community; the target audience of the thesis; and the outline of the thesis respectively.

1.1 Background and Motivation

Electricity has become a basic need for the mankind today. It is used to carry out a range of day to day work, such as operating domestic and industrial equipments, lighting, heating, air-conditioning, cooking, washing and many other tasks. We can visualize this fact of how it

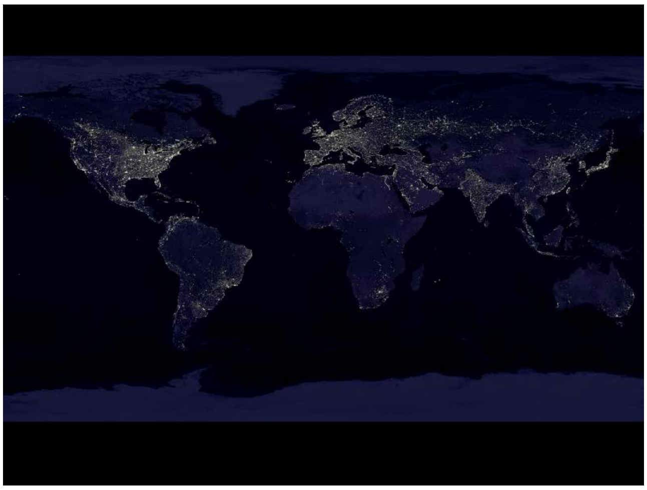


Figure 1.1: Earth at night. [Source : NASA].

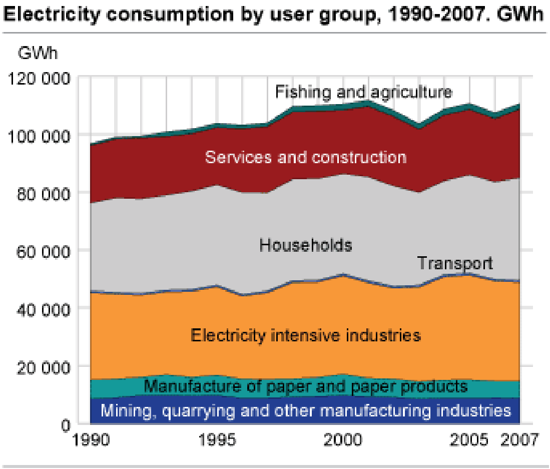


Figure 1.2: Electricity consumption by user group in Norway, 1990-2007 [1].

has become a basic need for mankind, by just looking at the earth at night in Figure 1.1, which illustrates the electricity required to light the earth at night.

This is true for the Norwegian community also, as we can see from Figure 1.2, which illustrates the electricity consumption among different user groups in Norway from 1990 to 2007. The figure indicates an increase in consumption for many user groups over the years. Therefore it is clear that electricity has become an important commodity for the people.

Forecasting of electricity consumption is necessary to manage the power system effectively and thereby fulfil this increasing demand. On a long-term basis, power companies would require the power consumption forecast in the next 10 or 20 years to plan their future activities properly, such as building adequate power plants and improve their transmission and distribution networks to meet the necessary demand. On a short-term basis, forecasting is required to perform daily operations such as unit commitment, energy transfer scheduling and load dispatch of a utility company [3]. Therefore, accurate prediction of electricity consumption is crucial for both, performing daily operations and making future power plans for a power supplying company.

Electricity forecasting can be done mainly in three ways, as *short-term*, *medium-term* and *long-term*. Short-term forecasting refers to an hour to a week prediction, while medium-term considers the range between a week and a year, and predictions that run more than a year are considered as long-term forecasts [4].

In the literature, we can observe various independent variables (predictors), that influence the consumption, have been used for forecasting under these three schemes, especially for short-term and long-term. In long-term forecasting, the most frequently used predictors are the *Gross Domestic Product (GDP)* [5–8] and the *population* [5, 6, 8]. In addition, factors such as *capital* [5], *per capita consumption* [6], *number of consumers* [6], *peak electricity demand* [6] and *electricity price* [8] have also been used for long-term forecasting. For short-term forecasting, *temperature* [9–11] and *humidity* [9,11] have been widely used and *wind-force* [10], *dampness* [10], *chilled water consumption* [9] and *gas fuel consumption* [9] have also been used for forecasting.

The most common approaches of electricity consumption forecasting includes, statistical techniques (multiple linear regression), artificial neural networks (ANN) , Genetic algorithms (GA) , grey methods (GM) and time series analysis. Here, we briefly outline some of those selected attempts as follows. A statistical technique - multiple regression analysis using least squares method has been used by Imitiaz et al. [6] to forecast annual electricity consumption for Malaysia. Fung and Tummala [8] have used ANN models to forecast annual consumption and compared the results with multiple linear regression models. Dedy et al [9] have also used ANN to forecast daily electricity consumption by optimizing its model using Taguchi [12] method. RBF neural networks combined with Genetic algorithms have been used by Zeng et al. [10] for short-term forecasting. Grey methods [13] have been applied to forecast electricity consumption in [5]. A time series analysis based on auto regressive model done by Wang Baosen et al. [7] has been carried out for long-term electricity consumption forecasting.

In this thesis our approach is to forecast short term electricity consumption on an hourly basis which extends up to predicting the behaviour of the next 24 hours, using a relatively novel approach : Gaussian Process (GP) .

GP possess several attributes that make it a potential model to solve supervised learning problems which emerged over the last decade. The main reason for this is its non-parametric, kernel based nature, which provides more flexibility in solving inferring problems. GP is more a data-oriented approach which is capable of learning functions from the data itself. Moreover, it can be used for non-linear regression problems which makes it more suitable for forecasting electricity which is non linear in nature.

Gaussian process has been used before in a range of fields especially for estimating purposes. Here we concentrate on several of these researches. Gaussian processes has been used by Luo and Qian [14] to estimate the colour values in Poly Ethylene Terephthalate (PET) . Vathsangam et al. [15] have used GPs to estimate walking speed of a human using on-body accelerometers and gyroscopes. As Pasolli et al. propose in [16], it has been also used for estimating chlorophyll concentration in subsurface waters by sensing data remotely. According to Bazi and Melgani [17], semi-supervised GPs has been used for the estimation of biophysical parameters from remote sensing data. Zhang and Yeung [18] have suggested, Multi-Task Warped Gaussian Process (MTWGP) , which is a variant of GP, for personalized age estimation of people.

GP has been used for short-term electricity forecasting on two occasions in the literature. Firstly, by Mori and Ohmi [11], and secondly by Alamaniotis et al. [4]. In the first, GP has been used to forecast daily consumption and hierarchical Bayesian model has been used to evaluate the hyperparameters of the covariance function. Moreover, they have compared GP with conventional methods such as MLP , RBFN and SVR . In the second, several kernel functions have been built and the predictors have been evaluated using Genetic algorithm.

In our approach, we use GP on its original context but more focusing on the data itself by modifying both feature vector and target variable. We use techniques such as reduction and normalization to modify the feature vector and test for different target variables that gives better results. Moreover, we compare GP with two traditional techniques: *multiple linear regression* (MLR) and *multiple back-propagation neural networks* (MBPNN) , which have not been used in [11]. More focus will be given to seasonal changes in consumption, as there is a tie between weather and consumption in a country like Norway. Therefore, it is interesting to see how GP cope with these variations as it has been found in [19] that, traditional techniques such as linear regression and NN are not much sensitive to variations in the local structure of the input space and hence we can examine the ability of GP in estimating varying electricity consumption. However as stated clearly by Imitiaz et.al [6], forecasts are never going to be perfect and the validity of the general rule '*The forecast is always wrong*' has to be assumed.

1.2 Thesis Definition

The thesis definition can be formulated as follows:

This thesis proposes Gaussian Processes (GP) as a novel approach for short term forecasting of electricity consumption. It includes designing the optimum feature vector by concerning various factors affecting the electricity consumption. This should consist of a data filtering strategy and data scaling technique. In addition, it is required to use a corresponding kernel function for achieving best predictions. The core of the thesis is the design of a technique for exploring trends of electricity consumption through cyclic patterns and hence normalize the data in such a way that accurate predictions could be obtained. Moreover, it is expected to compare the forecast results of GP with the traditional forecasting techniques.

1.3 Research Questions

Throughout this thesis, we will answer to the research questions outlined below.

What are the cyclic patterns that we can observe in electricity consumption and what are the influential factors causing these cycles?

Electricity consumption will be different from hour to hour, day to day, month to month, season to season and year to year. But they can have repeating cyclic patterns. For this purpose, we consider a dataset, provided by *Eidsiva Energy*[1] which contains electricity consumption data for 3 years. Using this dataset, we have to find out the patterns of electricity consumption and the factors that influence these patterns.

[1] http://www.eidsivaenergi.no/

What is the best feature vector possible for forecasting electricity consumption under GPs?

As we know, electricity consumption may depend on many factors as mentioned in section 1.1. In our approach we focus mainly on the weather factors(temperature, wind, cloud cover), season of the year (summer, autumn, spring, winter), days of the week, hours of the day and previous electricity consumption values. The most influencing factors could be identified once we answer the first research question about cyclic patterns, and studying more about the dataset. Hence we could identify a feature vector(s) that is more suitable to obtain better prediction results.

What type of target variables we can use for better prediction accuracy?

It is obvious that the electricity consumption is the default target variable. However, we need to explore another target variable(s) which is(are) capable of providing better results. This would require a thorough emphasis on the dataset.

What are the strategies used to improve the predictions?

We need to test different techniques, and features to find out the best possible combinations of the input feature space that produce the accurate results. This might include scaling, normalizing and reducing methods.

How good the predictions of GP compared with the traditional techniques?

It is necessary to compare the results of GP with the traditional techniques used to forecast electricity consumption. Here, we propose two traditional techniques : *multiple linear regression* and *multiple back-propagation neural networks.*

1.4 Research Approach

The main starting point for this thesis is the dataset obtained from Eidsiva Energy as mentioned above. This dataset consists hourly electricity consumption measured in three successive years 2008, 2009 and 2010. In addition to the electricity consumption, temperature, cloud cover, wind speed are also stated in this dataset.

Through extensive studies on the dataset, we answer the research questions outlined above one by one using GP. Dataset will be utilized under different categories based on their attributes and features.

GP testing is done using the pyXGPR [35] python code library developed by Marion Neumann et al. This tool is based on the GP theories explained by Rasmussen and Williams [26].

1.5 Contributions

In this thesis, we introduce Gaussian process as a novel approach for short-term electricity forecasting on a seasonal basis, for a region where electricity consumption is related with the weather condition. The best feature vectors for forecasting electricity consumption on seasonal basis are found using the weather factors, previous electricity consumptions, days of the week, and hour of the day.

In addition to the input feature space, this thesis tests for different target variable analysis, to obtain better results.

Moreover, we evaluate the results of GP when reduction and normalization methods are applied to the feature space, in terms of electricity forecasting.

Finally, GP is compared with two traditional techniques : multiple linear regression and multiple back-propagation neural networks, which have not been compared in [11]. This will allow us to examine how good GP process compared with the other techniques.

1.6 Target Audience

This thesis can be referred by anyone who is interested in the machine learning field. Specially for those who have interest in the emerging technique of Gaussian processes.

Those who are in electric utility companies might be very much interested in forecasting electricity consumption. Therefore, this thesis work will be beneficial for them who are seeking improvements in their electricity supplying process.

In general, anyone who is working with the computer science, electrical engineering and statistics can use this thesis as a reference. However the readers are assumed to have the background knowledge on statistics theory and machine learning.

1.7 Thesis Outline

Mainly, the thesis is organized into eight chapters as follows. Chapter 1 covers the basic introductory part of the thesis, the background and motivation ,the research questions that we should solve, the thesis definition, contributions to the knowledge and the target audience.

Chapter 2 illustrates how the electricity generation, transmission and distribution process works, the consumption patterns of the consumers and the previous approaches of forecasting.

Chapter 3 covers the basic theoretical background of GP and GP regression, with a short introduction to the random variable theories.

Chapter 4 is devoted to explore the proposed traditional forecasting techniques used for electricity forecasting: linear regression and artificial neural networks.

Chapter 5 is concerned with the proposed solution on how to forecast electricity consumption using Gaussian processes. Chapter 6 outlines the empirical results obtained from different tests including GP and traditional methods. Chapter 7 analyses the results obtained and compare GP with the traditional techniques. Finally, Chapter 8 concludes the findings by exploring the whole thesis work.

Chapter 2

Electricity Consumption and Forecasting

In chapter 1, we gave a brief introduction about the background of electricity consumption forecasting, with the traditional forecasting techniques and the need of Gaussian process. However, first we should understand the basics of an actual power system and how it processes in the practical situation and how forecasting is going to do any good for the improvement of the system. In section 2.1, we discuss some of the theoretical features of a typical power system and how it fits with our case. Then we move on to the consumption of electricity, produced by the system based on our dataset and the different features of consumption in section 2.2. In the last section we discuss the importance of forecasting and give some details about the previous approaches in forecasting.

2.1 Electricity Supplying Process

There is a complex process associated in delivering electricity safely and reliably to households and industries. This process involves three basic stages : generation, transmission and distribution. Figure 2.1 illustrates these three stages in a typical electric power system.

The root of power generation is the power plant, which includes turbines and generators. There are different types of power plants based on the source of energy, such as hydro power, thermal, wind, solar and so on. The main source of electricity in Norway is the hydro-power and it constitutes 99 % of the overall production [20].

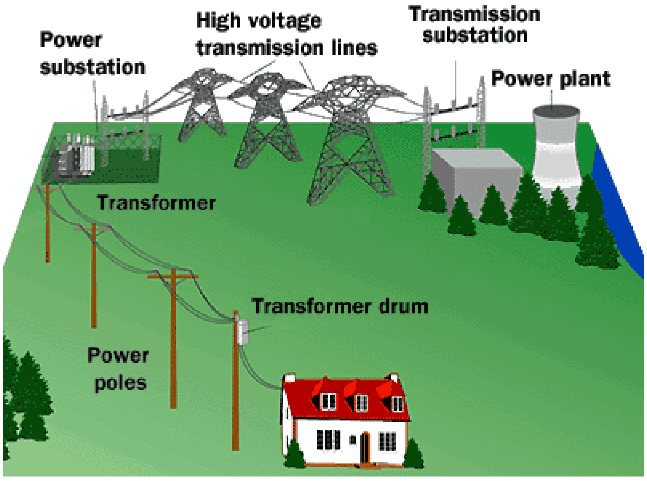


Figure 2.1: A typical power system with generation, transmission and distribution phases. [Source: http://science.howstuffworks.com/environmental/energy/power.htm].

In the next stage, three-phase power generated by this power plant is transmitted to a long distance through high voltage transmission lines as shown in the figure. Before transmitting, the voltage of the generated power, is increased (step-up) by the transformers in the transmission substation.

These transmission lines are then connected to power substations where the high voltage power is stepped down to a level that can be distributed across a certain region. This step is the last stage of the power system which is also known as the distribution stage. From the distribution transformer, the distribution bus lines carries electricity to the households and industries.

In short-term electricity forecasting, the main focus goes to the distribution stage. This is because, the controlling process could be only carried out on this stage by allocating proper resources on a short-term basis.

2.2 Electricity Consumption

In Norway, the authorities who distribute electricity from the final stage of the process are keen to provide better and continuous supply to their consumers all the time. In this process, they have to control and monitor the distribution of electricity through their grids according to the demand from the consumers.

Eidsiva Energy records the power consumption in the region on an hourly basis. In addition to the power consumption, the forecast temperature, actual temperature, wind speed and cloud cover are also measured with the consumption. These measurements which we found in the dataset are the basis for the forecasting carried out in this thesis.

Figure 2.2 depicts the electricity consumption for the three years: 2008, 2009 and 2010 in the considered dataset.

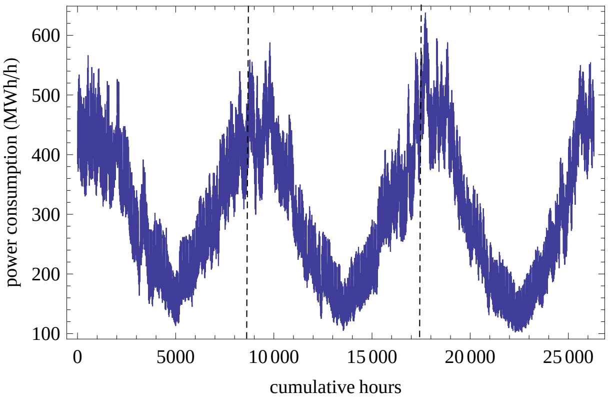


Figure 2.2: Electricity consumption for 2008, 2009 and 2010. The vertical dashed lines separate the three years consecutively, and the x axis shows the cumulative hours starting from January 1st 2008.

According to the figure we can see that, there is a clear pattern in the consumption throughout a year. Winter (rare ends of each graph) and summer (middle section of each graph) seasons could be clearly visible in the figure with identical consumptions. The average power consumption in 2008 is 319.344 MWh/h and for 2009 it is 300.442 MWh/h and 304.407 MWh/h for 2010. Therefore it seems to be in the range of early 300. Moreover, we can see from figure that the power normally ranges from 100 to 600.

The weather factors seem to be playing a big role in these consumptions, due to the seasonal variations in the graph. Temperature is the most influential factor in terms of seasonal changes. Therefore, we specially focus on the *forecast temperature* as it is the most influential environmental factor that affects the electricity consumption [21].

2.3 Electricity Consumption Forecasting

Forecasting of electricity consumption enables power utility companies to distribute electricity effectively. Electricity cannot be stored once it is generated. Therefore, it is very important to know how much electricity should be bought and distributed through the grid to the households and industries. Knowing the future consumption allows to manage this process in an effective way.

Therefore most power supplying companies and researchers have studied about different techniques to accurately forecast electricity consumption over the past few decades. Therefore we conclude this chapter by explaining how previous researches have been carried out regarding electricity consumption forecasting.

The research conducted by Imitiaz et al. [6] have used statistical analysis, which is basically linear regression, to evaluate and forecast long term electricity consumption demand for Malaysia. They have used factors such as *population*, *per capita electricity consumption*, *number of consumers*, *peak electricity demand* and *GDP* as the independent variables that affects the consumption. They have used training data of 10 years (from 1993-2003) to forecast for 7 years ahead (2004-2013).

As shown in [7], time series analysis based on autoregressive models has been used as another technique to predict power consumption. Time series analysis is a very popular technique for prediction. However it assumes that the past trends remain same for the future in estimating future values. The time series in this research has been based on the autoregressive model which express the value of a certain variable as a linear function of its previous values. They have used data from 1998 - 2005 as training, and 2006 - 2010 as test data. In our method also, we will adopt some of this technique in determining the feature vector as described in chapter 5.

From the range of methods used, grey methods [13] have also been a major consideration in terms of power prediction. Two models in grey methods: GM(1,1) and GM(0,N) have been used to forecast power consumption in a research done by Fang et.al [5].

In addition to these approaches, one of the most widely used techniques to predict power consumption is the neural networks as suggested in [8, 22, 23]. The research done by Fung and Tummala [8] compares the techniques, multiple linear regression and artificial neural network models, and concludes that ANN forecasts are at least as good as those generated by the multiple linear regression model. For ANN, they have used delta rule learning and error propagation methods to train the network. Different training and test datasets have been used in the span from 1970 - 1992.

Moreover, these two methods have also been used by Dhulst et al. [22] to predict the electric load consumed at a substation in Belgian electricity grid. In addition to that Quing et al. [23] suggests genetic algorithm and RBF neural network can be used to forecast power consumption, in which, GA optimizes the parameters of RBF neural network.

If we consider using of Gaussian processes in electricity consumption forecasting, we can find mainly two researches. The first one done by Alamaniotis and Ikonomopoulos [4], in which they apply genetic algorithm to determine the contribution from each independent predictor variable in order to compute a Pareto optimal solution. In this, they have used a set of kernels(will be defined in the next chapter) in the model. The kernel used in this technique are *Neural Net*, *Matérn*, and *Rational Quadratic*. The second, conducted by Mori and Ohmi [11] used Gaussian processes for daily power forecasting and a comparison has been done with MLP, RBFN and SVR.

Chapter 3

Gaussian Process and Regression

In this chapter we move on to the theoretical aspects of the main research area in this thesis - Gaussian Processes. Before explaining that, we point to give a basic understanding about the basics of random variables and Gaussian probability distribution. In section 3.3, we explain about GP and in the following sections detail how the regression process is done using GP.

3.1 Random Variables

The concept of random variable can be explained by means of an experiment specified by the space S. A random variable (RV) is a number $X(s)$ assigned to every outcome $s \epsilon S$ of that particular experiment [24]. In fact, the random variable X is a function whose domain is s and the range is the real numbers $\mathbb{R}$. It could be mathematically expressed as:

$$X : S \rightarrow \mathbb{R} \tag{3.1}$$

When we discuss about the distribution of a particular random variable, we consider two functions: *Cumulative distribution function* (CDF) and *probability density function* (PDF) . Therefore it is worth to know about these two functions before getting into the Gaussian probability distribution.

3.1.1 Cumulative Distribution Function

For any real number x from $-\infty$ to $+\infty$, the cumulative distribution function of a random variable X is given by:

$$F_X(x) = P\{X \leq x\} \tag{3.2}$$

It is the probability that the value of the particular random variable X is less than or equal to the considered value x. This scenario is visually illustrated in the diagram shown in Figure 3.1.

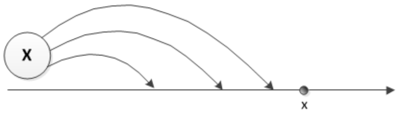


Figure 3.1: Illustration of Cumulative Distribution Function.

At $-\infty$ the CDF takes its minimum value (i.e zero) and at $+\infty$ it gets its maximum value(i.e one).

We categorize random variables into two based on the CDF. If the distribution function of a random variable is continuous, the random variable is said to be *continuous*, and if it is discrete (i.e staircase type) the random variable is said to be *discrete* [24].

If we consider more than one random variable (i.e multivariate random variables) the distribution function is called as joint *cumulative distribution function.* For two random variable X and Y, the joint distribution function is given as:

$$F_{XY}(x, y) = P\{X \leq x, Y \leq y\} \tag{3.3}$$

3.1.2 Probability Density Function

The next important function is the probability density function which is the derivative of the CDF and can be given as:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{3.4}$$

So as for CDF the density function can also be categorized into two types as continuous and discrete. For multivariate case, the joint probability density is given as:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \tag{3.5}$$

See Papolis [24] for more information about the theories of random variables, CDF and PDF.

3.2 Gaussian Probability Distribution

Gaussian distribution is also termed as Normal distribution. Typically, a Gaussian (or Normal) random variable X is denoted as $X \sim N(\mu, \sigma^2)$ where μ is the mean and σ is the standard deviation. The PDF of a Gaussian distribution is given by the equation:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{3.6}$$

The curve of the PDF against x takes a bell shape as shown in Figure 3.2. The bell shape curve is distributed symmetrically about the mean μ and the curve tends towards the x axis when the variance σ^2 is increased indicating more deviation from the mean. In a Gaussian distribution the total area under the curve always sums up to one.

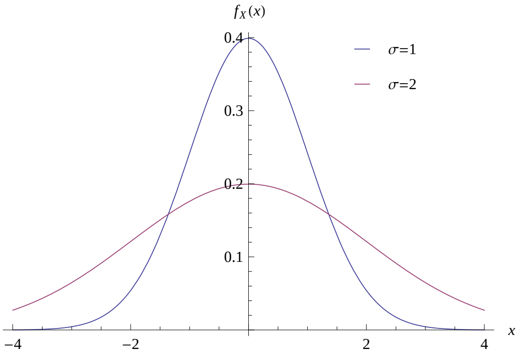


Figure 3.2: Gaussian Probability Distribution.

The concept of Gaussian distribution is extended to use in developing statistical models such as Gaussian processes as we will discuss in the next section.

3.3 Gaussian Process

With the basic introduction given in the preceding sections about random variables, this section outlines the theoretical aspects of Gaussian processes, which is the main area of focus in this thesis. GP has its basic foundations from statistics and machine learning, and it is considered as a general and rich framework which is related to a variety of other models such as Spline models, Support Vector Machines (SVM) , Least-Square methods, Relevance Vector Machines and Weiner filters [25].

Although GP has been in the use for a long time, it has not been extensively used in forecasting when compared with the other competitive techniques such as ANN or time series analysis. However, over the last decade it has become popular in the field of machine learning

[25].

Gaussian process is a generalization of the Gaussian probability distribution we discussed in section 3.2. Whereas a typical Gaussian distribution concerns about a single random variable, a Gaussian process is associated with a collection of random variables that produces a pool of functions relevant for prediction. In other words, Gaussian process is a distribution over functions. The definition of Gaussian process is as follows [26]:

Definition 1. A Gaussian Process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

In general, the definition means that the joint PDF of the selected finite number of random variables is normally distributed. The notation of Gaussian process is given as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{3.7}$$

where $m(x)$ is the mean function and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function. Instead of the mean and variance of the Gaussian distribution, Gaussian process is fully described by its mean function and covariance function.

3.3.1 Mean Function

The mean function m($\mathbf{x}$) is calculated as the expected value of f($\mathbf{x}$) as shown in equation 3.8.

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{3.8}$$

In most of the prediction scenarios, the mean function is assumed to be zero, in which the average value of the functions at each x in the prior Gaussian becomes zero. However, Rasmussen and Williams [26] states that it is not always necessary to have a Gaussian with a zero mean function. But in our case, we assume it to be zero for simplicity. More information about the non zero mean functions could be found in [26].

3.3.2 Covariance Function

With the assumption of a zero mean function, the whole focus shifts on to the covariance function $k(\mathbf{x}, \mathbf{x}')$, which is also known as the kernel function of the Gaussian process. Because of the existence of a kernel function, Gaussian process gets the name *kernel machines*. This kernel based non-parametric nature of GP makes it a more flexible model than the parametric models [27].

The covariance between the two random variables $f(x)$ and $f(x')$ is calculated as shown in equation 3.9.

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{3.9}$$

when $m(\mathbf{x}) = m(\mathbf{x}') = 0$, then the covariance function is given as:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}))(f(\mathbf{x}'))] \tag{3.10}$$

Note that, the covariance is measured between the function values $f(\mathbf{x})$ and $f(\mathbf{x}')$ although the notation is given as $k(\mathbf{x}, \mathbf{x}')$. Whereas the variance is measured for a single random variable, the covariance is a measure between two jointly distributed random variables. It evaluates how closely the two random variables are related. In other words, it finds the correlation between the two variables.

In GP learning, covariance function is the most significant function. The accuracy of the predictions mainly depends on the kernel that we choose. Asheri et al. [25] mention that GP can be made equivalent to the well known models such as large-scale neural networks, spline models and support vector machines by employing a suitable kernel function. Therefore, an appropriate function needs to be selected to approximate the kernel function.

This selected function should have certain characteristics as specified by Rasmussen and Williams [26]. Mainly, the selected function should be positive semi-definite and symmetric (i.e $k(x, x') = k(x', x)$).

In addition to these characteristics, most importantly, the function should be suitable for the particular application, such that we can obtain a smooth GP for the application.

There are some common kernel functions that have been used mostly in applications, such as *linear, γ-exponential*, *rational quadratic*, *Matérn*, *piecewise polynomial* and *squared exponential* [25].

3.4 Regression Analysis

In supervised learning, we infer a function from the labelled training examples (training data). Each of these training example is a pair consisting of an *input object* (typically a vector) and an *output value* (scalar value).

Depending on the type of the output values, the problem of inferring falls into two categories: *regression* and *classification.* When the output is continuous, the problem becomes regression and when it is discrete (output is a class label of the input) the problem is a classification problem. In this thesis, we focus on the problem of regression rather than classification.

Suppose we have a sample dataset with 5 inputs and their observed output values y, as shown in Table 3.1. The problem in regression is to predict the output value y for the new input value x=6 as shown in Figure 3.3. Each observation has a noise value as illustrated by the error bars. Similarly we should estimate the error bars (the confidence interval) for the predicted value in regression analysis.

Table 3.1: Sample dataset for regression.

x	1	2	3	4	5	6
y	2.0	3.0	4.25	5.5	5.75	?

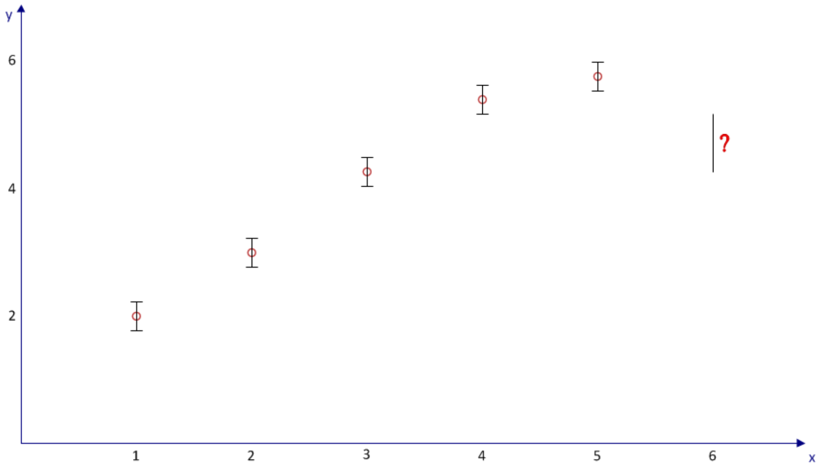


Figure 3.3: A set of sample training data points with one test data point whose target value is unknown.

Inputs can be of single dimension or multiple dimensions. In general, it is a vector and the output is a single scalar value.

As mentioned above, statistical inference can be performed by learning a function from the sample training dataset with different input-output patterns. However the accuracy of inference is solely based on the method we choose for the underlying function that maps the input to the correct outputs.

3.5 Gaussian Process Regression

In this section we look at how Gaussian process can be used to perform regression. Before seeing any data, first we have to assume the underlying function (Gaussian prior) and select a proper covariance function. However, this selection of a proper kernel should be done in accordance with the characteristics of the data considered. Then Gaussian process specified by the respective kernel function will produce a distribution of random functions.

After the training data points are introduced , it selects the best matching functions from the distribution that pass through or pass closely to the given data points. In this way it finds the best possible set of functions from the Gaussian prior. For better forecasting accuracy, GP requires adequate number of training samples.

However, the selection of the best functions solely depend on the choice of kernel function and its parameters. Although GP is a non parametric model, the kernel function is associated with some parameters such as signal variance and length-scale parameter. These parameters should be set properly for better learning experience.

Suppose we are given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, 2, ..., n\}$ with noisy observations, and we need to predict the target value y_* for the new input value x_*. The problem is to learn a function from the dataset, which involves an assumed Gaussian prior of functions. In fact, the observations and the underlying function f values are not the same due to the noisy measurements of the observations. Therefore the targets can be represented as:

$$y = f(\mathbf{x}) + \varepsilon, \tag{3.11}$$

with the assumption of a Gaussian noise model represented by $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. The underlying function f is approximated by a Gaussian process with zero mean function and a covariance function. The most commonly used covariance function is the *Squared Exponential* covariance function. Using this function we can express equation 3.7 as,

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{3.12}$$

$$k(x, x') = \sigma_f^2 \exp[\frac{(x - x')^2}{2l^2}] \tag{3.13}$$

According to equation 3.11, the actual observations (i.e y) can be specified by adding the noise model to the underlying function defined by equation 3.12. Then the covariance function related to the target values y, denoted as $cov(x, x')$, can be given by:

$$cov(x, x') = k(x, x') + \sigma_n^2 \delta(x, x') \tag{3.14}$$

where $\delta(x, x')$ is the Kronecker delta function which is equal to 1 iff $x = x'$ and 0 otherwise.

Using the kernel function, the correlation between each and every training data point can be measured. The matrix generated with each of these covariance value as elements is known as the *Covariance matrix* and denoted by K.

$$K = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \cdots & cov(x_n, x_n) \end{bmatrix} \tag{3.15}$$

The corresponding covariance matrix between the training data points and the test data points is given by K_*:

$$K_* = \begin{bmatrix} cov(x_*, x_1) & cov(x_*, x_2) & \cdots & cov(x_*, x_n) \end{bmatrix} \tag{3.16}$$

K_{**} specifies the covariance matrix between the test data points itself.

$$K_{**} = cov(x_*, x_*) \tag{3.17}$$

The joint distribution of the observations y and the predictions y_*, has been found as a multivariate Gaussian distribution [28]:

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right) \tag{3.18}$$

In fact, the actual prediction that we are interested in is given by the conditional distribution of y_* given y

$$y_*|y \sim \mathcal{N}(K_*K^{-1}y, K_{**} - K_*K^{-1}K_*^T) \tag{3.19}$$

This particular distribution is known as the posterior Gaussian distribution. According to Rasmussen and Williams [26], for any Gaussian posterior, the mean of the posterior distribution is called as the *Maximum a Posteriori* (MAP) value, which is the best estimate for the variable considered. In equation 3.19 the actual prediction y_* is its mean, which is the MAP estimate:

$$\bar{y}_* = K_*K^{-1}y \tag{3.20}$$

The variance of the estimation is given by

$$var(y_*) = K_{**} - K_*K^{-1}K_*^T \tag{3.21}$$

The variance is used to calculate the 95% confidence interval of the prediction as $\pm 1.96\sqrt{var(y_*)}$, which is approximately two times the standard deviation of the posterior distribution.

Now, let us apply these equations to the example shown in 3.3 and forecast the value of y at x_*=6 using Gaussian process. Suppose σ_f=3.35, l=2.95 and σ_n=0.08. Using these value we can calculate the covariance matrices K, K_* and K_{**} as follows:

$$K = \begin{bmatrix} 11.27 & 10.63 & 8.95 & 6.72 & 4.49 \\ 10.63 & 11.27 & 10.63 & 8.95 & 6.72 \\ 8.95 & 10.63 & 11.27 & 10.63 & 8.95 \\ 6.72 & 8.95 & 10.63 & 11.27 & 10.63 \\ 4.49 & 6.72 & 8.95 & 10.63 & 11.27 \end{bmatrix} \tag{3.22}$$

$$K_* = \left[\begin{array}{ccccc} 2.68 & 4.49 & 6.72 & 8.95 & 10.63 \end{array} \right] \tag{3.23}$$

$$K_{**} = 11.27 \tag{3.24}$$

Applying these values to equations 3.20 and 3.21, we get the prediction of y at x_*=6 and the variance of the prediction.

$$\bar{y}_* = 5.06 \tag{3.25}$$

$$var(y_*) = 0.14 \tag{3.26}$$

Now we can visually illustrate the results, with the predicted value and the 95% confidence interval as shown in Figure 3.4.

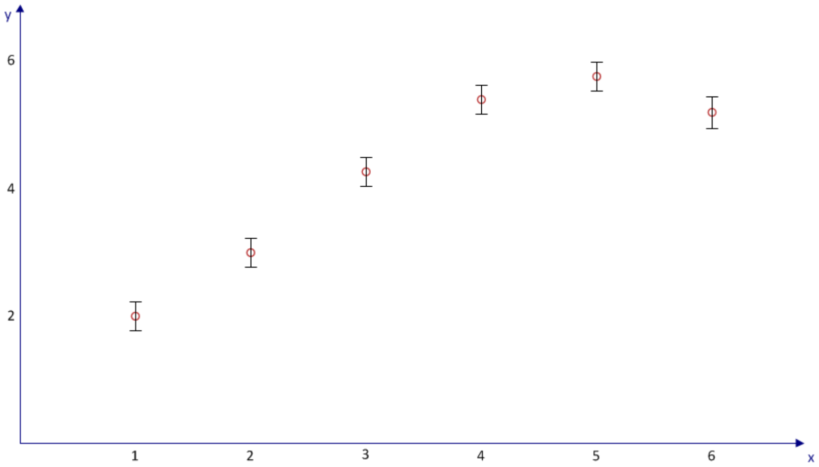


Figure 3.4: A set of sample training data points with the predicted target value of the test data point.

Equations 3.20 and 3.21 has a considerable computational issue because of the inverse operation of K of size n^2 [19]. Here n is the number of observations. Therefore, the complexity of Gaussian process is $O(n^3)$. However it can be reduced to a complexity of $O(nm^2)$ by selecting an active subset of columns of K. Here m is the rank of the matrix approximation [19]. Also cholesky decomposition can be used to factorize K, to get a numerically stable approximation [19].

3.5.1 Selection of Hyperparameters

The noise variance σ_n^2, and the parameters in the kernel function are taken as the free hyperparameters in Gaussian process. This is represented by the vector θ:

$$\theta = \{l, \sigma_f^2, \sigma_n^2\} \tag{3.27}$$

where l is the characteristic length scale, σ_f^2 is the signal variance and σ_n^2 is the noise variance.

In Gaussian Process Regression(GPR) , these three parameters are obtained by learning the data. In general, it uses Bayesian model to infer these parameters. This method is known as *marginal likelihood maximization* method. According to the Bayes rule, we can represent the posterior probability of the parameters as follows:

$$p(\theta|X, y) = \frac{p(y|X, \theta)p(\theta)}{p(y|X)} \tag{3.28}$$

$p(y|X, \theta)$ is the marginal likelihood which is to be maximized, and $p(\theta)$ is the prior probability of the parameters.

The marginal likelihood can be calculated by marginalizing the integral of the likelihood and the Gaussian prior, over the latent function f.

$$p(y|X, \theta) = \int p(y|f, X)p(f|X)df \tag{3.29}$$

Both $p(y|f, X)$ and $p(f|X)$ follow Gaussian distributions. The log value of the marginal likelihood gives the parameters which has been found as [26]:

$$L(\theta) = \log p(y|X, \theta) = -\frac{1}{2}y^T K^{-1} y - \frac{1}{2}log|K| - \frac{n}{2}log2\pi \tag{3.30}$$

The first term $-\frac{1}{2}y^T K^{-1} y$ represents the data-fit,which is the only term that contains the observed target values. The second term is the complexity penalty and the last term is a normalization constant.

Chapter 4

Traditional Approaches of Electricity Consumption Forecasting

In this thesis, results of Gaussian process are compared with the traditional techniques used for electricity consumption forecasting. We have selected two such traditional techniques : *Artificial Neural Networks* and *Linear regression.* The purpose of this chapter is to give the theoretical foundations of both ANN and linear regression. Section 4.1 introduces the concepts of ANN including its variants back-propagation and multiple back-propagation. Section 4.2 outlines the basics of linear regression and multiple linear regression explaining how regression is performed under different input variables.

4.1 Artificial Neural Networks

Artificial neural networks have been used for forecasting electricity consumption as we mentioned in chapter 1 and 2. In this section we go deeply into the theoretical aspects of ANN, and back-propagation based ANN which we use to test our dataset, to compare the results with Gaussian processes.

4.1.1 Introduction to ANN

The concept of ANN arises from the knowledge of the biological nervous systems [29]. The nervous system is constructed by a number of structural constituents known as *neurons*, which are connected to each other by links. This network of neurons connected through links, is referred to as a *neural network*.

A neuron is defined by Patterson [29] as follows:

> *A neuron is a small cell that receives electrochemical stimuli from multiple sources*

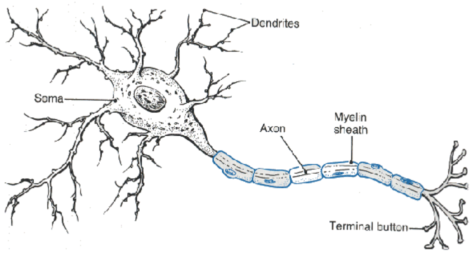


Figure 4.1: A typical neuron in a biological nervous system. [Source:http://www.mindcreators.com/NeuronBasics.htm]

> *and responds by generating electrical impulses that are transmitted to other neurons of effector cells.*

There are about 10^{10} to 10^{12} neurons in the human nervous system [29], which contains trillions of interconnections between them that makes it a highly complex system. Basic components of a typical neuron cell is illustrated in Figure 4.1.

The cell body is known as the Soma. A neuron cell has an input side and an output side. The input side is the one which is named as dendrites in the figure. Dendrites are connecting the outputs from the other neurons to this neuron through synapses. There are a number of various synaptic connections to the neuron from which it can receive input signals. The outputs are carried through the axon to other neurons (through dendrites) or directly to effector organs such as muscles and glands.

Neurons can be categorized into three as input, output and intermediate neurons. In the human body, input and output neurons constitute 10% of the neurons and the remaining 90% store informations and other signal transformations [29].

These concepts of neurons and biological nervous system have led scientists to develop the artificial neural networks.

4.1.2 Back-Propagation Neural Networks

Back Propagation Neural Network(BPNN) , one form of ANN, is a non parametric statistical modelling technique, which is used in regression. It is considered by Smith [30] as the only form of neural network which has produced a number of commercial applications. It is a feed-forward network which has the ability to propagate prediction error, back to the network as a feedback and improve its results.

In its simplest form, a BPNN basically composed of three layers of neurons(nodes) - *input layer*, *hidden layer* and *output layer*. Figure 4.2 depicts a typical feed-forward network with these three layers. It consist of 4 input nodes (representing four independent variables), 3 hidden nodes (which performs the basic calculations) and 2 output nodes(representing two

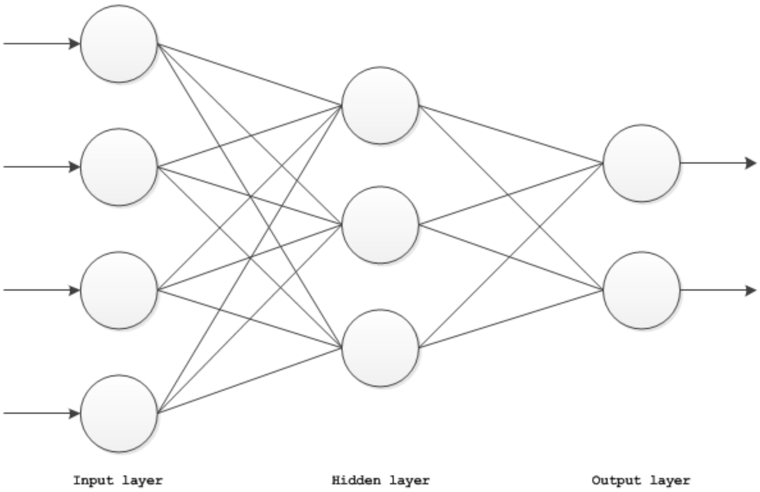


Figure 4.2: A typical feed-forward network used for back-propagation

output variables).

Each layer in the BPNN is associated with a particular functionality. Similar to the biological nervous system, the input nodes in the BPNN receives input values. But in this case, they receive the values of the independent variables used in the training process. If there are n number of variables in the feature vector, the network requires n input nodes in the input layer to accommodate the corresponding variables. The output nodes represent the dependant variables that needs to be estimated by the network. They output the estimated values of the dependant variables. Hidden layer is the intermediate layer which does the basic inner workings of the neural network, and also got its name *hidden* as it acts as a black box to the outside environment.

Every node in each layer is connected to the next layer by links ,which is analogous to the synaptic connections and dendrites in the biological nervous system. However, note that, there is no process like *back-propagation* in actual human brain [30].

Now we should pay our attention as to how BPNN estimates the values of the output variables given the training set of input-output pairs. Back-propagation involves two types of passes: a *forward pass*, which is referred to as a mapping from inputs to outputs, and a *backward pass*, which is referred to as the learning of the network.

In forward pass, a relationship, which is a mathematical equation, is generated (called the *mapping function*) between the input nodes and the output nodes across different connections in the network. Each connection between neurons has a certain strength which is termed as the *weight* between the two neurons. These weights play an important role in BPNN as they change their values in order to obtain better estimates.

The mapping function is built by using some standard functions. These functions used in the neural networks is so flexible that it can be configured to be close to any target function [30]. It

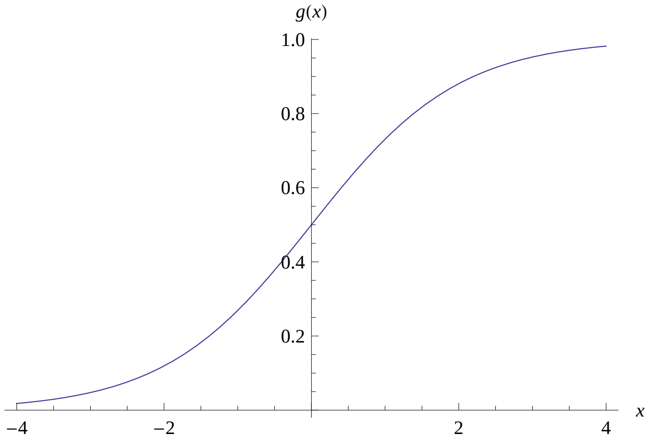


Figure 4.3: The logistic function which is one type of sigmoid functions.

achieves this flexibility from the weights of the equation. In general, this function is constructed basically by the sigmoid functions.

A sigmoid function takes the S shape. It should be bounded (i.e has an upper limit and a lower limit), monotonically increasing and differentiable. The most commonly used sigmoid function used in neural networks is the *logistic function* [30] depicted in equation 4.1.

$$g(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

Figure 4.3 illustrates the curve for the logistic function between the interval -4 and +4.

Now we must look at how the mapping function is constructed in a back-propagation neural network. Suppose the network has two input nodes, two hidden nodes and one output node as shown in Figure 4.4. x_1 and x_2 are the two input values. w_{ij} are the weights of each link from the input neurons and the hidden neurons where $i = 1, 2$ and $j = 1, 2$ for Figure 4.4. These weights are stored in the memory of hidden neurons. In addition to that, each hidden neuron stores a bias values which are shown as b_1 and b_2 in the figure. u_1 and u_2 are calculated in the hidden layer as a weighted sum of each input value as shown in equation 4.2 and 4.3.

$$u_1 = b_1 + x_1 w_{11} + x_2 w_{12} \tag{4.2}$$

$$u_2 = b_2 + x_1 w_{21} + x_2 w_{22} \tag{4.3}$$

Instead of outputting u_1 and u_2, the hidden neurons output the logistic values of them as the inputs to the next layer. Therefore the two outputs y_1 and y_2 can be expressed as in equation 4.4 and 4.5 using 4.1.

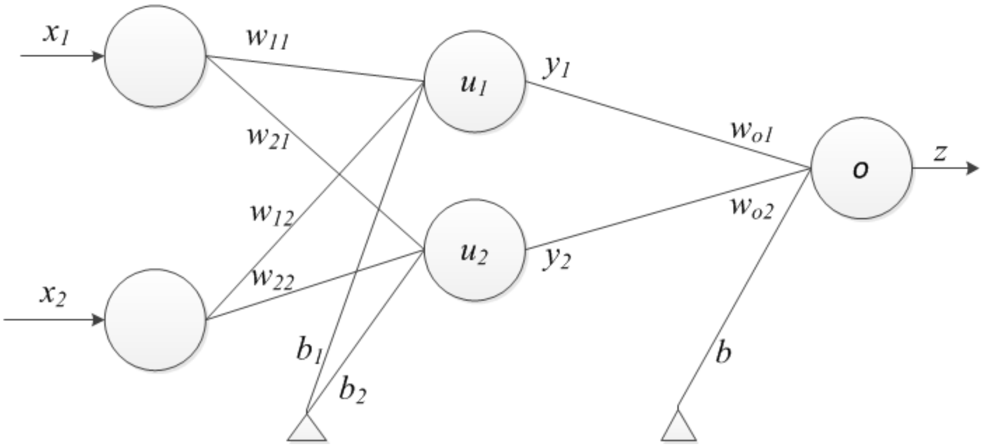


Figure 4.4: A BPNN with weights and bias values. Neurons are represented by circles and bias by triangles.

$$\begin{aligned} y_1 &= g(u_1) \\ &= g(b_1 + x_1 w_{11} + x_2 w_{12}) \\ &= \frac{1}{1 + e^{-(b_1 + x_1 w_{11} + x_2 w_{12})}} \end{aligned} \tag{4.4}$$

and

$$y_2 = \frac{1}{1 + e^{-(b_2 + x_1 w_{21} + x_2 w_{22})}} \tag{4.5}$$

Now y_1 and y_2 become the inputs to the output node o. The output node also perform a similar calculation to generate the final output value z. Similar to the hidden nodes, the output node also has a bias value b and weights for each connection link (i.e w_{o1} and w_{o2}). The output value z is given by

$$\begin{aligned} z &= g(o) \\ &= \frac{1}{1 + e^{-o}} \end{aligned} \tag{4.6}$$

where o is the weighted sum of the outputs coming from the hidden layer

$$o = b + y_1 w_{o1} + y_2 w_{o2} \tag{4.7}$$

In general, for a BPNN with K output nodes and J hidden nodes the output of the network can be given as [30]:

$$z_k = g(o_k), \quad k = 1, ..., K \tag{4.8}$$

where,

$$o_k = b_k + \sum_{j=1}^{J} w_{jk} y_j \tag{4.9}$$

In general, if we assume a multilayer network with n input nodes, J hidden nodes and one output node. The equation 4.2 is generalized to

$$u_j = b_j + \sum_{i=1}^{n} x_i w_{ji}, \qquad j = 1, ..., J \tag{4.10}$$

and the generalized output from the hidden nodes can be calculated using the logistic function.

$$\begin{aligned} y_j &= g(u_j) \quad j = 1, ..., J \\ &= g\left(b_j + \sum_{i=1}^{n} x_i w_{ji}\right) \end{aligned} \tag{4.11}$$

The final estimated output can be found by applying the weighted sum and subsequently applying the activation function to it as follows.

$$o = b + \sum_{j=1}^{J} y_j w_{oj} \tag{4.12}$$

and hence the final output z can be represented as a function of input x as follows,

$$\begin{aligned} z &= g(o) \\ &= g\left(b + \sum_{j=1}^{J} y_j w_{oj}\right) \\ &= g\left(b + \sum_{j=1}^{J}\left[g\left(b_j + \sum_{i=1}^{n} x_i w_{ji}\right)\ w_{oj}\right]\right) \end{aligned} \tag{4.13}$$

After the forward-pass is finished and the output is calculated, the backward-pass commences. This is also known as the learning process. The learning here refers to adjusting the weights we discussed above such that the mean squared error between the estimated and target values gets smaller. The method used to achieve this is *gradient descent*. The adjustment to the weights is done in the backward process by propagating the mean squared error from the output side to the hidden nodes. Therefore this method gets the name back-propagation . The training examples are feed to the network and possible output pattern is generated in forward pass. Then the estimated output pattern is compared with the desired output pattern and the difference is propagated in backward pass indicating the direction in which the correct adjustments should be made. In this way the weights between the hidden layer and the output layer are adjusted. This process is repeated a considerable number of iterations (known as *epochs*) until the total average error of the outputs converges to a minimum value.

4.1.3 Multiple Back-Propagation

In this thesis, we use the MBP Software[1] developed by Lopes and Ribeiro [2], to test for neural networks. In [2] they describe about Multiple Back-Propagation (MBP), which is a generalization of the BP algorithm.

Here, a Multiple Feed-Forward network is used, which is obtained by integrating two FF networks : *Main network* and *space network* as shown in Figure 4.5.

The output of selective activation neurons in the main network is given by [2]:

$$y_k^p = m_k^p \mathcal{F}_k(\sum_{j=1}^{N} w_{jk} y_j^p + \theta_k), \tag{4.14}$$

where y_k^p is the output of neuron k for pattern p, m_k^p the importance of the neuron for the output of the network, $\mathcal{F}_k$ the neuron activation function, θ_k the bias and w_{jk} the weight of the connection between neuron j and k. The main network can calculate its output only after the space network outputs are calculated. In the learning pass, the weights of both the networks

[1] This software could be downloaded for free from http://dit.ipg.pt/MBP/Download.aspx

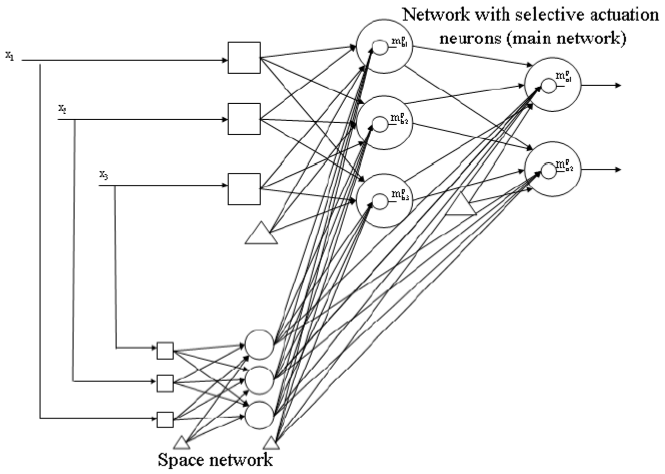


Figure 4.5: A Multiple feed-forward network. Hidden and output neurons are represented by circles, input neurons by squares, and bias by triangles [2].

should be adjusted. For more information about MBP networks see [2,31,32], and in Appendix A we have included some screen shots of the tool used for forecasting using MBP network.

4.2 Linear Regression

Linear regression is a common regression technique used in most inferring problems. Unlike neural networks, it lacks the ability to capture non-linear relationships between variables. However it is considered as one of the methods, which is easy to fit and highly scalable [19]. In the following sections we will describe basic theoretical aspects of linear regression and multiple linear regression and how we can use it to infer predictors using IBM SPSS ® Statistics software.

4.2.1 Statistical Significance

Before moving into the prediction, lets consider about the significance value which is mostly used in linear regression. It is also termed as significance or probability which is denoted by the letter p. The likelihood that a particular outcome may occur by chance is given by the p value.

It can be used to identify whether two or more variables are correlated to each other significantly. So we should always try to find a very smaller p value for valid results. Social scientists have accepted that a p value less than 0.05 is statistically a significant correlation [33].

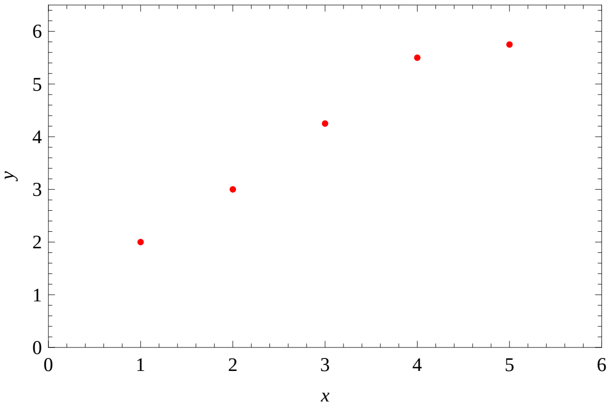


Figure 4.6: Scatter plot of the sample dataset.

4.2.2 Linear Regression

Linear regression analysis is a way of testing hypothesis concerning the relationship between two numerical variables and a way of estimating the specific nature of such relationships [34]. The relationship is expressed in the form of an equation or a model connecting the dependant variable and one or more independent variables depending on the problem of interest. The method of least squares is used most frequently in fitting a line in linear regression.

The simplest relationship between an independent variable x and a dependant variable y is represented as

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4.15}$$

where β_0 is the intercept and β_1 is the slope. The random error term is given by ϵ which should be normally distributed with 0 mean and at each possible value of x, the variance of $\epsilon | x_i$ should be constant and it should be independent of the other errors [34]. Normally we examine the residuals which are the differences between the observed values (y) and the estimated values to approximate this error term.

These unknowns have to be found using the samples in the training dataset. Lets consider the sample dataset found in Table 3.1 in chapter 3, which has an independent variable x and dependant variable y as shown in Figure 4.6.

SPSS generates four tables in linear regression analysis. *Table of variables* in the regression equation, a *model summary*, an *ANOVA table*, and a *table of coefficients*.

Table 4.1: Variables table.

Model	Variables Entered	Variables Removed	Method
1	x		Enter

A variables table with only one independent variable is shown in Table 4.1. We can have multiple linear regression models if we have multiple variables based on the variables in the Entered and Removed columns. However in this case there is only one model due to the availability of one independent variable.

Table 4.2: Model summary table.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984	.969	.958	.32914

The model summary table shown in Table 4.2 illustrates the goodness of fit in regression. Here, R is the correlation coefficient which ranges from -1 to +1 and R^2 is the coefficient of determination which is the squared of R. If R^2 is equal to 1, then it is a perfect fit. The value in the given table (i.e 0.969) is very close to one, meaning that the points in the dataset are experiencing a very good linear relationship. The adjusted R square value is a more better value than R square value which can be used for population estimates specially in multiple regression.

The third table is the ANOVA table. ANOVA is used to compare three or more means to one another. For a single independent variable it is called one-way ANOVA [34].

Table 4.3: ANOVA table.

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	10.000	1	10.000	92.308	.002
Residual	.325	3	.108		
Total	10.325	4			

The *Sig* value is also known as the *P-value* and, if it is less than 0.05 we say the ANOVA is significant(F value is significant) and it can be concluded that there is a regression in the model. The F value in the table is known as the *Levene statistic.* In Table 4.3, the p value is less than .05, and therefore we can conclude that the two variables are statistically significant.

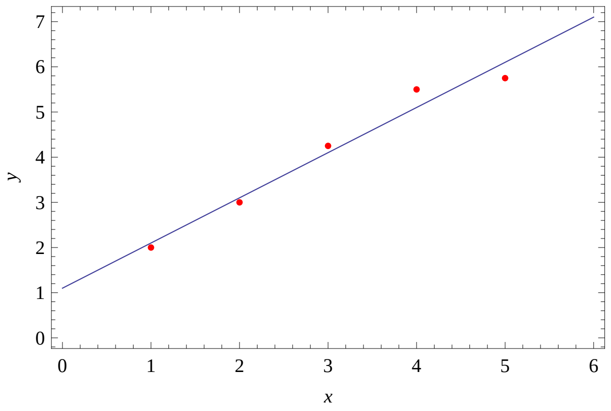


Figure 4.7: Fitted line for the sample dataset.

Table 4.4: Coefficients table.

Model		**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
		B	**Std.Error**	**Beta**		
1	(Constant)	1.100	.345		3.187	.050
	x	1.000	.104	.984	9.608	.002

The last table is the table of coefficients. We know that β_0 and β_1 are the coefficients in equation 4.15. The first value in column B (i.e 1.100) is the intercept β_0 and the second value 1.000 is the slope β_1. If we consider the t value and the significant value for the slope, the significance value is less than 0.05 meaning that there is a statistically significant relationship between x and y.

Using thes information we can interpret equation 4.15 as follows for the considered dataset.

$$y = 1.1 + x + \epsilon \tag{4.16}$$

The fitted line for the above discussed dataset could be illustrated as shown in Figure 4.7. Now we can estimate the output of the target variable when x=6, i.e y=7.1.

But now we should pay our attention to the error term (or disturbance) of the fitted line. For this, we have to look at Table 4.2. The standard error of the estimate .329, is a measure of the variability of the random error. This can be used to calculate the 95% confidence interval by multiplying by two. This can be considered as the residual error term for the regression line. Therefore our estimate for y should be as follows:

$$y = 7.1 \pm 2 \times .329$$

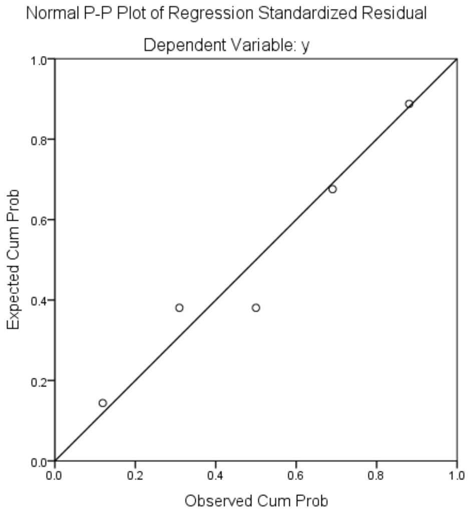


Figure 4.8: The normal P-P plot of regression, which is used for residual analysis.

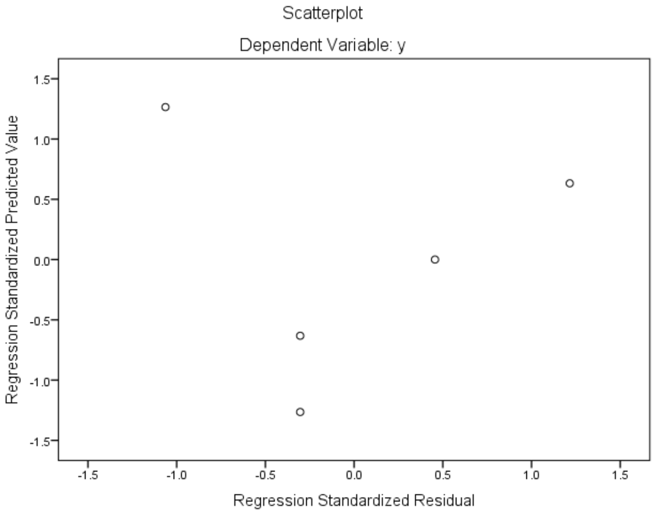


Figure 4.9: Graph of standardized predicted value versus standardized residual value, which is used for residual analysis.

$$y = 7.1 \pm .658$$

To find out the validity of the first assumption of the residuals (i.e normality) we look at the normal probability plot as shown in Figure 4.8.

The assumption of equal variances can be identified by the scattered plot of the standardized residuals versus the standardized fitted values. [34]. This is illustrated in Figure 4.9.

4.2.3 Multiple Linear Regression

In section 4.2.2, we discussed how one independent variable linearly related to a dependant variable. Now we discuss how several independent variables affect a dependant variable.

For example, suppose we have two independent variables and one dependant variable. Then the regression equation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{4.17}$$

If we look at Table 4.5, we can see the Pearson correlation coefficients for each variable. The coefficient for x_1 and y is -.165 and for x_2 and y it is -.018. Both have a negative value meaning that they have negative relationship with y. But x_1 is more correlated with y and it influence y more.

Table 4.5: Correlation for multiple variables.

	y	**x1**	**x2**
Pearson Correlation y	1.000	-.165	-.018
x1	-.165	1.000	-.048
x2	-.018	-.048	1.000
Sig. (1-tailed) y		.003	.383
x1	.003		.213
x2	.383	.213	
N y	279	279	279
x1	279	279	279
x2	279	279	279

Table 4.6: Coefficients for multiple variables.

Model	**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std.Error**	**Beta**		
1 (Constant)	23.195	8.227		2.819	.005
x1	-.057	.020	-.166	-2.802	.005
x2	-.106	.243	-.026	-.436	.663

We can generate the equation for the fitted line by referring to the B value in Table 4.6.

$$y = 23.195 - .057x_1 - .106x_2 + \epsilon \tag{4.18}$$

However when referring to Table 4.6 the Significance value of x_2 is greater than the P value .05. Therefore we can conclude that x_2 has no significant relationship to y. In such case, we can remove x_2 from the model as it does not affect the dependent variable. But there is a certain standard criteria for selecting and removing variables for a model.

Variable selection criteria

Although there are many predictor variables, all of them may not influence the behaviour of the response variable. Therefore, there should be a way of picking up the most significant variables. There are basically three methods for model selection: *forward selection, backward selection* and *stepwise selection.*

In forward selection, first the best predictor with the highest correlation with the dependant variable will be selected. In the next step, another variable is selected if it makes a significant variability to the model. Likewise, variables are added if they influence the behaviour of the dependant variable. This will end when there is no additional variables that significantly influence the behaviour. Normally the end condition will meet if the variable has a p value more than 0.05.

In backward selection, first add all the variables in the experiment to the model and remove one by one in each step until the stop condition is met. In general, if there are no more variable with a p value greater than or equal to 1.0 the process terminates.

Stepwise method combines both forward and backward methods, by dropping and adding variables at various steps. This is probably the most frequently used method of the three [33].

Chapter 5

Short-term Forecasting of Electricity Consumption using GP

In chapter 3, we discussed about the basic theoretical aspects of GP and GP regression. In this chapter we will explore how GP can be used to forecast electricity consumption based on the different factors influencing the consumption. We use pyXGPR [35] python code library to test GP regression. Section 5.1 finds the annual, seasonal and weekly cyclic patterns of consumption. With the idea of these cyclic patterns, section 5.2 designs the feature vector appropriate for forecasting electricity consumption. In section 5.3, we analyse two target variables and propose testing for them. Section 5.4 applies reduction and normalization to the feature space to improve results of forecasting. Finally in section 5.5 we discuss about the kernel function used in our work to test GP.

5.1 Cyclic Patterns of Consumption

As we mentioned in previous chapters, in this thesis, short term forecasting problem refers to; forecasting the consumption of the next hour and the next 24 hours. Our dataset consists of the power consumption data of three successive years: 2008, 2009 and 2010. In order to make the problem simple, we use the time horizon depicted in Figure 5.1 to find out cyclic patterns. We mainly focus on three stages in the hierarchy : *annual*, *seasonal* and *weekly* as described below.

5.1.1 Hourly Consumption on Annual Basis

In the process of searching for cyclic patterns we start from the top of the time horizon: *annual consumption.* We can observe how the hourly power consumption has behaved over the years considering 2008, 2009 and 2010. Figure 5.2 illustrates the mean hourly power consumption in

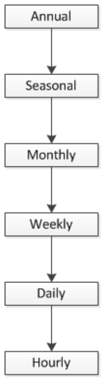


Figure 5.1: Time horizon hierarchy.

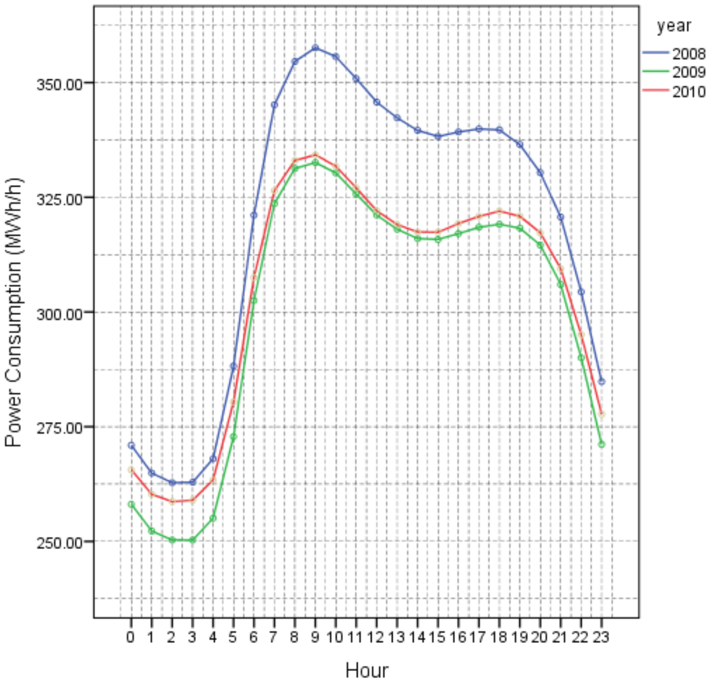


Figure 5.2: Mean hourly electricity consumption in 2008, 2009 and 2010.

all three years.

As can be seen from the figure, there is a certain pattern of mean power consumption. Higher power consumption could be observed at day time than at night which is quite acceptable because of the industrial power usage at day time. Hour 9 is the peak hour and hour 2 gives the lowest consumption for all 3 years. Although 2008 has a higher consumption compared to the other two, all of them displaying unique patterns in consumption. In 2008, the consumption ranges from 260 to 360 MWh/h , 250 to 340 MWh/h in 2009, and 250 to 340 MWh/h in 2010.

These unique patterns lead us to convincingly justify that the yearly behaviour of consumption is almost identical.

5.1.2 Hourly Consumption on Seasonal Basis

Next, we look at the seasonal behaviour of the dataset to analyse for the four seasons: *winter*, *spring*, *summer* and *autumn*. We consider the seasons according to the classification illustrated in Table 5.1.

Table 5.1: Seasons and their durations.

Season	**Duration**
Winter	December-January-February
Spring	March-April-May
Summer	June-July-August
Autumn	September-October-November

According to Figure 5.3, the four seasons have four different consumption patterns, where winter experiences the highest demand and summer the lowest. Spring and autumn have somewhat closer consumptions, but the two patterns are different from each other.

This is a clear indication that the electricity consumption has a very close relationship with the weather condition, especially for a country like Norway. During the winter season, we know that a lot of electricity is consumed for lighting and heating. This might be one of the main reasons why winter has much more consumption. This usage is completely turned upside down when summer comes.

For year 2009, the four seasons showing an identical behaviour as shown in Figure 5.4. In winter and summer it exhibits the same patterns like in 2008. However, there is a slight difference in spring and autumn.

Figure 5.5 depicts the electricity consumption for 2010 based on the four seasons, where we can observe a similar behaviour to 2009.

However, the overall patterns of all the three years are identical when we consider each season separately. Therefore, we can conclude that the seasonal cyclic pattern as a very good foundation for doing forecast.

5.1.3 Hourly Consumption on Weekly Basis

The weekly behaviour of power consumption can be considered as a decisive feature for prediction. In general, the weekday values are different from week end values as a result of the industrial power demand on weekdays.

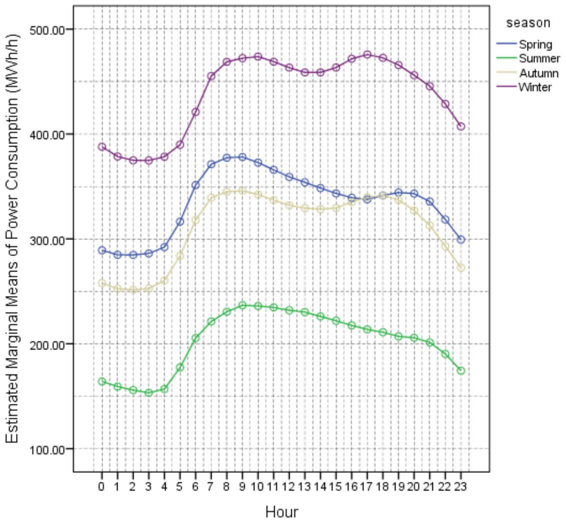


Figure 5.3: Mean hourly electricity consumption for different seasons in 2008.

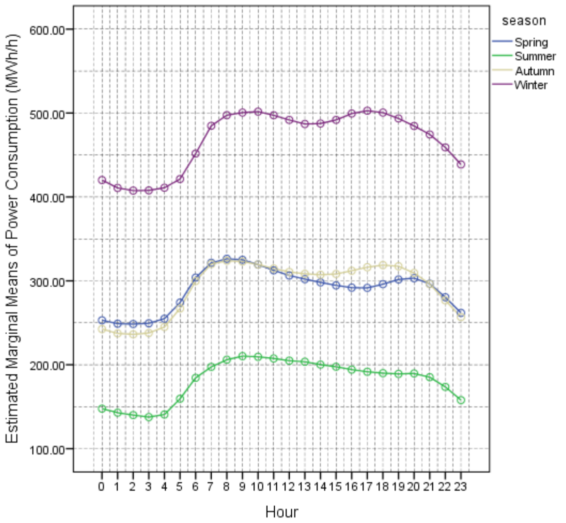


Figure 5.4: Mean hourly electricity consumption for different seasons in 2009.

Now let us examine how the different days of the week behave in consuming electricity. Figure 5.6 illustrates the mean power consumption on Mondays in year 2008, 2009 and 2010.

Figure 5.7 highlights the behaviour on Tuesdays which has a similar pattern like Monday, with almost similar magnitudes.

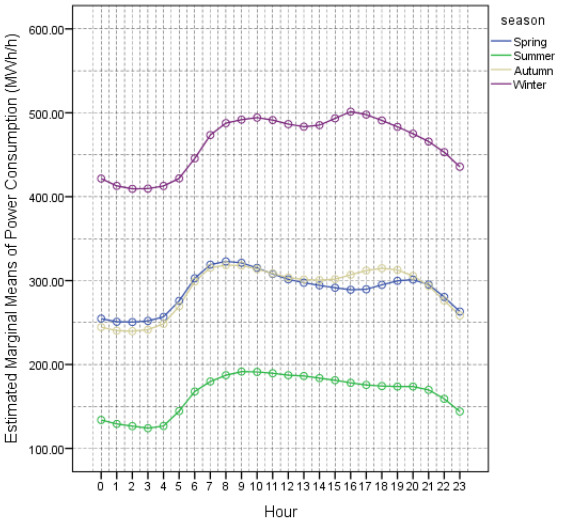
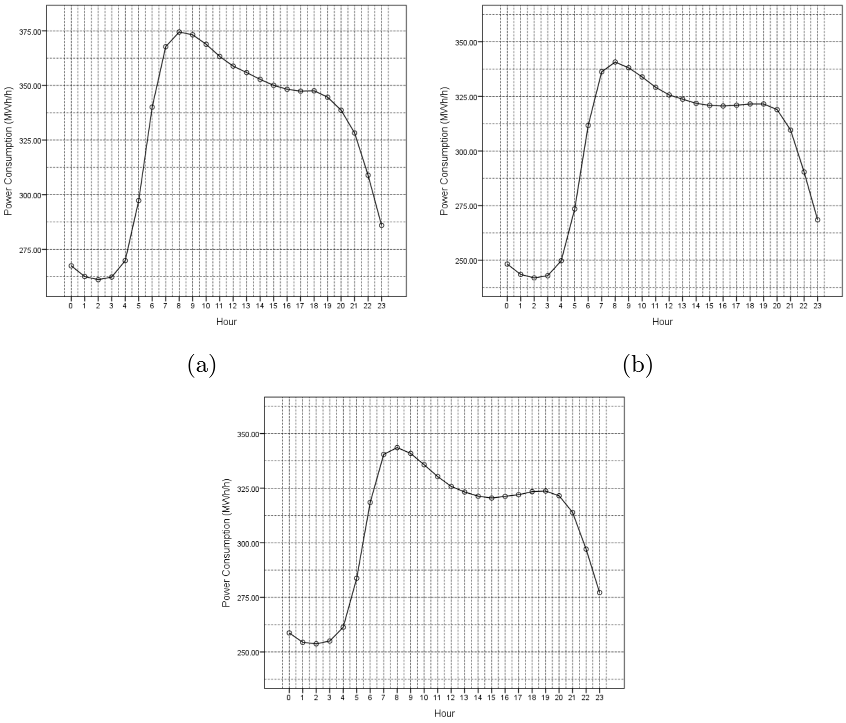


Figure 5.5: Mean hourly electricity consumption for different seasons in 2010.

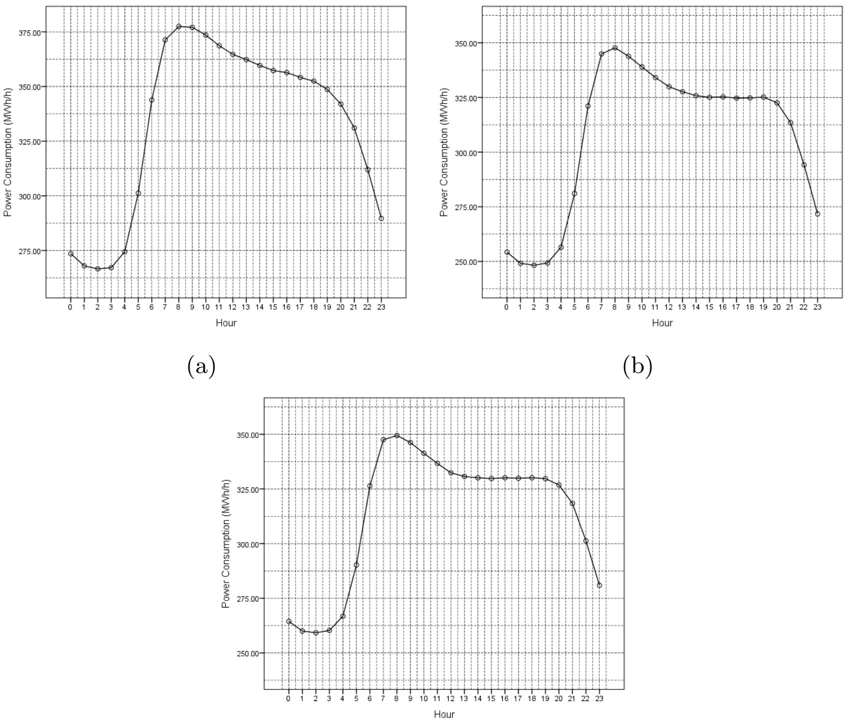


(a)

(b)

(c)

Figure 5.6: Mean hourly electricity consumption on Mondays in year (a) 2008 (b) 2009 and (c) 2010.



(a)

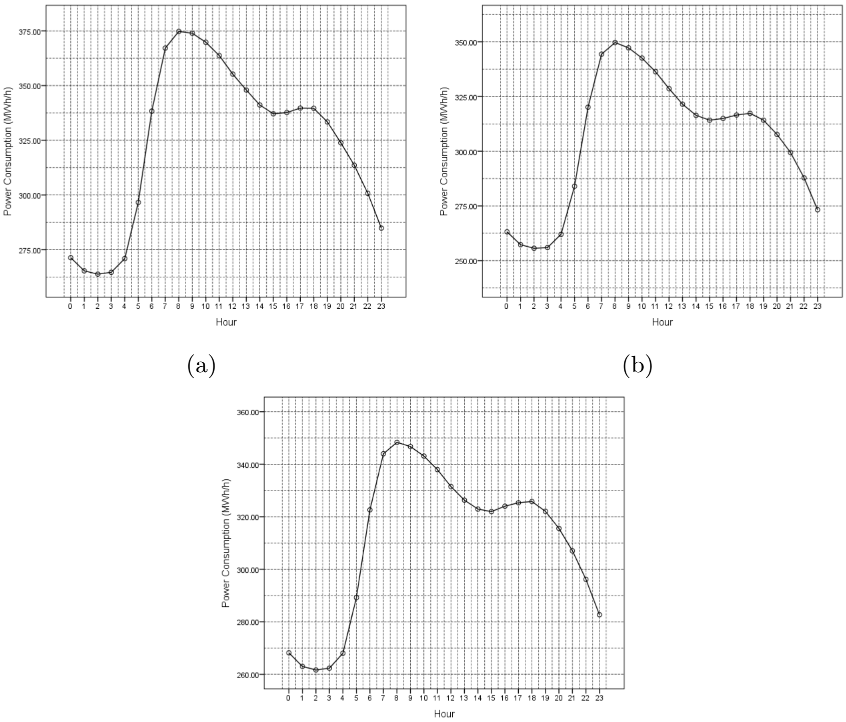
(b)

(c)

Figure 5.7: Mean hourly electricity consumption on Tuesdays in year (a) 2008 (b) 2009 and (c) 2010.

Wednesdays and Thursdays experience similar patterns like Mondays and Tuesdays. These figures are not illustrated here for clarity and refer Appendix for the corresponding figures.

However, Friday has a slightly deviated pattern from previous weekdays in between hours 15 to 18. But it does not affect very much for the rest of the hours of the day. The corresponding graph for Fridays is shown in Figure 5.8.

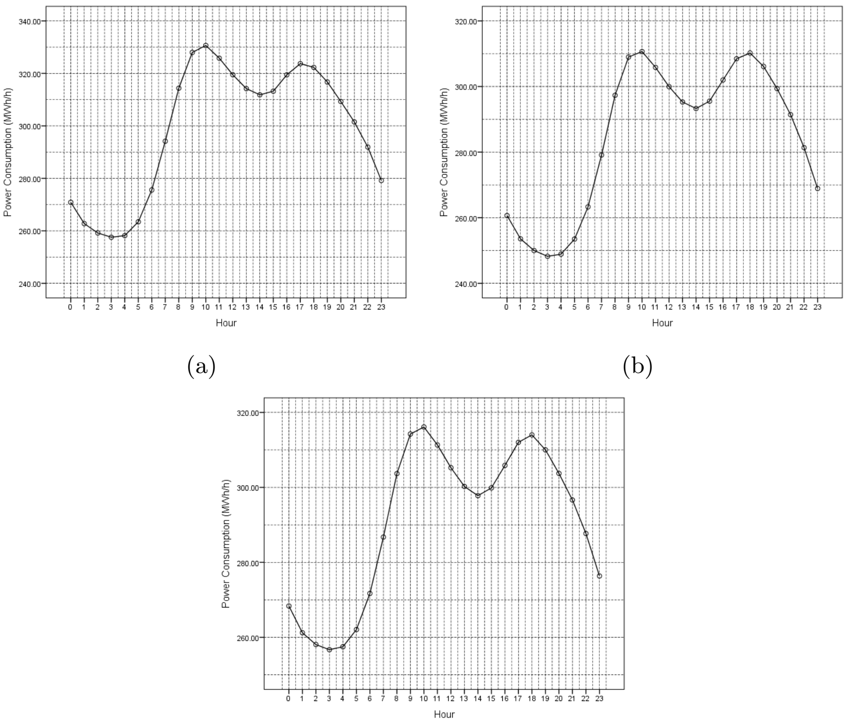


(a) (b)

(c)

Figure 5.8: Mean hourly electricity consumption on Fridays in year (a) 2008 (b) 2009 and (c) 2010.

Figure 5.9 illustrates the difference of power consumption on Saturdays compared to weekdays. Much apparent variations could be observed on Saturdays, especially in the day time and evening times.

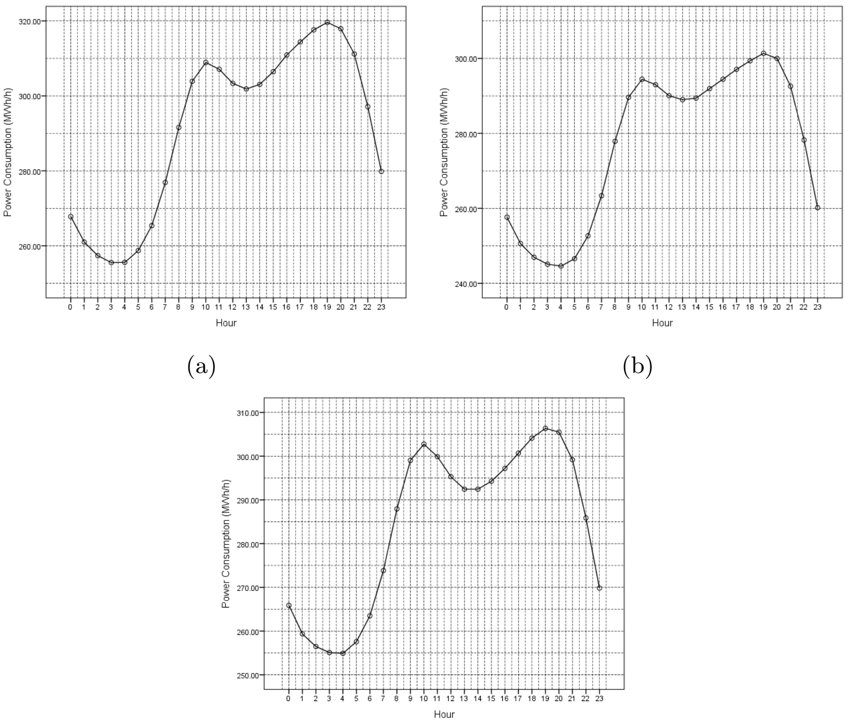


(a) (b)

(c)

Figure 5.9: Mean hourly electricity consumption on Saturdays in the year (a) 2008 (b) 2009 and (c) 2010.

Finally, the analysis of Sundays mean hourly consumption provides a graph with somewhat different but still similar pattern like Saturdays. Despite some discrepancies, the two behaviours could be taken as akin.



(a)

(b)

(c)

Figure 5.10: Mean hourly electricity consumption on Sundays in year (a) 2008 (b) 2009 and (c) 2010.

5.2 Feature Vector Design

From the preceding section, we identified the cyclic patterns associated with the electricity consumption, especially on annual, seasonal and weekly basis. When designing the feature vector we should take this into consideration. As we are focusing on short term forecasting on hourly basis and do testing for one year, we can ignore the annual pattern. However we have to deal with the other two in some way. Therefore we decide to divide test on the basis of seasonal prediction, which means we have to perform testing for each of the seasons separately.

When we look into the dataset, among all the independent variables - *temperature, wind speed, cloud cover* and *yesterday power consumption*, not all of them influence the consumption in considerable magnitude. For daily power consumption, Milindanath and Waseem [21] has found that *yesterday power consumption*, *temperature* and *wind speed* as the most powerful features which can be used to infer future power consumption. In general, *yesterday power consumption* has the highest influence whereas the other two have lower impact on the prediction. Therefore, these features could be taken as the foundation for the selection of a proper feature vector for the hourly prediction of power. In addition to that as we have found in section 5.1, different hours have different consumptions and different days especially weekends and weekdays have different consumptions. Therefore we need to include these two factors in the feature vector of independent variables.

Therefore we outline the following factors as the inherent features of the feature vector.

- Temperature
- Previous hour consumption
- hour
- day

In addition to these main factors, we also extend the feature vector to facilitate historical behaviour of temperature and power consumption to see how they affect to the future consumption. Therefore we include *power consumption before two hours* , *previous hour temperature* and *temperature before two hours* as additional features in the vector. Finally the feature vector is composed of the factors as summarized in Table 5.2.

Table 5.2: Variables considered in the feature vector and their notations.

Feature variable	**Notation**
previous hour consumption	P_{t-1}
hour before last hour consumption	P_{t-2}
current hour temperature	T_t
previous hour temperature	T_{t-1}
hour before last hour temperature	T_{t-2}
current hour	H_t
current day	D_t

5.3 Target Variable Analysis

It is obvious that in this analysis, the target variable is the power consumption of the next hour (P_t) and the next 24 hours ($P_t, P_{t+1}, P_{t+2}, ..., P_{t+23}$). But in order to perform the estimates, we model a different a target variable using the dataset.

It is not always true that the power consumption of a particular hour remains the same value in each year. There can be a significant deviation in the consumptions. This could be observed referring to Figure 5.2.

As an alternative approach we can think that the difference between two consecutive hours remains same as the patterns for the years remains almost same. For example in Figure 5.2, we measure the distances between successive hours from 0 to 23. Table 5.3 depicts the corresponding distances, according to which all the three years have quite similar distances for all the hours. However, the distances are different between different successive hours as it should normally be.

Table 5.3: Differences of mean power consumption (MWh/h) between successive hours of the 3 years

From-To (hour)	**2008**	**2009**	**2010**
0-1	-6.06	-5.84	-5.36
1-2	-2.07	-1.95	-1.60
2-3	0.12	-0.02	0.31
3-4	5.16	4.79	4.44
4-5	20.23	17.76	16.99
5-6	32.93	29.65	27.10
6-7	24.13	21.15	18.95
7-8	9.42	7.64	6.57
8-9	2.98	1.28	1.24
9-10	-1.98	-2.27	-2.48
10-11	-4.84	-4.60	-4.73
11-12	-5.14	-4.59	-4.97
12-13	-3.46	-3.08	-3.01
13-14	-2.74	-2.03	-1.61
14-15	-1.33	-0.20	-0.05
15-16	0.95	1.30	1.87
16-17	0.68	1.34	1.59
17-18	-0.19	0.66	1.14
18-19	-3.20	-0.85	-1.12
19-20	-6.12	-3.68	-3.60
20-21	-9.69	-8.59	-7.84
21-22	-16.28	-15.96	-14.38
22-23	-19.57	-18.90	-17.33
23-0	-13.92	-13.04	-12.09

Anyway this consistency can be used to model a new target variable, that is the power difference between the current hour and previous hour (for 24 hour case: between the considered hour and previous hour). Let us denote this as δP_t for the current hour and previous hour. This can be used as the new target variable and once the prediction is made, the actual prediction could be obtained as shown in the following equation generalized for 24 hours prediction.

$$P_{t+i} = P_{t-1} + \delta P_{t+i}, \quad i = 0, 1, ...23 \tag{5.1}$$

5.4 Feature Space Reduction and Normalization

After the feature vector and the target variable is fixed, we need to do some fine tuning to the dataset in order to see how these variations affect to the final prediction. In this section, we introduce two such techniques: *reduction* and *normalization.*

Kernel reduction is the technique we adopt here for space reduction. This is done by using the GP kernel $cov(x, x')$. Algorithm 1 shows how the reduction is done.

Algorithm 1 Feature space reduction using kernel function

```
$V$ = Feature Vector of $N$ data points
Calculate $K_*$ for $V$
$K_s$ =Sort($K_*$)
Select first $M$ points from $K_s$
```

In section 5.2, we developed a model for feature vector consisting of different predictor variables. But the values represented by these variables fall into different ranges. Therefore there is an inconsistency in their ranges. A unit increment in one variable may not be same as a unit increment in another variable. See Table 5.4 for the range of values in the predictor variables for 2008.

Table 5.4: Range of values of different predictor variables in 2008.

Feature variable	**Range**
previous hour consumption	112.785 MWh/h - 566.441 MWh/h
hour before last hour consumption	112.785 MWh/h - 566.441 MWh/h
current hour temperature	-16.5 ℃- 30.6 ℃
previous hour temperature	-16.5 ℃- 30.6 ℃
hour before last hour temperature	-16.5 ℃- 30.6 ℃
current hour	0 - 23
current day	1 - 7

To bring these different scales to a standard scale, we use normalization. Normalization has to be done for each variable based on their mean value (μ) and the standard deviation value (σ) of the distribution. Equation 5.2 can be used for normalizing a particular variable which has the value x.

$$Z_x = \frac{x - \mu}{\sigma} \tag{5.2}$$

5.5 Kernel Function

The squared exponential covariance function (kernel function) used in [28] is used in this research. The kernel function $k(x, x')$ is given by,

$$k(x, x') = \sigma_f^2 \exp[\frac{(x - x)^2}{2l^2}] + \sigma_n^2 \delta(x, x') \tag{5.3}$$

where σ_f^2 is the signal (noise free) variance, l is the length scale parameter, and $\delta(x, x')$ is the Kronecker delta function, which is equal to 1 iff $x = x'$ or 0 otherwise. The three parameters σ_f^2, l, and σ_n^2 are called hyperparameters (θ) which plays a very significant role in GP prediction. Optimum values set to this, would result in better prediction. Optimum values for these will be found using the maximum likelihood technique discussed in chapter 3

Using this kernel function, GP creates the covariance matrix K for all the points in the input feature vector. This covariance matrix is used later in predicting the corresponding target values given the input. This matrix is given by:

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix} \tag{5.4}$$

The maximum value an element in this matrix can obtain is $\sigma_f^2 + \sigma_n^2$, when the two points coincide each other. If the two points are far away (cannot see each other)from each other then the value of the element reaches 0. In this way GP identifies the best pairs as the ones with the highest element values. GP also needs to calculate the covariance matrices for the prediction point with the other points and with itself. These two matrices are given in equation 5.5 and 5.6.

$$K_* = \begin{bmatrix} k(x_*, x_1) & k(x_*, x_2) & \cdots & k(x_*, x_n) \end{bmatrix} \tag{5.5}$$

$$K_{**} = k(x_*, x_*) \tag{5.6}$$

Chapter 6

Experiments and Results

In this chapter we illustrate the empirical forecasting results obtained through Gaussian processes and other traditional methods. Here, we denote the actual consumption value at time t as A_t and the forecast value at the same instance as F_t. Based on this notation, error measurements are calculated in the following forms:

Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{t=0}^{N}(A_t - F_t)^2 \tag{6.1}$$

Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{t=0}^{N}|A_t - F_t| \tag{6.2}$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100}{N}\sum_{t=0}^{N}\left|\frac{A_t - F_t}{A_t}\right| \tag{6.3}$$

6.1 Gaussian Process

For Gaussian processes, we perform testing in two basic ways. First we predict the next hour power consumption taking 169 hours in the future as test data, starting from hour 23 of February 05^{th}, 2009. For this case we use the previous 800 data points (hours) for each hour as the training dataset. Therefore we name it as dynamic dataset as the dataset always changing for

Table 6.1: Fixed training and test datasets.

	Training dataset	**Test dataset**
Winter	15.12.2008 Hour 00 - 17.01.2009 Hour 08	17.01.2009 Hour 10 - 24.01.2009 Hour 11
Spring	31.03.2098 Hour 00 - 03.05.2009 Hour 08	03.05.2009 Hour 10 - 10.05.2009 Hour 11
Summer	05.06.2009 Hour 00 - 08.07.2009 Hour 08	08.07.2009 Hour 10 - 15.07.2009 Hour 11
Autumn	15.09.2008 Hour 00 - 18.10.2009 Hour 08	18.10.2009 Hour 10 - 25.10.2009 Hour 11

each hour. The second method is using a fixed dataset based on the four seasons in the year: winter, spring, summer and winter. The training and test datasets considered are shown in Table 6.1. In the next section, we start with the first method to see how the predictor variables behave in predicting δP_t and P_t, for the next hour. Forecast results for the next 24 hours will be given in section 6.1.7.

6.1.1 Dynamic Training Dataset - Original Data

First we will apply GP for the original dataset without any reduction or normalization. The results shown in Table 6.2, corresponds to the experiment where δP_t is used as the target variable. The best prediction result is given by the signature $P_{t-1}, T_t, T_{t-1}, H_t, D_t$ which has a MAE of 2.472 and a standard deviation 2.052. For all the combinations, the MAE value lies between 2 and 5, and eight combinations out of nine, have a MAPE of less than 1%, which is a promising figure. The first 5 combinations have excellent results with a consistent percentage error of approximately 0.5%.

Table 6.2: Results for the prediction of power difference for original dataset.

Signature	**MAE**	**MSE**	σ	**MAPE (%)**
	(MWh/h)	**(MWh/h)**	**(MWh/h)**	
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	2.472	10.300	2.052	0.508
P_{t-1}, T_t, H_t, D_t	2.629	11.770	2.211	0.536
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	2.652	12.000	2.236	0.541
$P_{t-1}, T_{t-1}, H_t, D_t$	2.678	12.061	2.219	0.546
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.826	14.408	2.542	0.576
$P_{t-1}, P_{t-2}, H_t, D_t$	3.039	23.691	3.814	0.617
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	3.244	22.88	3.526	0.668
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	4.365	37.027	4.252	0.908
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	4.948	43.952	4.426	1.024

Figure 6.1, illustrates the results for the best combination. We can observe how closely the actual and predicted are plotted. Slight deviations could be observed in the bends of the curve.

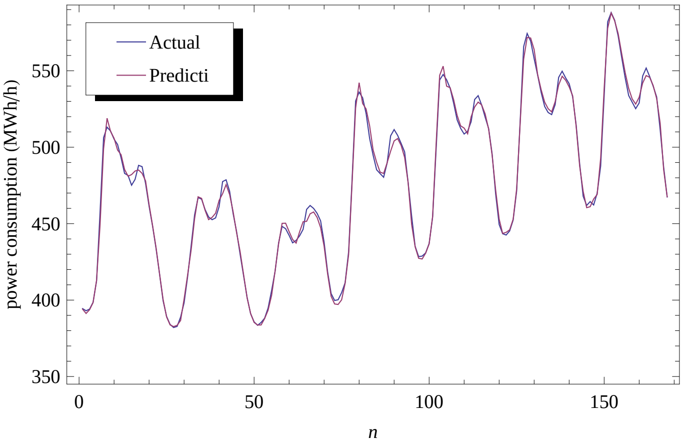


Figure 6.1: Actual and predicted electricity consumptions of the combination $P_{t-1}T_tT_{t-1}H_tD_t$ for the original data set with δP_t as the target variable.

The same experiment was done by just changing the target variable to actual power consumption P_t rather than the difference. Table 6.3 depicts the results for this case.

Table 6.3: Results for the prediction of power consumption for original dataset.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	σ **(MWh/h)**	**MAPE (%)**
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.926	48.455	5.766	0.838
$P_{t-1}, P_{t-2}, H_t, D_t$	5.288	134.092	10.333	1.108
$P_{t-1}, T_{t-1}, H_t, D_t$	8.265	1777.712	41.468	1.637
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	9.653	1691.942	40.104	1.978
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	12.682	4009.561	62.223	2.711
$\delta P_{t-1}, \delta P_{t-1}, T_t, T_{t-1}, H_t, D_t$	18.277	1339.698	31.807	3.673
$\delta P_{t-1}, T_t, T_{t-1}, H_t, D_t$	20.622	1656.488	35.194	4.206
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	21.310	4084.106	60.429	4.429
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	67.097	31974.840	166.245	13.324

The best prediction is given by the combination $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$, which has a MAE of 3.926. But we can see even the best result is not good like the previous case, where we can see the first seven results in Table 6.2 are better than the best result in this case. On the other hand, the standard deviation is also very high when we compare with the previous test. Therefore, we can see that there are predictions which are very far from the actual value when we use P_t as the target variable. The best prediction combination for the power difference case is not giving best results in this case (see the 5^{th} record from the top). Also note that in this test we have used power difference (δP_{t-1} and δP_{t-2}) as input variables as a trial to see how they behave. But the results are not that promising according to the table. The corresponding graph for the best prediction is illustrated in Figure 6.2. See how it deviates from the actual curve when compared with Figure 6.1.

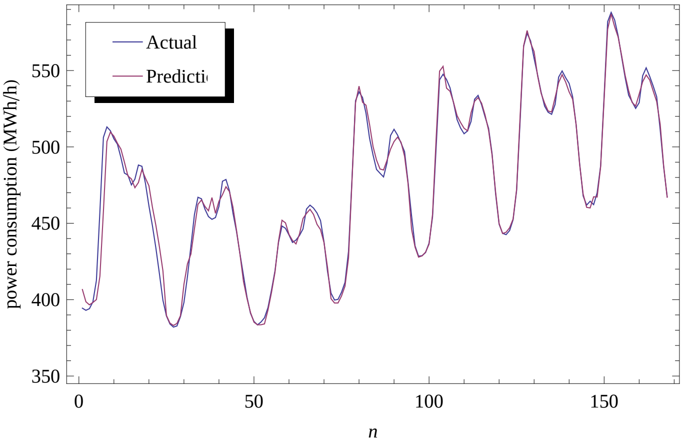


Figure 6.2: Actual and predicted electricity consumptions of the combination $P_{t-1}T_tT_{t-1}T_{t-2}H_tD_t$ for the original data set with P_t as the target variable.

6.1.2 Dynamic Training Dataset - Kernel Reduced Data

Now we use the kernel reduction method discussed in the previous chapter, for the previous 800 training data and select the best 700 values, based on the best kernel values. The results for the δP_t prediction is given in Table 6.4.

Table 6.4: Results for the prediction of power difference for kernel reduced dataset.

Signature	**MAE**	**MSE**	σ	**MAPE**
	(MWh/h)	**(MWh/h)**	**(MWh/h)**	**(%)**
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	2.473	10.259	2.042	0.508
P_{t-1}, T_t, H_t, D_t	2.682	12.273	2.261	0.547
$P_{t-1}, T_{t-1}, H_t, D_t$	2.693	12.160	2.222	0.550
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	2.776	13.094	2.328	0.574
P_{t-1}, H_t, D_t	2.877	14.646	2.531	0.586
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.005	17.236	2.873	0.609
$P_{t-1}, P_{t-2}, T_{t-1}, H_t, D_t$	4.021	36.744	4.549	0.829
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	4.088	37.976	4.625	0.840
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	11.707	278.298	11.920	2.433
$P_{t-1}, P_{t-2}, H_t, D_t$	9.702	179.315	9.258	2.025

The results are very much identical with the results for the original data in Table 6.2 in which, the best prediction is also $P_{t-1}, T_t, T_{t-1}, H_t, D_t$ with the same MAE, but a slightly better standard deviation. However, if we consider the full table, we can see the reduction does not produce very good results as we expected. As a whole, the results in Table 6.2 are slightly better. The predicted and actual curve for the best combination is given in Figure 6.3, which is much similar to Figure 6.1.

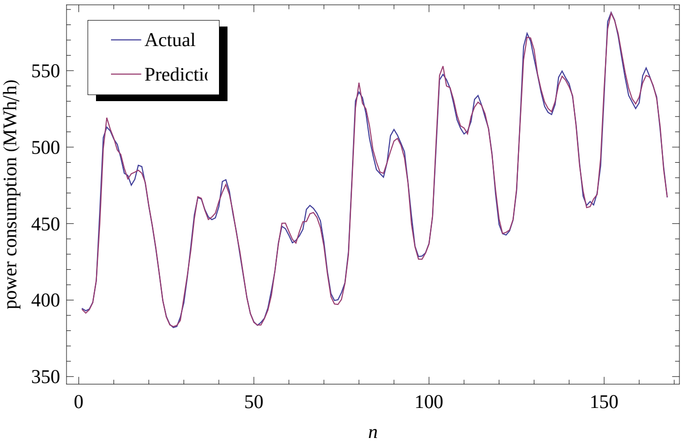


Figure 6.3: Actual and predicted electricity consumptions of the combination $P_{t-1}T_tT_{t-1}H_tD_t$ for the kernel reduced data set with δP_t as the target variable.

The results for the case where power consumption works as the target variable is depicted in Table 6.5. Here we can see better results than the results in Table 6.3, although not better than the power difference results. Here the best combination is P_{t-1}, H_t, D_t which has a MAE value of 3.531.

Table 6.5: Results for the prediction of power consumption for kernel reduced dataset.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	σ **(MWh/h)**	**MAPE (%)**
P_{t-1}, H_t, D_t	3.531	37.500	5.018	0.725
$P_{t-1}, P_{t-2}, T_{t-1}, H_t, D_t$	5.312	49.272	4.602	1.111
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	14.737	2599.815	48.958	2.932
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	39.330	16307.520	121.857	8.276
$P_{t-1}, T_{t-1}, H_t, D_t$	83.375	36542.751	172.536	17.632
P_{t-1}, T_t, H_t, D_t	128.017	57335.46	202.959	27.455
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	179.582	82746.61	225.387	38.855

Figure 6.4 shows the comparison of the two curves: actual and predicted for the best configuration for kernel reduction method, and power consumption as the target variable.

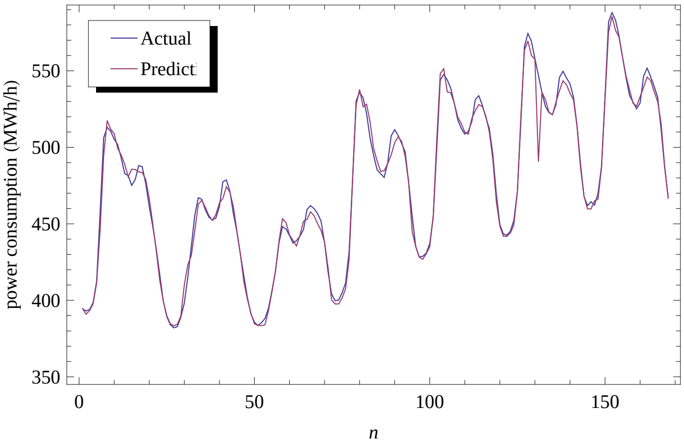


Figure 6.4: Actual and predicted electricity consumptions of the combination $P_{t-1}H_tD_t$ for the kernel reduced data set with P_t as the target variable.

6.1.3 Dynamic Training Dataset - Normalized Data

In this section, we present the results for the tests carried out for the normalized data. The normalization procedure was explained in chapter 5. Table 6.6 depicts the corresponding results for the δP_t predictions. According to the table, we can see that the first 5 results are very close to each other for every metric value considered. The MAE value is below 2.7 for these cases. The best combination is $P_{t-1}, T_t, T_{t-1}, H_t, D_t$, which is same as for the original data. The standard deviation and the MAPE values are very consistent compared with the other tests.

Table 6.6: Results for the prediction of power difference for normalized dataset.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	σ **(MWh/h)**	**MAPE (%)**
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	2.621	11.775	2.221	0.534
P_{t-1}, T_t, H_t, D_t	2.622	11.779	2.221	0.534
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.623	11.788	2.222	0.534
$P_{t-1}, T_{t-1}, H_t, D_t$	2.635	11.933	2.240	0.537
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	2.636	11.934	2.240	0.537
$P_{t-1}, P_{t-2}, H_t, D_t$	5.892	69.062	5.878	1.201
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	6.891	83.701	6.036	1.428
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	6.909	84.087	6.047	1.433
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	6.942	84.670	6.057	1.440

Figure 6.5, depicts the behaviour of actual consumption and the predicted consumption for the best signature. Most of the time the curves coincide each other except for some bends in

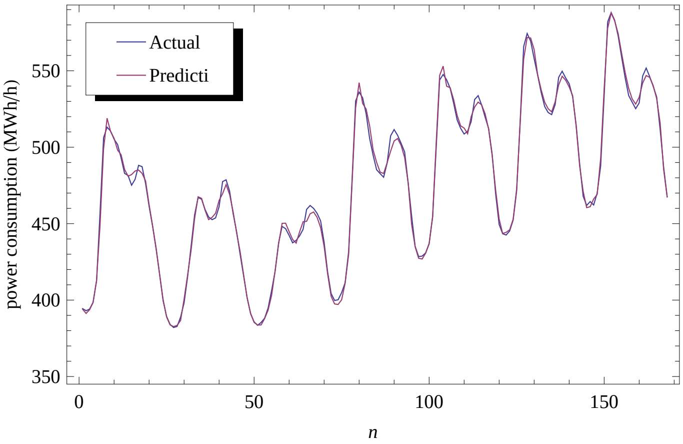


Figure 6.5: Actual and predicted electricity consumptions of the combination $P_{t-1}T_tT_{t-1}H_tD_t$ for the normalized data set with δP_t as the target variable.

the curve.

The prediction results for the target variable P_t is shown in Table 6.7. The best combination is $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ with a MAE of 2.768 and MAPE of 0.578. We can see that predictions for P_t are not as good as the predictions for δP_t. However in the table we can see P_{t-2} influences more when the target variable is P_t.

Table 6.7: Results for the prediction of power consumption for normalized dataset.

Signature	**MAE**	**MSE**	σ	**MAPE**
	(MWh/h)	**(MWh/h)**	**(MWh/h)**	**(%)**
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.768	13.089	2.336	0.578
$P_{t-1}, P_{t-2}, H_t, D_t$	2.930	15.587	2.654	0.612
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	6.932	138.182	9.522	1.454
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	7.034	228.146	13.407	1.516
$P_{t-1}, T_{t-1}, H_t, D_t$	7.970	296.798	15.319	1.645
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	11.377	571.961	21.099	2.363
P_{t-1}, T_t, H_t, D_t	23.482	6847.766	79.587	4.541

The corresponding curves for the actual power consumption and predicted consumptions for the best signature is shown in Figure 6.6.

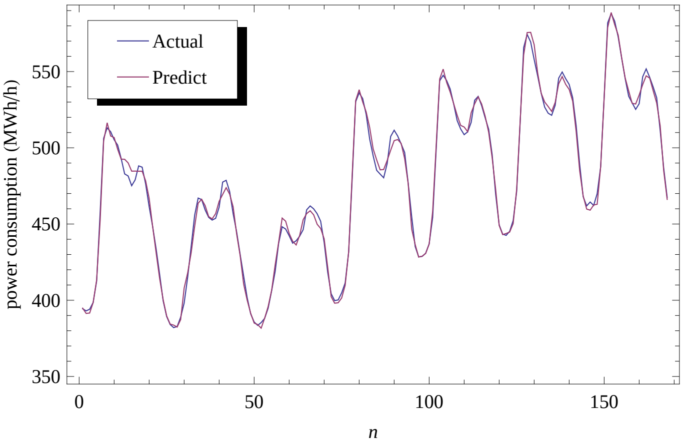


Figure 6.6: Actual and predicted electricity consumptions of the combination $P_{t-1}P_{t-2}T_tT_{t-1}T_{t-2}H_tD_t$ for the normalized data set with P_t as the target variable.

6.1.4 Fixed Training Dataset - Original Data

The second testing method with a fixed training and test dataset is a fast method than the dynamic dataset. Mainly the dynamic dataset results are important to identify which factors and combination have more effect on the final prediction. In section 6.1.1, 6.1.2 and 6.1.3 we found out that δP_t produces far better results as the target variable than the actual power consumption as the target variable. Therefore in this section and the following two sections we focus our study on the power difference to predict for seasonal variations including winter, spring , summer and autumn.

First consider the results for the original dataset without any normalization and reduction. The corresponding results are shown in Table 6.8. For the winter season, the best prediction is given by $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$, which has MAE of 3.010. The best combination for spring is $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$. It has a MAE of 2.991 which is a better result than winter. For summer, 1.862 is the best MAE given by the combination $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$. Summer result is the best of all four seasons as the autumn best value is 2.339 given by the combination $P_{t-1}, P_{t-2}, T_{t-1}, T_t, H_t, D_t$. But if we take the total mean MAE of all the four seasons, we can select it as the best combination.

Table 6.8: Results for the prediction of power difference for original dataset on seasonal basis.

Signature	**MAE (MWh/h)**				**Total Mean Error (MWh/h)**
	Winter	**Spring**	**Summer**	**Autumn**	
$P_{t-1}, P_{t-2}, H_t, D_t$	3.293	3.028	1.960	2.511	2.698
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.010	3.138	2.180	2.562	2.722
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	3.027	3.304	2.260	2.635	2.806
P_{t-1}, T_t, H_t, D_t	3.651	3.151	1.866	2.565	2.808
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	4.117	3.018	1.862	2.339	2.834
$P_{t-1}, T_{t-1}, H_t, D_t$	3.651	3.248	1.901	2.652	2.863
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	4.137	3.118	1.912	2.347	2.878
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	3.015	3.150	2.258	3.109	2.883
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	4.106	2.991	1.905	5.105	3.527

6.1.5 Fixed Training Dataset - Kernel Reduced Data

Next, consider the results for the reduced fixed training dataset using the kernel reduction method. The results shown in Table 6.9 are not as good as the first case. The best prediction for winter is given by $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$. But its MAE value is larger than 3.5. We can observe that $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ as the best combination altogether because it produces the best results for spring(3.539), summer(2.276) and autumn(2.526). Summer has the best prediction of all.

Table 6.9: Results for the prediction of power difference for kernel reduced dataset on seasonal basis.

Signature	**MAE (MWh/h)**				**Total Mean Error (MWh/h)**
	Winter	**Spring**	**Summer**	**Autumn**	
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	5.234	3.539	2.276	2.526	3.394
$P_{t-1}, P_{t-2}, H_t, D_t$	4.022	3.735	2.651	3.223	3.408
P_{t-1}, T_t, H_t, D_t	4.275	3.855	2.564	3.347	3.510
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	4.255	3.862	2.894	3.172	3.546
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.627	3.776	2.811	4.046	3.565
$P_{t-1}, T_{t-1}, H_t, D_t$	4.271	3.963	2.601	3.429	3.566
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	4.273	3.961	2.885	3.386	3.626
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	5.416	3.622	2.600	3.102	3.685
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	5.392	3.639	2.310	5.304	4.161

6.1.6 Fixed Training Dataset - Normalized Data

Finally we do the testing for the normalized data for the four seasons. Table 6.10 illustrates the prediction results. The best combination is $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ which has 2.952 for winter, 2.999 for spring, 1.852 for summer, and 2.404 for autumn. However the best prediction for winter (i.e 2.950) is given by $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$ and for summer it is 1.848 given by $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$.

Table 6.10: Results for the prediction of power difference for normalized dataset on seasonal basis.

Signature	**MAE (MWh/h)**				**Total Mean Error (MWh/h)**
	Winter	**Spring**	**Summer**	**Autumn**	
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.952	2.999	1.852	2.404	2.552
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.950	3.017	1.862	2.451	2.570
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	2.996	3.052	1.893	2.347	2.572
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.010	3.138	1.848	2.505	2.625
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	3.015	3.150	1.866	2.562	2.648
P_{t-1}, T_t, H_t, D_t	3.044	3.151	1.866	2.565	2.656
$P_{t-1}, T_{t-1}, H_t, D_t$	3.031	3.248	1.900	2.652	2.708
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	3.027	3.304	1.893	2.635	2.715
$P_{t-1}, P_{t-2}, H_t, D_t$	3.449	3.028	1.960	2.511	2.737

6.1.7 Next 24 Hours Prediction

The best signatures found in preceding sections is used here to predict for the next 24 hours. For the winter season the best combination was $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$, and $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$, $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ and $P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$ were the best signatures for spring, summer and autumn respectively. Table 6.11 illustrates the next 24 hour prediction for the power difference prediction for the original data, in winter season.

From the table we can observe that the accuracy of the predictions reduces as we go away from the current hour.

Table 6.11: Results for the prediction of power difference for next 24 hours in winter.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}
MAE	3.0	5.3	6.1	6.9	8.3	9.8	9.0	9.8	9.7	10.2	11.9	11.3	14.4	12.0	15.2	14.8	12.1	12.0	15.0	15.4	14.9	13.6	14.2	14.4
MAPE	0.7	1.3	1.5	1.7	2.0	2.4	2.2	2.4	2.4	2.5	2.9	2.8	3.5	3.1	3.7	3.6	3.2	3.2	3.8	3.9	3.8	3.4	3.6	3.6

The corresponding 24 hour prediction for the spring season is illustrated in Table 6.12. A similar behaviour could be observed for this case also. Accuracy of winter is better than spring.

Table 6.12: Results for the prediction of power difference for next 24 hours in spring.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}
MAE	3.0	5.8	7.8	12.1	13.5	12.7	14.9	16.1	17.4	18.1	18.6	18.4	20.4	22.9	26.7	21.1	19.4	20.6	20.8	21.4	18.6	17.8	17.3	16.8
MAPE	1.2	2.4	3.2	5.0	5.5	5.1	6.0	6.6	7.1	7.3	7.5	7.5	8.4	9.5	11.2	8.5	7.8	8.3	8.4	8.8	7.5	7.1	6.8	6.7

In the Summer, as illustrated in Table 6.13 we can see somewhat different behaviour, but almost similar. When we go far from the current hour the accuracy reduces until about t_{17} and then the accuracy starts to improve.

Table 6.13: Results for the prediction of power difference for next 24 hours in summer.

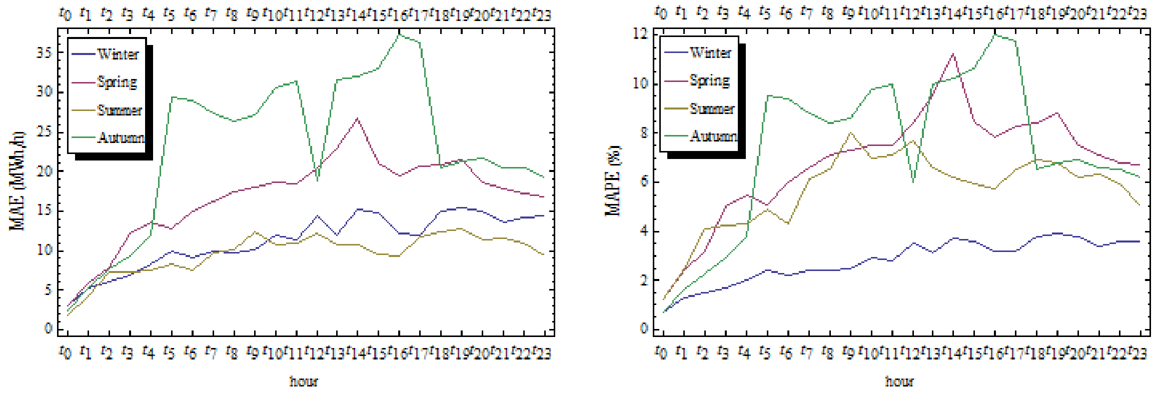
	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}
MAE	1.8	4.2	7.2	7.2	7.4	8.3	7.4	9.7	10.1	12.4	10.8	10.9	12.1	10.8	10.8	9.5	9.2	11.7	12.4	12.8	11.4	11.5	10.9	9.2
MAPE	1.2	2.5	4.1	4.2	4.3	4.9	4.3	6.1	6.5	8.0	7.0	7.1	7.7	6.6	6.2	5.9	5.7	6.5	6.9	6.8	6.2	6.3	5.9	5.0

The accuracy of the results of autumn is worse when considering the other three seasons. The behaviour is very similar to winter and spring. The results are given in Table 6.14.

Table 6.14: Results for the prediction of power difference for next 24 hours in autumn.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}
MAE	2.3	5.2	7.6	9.3	12.0	29.3	29.0	27.4	26.4	27.2	30.6	31.3	18.8	31.5	32.0	33.0	37.3	36.3	20.4	21.2	21.6	20.5	20.4	19.3
MAPE	0.7	1.6	2.3	2.9	3.8	9.5	9.4	8.8	8.4	8.6	9.8	10.0	6.0	10.0	10.2	10.6	12.0	11.7	6.5	6.8	6.9	6.6	6.5	6.2

Figure 6.7 depicts the MAE and MAPE summaries of the 24 hour predictions, where we can observe better accuracy for summer and worse and inconsistent pattern for winter in the first graph. However the second graph gives very interesting result. Although winter is bad in terms of absolute error, its error percentage is very good meaning it has a better prediction accuracy compared to its actual power consumption values. But it is opposite for the summer season, as its percentage error is higher compared to winter.



(a)

(b)

Figure 6.7: The (a) MAE and (b) MAPE curves for the next 24 hours predictions for the best combinations in the four seasons. $P_{t-1}T_tT_{t-1}T_{t-2}H_tD_t$ for winter, $P_{t-1}P_{t-2}T_tT_{t-1}T_{t-2}H_tD_t$ for spring, $P_{t-1}P_{t-2}T_tT_{t-1}H_tD_t$ for summer and $P_{t-1}P_{t-2}T_tT_{t-1}H_tD_t$ for autumn.

6.2 Traditional Approaches

In this section we look at the results of the two traditional techniques of electricity forecasting : MBPNN which is a variant of Artificial Neural Networks and MLR. We used the same datasets used for the fixed dataset method in these two cases to predict power difference. But in these cases we ignore the kernel reduction method as it is not related to these techniques. However we focus on the normalization.

6.2.1 MBPNN - Original Data

We use the tool developed by Noel Lopes and Bernardete Ribeiro [2] for multiple back-propagation to perform the tests on Neural networks. It is interesting to see how neural networks works as it has been observed that a neural network model with a good fitting performance for the past data may not give as good forecasting performance for the future [36].

Table 6.15 depicts the results obtained for the original data using the MBP network with 10 hidden layer neuron and 10000 epochs. According to the table, we can observe that the best predictions given for the summer season like GP. For the winter, the best result is 2.678 which is a better result than the GP value (3.010). But for the spring, GP has a better value (2.991) than MBP which is 3.584 according to Table 6.15. The best prediction for summer is 1.820 which is slightly better than GP. For the autumn, the best prediction is 2.750.

Table 6.15: Results for the prediction of power difference in MBPNN with 10 hidden layer neurons and 10000 epochs.

Signature	**MAE (MWh/h)**				**Total Mean Error (MWh/h)**
	Winter	**Spring**	**Summer**	**Autumn**	
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	2.876	3.714	1.820	2.844	2.814
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.678	3.885	2.100	2.750	2.853
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.788	4.474	2.180	2.777	3.055
$P_{t-1}, P_{t-2}, H_t, D_t$	3.428	3.584	1.932	3.450	3.098
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	3.153	4.054	2.052	3.627	3.222
P_{t-1}, T_t, H_t, D_t	3.607	4.248	2.159	3.330	3.336
$P_{t-1}, T_{t-1}, H_t, D_t$	3.329	4.415	2.320	3.408	3.368
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	3.822	4.451	2.285	3.162	3.430
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.446	4.334	2.653	3.394	3.457

6.2.2 MBPNN - Normalized Data

When we consider normalized data, winter produces an improved result of 2.378. Other than that all the others do not have a better prediction than for the original data.

Table 6.16: Results for the prediction of power difference in MBPNN with 10 hidden neurons and 10000 epochs.

Signature	**MAE (MWh/h)**				**Total Mean Error (MWh/h)**
	Winter	**Spring**	**Summer**	**Autumn**	
$P_{t-1}, T_{t-1}, H_t, D_t$	2.378	4.696	2.195	3.958	3.307
$P_{t-1}, P_{t-2}, H_t, D_t$	3.631	3.804	2.089	3.818	3.336
$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	3.946	4.057	2.982	4.077	3.766
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	4.037	4.443	2.765	3.906	3.788
$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	5.098	5.260	2.455	3.606	4.105
P_{t-1}, T_t, H_t, D_t	5.171	4.942	2.288	4.203	4.151
$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	4.592	4.832	2.447	5.105	4.244
$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	4.360	4.903	5.381	3.870	4.628
$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	4.168	4.848	2.139	8.280	4.859

6.2.3 Multiple Linear Regression

Linear regression is very popular method in regression as we mentioned in Chapter 4. We perform multiple linear regression to predict the power difference for the original data using SPSS[1].

[1] http://www-01.ibm.com/software/analytics/spss/products/statistics/

First, for the winter training dataset the multiple linear regression method was applied using the forward selection method. Four models are obtained. According to Table 6.17 and Table 6.18, all the models are significant with their parameters. By considering Table 6.19 , it can be seen that model 4 has higher R square value than others and lower standard deviation (i.e standard error). Therefore it can be selected as the best model to describe the situation.

Table 6.17: Coefficients table for winter.

Model		**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
		B	**Std.Error**	**Beta**		
1	(Constant)	35.763	3.715		9.626	.000
	P_{t-2}	-.082	.008	-.325	-9.693	.000
2	(Constant)	25.378	2.356		10.772	.000
	P_{t-2}	-.799	.021	-3.157	-37.675	.000
	P_{t-1}	.741	.021	2.927	34.923	.000
3	(Constant)	35.839	2.614		13.708	.000
	P_{t-2}	-.811	.020	-3.203	-39.619	.000
	P_{t-1}	.724	.021	2.860	35.272	.000
	T_{t-1}	-.483	.060	-.198	-8.036	.000
4	(Constant)	31.747	3.000		10.583	.000
	P_{t-2}	-.798	.021	-3.154	-38.236	.000
	P_{t-1}	.726	.020	2.866	35.481	.000
	T_{t-1}	-.402	.067	-.165	-6.032	.000
	H_t	-.148	.054	-.072	-2.748	.006

Table 6.18: ANOVA table for winter.

Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
1	Regression	17264.905	1	17264.905	93.951	.000
	Residual	146644.945	798	183.766		
	Total	163909.851	799			
2	Regression	105954.008	2	52977.004	728.532	.000
	Residual	57955.843	797	72.717		
	Total	163909.851	799			
3	Regression	110302.811	3	36767.004	545.955	.000
	Residual	53607.040	796	67.346		
	Total	163909.851	799			
4	Regression	110807.043	4	27701.761	414.722	.000
	Residual	53102.807	795	66.796		
	Total	163909.851	799			

Table 6.19: Model summary table for winter.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.325	.105	.104	13.556017
2	.804	.646	.646	8.527455
3	.820	.673	.672	8.206432
4	.822	.676	.674	8.172881

Therefore the regression model can be given by the following equation,

$$\delta P_t = 31.747 + (.726)P_{t-1} - (.798)P_{t-2} - (.402)T_{t-1} - (.148)H_t \tag{6.4}$$

After substituting this equation for the training dataset, the predictions could be obtained as shown in Table 6.20.

Table 6.20: Results for the prediction of power consumption for winter using MLR.

Signature	MAE (MWh/h)	MSE (MWh/h)	σ (MWh/h)
$P_{t-1}, P_{t-2}, T_{t-1}, H_t$	6.147	76.712	6.258

For the spring training dataset we can observe coefficient table and the ANOVA tables as shown in Table 6.21 and .6.22.

Table 6.21: Coefficients table for spring.

Model	**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std.Error**	**Beta**		
1 (Constant)	6.733	.787		8.557	.000
H_t	-.602	.059	-.340	-10.228	.000
2 (Constant)	24.996	2.471		10.116	.000
H_t	-.485	.059	-.274	-8.241	.000
P_{t-2}	-.072	.009	-.258	-7.766	.000
3 (Constant)	18.417	1.797		10.252	.000
H_t	-.350	.043	-.198	-8.206	.000
P_{t-2}	-.691	.024	-2.490	-29.129	.000
P_{t-1}	-.639	.023	2.301	27.198	.000
4 (Constant)	20.973	1.994		10.516	.000
H_t	-.290	.047	-.164	-6.140	.000
P_{t-1}	-.690	.024	-2.484	-29.186	.000
P_{t-2}	.629	.024	2.267	26.661	.000
T_t	-.235	.081	-.077	-2.894	.004
5 (Constant)	24.771	2.339		10.590	.000
H_t	-.275	.047	-.156	-5.833	.000
P_{t-1}	-.692	.024	-2.492	-29.419	.000
P_{t-2}	.624	.024	2.250	26.532	.000
T_t	-.267	.082	-.088	-.2673	.001
D_t	-.468	.153	-.074	-3.066	.002
6 (Constant)	22.287	2.624		8.492	.000
H_t	-.332	.054	-.188	-6.095	.000
P_{t-1}	-.693	.023	-2.498	-29.532	.000
P_{t-2}	.635	.024	2.287	26.437	.000
T_t	-.547	.158	-.180	-3.463	.001
D_t	-.405	.155	-.064	-2.609	.009
T_{t-2}	.385	.186	.127	2.071	.039

Table 6.22: ANOVA table for spring.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13962.657	1	13962.657	104.608	.000
	Residual	106513.918	798	133.476		
	Total	120476.575	799			
2	Regression	21456.651	2	10728.325	86.351	.000
	Residual	99019.924	797	124.241		
	Total	120476.575	799			
3	Regression	69153.4461	3	23051.149	357.514	.000
	Residual	51323.129	796	64.476		
	Total	120476.575	799			
4	Regression	69688.446	4	17422.149	272.714	.000
	Residual	51323.129	795	63.884		
	Total	120476.575	799			
5	Regression	70282.760	5	14056.552	222.356	.000
	Residual	50193.815	794	63.216		
	Total	120476.575	799			
6	Regression	70552.760	6	11758.793	186.779	.000
	Residual	49923.814	793	62.956		
	Total	120476.575	799			

According to table 6.21, we can see that all the coefficient of the models are significant and by considering table 6.22, it can be obtained that all the test are also significant. From figure 6.23 we can observe that R square values are higher and close from model 3 to 6. Although the model 4 and 5 and 6 have higher R square value than 3, we try for ease of understanding and interpretation to describe the process with as few variables as possible. Since the model 6 has 6 predictor variables and model 5 has 5 predictor variables and the model 4 has 4 predictor variables , model 3 with 3 predictor variables can be taken as the suitable model to describe the spring season.

Table 6.23: Model summary table for spring.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.340	.116	.115	11.553185
2	.422	.178	.176	11.146336
3	.758	.574	.572	8.029713
4	.761	.578	.576	7.992762
5	.764	.583	.581	7.959874
6	.765	.586	.582	7.934458

The corresponding regression equation can be written as follows:

$$\delta P_t = 18.417 + (.639)P_{t-1} - (.691)P_{t-2} - (.350)H_t \tag{6.5}$$

Table 6.24: Results for the prediction of power consumption for spring using MLR.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	**σ (MWh/h)**
P_{t-1}, P_{t-2}, H_t	5.934	71.512	6.042

The results for the testing is as shown in Table 6.24.

For the summer dataset the corresponding coefficient and ANOVA tables are depicted in Table 6.25 and 6.26 respectively.

Table 6.25: Coefficients table for summer.

Model	**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std.Error**	**Beta**		
1 (Constant)	21.090	1.886		11.180	.000
P_{t-2}	-.111	.010	-.374	-11.376	.000
2 (Constant)	15.048	1.235		12.183	.000
P_{t-2}	-.794	.021	-2.671	-36.945	.000
P_{t-1}	.715	.021	2.404	33.249	.000
3 (Constant)	14.358	1.207		11.895	.000
P_{t-2}	-.768	.021	-2.585	-36.117	.000
P_{t-1}	.706	.021	2.376	33.686	.000
H_t	-.226	.034	-.151	-6.687	.000
4 (Constant)	16.988	1.436		11.829	.000
P_{t-2}	-.767	.021	-2.580	-36.270	.000
P_{t-1}	.6990	.021	2.351	33.363	.000
H_t	-.169	.038	-.112	-4.461	.000
T_{t-2}	-.145	.044	-.078	-3.330	.001
5 (Constant)	18.161	1.426		12.737	.000
P_{t-2}	-.728	.022	-2.449	-33.203	.000
P_{t-1}	-.649	.022	2.183	28.895	.000
H_t	-.071	.041	-.048	-1.740	.082
T_{t-2}	-.864	.137	-.463	-6.326	.000
T_t	.705	.127	.376	5.542	.000
6 (Constant)	21.542	1.650		13.058	.000
P_{t-2}	-.729	.022	-2.455	-33.589	.000
P_{t-1}	.641	.022	2.157	28.719	.000
H_t	-.044	.041	-.030	-1.075	.283
T_{t-2}	-.921	.136	-.493	-6.766	.000
T_t	-.749	.126	.400	5.924	.000
D_t	.427	.108	-.084	-3.969	.000
7 (Constant)	22.211	1.528		14.534	.000
P_{t-2}	-.729	.022	-2.455	-33.581	.000
P_{t-1}	.637	.022	2.143	28.990	.000
T_{t-2}	-1.000	.114	-.536	-8.746	.000
T_t	.808	.114	.431	7.094	.000
D_t	-.446	.106	-.088	-4.205	.000

Table 6.26: ANOVA table for summer.

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	12104.780	1	12104.780	129.414	.000
Residual	74641.427	798	93.536		
Total	86746.207	799			
2 Regression	55477.196	2	27738.598	707.015	.000
Residual	31269.011	797	39.233		
Total	86746.207	799			
3 Regression	57140.419	3	23051.149	357.514	.000
Residual	29605.788	796	64.476		
Total	86746.207	799			
4 Regression	57547.679	4	14386.920	391.718	.000
Residual	29198.528	795	36.728		
Total	86746.207	799			
5 Regression	58635.246	5	11727.049	331.233	.000
Residual	28110.961	794	35.404		
Total	86746.207	799			
6 Regression	59182.838	6	9863.806	283.782	.000
Residual	27563.369	793	34.758		
Total	86746.207	799			
7 Regression	59142.686	5	11828.537	340.241	.000
Residual	27603.521	794	34.765		
Total	86746.207	799			

According to table 6.25 we can see that all the parameters of the models are significant except the models 5 and 6. By considering table 6.26 it can be obtained that all the tests are also significant. Furthermore, we can obtain that R square values are close from model 2 to 7. Although the model 7 has higher R square value we select the best one with as few variables as possible. Since the model 7 has 5 predictor variables and model 4 has 4 predictor variables , model 4 can be taken as the suitable model to describe the summer season.

Table 6.27: Model summary table for summer.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.374	.140	.138	9.671382
2	.800	.640	.639	6.263656
3	.812	.659	.657	6.098623
4	.814	.663	.662	6.060339
5	.822	.676	.674	5.950146
6	.826	.682	.680	5.895621
7	.826	.682	.680	7.896197

The corresponding regression equation is as follows.

$$\delta P_t = 16.988 + (.691)P_{t-1} - (.767)P_{t-2} - (.145)T_{t-2} - (.350)H_t \tag{6.6}$$

The results for the testing is as shown in Table 6.28.

Table 6.28: Results for the prediction of power consumption in summer using MLR.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	σ **(MWh/h)**
P_{t-1}, P_{t-2}, H_t	3.350	19.581	2.900

For the autumn dataset the corresponding coefficient and ANOVA tables are depicted in Table 6.29 and 6.30 respectively.

Table 6.29: Coefficients table for summer.

Model		**Unstandardised Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
		B	**Std.Error**	**Beta**		
1	(Constant)	9.462	.883		10.711	.000
	H_t	-.817	.066	-.401	-12.359	.000
2	(Constant)	23.972	2.351		10.195	.000
	H_t	-.654	.069	-.321	-9.494	.000
	P_{t-2}	-.061	.009	-.224	-6.631	.000
3	(Constant)	17.585	1.653		10.639	.000
	H_t	-.402	.049	-.197	-8.242	.000
	P_{t-2}	.704	.023	-2.581	-30.582	.000
	P_{t-1}	.656	.023	2.402	29.081	.000
4	(Constant)	25.121	2.082		12.065	.000
	H_t	-.295	.051	-.145	-5.758	.000
	P_{t-2}	-.708	.023	-2.581	-30.582	.000
	P_{t-1}	.637	.022	2.333	28.491	.000
	T_t	-.400	.069	-.149	-5.764	.000
5	(Constant)	18.255	2.344		7.786	.000
	H_t	-.472	.058	-.232	-8.084	.000
	P_{t-2}	-.691	.022	-2.532	-30.980	.000
	P_{t-1}	.647	.022	2.370	29.470	.000
	T_t	-1.303	.167	-.487	-7.814	.000
	T_{t-2}	1.139	.192	.425	5.932	.000

Table 6.30: ANOVA table for autumn.

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	25700.814	1	25700.814	152.744	.000
Residual	134272.098	798	168.261		
Total	159972.912	799			
2 Regression	32721.742	2	16360.871	102.471	.000
Residual	127251.170	797	159.663		
Total	159972.912	799			
3 Regression	98273.988	3	32757.996	422.623	.000
Residual	61698.923	796	77.511		
Total	159972.912	799			
4 Regression	100748.704	4	25187.176	338.102	.000
Residual	59224.207	795	74.496		
Total	159972.912	799			
5 Regression	103261.645	5	20652.329	289.148	.000
Residual	56711.267	794	71.425		
Total 159972.912	799				

According to table 6.29 it can be seen that all the coefficient of the models are significant. By considering table 6.30 we can obtain that all the test are significant too. From Figure 6.31 we can observe that R square values are higher and close from model 3 to 5. Although the model 4 and 5 have higher R square values than model 3 we select the model with as few variables as possible. Since model 5 has 5 predictor variables and model 4 has 4 predictor variables , model 3 with 3 predictor variables can be taken as the suitable model to describe the autumn season.

Table 6.31: Model summary table for autumn

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.401	.161	.160	12.971537
2	.452	.205	.203	12.635771
3	.784	.614	.613	8.804045
4	.794	.630	.628	8.631098
5	.803	.645	.643	8.451318

The corresponding regression equation is as follows.

$$\delta P_t = 17.585 + (.656)P_{t-1} - (.704)P_{t-2} - (.402)H_t \tag{6.7}$$

The results for the testing is as shown in Table 6.32.

Table 6.32: Results for the prediction of power consumption in autumn using MLR.

Signature	**MAE (MWh/h)**	**MSE (MWh/h)**	σ **(MWh/h)**
P_{t-1}, P_{t-2}, H_t	6.061	84.424	6.926

Chapter 7

Discussion and Summary of Results

In this chapter, we will discuss about the results we found earlier and answer the research questions outlined in chapter 1. Section 7.1 discusses the first research question about the cyclic patterns of electricity consumption. Then, in section 7.2 and section 7.3, we answer research questions 2, 3 and 4 and finally the results of GP will be compared with MBPNN and MLR in section 7.4 to answer the last research question.

7.1 Cyclic Patterns

Electricity consumption of three successive years were evaluated to check for cyclic patterns in chapter 5. We found out that the patterns could be mainly divided into three types as *annual*, *seasonal* and *monthly*.

On annual context, all the three years showed a similar mean hourly electricity consumption trend, although there were slight deviations in the exact values. Both 2009 and 2010 showed similar cyclic patterns in terms of hourly electricity consumption.

On seasonal context, we considered the electricity consumption of the four seasons : winter, spring, summer and autumn in the three years. We could see exactly the same cyclic pattern of 2008 is repeated in 2009 and 2010 for all four seasons.

On weekly context, we considered weekdays and weekends in our analysis, where we could observe that Mondays, Tuesdays, Wednesdays and Thursdays illustrated an identical pattern while Fridays showed slightly deviated pattern. Saturdays and Sundays had almost similar patterns too.

Table 7.1: Best forecasting results of the six categories for the dynamic dataset.

No	Method	Signature	Target variable	MAE (MWh/h)	MAPE(%)
1	Original	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.472	0.508
2	Original	P_{t-1}, H_t, D_t	P_t	3.926	0.838
3	Normalized	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.621	0.534
4	Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	P_t	2.768	0.578
5	Kernel	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.473	0.508
6	Kernel	P_{t-1}, H_t, D_t	P_t	3.531	0.752

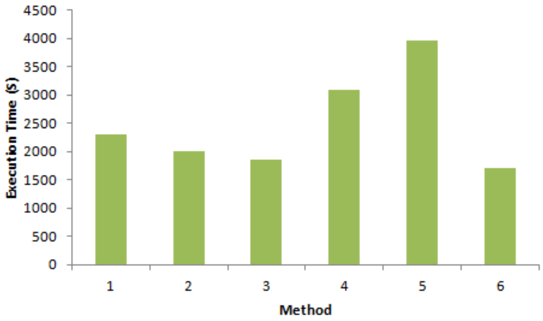


Figure 7.1: Execution time for each best prediction in the six categories.

7.2 Dynamic Training Dataset

Testing for this dataset was done using three types of feature spaces : *Original*, *Kernel Reduction* and *Normalized*, for two target variables δP_t and P_t. Altogether there were six different combinations of testing for this dataset.

Table 7.1 illustrates the best forecasts of each of these six combinations for different feature vectors. The first method which uses the original data as the feature space with the signature $P_{t-1}, T_t, T_{t-1}, H_t, D_t$, and δP_t as the target variable produces the best result, whereas method 2 which uses original dataset, but P_{t-1}, H_t, D_t as the signature and P_t as the target variable gives the lowest MAE and MAPE values of all the best results.

The second best forecast is given by the kernel reduction method with signature $P_{t-1}, T_t, T_{t-1}, H_t, D_t$ and δP_t as the target variable which has a MAE of 2.473 MWh/h and MAPE of 0.508%.

Furthermore, we can observe that δP_t as the target variable which gives better results than P_t as the target variable.

However, the effectiveness of these results could be evaluated by considering the execution time of each result in Table 7.1, as indicated in Figure 7.1. Although method 5 has an MAE of 2.473 MWh/h, it tends to be more time consuming as it has the highest execution time of

Table 7.2: Top 10 forecasting results for dynamic dataset.

Method	Signature	Target variable	MAE(MWh/h)	MAPE(%)
Original	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.472	0.508
Kernel	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.473	0.508
Normalized	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	δP_t	2.621	0.534
Normalized	P_{t-1}, T_t, H_t, D_t	δP_t	2.622	0.534
Normalized	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	δP_t	2.623	0.534
Original	P_{t-1}, T_t, H_t, D_t	δP_t	2.629	0.508
Normalized	$P_{t-1}, T_{t-1}, H_t, D_t$	δP_t	2.635	0.537
Normalized	$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	δP_t	2.636	0.537
Original	$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	δP_t	2.652	0.541
Original	$P_{t-1}, T_{t-1}, H_t, D_t$	δP_t	2.678	0.546

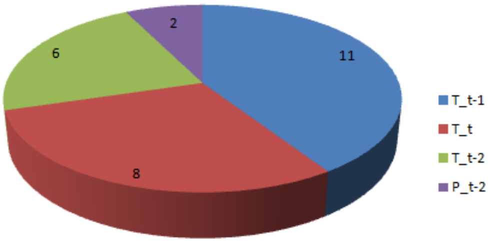


Figure 7.2: Usage of the variables P_{t-2}, T_t, T_{t-1} and T_{t-2} for the best results with an MAE less than 3.0.

3966.4 seconds. The reason for this is that kernel reduction method trains the dataset twice; one for the selection of the best kernel values and the other to generate the predictions. The $O(n^3)$ complexity is doubled due to these two training processes. Therefore, we can consider method 1 in Table 7.1 as the best combination of all six results, in terms of both forecasting accuracy and execution time in dynamic datasets.

Moreover, we have list the top 10 forecasts generated for the dynamic dataset in Table 7.2. All the results in the table can be considered as good, because they all have a MAPE of 0.5%. Moreover, note that all of them have δP_t as their target variable indicating how good the prediction of power difference than the power consumption itself. In addition to that, 50% of the results belong to the *Normalized* category, 40% to *Original* and 10% to *Kernel reduction.*

The contributions of the independent variables P_{t-2}, T_t, T_{t-1} and T_{t-2} for the best predictions which have a MAE of less than 3.0 are visualized in Figure 7.2. Figure highlights T_{t-1} as the best factor which has been repeatedly used in 11 times out of 15 best predictions and P_{t-2} as the worst factor which has been involved in only 2 times out of 15.

Table 7.3: Top 10 forecasting results for winter.

Method	Signature	MAE (MWh/h)
Normalized	$P_{t-1}, P_{t-2}T_t, T_{t-1}, H_t, D_t$	2.950
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.952
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.996
Normalized	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.010
Original	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.010
Normalized	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	3.015
Original	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	3.015
Normalized	$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	3.027
Original	$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	3.027
Normalized	$P_{t-1}, T_{t-1}, H_t, D_t$	3.031

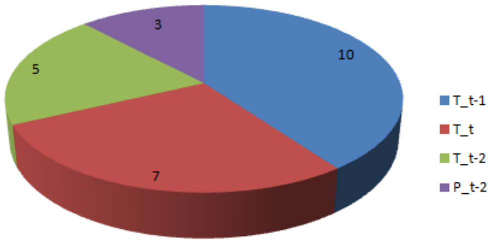


Figure 7.3: Usage of the variables P_{t-2}, T_t, T_{t-1} and T_{t-2} in the top 10 predictions of winter season.

7.3 Fixed Training Dataset

In this section, we analyse the results for the four seasons which was done using fixed training and test datasets. Here, the six combinations in section 7.2 was reduced to three, as the tests for the target variable P_t was removed from the tests due to its inaccuracy compared to δP_t. The results for each season are summarized in the following.

The top 10 forecasting results for winter season are illustrated in Table 7.3. We can observe that *Normalized* method dominates the winter season in terms of forecasting. The best combination is $P_{t-1}, P_{t-2}T_t, T_{t-1}, H_t, D_t$, which has 2.950 MWh/h as the MAE. Most importantly, we cannot find the *Kernel reduction* method in the table indicating its inappropriateness of forecasting the consumption in winter season. Therefore, we can identify that normalizing the feature vector allows for better results in forecasting for the winter season.

Figure 7.3 illustrates the frequency of the four variables : P_{t-2}, T_t, T_{t-1} and T_{t-2}, used in the best forecasting combinations of winter. We can observe that T_{t-1} is involved in all 10 predictions, and T_t is used in 7 occasions. This indicates that temperature of last hour and forecast temperature of current hour affects the consumption of current hour very much.

Table 7.4: Top 10 forecasting results for spring.

Method	Signature	MAE(MWh/h)
Original	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.991
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.999
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	3.017
Original	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	3.018
Normalized	$P_{t-1}, P_{t-2}, H_t, D_t$	3.028
Original	$P_{t-1}, P_{t-2}, H_t, D_t$	3.028
Normalized	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	3.052
Original	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	3.118
Normalized	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.138
Original	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	3.138

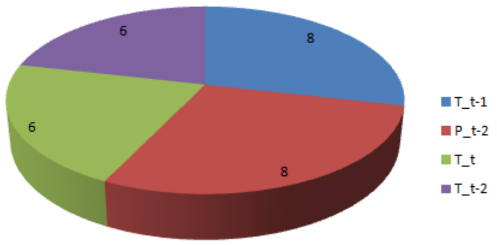


Figure 7.4: Usage of the variables P_{t-2}, T_t, T_{t-1} and T_{t-2} in the top 10 predictions of spring season.

For the spring season, both the *Original* and *Normalized* methods equally affect the forecasts. We can notice from Table 7.4, that there are pairs from both methods which give the same results for the same combination of variables. However, a slightly better result is produced by *Original* method for the combination $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$.

When we compare the results of winter and spring, we can conclude that the results of winter is better than spring.

The variable contributions illustrated in Figure 7.4 indicate that both T_{t-1} and P_{t-2} are equally important variables for spring predictions. This is the first time we have seen P_{t-2} affecting significantly for forecasts.

Similar to the winter season, *Normalized* method dominates in summer, as we can see from Table 7.5, where 80% of the top 10 results associated with the normalized method, and the best combination is given by $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$, which has MAE of 1.848 MWh/h.

Results for summer are better than both winter and spring. All top 10 results in summer have MAE of less than 2.0 MWh/h.

Table 7.5: Top 10 forecasting results for summer.

Method	Signature	MAE(MWh/h)
Normalized	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	1.848
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	1.852
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	1.862
Original	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	1.862
Normalized	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	1.866
Normalized	P_{t-1}, T_t, H_t, D_t	1.866
Original	P_{t-1}, T_t, H_t, D_t	1.866
Normalized	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	1.893
Normalized	$P_{t-1}, T_{t-1}, T_{t-2}, H_t, D_t$	1.893
Normalized	$P_{t-1}, T_{t-1}, H_t, D_t$	1.900

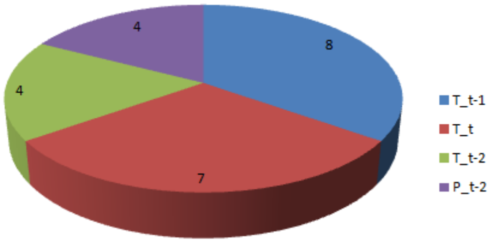


Figure 7.5: Usage of the variables P_{t-2}, T_t, T_{t-1} and T_{t-2} in the top 10 predictions of summer season.

Similarly to the winter season, the effect of temperature can be seen for summer also. According to Figure 7.5 T_{t-1} and T_t have been used most frequently among the top 10 combinations.

Table 7.6: Top 10 forecasting results for autumn.

Method	Signature	MAE(MWh/h)
Original	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.339
Normalized	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	2.347
Original	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	2.347
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.404
Normalized	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.451
Normalized	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.505
Normalized	$P_{t-1}, P_{t-2}, H_t, D_t$	2.511
Original	$P_{t-1}, P_{t-2}, H_t, D_t$	2.511
Kernel	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.526
Normalized	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$	2.562

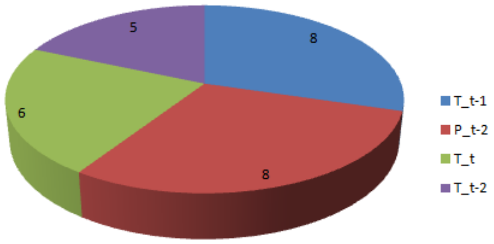


Figure 7.6: Usage of the variables P_{t-2}, T_t, T_{t-1} and T_{t-2} in the top 10 predictions of autumn season.

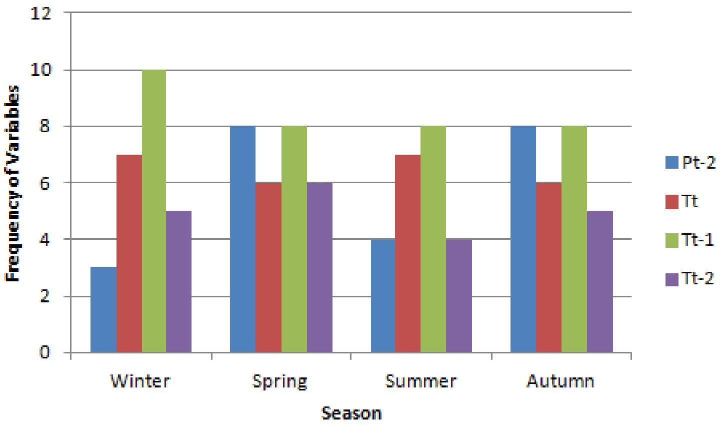


Figure 7.7: Usage of variables $P_{t-2}T_t, T_{t-1}$ and T_{t-2} in top 10 predictions for all the seasons.

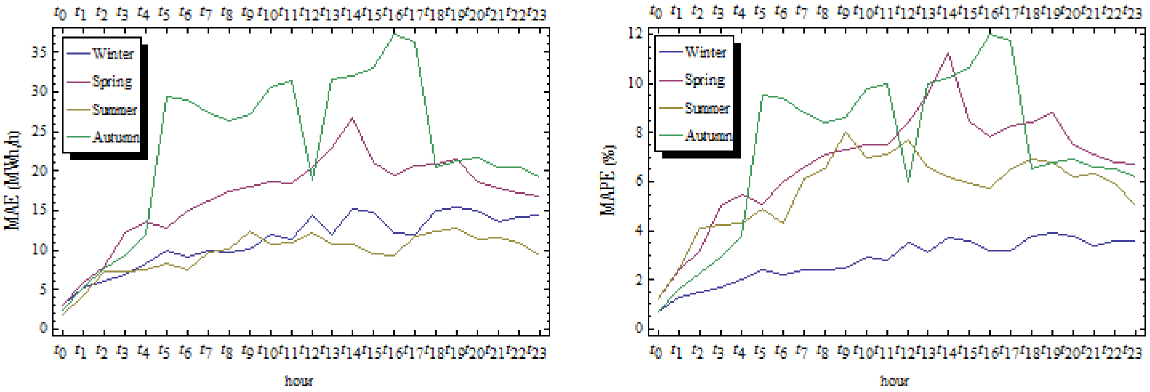
The top 10 forecasting results in autumn are depicted in Table 7.6. Autumn results are little bit similar to spring results in terms of the method. Both *Original* and *Normalized* methods involved in producing best results. The best combination is the same combination given for winter : $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$. But in this case the method is *Original*.

If we consider the variable usage, it is quite similar to spring which has equal contributions from T_{t-1} and P_{t-2}. Therefore, we can see that *previous hour temperature* and *electricity consumption before two hours* are affecting the electricity consumption for these two seasons. The corresponding graph of the usage of variables in autumn is shown in Figure 7.6.

Figure 7.7 summarizes the usage of variables in all four seasons. We can see winter and summer have almost similar variable usage, and on the other hand spring and autumn have similar usage too. T_{t-1} has been the dominant feature variable among the four variables for all seasons.

7.3.1 Next 24 Hours Prediction

For the next 24 hour predictions, lets summarize Table 6.11 to 6.14 in chapter 6, and Figure 6.7 is also repeated here for clarity of the explanation.



(a)

(b)

Figure 7.8: The (a) MAE and (b) MAPE curves for the next 24 hours predictions for the best combinations in the four seasons. $P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ for winter, $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ for spring, $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$ for Summer and $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$ for autumn.

For winter season, the MAPE does not exceed 4.0%, which means the 24 hours predictions are excellent for winter for the best combination $P_{t-1}, P_{t-2}T_t, T_{t-1}, H_t, D_t$. The first 10 predictions are very good compared to the remaining 14 hours and there is an upward trend for winter.

Although it seems that the results for summer is better than winter in Figure 7.8(a), the graph of MAPE values given in Figure 7.8(b) indicate that winter is better than summer in terms of next 24 hour predictions.

Both spring and autumn shows similar behaviours in forecasting the next 24 hour consumption of electricity using their best signatures $P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$ and $P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$ respectively.

7.4 Comparison between GP and Traditional Approaches

In this section, we will compare the results of GP for the four seasons with MBPNN and MLR methods.

Table 7.7 illustrates the comparison of the results of the three methods for the winter season. From the table, we can see that MBPNN (MAE of 2.378 MWh/h) produces better results than

GP (MAE of 2.950 MWh/h). However the results of MLR is not even closer to the other two methods.

Table 7.7: Comparison of the forecasting results for winter.

	Signature	**MAE(MWh/h)**
GP	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.950
MBPNN	$P_{t-1}, T_{t-1}, H_t, D_t$	2.378
MLR	$P_{t-1}, P_{t-2}, T_{t-1}, H_t$	6.147

For spring, we can see our method of GP gives the best results of 2.991 MWh/h for MAE, from Table 7.8. MBPNN gives a result of 3.584 MWh/h and we can see that MLR (3.350 MWh/h) is better than MBPNN in this case.

Table 7.8: Comparison of the forecasting results for spring.

	Signature	**MAE(MWh/h)**
GP	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.991
MBPNN	$P_{t-1}, P_{t-2}, H_t, D_t$	3.584
MLR	P_{t-1}, P_{t-2}, H_t	3.350

In summer, we can observe MBPNN again gives the best result of 1.820 MWh/h, and GP has a closer value of 1.848 MWh/h. This is illustrated in Table 7.9. The result of MLR is not better than the other two methods.

Table 7.9: Comparison of the forecasting results for summer.

	Signature	**MAE(MWh/h)**
GP	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	1.848
MBPNN	$P_{t-1}, P_{t-2}, T_{t-1}, T_{t-2}, H_t, D_t$	1.820
MLR	P_{t-1}, P_{t-2}, H_t	5.934

GP gives the best result for the forecast of autumn, which is 2.339 MWh/h and MBPNN has 2.750 MWh/h. MLR is not giving good results for this case also.

Table 7.10: Comparison of the forecasting results for autumn.

	Signature	**MAE(MWh/h)**
GP	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.339
MBPNN	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$	2.750
MLR	P_{t-1}, P_{t-2}, H_t	6.061

From the above comparison, we can see that while GP is producing best results for spring and autumn , MBPNN have produced best forecasts for winter and summer. This indicates that GP is as better approach as MBPNN in forecasting electricity consumption.

Chapter 8

Conclusion and Further Work

8.1 Conclusion

In this thesis, we have examined how GP can be used to forecast electricity consumption on an hourly basis, by employing different feature vectors, against two target variables, under three different methods : original, normalized and kernel reduction. Based on the results in the preceding chapters we come to the following conclusions.

When forecasting the next hour consumption and next 24 hour consumptions, we can see that, previous hour temperature has a major effect on prediction, in addition to the previous hour consumption. This is true for both dynamic and fixed training datasets. The effect is very high for winter and summer seasons, indicating their relationship with the temperature.

Among other factors, consumption of the hour before last hour P_{t-2}, influences the consumption differently on different seasons. The effect is very low for the dynamic dataset, but high effect on spring and autumn predictions in the fixed dataset. However, it is not good for winter and summer.

When considering the best feature vectors for predictions, we can come to the conclusion illustrated in Table 8.1. In addition to the best feature vectors, we can come to another conclusion by examining these best forecasts. The target variable used in all these cases, is the power difference δP_t. Therefore we can conclude that δP_t is better than using P_t as the target variable in electricity consumption forecasting.

If we consider the three feature space modification techniques, normalizing the feature variables produce better results for winter and summer than both original and kernel reduction. Original and normalizing methods equally contribute for best results in spring and autumn. However majority of the best results are given by normalizing. Therefore we can conclude that both methods are equally good in forecasting electricity consumption.

However we can conclude that kernel reduction method is not a suitable reduction method in

Table 8.1: Best feature vectors.

Dataset/Season	Feature vector
Dynamic dataset	$P_{t-1}, T_t, T_{t-1}, H_t, D_t$
Winter	$P_{t-1}, P_{t-2}T_t, T_{t-1}, H_t, D_t$
Spring	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$
Summer	$P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$
Autumn	$P_{t-1}, P_{t-2}, T_t, T_{t-1}, H_t, D_t$

this context. Not only it produces bad forecasting results, but also using much of the computer resources due to its increasing complexity.

For the best feature vectors in Table 8.1,the next 24 hour prediction is best in winter than the other 3 seasons. Altogether GP produces sufficiently good results for the next 24 hour predictions.

Finally if we consider the results between GP and the two traditional methods : MBPNN and MLR, we can conclude that Gaussian processes is as better as Multiple Back-Propagation Neural networks in terms of short-term electricity forecasting, and it is far better than Multiple linear regression.

8.2 Further Work

In the later stages of the thesis, we could find some interesting results which could lead to future extensions of this work.

Table 8.2: Effect of previous consumption differences in the feature vector of the original dataset.

Method	**Signature**	**MAE(MWh/h)**
Original	$\delta P_t, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.501
Original	$\delta P_t, \delta P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.186
Original	$\delta P_t, \delta P_{t-1}, \delta P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.143
Original	$\delta P_t, \delta P_{t-1}, \delta P_{t-2}, \delta P_{t-3}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.121
Normalized	$\delta P_t, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.540
Normalized	$\delta P_t, \delta P_{t-1}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.194
Normalized	$\delta P_t, \delta P_{t-1}, \delta P_{t-2}, T_t, T_{t-1}, T_{t-2}, H_t, D_t$	2.090

If we use previous electricity consumption differences such as $\delta P_t, \delta P_{t-1}, \delta P_{t-2}$ and so on, for the dynamic dataset, we could see improvements of the results in both original and normalized datasets. Some of those results are given below in Table 8.2:

If you observe closely we can see a pattern, that the error is reducing once we increase the number of power differences variables in the feature vector. This is an exciting feature which could be extended more in the future extension of this work.

Moreover, it could be developed a new kernel function that best describes the actual electricity consumption.

In addition to that, more feature space reduction and scaling methods could be tested to see how GP behaves in forecasting future electricity consumption.

Finally, we can compare GP with more traditional techniques such as genetic algorithms, grey methods and time series analysis.

Bibliography

[1] "Focus on Energy : Statistics Norway." [Online]. Available: http://www.ssb.no/english/subjects/01/03/10/energi_en/fig03-elektrisitet1990-2007-en.gif

[2] N. Lopes and B. Ribeiro, "An efficient gradient-based learning algorithm applied to neural networks with selective actuation neurons," *Neural, Parallel & Scientific Computations*, vol. 11, no. 3, pp. 253–272, 2003.

[3] M. Hayati and Y. Shirvany, "Atrificial neural network approach for short term load forecasting for illam region," *World Academy of Science, Engineering and Technology*, vol. 28, pp. 280–284, 2007.

[4] M. Alamaniotis, A. Ikonomopoulos, and L. Tsoukalas, "A pareto optimization approach of a gaussian process ensemble for short-term load forecasting," in *Intelligent System Application to Power Systems (ISAP), 2011 16th International Conference on*, sept. 2011, pp. 1 –6.

[5] Z. Fang, S. Liu, C. Yuan, and C. Mi, "Forecast of electricity consumption of jiangsu province by grey methods," in *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, june 2008, pp. 1759 –1762.

[6] A. Imtiaz, N. Mariun, M. Amran, M. Saleem, N. Wahab, and Mohibullah, "Evaluation and forecasting of long term electricity consumption demand for malaysia by statistical analysis," in *Power and Energy Conference, 2006. PECon '06. IEEE International*, nov. 2006, pp. 257 –261.

[7] W. Baosen, H. Dawei, C. Yi, and Z. Yizhe, "Research on the forecast of electricity consumption based on autoregressive model," in *Challenges in Environmental Science and Computer Engineering (CESCE), 2010 International Conference on*, vol. 2, march 2010, pp. 166 –169.

[8] Y. Fung and V. Rao Tummala, "Forecasting of electricity consumption: a comparative analysis of regression and artificial neural network models," in *Advances in Power System Control, Operation and Management, 1993. APSCOM-93., 2nd International Conference on*, dec 1993, pp. 782 –787 vol.2.

[9] D. Purwanto, H. Agustiawan, and M. Romlie, "The taguchi-neural networks approach to forecast electricity consumption," in *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*, may 2008, pp. 001 941 –001 944.

[10] Z. Qing-wei, X. Zhi-Hai, and W. Jian, "Prediction of electricity consumption based on genetic algorithm - rbf neural network," in *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, vol. 5, march 2010, pp. 339 –342.

[11] H. Mori and M. Ohmi, "Probabilistic short-term load forecasting with gaussian processes," in *Intelligent Systems Application to Power Systems, 2005. Proceedings of the 13th International Conference on*, nov. 2005, p. 6 pp.

[12] "Introduction to Taguchi Method." [Online]. Available: http://www.ee.iitb.ac.in/~apte/CV_PRA_TAGUCHI_INTRO.htm

[13] L. Sifeng and L. Yi, "An introduction to grey systems: Foundations, methodology and applications," *Kybernetes*, vol. 32, no. 24, pp. 178–190, 1998.

[14] N. Luo and F. Qian, "On line estimation of color values (b*) in pet process using gaussian process regression," in *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, july 2010, pp. 5842 –5845.

[15] H. Vathsangam, A. Emken, D. Spruijt-Metz, and G. Sukhatme, "Toward free-living walking speed estimation using gaussian process-based regression with on-body accelerometers and gyroscopes," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on*, march 2010, pp. 1 –8.

[16] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data," *Geoscience and Remote Sensing Letters, IEEE*, vol. 7, no. 3, pp. 464 –468, july 2010.

[17] Y. Bazi and F. Melgani, "Semisupervised gaussian process regression for biophysical parameter estimation," in *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, july 2010, pp. 4248 –4251.

[18] Y. Zhang and D.-Y. Yeung, "Multi-task warped gaussian process for personalized age estimation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 2622 –2629.

[19] K. Das and A. Srivastava, "Block-gp: Scalable gaussian process regression for multimodal data," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, dec. 2010, pp. 791 –796.

[20] "Supplying Europe with renewable energy: A brief introduction to the Norwegian energy group Agder Energy." [Online]. Available: www.ae.no/ae/multimedia/archive/00025/Supplying_Europe_wit_25082a.pdf

[21] M. Samarasinghe and W. Al-Hawani, “Gaussian processes for multivariate prediction of power consumption,” Faculty of Engineering and Science, University of Agder, IKT 508 Specialization Project Report.

[22] R. D’hulst, K. Verhaegen, T. Blanco, J. Suykens, J. Driesen, J. Vandewalle, and R. Belmans, “Project-oriented approach to undergraduate teaching in green power production and prediction of electricity consumption,” in *Power Engineering Society General Meeting, 2005. IEEE*, june 2005, pp. 157 – 163 Vol. 1.

[23] Z. Qing-wei, X. Zhi-Hai, and W. Jian, “Prediction of electricity consumption based on genetic algorithm - rbf neural network,” in *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, vol. 5, march 2010, pp. 339 –342.

[24] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York, USA: McGraw-Hill, 1991.

[25] H. Asheri, H. Rabiee, N. Pourdamghani, and M. Rohban, “A gaussian process regression framework for spatial error concealment with adaptive kernels,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, aug. 2010, pp. 4541 –4544.

[26] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning.* Cambridge, Massachusetts, London, England: MIT Press, 2006.

[27] H. Zhou and D. Suter, “Fast sparse gaussian processes learning for man-made structure classification,” in *Computer Vision and Pattern Recognition, 2007. CVPR ’07. IEEE Conference on*, june 2007, pp. 1 –6.

[28] M. Ebden, “Gaussian processes for regression: A quick introduction,” Robotics Research Group,Department of Engineering Science, University of Oxford, Tech. Rep.

[29] D. W. Patterson, *Artificial Neural Networks: Theory and Applications.* Singapore: Prentice Hall, 1996.

[30] M. Smith, *Neural Networks for Statistical Modelling.* New Your, USA: Van Nostrand Reinhold, 1993.

[31] N. Lopes and B. Ribeiro, “Hybrid learning in a multi-neural network architecture,” in *INNS-IEEE International Joint Conference on Neural Networks (IJCNN 2001)*, vol. 4, 2001, pp. 2788–2793.

[32] ——, “GPU implementation of the multiple back-propagation algorithm,” in *Intelligent Data Engineering and Automated Learning (IDEAL 2009)*, ser. Lecture Notes in Computer Science, E. Corchado and H. Yin, Eds., vol. 5788. Springer Berlin / Heidelberg, 2009, pp. 449–456.

[33] D. George and P. Mallery, *SPSS for Windows Step by Step: A Simple Guide and Reference 16.0 Update.* USA: Pearson, 2009.

[34] R. H. Carver and J. G. Nash, *Doing Data Analysis with SPSS Version 16.* USA: Brooks/Cole, 2009.

[35] "pyXGPR," October 2011. [Online]. Available: http://www-kd.iai.uni-bonn.de/index.php?page=software_details&id=19

[36] X. Liu, B. Ang, and T. Goh, "Forecasting of electricity consumption: a comparison between an econometric model and a neural network model," in *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, nov 1991, pp. 1254 –1259 vol.2.

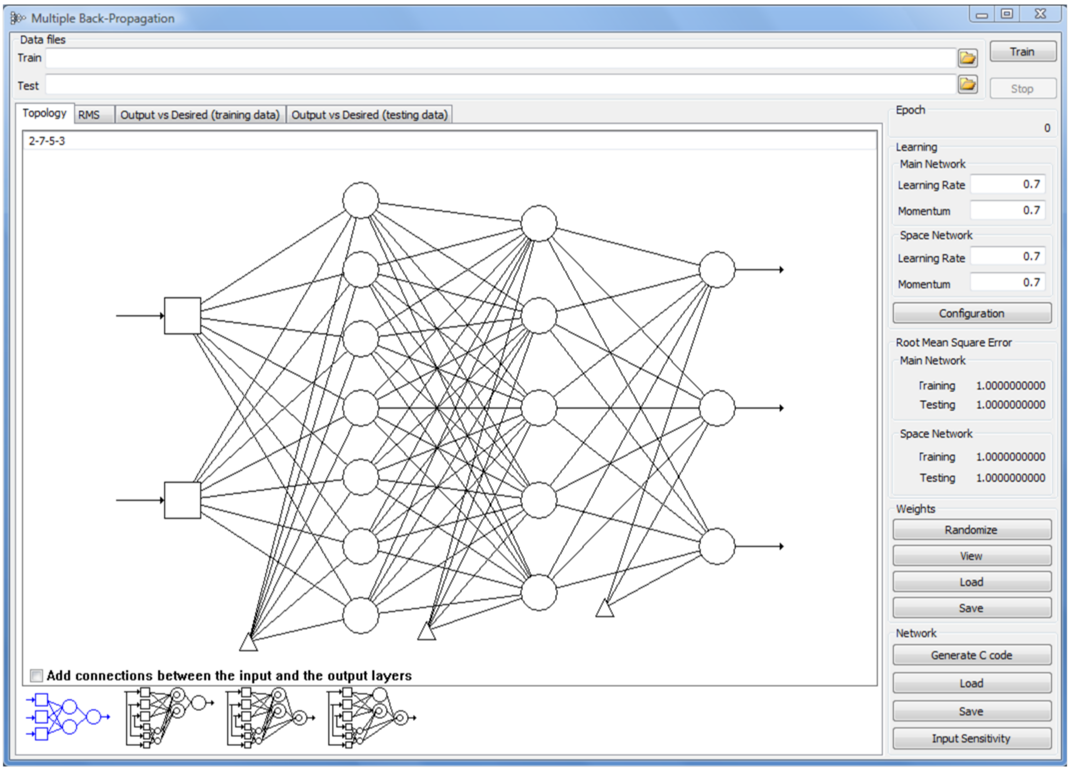
Glossary and Abbreviation

$\delta(x, x')$	Kronecker delta function
$\mathbb{E}$	Expectation
μ	mean
σ	Standard deviation
σ_f^2	variance of the (noise free) signal
σ_n^2	noise variance
θ	vector of hyperparameters
$f(x)$	Gaussian process latent function
$k(x, x')$	covariance function evaluated at x and x'
l	characteristic length-scale
$m(x)$	Gaussian process mean function
ANN	Artificial Neural Network
ANOVA	ANalysis Of VAriance
BPNN	Back-Propagation Neural Network
CDF	Cumulative Distribution Function
FF	Feed Forward
GA	Genetic Algorithm
GDP	Gross Domestic Product
GM	Grey Method
GPR	Gaussian Process Regression

GP	Gaussian Process
ICT	Information and Communication Technology
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MAP	Maximum a Posteriori
MBPNN	Multiple Back-Propagation Neural Network
MBP	Multiple Back-Propagation
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MTWGP	Multi-Task Warped Gaussian Process
NASA	National Aeronautics and Space Administration (USA)
PDF	Probability Density Function
PET	Poly Ethylene Terephthalate
RBFN	Radial Basis Function Network
RBF	Radial Basis Function
RV	Random Variable
SVM	Support Vector Machines
SVR	Support Vector Regression

Appendix A

Screen-shots of MBP Software



2-7-5-3

☐ Add connections between the input and the output layers

Epoch 0

Learning

Main Network

Learning Rate 0.7

Momentum 0.7

Space Network

Learning Rate 0.7

Momentum 0.7

Root Mean Square Error

Main Network

Training 1.0000000000

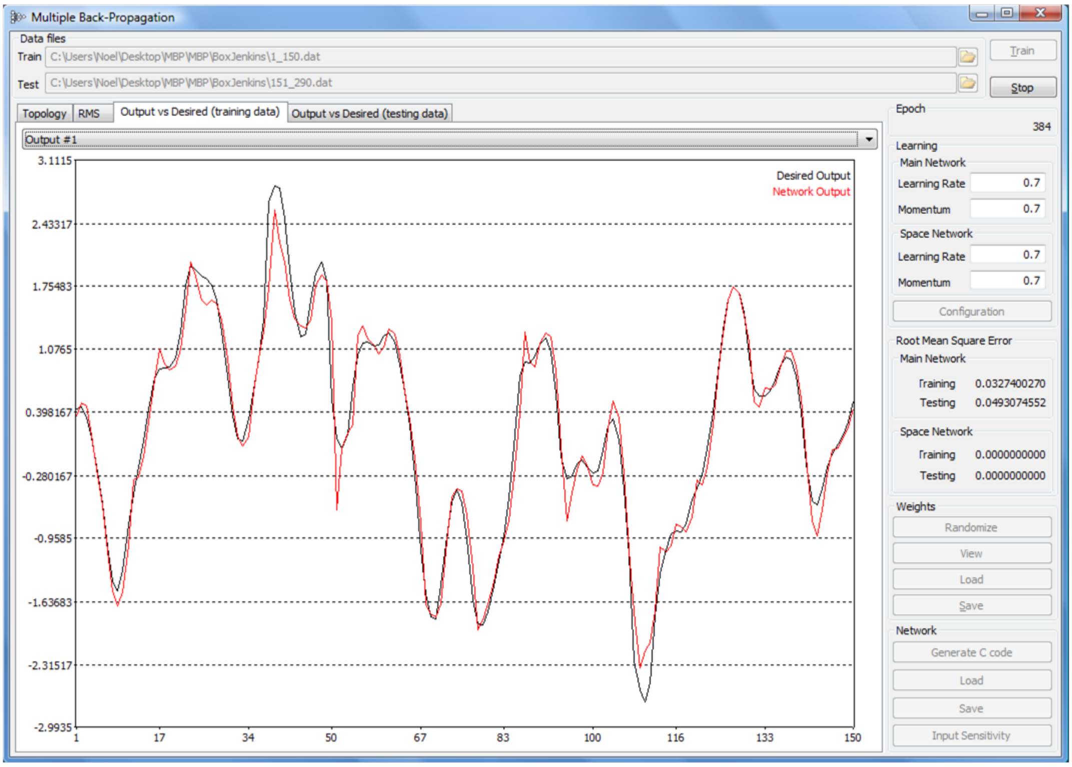
Testing 1.0000000000

Space Network

Training 1.0000000000

Testing 1.0000000000

Figure A.1: The main screen of defining the neural network to be tested



Train C:\Users\Noel\Desktop\MBP\MBP\BoxJenkins\1_150.dat

Test C:\Users\Noel\Desktop\MBP\MBP\BoxJenkins\151_290.dat

Epoch 384

Learning

Main Network

Learning Rate 0.7

Momentum 0.7

Space Network

Learning Rate 0.7

Momentum 0.7

Root Mean Square Error

Main Network

Training 0.0327400270

Testing 0.0493074552

Space Network

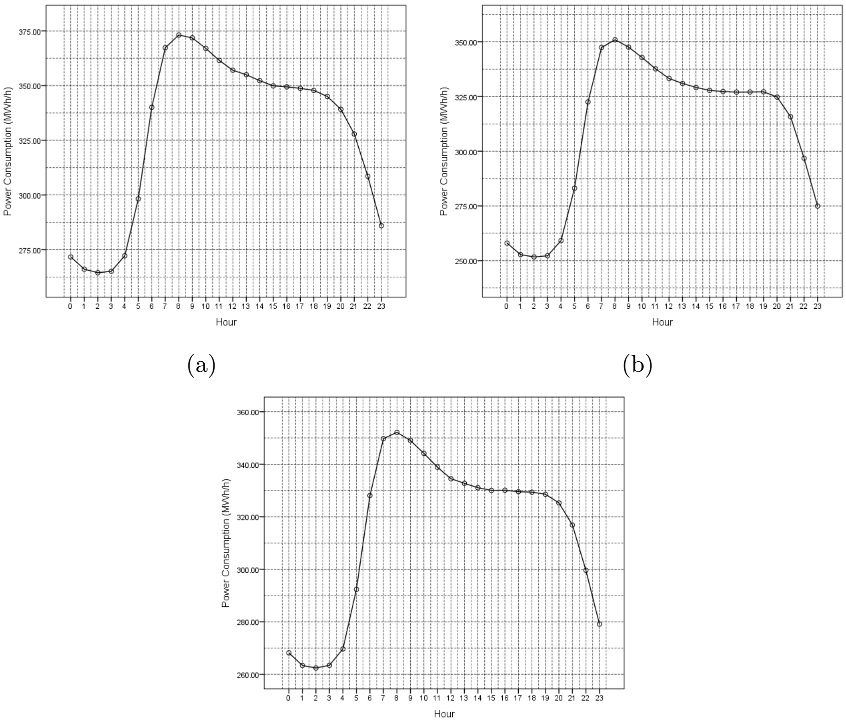
Training 0.0000000000

Testing 0.0000000000

Figure A.2: Output of the training dataset

Appendix B

Graphs of Mean Hourly Consumption

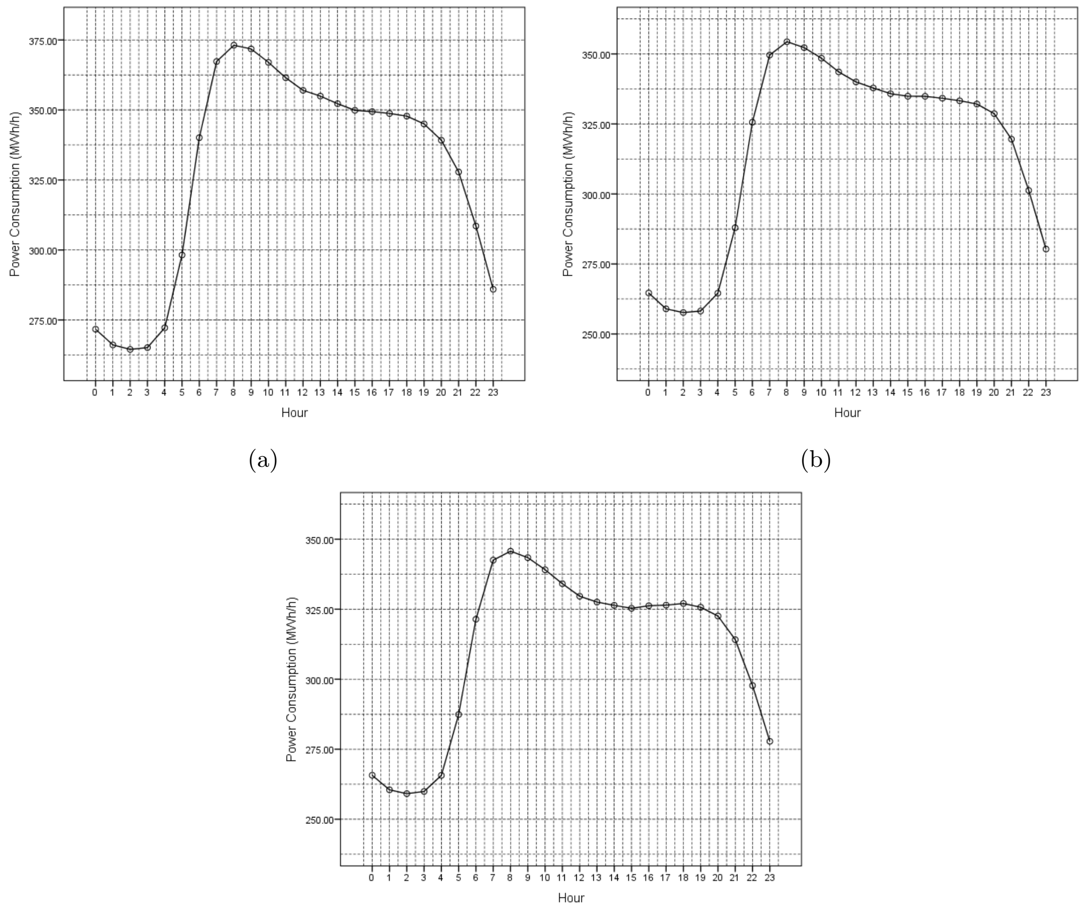


(a)

(b)

(c)

Figure B.1: The mean hourly electricity consumption on Wednesdays in year (a) 2008 (b) 2009 and (c) 2010



(a)

(b)

(c)

Figure B.2: The mean hourly electricity consumption on Thursdays in year (a) 2008 (b) 2009 and (c) 2010