ELECTRICITY DEMAND PREDICTION OF LARGE COMMERCIAL BUILDINGS USING SUPPORT VECTOR MACHINE

N.G. Indika Sujeewa Samarawickrama

(108863K)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science

Department of Electrical Engineering

University of Moratuwa
Sri Lanka

July 2014
DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor: Date
ACKNOWLEDGEMENTS

First of all, I would like to express my deepest sense of gratitude to my supervisors, Dr. K.T.M.U. Hemapala and Dr. B. Jayasekara, for their support, guidance and valuable advice throughout these academic years. I would like to thank University of Moratuwa for giving me the opportunity to study in the institute. My appreciation and thanks to the following people who have been very helpful and for their great contribution in the accomplishment of the study:

- All the building owners and managers for their help in the data collection.
- The members of Metrological Department, for sharing their knowledge and experiences and providing all weather data to the research.

Finally, thanks to all lecturer & my friends whom I have been working with throughout the period of study in University of Moratuwa.
ABSTRACT

In the framework of a competitive commercial world, having accurate energy forecasting tools becomes a Key Performance Indicator (KPI) to the building owners. Energy forecasting plays a crucial role for any building when it undergoes the retrofitting works in order to maximize the benefits and utilities. This paper provides accurate energy forecasting based on Support Vector Machine Regression (SVMR). Results and discussions from real-world case studies of commercial markets of Colombo, Sri Lanka are presented.

In the case study, four commercial buildings are randomly selected and the models are tried and tested using monthly landlord utility bills. Careful analysis of available data reveals the most influential parameters to the model and these are as follows: Mean outdoor dry-bulb temperature (T), Solar radiation (SR) and Relative humidity (RH). Selection of the kernel with radial-basis function (RBF) is based on stepwise searching method to investigate the performance of SVM with respect to the three parameters such as C, γ and ε.

The results showed that the structure of the training set has significant effect to the accuracy of the prediction. The analysis of the experimental results reveal that all the forecasting models give an acceptable result for all four commercials buildings with low coefficient of variance with a low percentage error (% error).
# TABLE OF CONTENTS

Declaration ........................................... i  
Acknowledgement .................................... ii  
Abstract .............................................. iii  
Table of Contents .................................... iv  
List of Figures ....................................... viii  
List of Tables ........................................ ix  
List of Abbreviations ................................. x  
List of Appendices ................................... xii  

## 1. INTRODUCTION

1.1 Background ........................................ 1  
1.2 Definition of Baseline Model ..................... 4  
1.3 Research Objectives ................................ 5  
1.4 Scope and Limitations ............................. 5  
1.5 Organization of Thesis ............................ 6  

## 2. LITERATURE REVIEW

2.1 Introduction ....................................... 7  
2.2 Classification of Baseline Models .............. 7  

    2.2.1 Regression-based models .................. 7  
        2.2.1.1 Variable-based degree-day models .... 8  
        2.2.1.2 Linear regression models .......... 9  
            2.2.1.2.1 Single-variant linear regression model ... 10  
            2.2.1.2.2 Multivariate linear regression model ... 10  
            2.2.1.2.3 Integrated model ............. 11  
        2.2.1.3 Change-point models ............. 11  

2.2.2 Calibrated simulation ........................ 12  
2.2.3 Artificial neural networks .................. 13  
2.2.4 Fourier series ................................ 14
2.2.5 Bin method
2.2.6 Support vector machines
2.2.6.1 Characters of SVMs for regression estimation
2.2.6.2 Kernel selection
2.2.6.3 Modification of parameters
2.2.6.3.1 Selection parameter C
2.2.6.3.2 Selection parameter $\varepsilon$
2.2.6.3.3 Stepwise search
2.3 Discussion

3. METHODOLOGY FOR EVALUATION AND PRE-PREPARATION TECHNIQUES

3.1 Introduction
3.2 Errors and Bias Inside the Model
3.3 Normalize Energy Use
3.4 Percentage Changes Based on Annul Energy Use
3.5 Possibilities of Utility Bill Reading Dates
3.6 Primary Modeling of Whole Building Energy Consumption
3.7 Selection of Weather Variables
3.7.1 Data collection
3.7.2 Discussion

4. BASLINE MODELS OF BUILDING LANDLORD ENERGY CONSUMPTION

4.1 Introduction
4.1.1 Building Landlord Energy Consumption
4.2 Motivation
4.3 Selection of Independent Variables
4.4 Methodology for the Application of SVM in Baseline Monthly Model Development
4.4.1 The objectives of the SVMs investigations
4.4.2 Model development for forecasting of building electricity load

4.4.3 Data collection

4.5 Support Vector Machine Regression

4.5.1 Theory support vector regression (SVR)

4.5.2 Kernel selection

4.5.3 Normalizing parameters

4.6 Model Development for Monthly & Weekly Load Prediction

4.6.1 Preprocessing of collected data

4.6.1.1 Weather data

4.6.1.2 Model identification

4.6.2 Formatting data

4.6.3 Evaluation indices

4.6.4. SVM algorithm development

4.6.4.1 Variable selection

4.6.4.2 Parameter characteristics of SVM

4.6.4.3 Modification of the width parameter $\gamma$ of Gaussian kernel

4.6.5 Storage effects analysis to prediction

4.7 Results of SVM Prediction for Energy Load Forecasting

4.7.1 Results of SVM for prediction of landlord energy consumption

4.7.2 Validation of the results

4.7.3 Residual values calculation

4.7.4 Discussion

4.8 Baseline Models of Building Weekly Landlord Energy Consumption

4.8.1 Data collection

4.8.2 Preprocessing of collected data

4.8.3 Weather data

4.8.4 Model identification

4.8.4.1 Formatting data

4.8.4.2 Evaluation indices
4.8.4.3 Parameter characteristics of SVM

4.8.4.4 Modification of the width parameter $\gamma$ of Gaussian kernel

4.8.4.5 Storage effects analysis to prediction

4.8.5 Results of SVM for prediction of landlord energy consumption

4.8.6 Residual values calculation

4.9 Observations of the Experiment on Both Daily & Monthly Load Forecasting

4.10 Implication

5. CONCLUSION & RECOMMENDATIONS

5.1 Review and Achievement of Research Objectives

5.2 Contribution of the Study

5.3 Recommendations for Further Research

REFERENCES

APPENDIX A: An Overview of Statistical Learning Theory Used By SVMs

APPENDIX B: An Introduction to LIBSVM 2.6 Program

APPENDIX C: Matlab Code
LIST OF FIGURES

Figure 1.1: Saving determination using the baseline model 4
Figure 4.1: Energy demand data profile of four years. 34
Figure 4.2: Weather data profile of four years. 35
Figure 4.3: Parameter for the support vector regression. 39
Figure 4.4: Proposed model to energy forecasting. 45
Figure 4.5: Results of C value with MSE. 52
Figure 4.6: Results of C value with NSV. 52
Figure 4.7: Results of ε value with RMSE for all four buildings. 53
Figure 4.8: NSV with ε value for all four buildings. 53
Figure 4.9: Results of γ value with MSE. 54
Figure 4.10: Results of γ value with NSV. 55
Figure 4.11: Results of SVM prediction for Building A. 57
Figure 4.12: Results of SVM prediction for Building B. 57
Figure 4.13: SVM prediction for Building C. 58
Figure 4.14: SVM prediction for Building D. 58
Figure 4.15: Residues of estimated landlord energy consumption. 60
Figure 4.16: Energy demand data profile of four years. 61
Figure 4.17: Weather data profile of four years. 62
Figure 4.18: Results of SVM prediction for Building E 65
Figure 4.19: Results of SVM prediction for Building E 65
Figure 4.20: Results of SVM prediction for Building E 66
Figure 4.21: Results of SVM prediction for Building E 66
Figure 4.22: Results of SVM prediction for Building E 67
Figure 4.23: Results of SVM prediction for Building E 68
Figure 4.24: Results of SVM prediction for Building E 69
Figure 4.25: Residues of estimated landlord energy consumption. 70
LIST OF TABLES

Table 2.1: Comparison of different models in the literature. 21
Table 2.2: Summary for some commonly used baseline models. 22
Table 4.1: Total energy consumption of four buildings. 44
Table 4.2: Results for C, ε, γ after SVMR–Buildings (A, B, C, D). 52
Table 4.3: Historical monthly data for Input parameters setting. 56
Table 4.4: The prediction errors of SVM models in different cases. 56
Table 4.5: Comparison of results with standards. 59
Table 4.6: Total energy consumption of four buildings. 63
Table 4.7: Results for C, ε, γ after SVMR – Building (E). 67
Table 4.8: Results for Storage effect to Building (E). 68
Table 4.9: Results for storage effect to Building (E). 69
Table 4.10: Results for short term and medium term load forecasting. 72
LIST OF ABBREVIATIONS

$a_i(*)$  Lagrange multipliers
$C$  regularization constant value
$f(x)$  general real-valued function
$h$  VC dimension
$H$  high dimensional feature space
$L(y,f(x))$  loss function
$K(x_i,x_j)$  kernel function
$N$  dimension of high dimensional feature space; natural number
$n$  number of observations
$m$  number of the month
$p$  number of parameters in the model
$q_{sol}$  total global horizontal solar radiation
$q_{i,s}$  internal sensible heat gains
$R^2$  coefficient of determination
$R(f)$  generalization error
$R_{emp}(f)$  empirical error
$T$  outdoor dry-bulb temperature
$Tc$  cooling coil leaving air dry-bulb temperature
$Th$  heating coil leaving air dry-bulb temperature
(*)  variables with and without *
$x_i,x_j$  input and input space
$Y$  measured total energy consumption
$Y_i$  measured energy consumption of unit $i$
$\bar{Y}$  mean of measured energy consumption
$\hat{Y}$  predicted energy consumption
$\xi_i(*)$  slack variables
$\omega$  weight vector
$\varepsilon$  tube size of the $\varepsilon$-insensitive loss function
ERM  empirical risk minimization
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>KKT</td>
<td>Karush-Kuhn-Tucher</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>OLS</td>
<td>ordinary least squares</td>
</tr>
<tr>
<td>QP</td>
<td>quadratic programming</td>
</tr>
<tr>
<td>SRM</td>
<td>structural risk minimization</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>VC</td>
<td>Vapnik-Chervonenkis</td>
</tr>
<tr>
<td>Z</td>
<td>general dependent variable</td>
</tr>
</tbody>
</table>
1.1 Background

It is important to understand how technological development has negatively impacted the environment while positive impact becomes more dominant among the society. Current researches have been mainly focused on this concern and address the solution for minimizing environmental pollution and overall energy consumption. Reducing energy consumption & waste is critically affecting mankind as energy becoming a limiting factor to the world. Therefore, it is much beneficial to find out the more efficient ways to reduce the energy consumption.

After the energy crisis in 1970s, people started thinking the need of efficient energy use. As only limited natural resources are used in the generation of electricity, energy remains the critical factor for the success of economy in both short- and long-term future. Moreover, it has direct connection to the design of the building and hence it is important to consider the cost effectiveness in operating and maintaining the building in terms of energy. Energy in the form of electricity is used in the building to operate equipment such as emergency systems, air-conditioning, lighting, office systems and other appliances. High energy consumption can be resulted not only from the different forms of energy use stated above but also from some common deficiencies in the building design including outdated and inefficient equipment, improper equipment selection and installation, and inadequate maintenance. One of the cheapest and most effective ways of minimizing this high consumption is by enhancing energy efficiency through the application of energy conservation measures (ECMs). For example, United States of America has initiated different building energy conservation programs such as, Federal Energy Management Program (FEMP, 1996) and the International Performance Measurement and Verification Protocol, [48].
One of the important elements in the implementation of energy conservation program is the ability to verify the savings from measured energy use data; Fels and Keating, 1993, [41]. However, there is no direct way of measuring energy use or savings since instruments cannot measure the absence of energy use or demand after retrofitting. Determination of energy savings requires both accurate measurements and replicable methodology.

Many past researches were carried out to find out solutions to reduce the energy consumptions in buildings and found that the simplest way of doing this is making current buildings less wasteful. Prediction of energy consumption of a building is a key point needs to be considered while building undergoes its retrofitting & designing of a building. Furthermore, Fels [1]; Kissock [2]; Krarti et al [3]; Dhar et al [4] & Dong et al [5,6]; had done similar researchers on predictions of consumption of energy in buildings. The main factors affecting to the energy consumption in a building can be classified into its physical properties, (optical and thermal properties) and environmental data which affecting to the building performances. Modeling with all affecting parameter for a building is very difficult to consider during the process of model development to building energy prediction. When increasing the input parameters its increase the complexity of the model resulting a misleading model with low accuracy. Some years ago, researchers began to predict the building energy consumption with the raising of importance to saving energy usage, [7–14]. With the raising interest to energy forecasting for day to day life, new innovations came to the field, such as the correlation method, [15,16]; admittance and Fourier methods, [17]; the neural networks method, [18]; and so on. Furthermore, new software packages for analysis the building energy consumption, such as Energy Plus, DeST (Designer’s Simulation Toolkit) and ESP-r, also used for better prediction of energy load consumption. Most of these tools were used in several projects successfully to predict the building load, [19–22]. However, there are negative feedbacks from using such software such as need of high levels of skills to operation with time-consuming factor.
Therefore, its make additional effect to lower grade technician to predict building load using such kind of highly advanced technical products, which may also lead certain errors in the readings. However, when it comes to, building load forecasting techniques, the artificial neural networks (ANNs) become an efficient and gain a higher popularity by overcoming most of the problems mentioned above. When using NNs, after the model has been established, it becomes simpler so that even for an ordinary lower grade technician could handle it. Moreover, the properties of ANN such as competitive non-linear mapping ability makes more handy when it comes to handling complex problems such as number of inputs data as matrix. Therefore, the applications of ANNs play a significant role in predicting building energy consumption, [23]. However, there are certain short coming of ANNs which affect to the accuracy of the prediction.

In order to overcome the above all problems, further researchers were conducted to explore more accurate systems. In Min-Yuan et al., [24]; in his research, he utilized multivariate adaptive regression splines to predict the residential building energy consumption and the result shows a high level of accuracy. Moreover, Risheek. Jain et al., [25]; developed a forecasting tool by utilizing novel method of Support Vector Machines (SVM), to predict the energy consumption of multifamily residential buildings, with high prediction accuracy. However, when it comes to accurate prediction of energy consumption of the building it’s a challenging and time consuming event, which involves gathering of larger number of data. Chunlei Zeng et al., [26]; observed the energy consumption of a multiproduct pipeline to develop a forecasting model with the help of artificial neural networks. The research carried out by Bing Dong et al., [27]; studied the applicability of SVM to building prediction shows high accuracy levels with compared to Neural Networks and other prediction methods such as regression methods. However, Bing Dong et al., [27]; it’s not address about the Kernel parameter, and storage effect analysis optimization processes to energy prediction which affecting to overall performance of the model. In this research these two areas, (kernel parameter optimization & storage effect analysis) are evaluated and developed the model.
1.2 Definition of Baseline Model

The baseline model provides a way to compute energy savings. In ASHRAE Guideline 14P, 2002, [39]; it is defined as the set of arithmetic factors, equations, or data used to describe the relationship between energy use or demand and other baseline data. According to Reddy et al.,[42]; baseline methodology helps to verify savings from energy conservation programs and determines progress toward present energy-efficiency goals. Furthermore, the baseline model can tell how much energy the building would have used if the retrofit had not been made; Chen, 2003, [46]. Overall, the baseline model defines as a method to measure savings by comparing pre-retrofit and after retrofit energy use by considering the complicated impacts of weather conditions and other usage factors. Thus, an accurate baseline model helps to measure the true energy savings.

Figure 1.1 shows the importance of the baseline model. The bold line represents the measured energy use while dotted line represents the adjusted energy use based on baseline model calculation. This is calculated by considering the changes of weather conditions and usage factors. The area in between in bold line and dot line represents the actual energy savings during the post-retrofit period.

Figure 1.1: Saving determination using the baseline model (Source: ASHRAE Guideline 14P, 2002), [39].
The basic energy saving equation used in base line method is given below.

\[ E_s = E_b - E_p \pm d \]  \hspace{1cm} (1.1)

Where, \( E_s \) is energy savings, \( E_b \) is base year energy use, \( E_p \) is post-retrofit energy use and \( d \) is an adjustments. The base year is a period of time, which is prior to the implementation of energy conservation measures (ECMs). The base year energy usage is determined using the measured equipment performance data prior to the ECM coupled with the assumptions about how that equipment would have operated in the post-retrofit period; IPMVP, 2001.[48]. Thus, baseline model used as a tool to develop and verify the base year energy usage.

1.3 Research Objectives

This research is mainly focused on studying base line landlord building energy consumption and the objectives are given below.

1) Evaluate the existing methods for baseline landlord building energy consumption.
2) Explore and establish Support Vector Machine incorporating electricity bills and weather factors for analysis the baseline building monthly & daily landlord electricity consumption.

1.4 Scope and Limitations

Five commercial buildings in the central business area of Sri Lanka, which are considered to be tropics, have been selected randomly. The base line model of landlord energy consumption was developed using the monthly utility bills collected from the owner of the building or the power supply companies. The weather data including outdoor temperature, relative humidity and solar radiation are considered as affecting variables to the energy consumption. Furthermore, the limitations and assumptions of this study are summarized below.
(1) Due to the limitations to get into the building and time constraint, only a small sample of buildings; 5 buildings, have been studied to develop the baseline model of landlord energy consumption assuming that all the five buildings are retrofitted on the landlord level. Every building is taken as one package and the method of retrofitting has not been considered.

(2) All the utility bills are assumed to be representing the electricity consumption of the selected buildings.

(3) Occupancy rates are assumed to be constant during the baseline year.

(4) Factors related to the building performance such as thermal comfort inside the building have not been considered.

1.5 Organization of Thesis

This thesis contains five chapters. The chapter 1 summarizes the background and definition of baseline model. It is further extended to explain the research objectives, scope and limitations. Chapter 2 describes the development and applications of the various baseline models used in the measurement and verification and retrofitting projects. Differences and applicability of different methods have also been discussed in this chapter. Chapter 3 used to explain the methodology for performance evaluation and pre-preparation techniques in the development of baseline model of landlord building energy consumption for the tropics. Research design and results have been explained in detail under the chapter 4. Moreover, Support Vector Mechanics (SVM) and stepwise search to optimize new parameters are introduced in this chapter to predict building landlord energy consumption. The last chapter of the thesis, chapter 5, contains conclusion and recommendations for future research.
CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter reviews the various methods used to establish the baseline models for landlord energy consumption. Moreover, importance and performance characteristics of different models have been discussed in detail.

2.2 Classification of Baseline Models

The baseline model provides a way to compute energy savings. It is the methodology used in the analysis of measured energy data for a building. Various existing methodologies are explained in this chapter to understand the function of these methodologies, which helps in creating a new successful method.

2.2.1 Regression-based models

Regression analysis is one of the important methods which statistical technique is used to build a mathematical model to relate dependent variables to independent variables, [51]. In general, a regression model is defined as a single algebraic equation with the form similar to equation 2.1, [52].

\[
Z = f (X_1, X_2, X_3, \ldots, X_k) + u
\]  

(2.1)

where, \(Z\) is a variable whose movements and values may be described by the variables \(X_1, X_2, \ldots, X_k\). The letters are known as regressors and have relationship to the dependent variable \(Z\). The additional term \(u\) is a random variable, which is included to account for the fact that movements in \(Z\) are not completely explained by the variables. The building energy consumption is considered to be the dependent variable, while other parameters such as weather and non-weather data are taken as independent variables.
There are three different types of regression models called Variable-based Degree-Day Model (VBDD), Linear Regression Model, and Change-Point models, which used generalized least squares regression to determine the model coefficients.

### 2.2.1.1 Variable-base degree-day model

This model has been originated from the fixed-base-temperature degree-day model. This allows the degree-day base temperature to be a variable, which can account for differing insulation level, thermostat settings, solar gains, and internal heat gains. The degree-day base temperature should equal to the building’s balance point temperature, which is dependent on thermostat settings, solar gains, internal heat gains and insulation levels. The balance point temperature is the outdoor ambient temperature at which heat losses through the envelope exactly balance the internal and solar gains so that contribution from the heating or the cooling system is necessary to maintain the interior temperature.

In 1978, Arens and Carrol, [54]; considered 53 °F or 11.7 °C as the appropriate lower base temperature for the newer and better-insulated building block. Moreover, VBDD has been found to be appropriate for determining energy savings in residential conservation programs, [38]. It is also believed that this method is most suitable for shell-dominated buildings such as residences and small commercial buildings, [42].

In 1980s, Fels [1]; adapted VBDD method to measure the savings as the Princeton Scorekeeping Method (PRISM). PRISM has been widely used to fit billing data in commercial buildings, [55]. Day et al.,[53]; developed a new degree-day methodology, which was demonstrated to be more accurate than the previous methods.

There are basic functional forms of the VBDD models, [42].
1) For electricity use, electricity demand and water use (increase with outdoor temperature, $T$)

$$Y = \alpha + \beta_c \cdot DD(\tau_c)$$  \hspace{1cm} (2.2)

2) For gas use (increases with decreasing $T$):

$$Y = \alpha + \beta_h \cdot DD(\tau_h)$$  \hspace{1cm} (2.3)

3) For electricity use that increases with both increasing and decreasing $T$ (eg. heat pumps)

$$Y = \alpha + \beta_h \cdot DD(\tau_h) + \beta_c \cdot DD(\tau_c)$$  \hspace{1cm} (2.4)

Where $\alpha$ is the base energy use of the VBDD model, $\beta_c$ is the slope for the VBDD cooling model, $\beta_h$ is the slope for the VBDD heating model. $DD(\tau)$ are the degree-days to the base $\tau$, and the subscripts $c$ and $h$ stand for cooling and heating, respectively.

The best-fit VBDD model is identified using a search method by regressing equation (2.2), (2.3) and (2.4) using $DD(\tau_c)$ and $DD(\tau_h)$ in each energy period for successive base temperature, $\tau$, from 41 °F to 80 °F. The base temperature that results in the model with the highest $R^2$ is recorded; [2].

In addition, Day et al., [53]; found that the bias error of degree-day method for estimation energy consumption in buildings can be ranged from $\pm$ 4.5% for relatively high degree-day values (seasonally and yearly values) to $\pm$10% for low degree-day values (monthly values).

2.2.1.2 Linear regression models

Single-variant linear regression model and the multivariate linear regression are the two different types of linear regression models.
Single-variant linear regression model

In this model, mean temperature is considered as the model independent variable. Reddy et al., [42]; used this method in their studies and whole building energy consumption is estimated within 90% confidence level. The variance of forecast error is within 9%. Kissock et al., [55]; took ambient-temperature as the sole independent variable to baseline cooling and heating energy use in an engineering center of Texas A&M university and results presented high coefficient of variance more than 10%. Reddy et al., [42]; two-parameter model defined as below equations, (2.5),(2.6).

1) For energy use that increases with increasing outdoor temperature $T_0$ (eg: electricity use for air conditioning):
   \[ Y = Y_0 + RS.T_0 \]  \hspace{1cm} (2.5)

2) For energy use that increases with decreasing outdoor temperature $0T$ (eg: gas use)
   \[ Y = Y_0 - LS.T_0 \]  \hspace{1cm} (2.6)

Where, $Y_0$ is the intercept that represents the value of energy use when $T_0 = 0^\circ\text{F}$. RS means right-hand slope and LS means left-hand slope.

Multivariate linear regression model

If there are more than one independent variables, the regression model is called a multiple or multivariate linear regression model. Katipamula et al., [57]; used multiple linear regressions with internal gain, solar radiation, and humidity ratio as independent variables based on hourly and daily basis in addition to temperature. This model was specifically used in dual-duct constant volume (DDCV) and variable air volume (VAV) systems and derived equation is given as:

\[ \hat{E} = a + bT_0 + cI + dIT_0 + eR_{dp} + f q_{\text{sol}} + g q_i \]  \hspace{1cm} (2.7)
Where $a, b, c, d, e, f$ and $g$ are regression coefficients, and $T_0$ is the outdoor dry-bulb temperature, $T_{dp}$ is the outdoor dew point temperature, superscript “+” means $T_{dp}$ is set to zero when it is negative, $q_{sol}$ is total global horizontal solar radiation, $q_i$ is the internal sensible heat gains and $i$ is an indicator variable which is 1 when $T_0$ is greater than the change point temperature and 0 otherwise. According to their studies, for some buildings, internal gains had a modest impact on consumption, while the impact was negligible for others. Thus, even with a multiple regression model, $T_0$ and $T_{dp}$ accounts for more than 90% of the variation in the cooling energy consumption on all time scales for both DDCV and DDVAV systems.

**Integrated model**

Sonderegger, [58]; created a baseline equation with the combination of multivariate regression and degree-day methods. This resultant model considered both non-weather-related and weather related independents. Moreover, this can accommodate up to five simultaneous independent variables for a maximum of eight free parameters. The results showed that with different baseline year period $R^2$ ranges from 0.74 to 0.95. However, too many independent variables may cause unexpected noises during regression process.

All the different kinds of linear regression models presented above are simpler, easier and more practical compared to other baseline models, which are explained in the next section.

**2.2.1.3 Change-point models**

This model shows a nonlinear relationship between heating and cooling energy and ambient temperature caused by system effects. The independent variable can only be the outdoor temperature and has been applied to many commercial buildings as explained in previous literature such as Reddy et al., [42].
2.2.2 Calibrated simulation

Literature illustrates several calibration procedures developed by different research groups to get better performance. Procedures for calibrating hourly simulation programs to create baseline models for building were developed in the 1990s. After that, developed a hybrid calibration procedure for taking an hourly simulation model and applied it to data obtained from whole building electricity consumptions and system level consumptions. Bronson et al., [59]; developed graphical procedures to permit visual based analysis on hourly data based on DOE-2 computer simulation. DOE-2 simulations were significantly improved when schedules based on measured data were introduced. Interestingly, the availability of comparative three-dimensional surface plots significantly improved the ability to view small differences between the simulated and measured data. In 1993, Katipamula and Claridge, [60]; developed a simplified version of hourly simulation, which is based on the ASHRAE TC 4.7. Moreover, calibrated DOE-2.1 program developed by U.S. Department of Energy has been widely used. In IPMVP, calibrated simulation was recommended as option D situations where calibrated simulation approaches are used as given below; IPMVP, [48].

1) Either base year or post-retrofit energy data unavailable or unreliable.
2) The energy conservation measures (ECMs) involve diffuse activities, which cannot easily be isolated for the rest of the facility.
3) The facility and the ECMs can be modeled by well-documented simulation software, and reasonable calibration can be achieved against actual metered energy and demand data.
4) The impact of each ECM on its own is to be estimated within a multiple ECM project and the costs of options A or B are excessive. Option A is referred to partially measured retrofit isolation which, savings are determined by partial field measurement of the energy use of the systems to which an ECM was applied. Option B is referred to retrofit isolation which, savings are determined by field measurement of the energy use of the systems to which an ECM was applied.
5) Major future changes to the facility are expected during the period of savings determination.

6) An experienced energy simulation professional is available and adequately funded for gathering suitable input data and calibrating the simulation model.

The accuracy of the energy retrofitting savings is completely dependent on how well the simulation models actual performance and how well calibrated it is to actual performance.

2.2.3 Artificial neural networks

The Artificial Neural Networks or ANN is used in many different fields of forecasting of building energy use for both short and long term periods. They provide an attractive way for determining the dependence of energy consumption on occupancy dependent factors as well as weather variables. It is appropriate to view neural networks as a set of powerful non-linear regression tools. The early application of neural networks for the prediction of building energy consumption utilized feed-forward networks, which require the use of immediate past consumption as an input. Kyung-Jin Jang et al., [61]; utilized auto associative neural network as a preprocessor to replace the missing data and applied a standard feed-forward artificial neural network to predict building energy consumption in two different buildings based on hourly data.

The results showed that all the R² values are all more than 0.9 and CV of the predicted results are within 10%. Krarti et al., [3]; identified the different features of neural network applications for the evaluation of ECM retrofits. Actual building data can be readily used for pre-retrofit modeling of building and established building physics principles can be used along with the pre-retrofit networks to estimate electricity and thermal energy saving, [3]. Afterwards, he commented on the application of AI-based techniques as shown: (Neural networks: Short-term load, weather forecasting and systems modeling, neural networks and genetic algorithms: Controls of thermal energy storage, Fuzzy logic model-based approach: Fault
detection and diagnostic However, all these methods are suitable only for the systems with large pool of data. Overall, the NN method can predict the annual landlord energy consumption reasonably well, but not the monthly, Dong et al., [27]. The possible reasons for this could be;

1) Small number of training data: Generally, NN needs a large pool of data for training.

2) Limited input variables: Not only has the climate variables but also other factors (such as human and management) had certain contribution to the changes of building energy consumption.

3) Difficulty in optimizing network-controlling parameters: There are lots of other parameters in NN (number of hidden nodes and hidden layers, the transfer function, the learning rate, the momentum term etc.), which create errors in the results.

2.2.4 Fourier series

The Fourier series is a classic technique for modeling time series data with periodicity. Dhar et al., [4]; developed a related method called Generalized Fourier Series (GFS) using classic periodic behavior of both non-weather and weather dependent energy use. They showed how the physically expected dependence on outdoor dry-bulb temperature and other variables (relative humidity and solar radiation) can be combined with Fourier series techniques to capture the hourly, daily and seasonal periodicity. After that, they developed a model called Temperature based Fourier Series (TFS). In this model, they selected outdoor temperature as the only weather variable to model hourly heating and cooling use in commercial buildings. It has been found to model heating and cooling energy use in commercial buildings accurately.

A comparative study between GFS model and TFS model showed that in terms of heating energy use all $R^2$ values are more than 0.73, while the CV ranges from
14.57% to 24.55%, and the TFS model is slightly better than GFS model, and in terms of cooling energy use all $R^2$ values are more than 0.8, while the CV ranges from 6.3% to 18.75%, and the GFS model is slightly better than TFS model.

2.2.5 Bin method

This method is based on steady-state modeling of building energy systems, [62]. The classical bin method takes outdoor temperatures into bin groups of equal size; typically 5°F (2.8 °C) bins and the bins are separated into three daily, eight-hour groups. This method takes into account both occupied and unoccupied conditions and accounts for internal loads by adjustment of the building balance point.

However, the classical bin method may not provide accurate energy predictions for building with high latent heat loads. Knebel [62] extends the basic bin method to account for weekday/weekend and partial-day occupancy effects, to calculate building loads at four temperatures and to better describe secondary and primary equipment performance, [38].

2.2.6 Support vector machines

Support vector machines (SVMs) is one of the important method developed by Vapnik and his co-workers in 1995, [28]; and it has been widely applied in different literature in classification, forecasting and regression of random datasets, [50, 37]. One of its main application fields in regression modeling is the time series financial forecasting. The Vapnik-Chervonenkis theory is developed from the statistical learning theory [63].

Characters of SVMS for regression estimation

Characteristics of SVM are given below.
Chapter 2 Literature review

1) SVMs estimate the regression using kernel functions, a set of linear functions that are defined in a high-dimensional feature space and inputs have nonlinear performance.

2) SVMs carry out the regression estimation by risk minimization, where the risk is measured using Vapnik’s ε-insensitive loss function.

3) SVMs implement the SRM principle, minimizes the risk function consisting of the empirical error and the value of confidence level.

4) Training SVMs is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVMs is always unique and globally optimal while network’s training requires nonlinear optimization with the danger of getting stuck into local minima.

5) The solution to the problem is only dependent on a subset of training data points which are referred to as support vectors.

One disadvantage of SVMs is that the training time scales somewhere between quadratic and cubic with respect to the number of training samples. According to Cao et al., [37]; a large amount of computation time is required in solving large-size problems. In SVM model development, the data set divided to two sets of inputs as given in equation (2.9). A2 is for training the data set while A3 is for testing &validation on empirical data set. The data arrangement for feeding to the inputs matrix (A1) is presented in equation (2.8) [29]:

\[ A1=\text{input Matrix} = [\text{timeDelayInput1, timeDelayInput11, timeDelayInput111, MaxTemperature\_norm, Humidity\_norm, SolarRadiation\_norm}] \]  

Where the time delay inputs represent electrical consumption values for the previous three time steps, as a solution to storage effect & MaxTemperature\_norm, Humidity\_norm, SolarRadiation\_norm indicator variable that visualize normalizes data of temperature, relative humidity & solar radiation respectively.

The forecasting model for the training inputs is presented by equation (2.9).
\[ A2 = \text{ForecastingModel} = \text{svmtrain(TrainOutput, TrainInput,'-s -t -ε -c -g')} \]  

(2.9)
In equation (2.9), the Train Output, Train Input are the training set files & from the symbols (s - t - ε - c - g), its demonstrate the model file. Model_file is the file generated by svm-train & test file is the data need to predict. And after running the package, svm-predict will produce output in the output_file as per the equation (2.10), [30]

\[ A_3 = \text{[prediction]} = \text{svmpredict (TestOutput, TestInput, Forecasting Model)} \]  

Corresponding symbols are denoted as follows.

-s SVM_type : set type of SVM

0 -- C-SVC
1 -- nu-SVC
2 -- one-class SVM
3 -- epsilon-SVR
4 -- nu-SVR

-t kernel_type : set type of kernel function

0 -- linear: u'*v
1 -- polynomial: \((gamma * u' * v + coef0)^\text{degree}\)
2 -- radial basis function: \(\text{exp}(-gamma * |u - v|^2)\)
3 -- sigmoid: \(\text{tanh}(gamma * u' * v + coef0)\)
-g-- gamma : set gamma in kernel function
-p-- epsilon : set the epsilon in loss function of epsilon-SVR
-c --cost : set the parameter C of C-SVC, epsilon-SVR, and nu-SVR

2.2.6.2 Kernel selection

The kernels can perform all the necessary computations in input space, without computing the map to high dimensional feature space, \(\Phi(x)\). Most commonly used kernels for nonlinear regressions are, linear kernel \(K(x_i,x_j) = x_i.x_j\), polynomial kernel
K(xᵢ, xⱼ) = (xᵢ − xⱼ + 1)^d and the radial-basis function (RBF) kernel, K(xᵢ, xⱼ)= exp(-γ ||xᵢ−xⱼ||^2)^γ>0, where d and γ are defined as kernel parameters.

The RBF kernel, which is based on Gaussian function, nonlinerly maps samples into a higher dimensional space and handle the case when the relation between class labels and attributes is nonlinear. Linear kernel which is a special case of RBF, showed that the linear kernel with a penalty parameter C had the same performance as the RBF kernel with some parameters (C, γ), [31]. The polynomial kernel has more hyper-parameters than the RBF kernel and hence RBF kernel has less numerical difficulties in contrast to polynomial kernels. Moreover, it is found that the sigmoid kernel is not valid under some parameters, [28]; and hence RBF kernel is selected in this study. According to the definition of −γ =1/k by Limsvm-2.6 where k is the number of attributes in the input data, −γ is constantly set to 1/3 in the future modeling. Finally, all the training and test data sets are scaled to [0,1].

2.2.6.3 Modification of kernel parameters

C and ε are the two important parameters (except γ) used in RBF kernels. ε is the key parameter in the ε -insensitive loss function. The key point is to select C and ε so that the regression can accurately predict unknown data such as testing data. The cross-validation approach is used to determine performance on regressing and prevent the over-fitting problem. In v-fold cross-validation, the training set is divided into v subsets of equal size.

Then one subset is tested using the regression trained on the remaining (v − 1) subsets and hence each instance of the whole training set is predicted once. The accuracy of cross-validation is shown as the average S-MSE.

In this study, v equals four, which means that three-fold-validation is conducted. After selecting proper parameters, one-time search method developed by Francis et al., 2001, was performed. The stepwise search method is used to measure the performance as explained in Dong et al., [27].
Selection of parameter $c$

According equation 4.2 (chapter 4), parameter $C$ determines the tradeoff between the model complexity and the degree to which deviations larger than $\varepsilon$ are tolerated in optimization formulation. Moreover, the regularization parameter $C$ decides the range of values $0 \leq (a \cdot a^*) \leq C$, $i=1,...,l$ assumed by dual variables used as linear coefficients in SVMs solution (equation 4.5). Thus, a “good” value for $C$ can be chosen equal to the range of output values of training data, Dong et al., [27].

Theoretically,

1) When $C$ is small: It will under-fit the training data because the weight placed on the training data is too small thus resulting in large values of MSE on the test sets.

2) When $C$ is too large: SVM will over-fit the training set, which means that $\frac{1}{2} \| \omega \|^2$ will lose its meaning and the objective goes back to minimize the empirical risk only.

Number of support vectors increases slightly as $C$ increases. When $C$ gets larger, the optimization formulation (4.2) emphasized more on the empirical risk and makes the model fit the training data better at the cost of larger model complexity. Hence more support vector numbers are needed to determine the model and more data points can be selected as the support vectors in the optimization formulation.

Selection of parameter $\varepsilon$

According to figure 4.3 (chapter 4), parameter $\varepsilon$ is used to fit the training data and it controls the width of the $\varepsilon$-insensitive zone. The $\varepsilon$ does not affect the performance of SVMs much, while the number of support vectors shows a decreasing function of $\varepsilon$. Generally, the larger the $\varepsilon$, the fewer number of support vectors and thus the sparser
the representation of the solution (27). If the $\varepsilon$ is too large, it can be worsened the accuracy on the training data. The optimization for parameter $\gamma$ (kernel parameter) also same as for parameter C & $\varepsilon$, which describe in chapter for section 4.8.4.5.

**Stepwise search**

Apart from the method explained above, grid-search and stepwise method are the most common methods used in identifying best C and $\varepsilon$.

1) Grid-search: Frequently used and the most complex and reliable one. All pairs of (C, $\varepsilon$) are tried and the one with the best performance is picked up. However the efficiency of the grid-search is low because it computes the performance at all pairs of C and $\varepsilon$ to get the performance surface.

2) Stepwise Method: More efficient to quickly search the peak point of the performance surface. It is more accurate than one-time search, which is conducted only once on every parameter.

One-time search is first conducted to get MSE1 and then the same selection process is conducted again on parameter C (fixing the first result of $\varepsilon$) and $\varepsilon$ (fixing the second result of C), to get lowest MSE2. The one-time search continues until $n \ MSE - (n-1) \ MSE < 0.00001$ and then the training is stopped. After finding the best (C, $\varepsilon$), the whole training set is trained again to generate the final regression.

**2.3 Discussion**

Statistical regression model is the easiest and most common way to establish baseline models and it shows fairly good accuracy depending on the important independent parameters considered. In contrast, computer simulation methods (DOE-2, Energy Plus, BEST) are time consuming and need detailed information of building operation characteristics, physical conditions and large input of energy consumption data.
Other models such as ANN and Fourier series models are mostly simulated based on hourly data and data inputs, processing and results rely on certain software and have quite high accuracy. Nevertheless, a very few models have been actually implemented in buildings as they need more improvement and computational efficiency to get widespread uses. The differences between several baseline-modeling methods used in literature are summarized in table 2.1.

Table 2.1 Comparison of models discussed in literature

<table>
<thead>
<tr>
<th>Models</th>
<th>Examples</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>Linear,</td>
<td>Easy to establish</td>
<td>Fair accuracy</td>
</tr>
<tr>
<td>Regression models</td>
<td>Multiple-linear,</td>
<td>Most commonly used</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change-point,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>degree-day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other models</td>
<td>Support Vector Machine,</td>
<td>High accuracy</td>
<td>Large pool data, however, SVM show high accuracy</td>
</tr>
<tr>
<td></td>
<td>Neural Network,</td>
<td></td>
<td>with small pool of data</td>
</tr>
<tr>
<td></td>
<td>Fourier Series</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This chapter mainly focused on comparing the previous methodologies and studies used in building energy analysis and Table 2.2 summarizes the methodologies used in baseline modeling.

Choosing a proper and practical methodology is important for baseline building energy consumption and energy saving estimation and this is directly related to the time scale such as hourly, daily and monthly or the level such as system and facility. Regression based model is considered to be more practical for all time periods’ data, while neural networks and Fourier Series are more suitable for modeling hourly building cooling and heating energy consumption.
Table 2.2 Summary for some baseline models.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Types of Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kusuda et al. (1981)</td>
<td>Variable-base degree-day</td>
</tr>
<tr>
<td>Fels et al. (1986)</td>
<td>Single-variable regression model</td>
</tr>
<tr>
<td>Pope (1987)</td>
<td>Modified Bin</td>
</tr>
<tr>
<td>Kissock et al. (1993)</td>
<td>Single-variable regression model</td>
</tr>
<tr>
<td>Katipamula et al. (1994)</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>Kissock et al. (1998)</td>
<td>Linear and change-point models</td>
</tr>
<tr>
<td>Krarti et al. (1998)</td>
<td>Neural networks</td>
</tr>
<tr>
<td>Reddy et al. (1998)</td>
<td>Multiple linear regression</td>
</tr>
<tr>
<td>Dong et al (2005)</td>
<td>Support vector machine</td>
</tr>
</tbody>
</table>
CHAPTER 3
METHODOLOGY FOR ERROR EVALUATION AND PRE-PREPARATION TECHNIQUES

3.1 Introduction

In this chapter, methodologies, which used in performance evaluation and pre-preparation for baseline model development, are explained. Daily & Monthly utility electricity bills of buildings in Sri Lanka, are collected for modeling and analysis. According to the literature analysis, coefficient of determination ($R^2$), the coefficient of variance of the root-mean-square error (CV-RMSE), Mean Squared Error, (MSE) and % error are used in evaluating the goodness of fit of a model. The value of $R^2$ is defined as the Pearson correlation coefficient between the observed and fitted values. The CV-RMSE is a non-dimensional measure that that can be calculated by dividing root-mean-square error (RMSE) by the mean value of total energy consumption $Y$ and usually this is given as a percentage. The CV-RMSE defined below equation (3.1):

$$CV - RMSE = \frac{RMSE}{\bar{Y}} \times 100$$

(3.1)

Where, $RMSE = [MSE]^{\frac{1}{2}} = \left[ \frac{\sum_{i=1}^{N} (Y_i - \hat{Y})^2}{N} \right]^{\frac{1}{2}}$

$Y_i$ is the actual energy consumption of single month/day/hour, $i (i=1,...N)$. $\bar{Y}$ is the mean value of $Y$. $\hat{Y}$ is the value of $Y$ predicted by the regression model. $N$ is the number of observations.

$R^2$ and CV-RMSE have been used as the important parameters in determining the goodness of fitting. According to Reddy et al., [42]; CV-RMSE of less than 5% are considered excellent models, those less than 10% are considered good models, and those less than 20% are taken to be mediocre models and those greater than 20% are considered to be poor model. In this thesis, the values of $R^2$, CV-RMSE, Mean
Squared Error, (MSE) and % error follow the criteria pointed out by Reddy et al., [42]; as it applies to the whole building energy consumption analysis, and it may also be considered to become rigorous than normal industry models.

3.2 Errors and Bias of Baseline Model

It is very important to analyze the errors and bias accompanied with designed baseline model. There are different types of errors including sampling error. In ASHARE, [39]; sampling error referred to the errors resulting from the fact that a sample of units was observed rather than observing the entire set of units. Thus, no sampling errors need to be considered in this study due to all the samples are used with the entire unit. However, other sources of uncertainty within the model have been explored in Reddy et al., [43]; and it is categories under the regression model error.

1) Model misspecification error: This is due to approximation of the true driving function of the response variable. This error caused by, (a) the inclusion of irrelevant regression variables or non-inclusion of important regression variables (b) the assumption of a linear relationship, when the physical equations suggest nonlinear interaction among the regression variables, and (c) the incorrect order of the model, i.e. either a lower order or a higher order model than that suggested by the physical equations.

(2) Model extrapolation error: This error arises when the prediction carried out in the outside the domain of the original data. Models identified from short datasets, which do not satisfactorily represent the annual behavior of the system and thus this will lead an error. Moreover, prediction of long-term building loads from short-term in-situ tests will also cause this type of error.

(3) Model prediction errors: This error arises because a model is never “perfect”, certain amount of the observed variance in the response variable is unexplained by the model. This variance introduces an uncertainty in prediction and hence $R^2$ will
never be 1.0. This uncertainty may occur even when the “exact” functional form of the regression model is known.

### 3.3 Normalized Energy Use

The baseline model is used to correct for changes in energy use due to changes in weather data from year to year, month to month and day to day. However, fluctuations in the conditioned area and the population need to be considered equally and the adjustments can be made to remove such effects and hence by giving an accurate estimate to the energy efficiency measures. Fels and Keating et al., [41]; assumed a proportional relationship between energy use and changes in the conditioned area. Hence, normalized area-change energy use is merely the annual mean monthly energy use divided by the conditioned area for that particular year. Moreover, modeling with population-varied energy use is really challenging and researches still moving on this area. Here, it is assumed that normalizing energy use by conditioned area would sufficient to normalize energy use for population changes, if it is speculated that population could be related to conditioned area, [42].

### 3.4 Percentage Changes Based on Annual Energy Use

According to Reddy et al., [42]; the equation for percentage change between actual energy use and projected energy use is given as following equation, (3.2):

\[
\Delta E(\%) = \frac{E_{\text{measured}} - E_{\text{projected}}}{E_{\text{projected}}} \tag{3.2}
\]

Where, \(E_{\text{measured}}\) is the annual-based mean monthly energy use determined by simply averaging 12 monthly utility bills for a certain year. \(E_{\text{baseline projected}}\) is the annual energy use predicted by the baseline model using corresponding weather data for the corresponding year. The direct deviation between actual (\(Y\)) and predicted (\(\hat{Y}\)) energy consumption, which is also called percentage error (% error), is defined simply as follows in equation (3.3).
\[
\% \text{error} = \frac{Y - \hat{Y}}{Y} \times 100
\]  
(3.3)

### 3.5 Possibilities of Utility Bill Reading Dates

As building owners of Sri Lanka do not provide detailed information on utility bill reading dates, it is very important to consider the uncertainty in the billing reading dates in the process of model development. Moreover, every building may have its own policy for recording electricity use and hence it is more challenging to verify the specific utility reading period. Reddy et al., [42]; showed two possibilities in recording building energy use.

1. The utility bill period is correspondent with weather data period. It means that the utility bill reading dates are the same dates as weather data dates.

2. The utility bill period is 15 days later than the weather data period. It is decided to take the average value of the temperature of the present month and that of the previous month and associate those values to the particular utility bill. If the model turns out to be substantially better than the previous case 1, above, this would imply that the utility bill was read sometime around the middle of the month as against the beginning.

For the baseline model, all these two possibilities are performed and then the one, which has the best-fit regression, is selected.

### 3.6 Primary Modeling of Whole Building Energy Consumption

The energy consumption of a building is a complex function with climatic conditions such as dry-bulb and dew-point temperatures, building characteristics such as loss coefficients and internal loads, building usage such as for commercial or residential use, type of heating, ventilation and air conditioning equipment used. Since some of
these parameters are difficult to estimate accurately, they are impractical to be considered as model variables. According to the literature, the most important indicator for baseline building consumption is outdoor dry-bulb temperature with a considerable effect from other weather factors like humidity for countries in tropical region.

### 3.7 Selection of Weather Variables

In addition to the temperature (T); relative humidity (RH) and solar radiation (SR) are chosen as two additional variables relevant to local conditions. It is well known that the cooling energy consumption accounts the most in the whole commercial building energy consumption. When it comes to cooling load the main affecting parameter for cooling loads of a building are weather parameters. Therefore in order to improve the model performances the above mentioned environmental factors were selected other than the temperature. Moreover, previous studies have shown that buildings can be physically modeled into two zone buildings, exterior zone and interior zone, with adequate accuracy, [35].

#### 3.7.1 Data collection

In chapter 4, weather data profiles are presented including temperature (T), relative humidity (RH) and solar radiation (SR). Figure 4.2 shows the monthly average weather data from year 2010 to 2013. As SR is a huge number, it has been plotted as monthly value divided by ten. Figure 4.2 shows that there is no significant variation of weather details over the four years. For example, the highest average temperature appeared in May 2000 as 29.4°C, while the lowest average temperature appeared in January 2000 as 26.5 °C. Hence, the overall difference of average temperature is only 2.9 °C. Similar situation observed in RH and SR data.

The utility bills of these four buildings were collected through surveys which were carried by on buildings energy efficiency. The highest value, in terms of
consumption, appears in building A, which is 27,327 MWh/yr, while the lowest value appears in Building B, which is 594 MWh/yr.

Moreover, seasonal variation of energy usage are monitored and the results shows similar variation of energy consumption over the four year periods. The figure 3.1 to 3.4 shows the seasonal variation of energy data of the selected four buildings.

Figure 3.1: Seasonal variation of energy of Townhall Branch from 2010 - 2014

Figure 3.2: Seasonal variation of energy of Pettah Branch from 2010 - 2014
Chapter 3 Methodology for error evaluation and pre-preparation techniques

Figure 3.3: Seasonal variation of energy of Bambalapitiya Branch from 2010 – 2014

Figure 3.4: Seasonal variation of energy of City Branch from 2010 - 2014

Therefore, four buildings, namely building locations, Townhall, Pettah, Bambalapitiya, City in Sri Lanka labeling A, B C and D, are taken as case studies in the further research.
3.7.2 Discussion

The effect of independent variables for these building may be different due to different reasons. Physical surrounding (location and height of buildings have important effect on solar radiation gains and heat gains) of the building is totally different though they have selected within the same area. Furthermore, different building has different shape, materials and orientation, which also cause different OTTV (overall thermal transmit value). Overall, these arrangements significantly affect the total load consumption of HVAC systems.

In conclusion, regression models based on utility bills and weather data are enough for establishing baseline models for the building energy use on monthly & daily basis. However, due to the little changes in building energy consumption from month to month, with seasonal changes in the tropical area, more accurate baseline model may be expected on daily or hourly basis.
CHAPTER 4
BASELINE MODELS OF BUILDING MONTHLY & DAILY LANDLORD ENERGY CONSUMPTION

4.1 Introduction

This chapter describes the research design, methodology and results for the baseline model development of building monthly & daily landlord energy consumption. While it is useful and convenient to obtain a baseline model for the whole building energy consumption, in practice, most buildings are managed by the landlord and most of the energy efficient measures implemented are related to part of the landlord consumption.

4.1.1 Building Landlord Energy Consumption

It is well known that for many buildings energy consumption there are certain constant loads such as lighting, fan and plug loads, which do not change with weather variables. A building’s landlord energy consumption refers to the energy utilized by the common facilities, systems, services and space provided by the landlord. They typically comprise:

a) Air-conditioner central plant system which supply air-conditioning inside the building;
b) Vertical transportation service such as escalator and lift;
c) Ventilation system such as exhaust fan and ventilator;
d) Artificial lighting system in the common area i.e. corridor or public commonservice area such as toilets, services lifts, car park and so on.
e) Cleaning services, decorations, and external gardens and flood lighting.

Obviously, the usages of these systems present certain non-linear performance between building energy use and weather data.
4.2 Motivation

The past researchers in [1, 2, 3, 4, 5]; of landlords’ energy consumption shows that landlord energy consumption method is significantly inferior as compared with earlier methods of whole building energy consumption as given in Chapter 3 (Section 3.6). In addition, the building owner or managers often receive the landlord bills only. It seems more meaningful to analysis baseline landlord energy consumption rather than the whole building energy use for the benefits of both building owners and the energy efficiency measures retrofitting contracts. For these reasons further many researches were conducted to examine various non-linear methods and determine this suitability as a baseline model for building landlords’ energy consumption. However, the nonlinear training needs a large pool of data for the model construction, at least ten times of the inputs parameters [37]; which are weather data in this study.

4.3 Selection of Independent Variables

Selection of independent variables addresses the problem of choosing a small subset of most important features from available features. The goal is to improve the speed and generalization performance of SVM model. There are many independent variable selection methods when it’s going to develop prediction systems. Normally, they are based on statistical methods such as the stepwise method, Erivastava,1990. However, there are some shortcomings if based on such kind of methods. Mainly, it cannot be guaranteed that the selected independent variable or feature set is optimal.

However, for the purposes of this application, many previous studies have shown that weather data are the most important elements affecting the building energy consumption, and it is sufficient to set up an accurate baseline model for residential commercial buildings , Kreider et al., [47]. In addition, Dong et al., [27]; in his research shows that cooling energy consumption has significant relationships with outdoor temperature, solar radiation, internal sensible heart gains, and specific humidity of air. Hence, based on the rule of exploration, weather data including
outdoor temperature (°C), relative humidity (%) and solar radiation is selected as independent variables for model inputs.

### 4.4 Methodology for the Application of SVM in Baseline Monthly Model Development

Firstly, five commercial buildings were selected, located in the business area from the buildings in Colombo region. For each building, up to four years of daily and monthly building energy consumption data as utility bills were collected. The first three years’ data was used for training and establishing the baseline model for monthly model & another one year utility bill was used for testing and verification. The methodology to short term prediction is mentioned in section 4.8.

At the same time, the corresponding weather data were collected from metallurgical department Sri Lanka. The weather data period used was the same as the energy consumption data period. All the weather data are monthly mean values. Secondly, when all the data are ready; the SVM was applied to set up the baseline model. The weather data include the outdoor temperature, relative humidity and solar radiation. The total number of input parameters are four, including three weather data and the time tag. The output is the building landlord’s energy consumption. Thirdly, after training, the baseline model, it used to predict the energy consumption for the projected period. In this study, the projected year was selected as year 2013. Finally, the predicted annual building monthly landlord energy consumption was compared with the actual consumption value.

In addition to the weather data inputs, it is necessary to remove the effects of year-to-year changes in conditioned areas and population as mentioned in Chapter 3, section 3.3. Here, landlord energy consumption is considered, it assumed that, this feature potentially removes the effects of population changes. Hence, it is assumed that normalizing energy use by landlord area would be implicit enough for such changes. However, it should be noted that most of the commercial buildings are with central HVAC systems, which supplies cooled air to all tenant spaces and some common
spaces. Hence, normalizing landlord energy consumption should be based on the whole building gross floor area (GFA) rather than the landlord’s common areas only. Finally, the performance measurement criteria are CV-RMSE as defined in equation (3.1), % error, which is defined in equation (3.2 & 3.3) and mean square error (MSE).

4.4.1 The objectives of the SVMs investigations

1) To examine the feasibility of applying SVMs in predicting building energy consumption.

2) To investigate the effect of different parameters of SVMs on the medium & short term, (section 4.8); prediction accuracy.

The selected four commercial buildings presented are used for detailed analysis in this section. The building landlord energy consumption is chosen to build and test models. The weather parameters, namely, outdoor temperature (T), relative humidity (RH) and solar radiation (SR) are taken as three inputs, the namely three features.

3) Libsvm-2.6 developed by Chih-Chung Chang and Chih-Jen Lin, [45]; is applied in this study to produce and test the application of SVMs in the development of baselining models. The detailed command of the software is introduced in Appendix B. A detailed development and theory of SVMs for regression is introduced in the following section.

4.4.2 Model development for forecasting of building electricity load

This research describes how computational models can be used to predict the amount of electricity will consume by a building. The computational model is developed using a regression method called Support Vector Machine Regression (SVMR) which shows superior performance than all machine learning techniques used in
energy forecasting, Dong et al., [27]. This method is solely based on historical data of the building while no prior knowledge of its physical properties is required to the prediction process. Thus, it can be easily applicable to different types of buildings only with slight modifications to the system parameters. In this research, the SVM regression models were built to predict medium term future electricity usage based on past energy, monthly mean temperature, relative humidity & solar radiation data as they are the significantly affecting parameters to energy prediction, [27].

Large commercial buildings specifically, such as HNB tower, WTC, BOC buildings consume heavy amount of energy in the Colombo city, Sri Lanka. Therefore, one of the greatest opportunities to reduce the energy consumption in Colombo city, Sri Lanka, is through reducing energy consumption in buildings. The large commercial buildings with tenants used to this research are with regular working hours from m 8:00 a.m. to 8:00 p.m. throughout the working days.

In this research, the total energy consumption is measured for a defined period of time, for example the number of kilowatt-hours per square meter consumed in a month. Most residential buildings are charged solely based on their total energy consumption in a given billing cycle (per month), and have no need for measuring their energy usage in any other way. However, when it comes to large commercial buildings, the total amount of consumption over these large blocks of time does not give an accurate depiction of the energy usage trends. This is because typical office buildings consume most of their energy during regular weekday office hours and use relatively low consumption of energy at night and on weekends, when the building is in off line.

This requires the main grid to supply these buildings with a large amount of energy at certain times, instead of an average amount constantly, making the total monthly consumption value a somewhat misleading indicator from the true energy requirements of the building. Furthermore, its make the main grid to face for an uncertainty with the fluctuation of energy demand which affecting negatively for a
developing country like Sri Lanka. Again it implies the needs of accurate energy prediction system to deal with a steady energy supply to a country like Sri Lanka as the whole country economy is depends on the energy consumption which leads to determine the market price of every goods. Therefore, when it comes to feeding the data to SVM model, these uncertainties of energy consumptions need to be avoided in order to develop an accurate model. In this study to avoid such unsteadiness of the energy requirement of buildings, the rate of energy consumption, called demand, is used to show the amount of energy the grid must supply the building at a given time.

4.4.3 Data collection

To develop an accurate model for prediction process four commercial building were selected, among the Colombo city, Sri Lanka and figure 4.1 shows monthly energy demand values for selected four commercial buildings for the modeling process. The energy demand of a building is determined by several factors. There are fixed building characteristics, such as size, materials, location, orientation, design that contribute to the energy consumption of a building.

![Energy demand data profile of four years.](image)

Figure 4.1: Energy demand data profile of four years.

Moreover, there are some dynamic characteristics that affect the energy usage of a
building. Weather has a significant effect on the amount of energy required to functions like air-conditioning heating and cooling system of a building. Figure 4.2, shows, the selected weather parameters (temperature, relative humidity, solar radiation) in Colombo region, Sri Lanka.

Figure 4.2: Weather data profile of four years

The utility bills of these four buildings were collected through surveys which were carried by on buildings energy efficiency. The highest value, in terms of consumption, appears in building A, which is 27,327 MWh/yr, while the lowest value appears in Building B, which is 594 MWh/yr.

In addition, the weather data was collected through the same process described in chapter 3, Section 3.7.1. Meanwhile, some building owners and managers did not provide complete and detailed billinformation in the survey carried, which leads to the limited utility bills up to four years.

Therefore, only four buildings, namely building locations, Townhall, Pettah, Bambalapitiya, City in Sri Lanka labeling A, B C and D, are taken as case studies in the further research.
4.5 Support Vector Machine Regression

Following all the stages of developments, researchers were more eager to find out more advanced ways of load forecasting systems. As a result a new technological development called support vector machine (SVM) came to play its role of predicting energy consumption of buildings. Vapnik and co-workers, [28]; first developed the support vector mechanics in mid of nineties. The applications of SVM are mainly used in areas such as classification, forecasting and regression problems which visualize superior performances compared to other model. Regression modeling has obtained its practical success with the basis of Vapnik-Chervonenkis (VC),[28]; theory, which is derived from statistical learning theory. SVM became very popular owning its superior performance in building field by resulting accurate & efficient prediction process. Furthermore, Dong and his coworkers, [27]; used SVM to predict the monthly landlord energy consumption of four real buildings in the tropical region which again proof the superiority of SVM techniques compared to other prediction models.

In this report, it is outlines the SVM as a tool to predict monthly & weekly building energy consumption and how it perform a country like Sri Lanka. Firstly, building energy consumption prediction models are established using on the SVM theory. Then the optimal parameter settings which directly affects to the performances of the SVM model, were elaborated to develop accurate and efficient forecasting system. Resulting SVM models are used to analysis the prediction capabilities of selected locations for energy prediction in Colombo, Sri Lanka & then prediction accuracies are monitored with the real data to monitor the performances of developed models.

SVM gained its popularity due to its auspicious performance and many more appealing features. Superior empirical performances of SVM can be pointed out for such raising popularity among researchers. There are many advantages of using SVM in building prediction process. Among them, the ability of SVM to embraces the principle of structure risk minimization (SRM), which engaged with the superior
qualities than the traditional empirical risk minimization (ERM) principle can pointed out. As mentioned in this report SRM principle helps in reducing an upper level of the generalization error, which consists of training error and a confident level (expected risk) based on VC dimension. SRM principle is differing from ERM principle which considers only minimizing training the error. Applications of SVR show very much efficient with effective results when tackling classification and regression problems. Furthermore, application of the kernel function in the SVMR depicts the input space into a high-dimensional feature space with its nonlinear mapping ability & performs a linear regression in the feature space, Vapnik V, [28].

SVM is based on the structural risk minimization (SRM) inductive principle which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence level. This is the difference from commonly used empirical risk minimization (ERM) principle which only minimizes the training error. Based on such induction principle, SVMs usually achieves higher generalization performance than the traditional neural networks that implement the ERM principle in solving many data mining problems.

In this study, SVMs is extended to estimate non-linear functions for regression estimation problems. The idea of generating nonlinear SVMs is to map the original input space \( X \) into a high dimensional feature space \( H \) by some function \( \phi (x) \) and then to construct a linear function in the high dimensional feature space which corresponds to a nonlinear function in the original input space.

### 4.5.1 Theory Support Vector Regression (SVR)

The ultimate goal of using SVR, to developed a model to predict the future values (Unknown outputs) by employing the know inputs. When it comes to training the SVM model, the formation of the training model is based on training data, \((X1,Y1), (X2,Y2), ..., (Xn,Yn)\), where, \( Xn \) are the input vectors, \( Yn \) is the output scalars associated with the input data. After training the model the next step is testing the model. Thereby, the SVM predict the new inputs which feed to the systems \( x_1,x_2,...,x_n \) to predict the unknown outputs \( y_1,y_2,...,y_n \). For example, let’s consider a known data
set, which are labeled as training data \( \{X_a, Y_a\} \). \( A = 1 \ldots P \), with the input data \( X_k \in \mathbb{R}^n \), and the output scalars \( Y_k \in \mathbb{R} \). With these two types of data the regression model can generate using the non-linear mapping function \( \varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^p \). This nonlinear mapping function maps the input space in to a high-dimensional feature space and constructs the linear regression in it. Below (4.1) describe the formula, [37].

\[
f(x) = \omega \cdot \varphi(x) + b
\]  

(4.1)

In this equation (4.1), the \( \omega \) represent the weight vector and \( b \) is the bias term, and \( \varphi(x) \) represents the high-dimensional feature spaces that nonlinearily mapped from the input space \( x \). In order to estimate the coefficients \( \omega \) and \( b \) its need to minimize the regularized risk function, [28]; which expressed below, (4.2).

\[
\frac{1}{2} \| \omega \|^2 + C \frac{1}{l} \sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i))
\]  

(4.2)

\( \| \omega \|^2 \) named as the regularized term which controlling the function capacity. The empirical error expressed by the second term of the equation, which measured by the \( \varepsilon \)-insensitive loss function, which is defined below (4.3);

\[
L_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i = f(x_i)| = \varepsilon, & |y_i = f(x_i)| \geq \varepsilon \\ 0 & \end{cases}
\]  

(4.3)

\( C \) is called as the cost of error (regularization constant). \( \varepsilon \) introduce a \( \varepsilon \) tube which demonstrate in the Fig. 3. Therefore, when the predicted value falls inside the tube the loss is calculated as zero.

However, if the prediction laid outside the tube, the loss is magnitude of the difference between the predicted value and the radius \( \varepsilon \) of the tube. When it comes to practice experiment these both values (\( C \& \varepsilon \)) were determined by users.
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

Figure 4.3: Parameter for the support vector regression. [28, 31]

The equation (4.4); expressed the transformation of the primal objective function in order to get the values of ω and b by introducing the positive slack variables \( \xi_i(\ast) \) denotes variables with and without (\( \ast \)).

\[
\text{Minimize} \quad \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{l} (\xi_i + \xi^*_i) \tag{4.4}
\]

Where,
\[
\begin{align*}
    y_i - \omega \cdot \varphi(x_i) - b & \leq \epsilon + \xi_i \\
    \omega \cdot \varphi(x_i) + b - y_i & \leq \epsilon + \xi^*_i, \quad i = 1, ..., l
\end{align*}
\]

\[\xi^*_i \geq 0\]

The optimization problem defined in equation, (4.4); need to transforms in to its dual formulation by using the Lagrange which expresses in equation (4.5 & 4.6); to solve more efficient way.

\[
L = \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{l} (\xi_i + \xi^*_i) - \sum_{i=1}^{l} (n_i \xi_i + \xi^*_i \eta_i) - \sum_{i=1}^{l} a_i (\epsilon + \xi_i - y_i + \omega \cdot \varphi(x_i) + b) - \sum_{i=1}^{l} a_i (\epsilon + \xi^*_i - y_i - \omega \cdot \varphi(x_i) + b) \tag{4.5}
\]

Here L means Lagrange and \( \eta, \eta^* \) are Lagrange multipliers. Hence the dual variables in (4.5) have to satisfy positive constraints, \( \eta^* , a^* \geq 0 \).
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

The resulting model for SVM regression is expressed in the equation (4.5), and $a_i, a_i^*$ represent the Lagrange multipliers.

$$f(x) = \sum_{i=1}^{l}(a_i-a_i^*)\phi(x_i) \cdot \phi(x) + b$$  \hspace{1cm} (4.6)

By introducing kernel function $K(x_i, x_j)$ which defines the dot product between $\Phi(x_i)$ and $\Phi(x_j)$ above equation (4.6), can be rewritten as follows, (4.7).

$$f(x) = \sum_{i=1}^{l}(a_i-a_i^*)K(x_i, x) + b$$  \hspace{1cm} (4.7)

Kernel functions has the capability of computing the dot product in a high-dimensional feature space by using original input data, without computing the $\phi(x)$. $K(x_i, x_j)$, the kernel function represent the inner product of two vectors in the feature space as in the equation (4.8).

The kernel, $K(x_i, x_j) = \Phi(x_i), \Phi(x_j)$  \hspace{1cm} (4.8)

The main advantage of use of kernels, they can perform all the necessary computations in input space, without mapping the $\Phi(x)$. Numbers of kernels are there, that can be used to such kind of energy prediction methods. Commonly used kernels for nonlinear regressions are, linear kernel $K(x_i, x_j) = x_i \cdot x_j$, polynomial kernel $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$ and the radial-basis function (RBF) kernel, $K(x_i, x_j) = \exp(-\gamma ||x_i-x_j||^2), \gamma>0$, where $d$ and $\gamma$ are defined as kernel parameters, Dong et al., [27].

Based on the Karush–Kuhn–Tucker (KKT), (Kuhn and tucker); conditions of quadratic programming, $b$ is calculated as follows, (4.8.1):

$$b = y_i - \sum_{i=1}^{l}(a_i-a_i^*)K(x_i, x) - \varepsilon \text{ if } 0 < a_i < C$$

$$b = y_i - \sum_{i=1}^{l}(a_i-a_i^*)K(x_i, x) + \varepsilon \text{ if } 0 < a_i^* < C$$  \hspace{1cm} (4.8.1)

The kernel, $K(x_i, x_j) = \Phi(x_i) \ast \Phi(x_j)$  \hspace{1cm} (4.9)
In addition, based on the KKT conditions of quadratic programming, only a certain number of coefficients \((a_i, a^*_i)\) in equation (4.5) will assume non-zero values. The data points associated with them have approximate errors equal to or larger than \(\varepsilon\) and are referred to as support vectors. These are the data points lying on or outside the \(\varepsilon\)-bound of the decision function. According to equation (4.5), it is evident that support vectors are the only elements of the data points that are used in determining the decision function as the coefficients \((a_i, a^*_i)\) of other data points are all equal to zero.

Generally, the larger the \(\varepsilon\), the fewer the number of support vectors and thus the sparser the representation of the solution. However, a larger \(\varepsilon\) can also depreciate the approximation accuracy placed on the training points. In this sense, \(\varepsilon\) is a trade-off between the sparseness of the representation and closeness to the data, [28].

### 4.5.2 Kernel selection

The kernel selection is a most important step in developing of the appropriate Support Vector Machine model for regression. As mentioned above there are several kernel types that can be employed when conducting such kind of energy prediction. However, when it comes to prediction of energy demand and temperature the most popular kernel is the Gaussian function, which is the function included in the RBF kernel. Many researchers including the Dong et al., has employed the RBF kernel in to their experiments. “The RBF kernel nonlinearly maps samples into a higher dimensional space, and unlike the linear kernel, it can handle the case when the relation between class labels and attributes is non-linear”. In such kind of energy prediction model the non-linear, dynamical nature of the weather on energy demand in heating, ventilation, and air conditioning systems, excludes the possibility of using the linear kernel. Another type of kernel that can be employed in such kind of energy prediction is the polynomial kernel. When it comes to polynomial kernel it has many more hyper-parameters than the RBF kernel which leads to an end result of increasing the complexity of the forecasting model.
Therefore, when considering the kernel selection to such kind of energy prediction, the left best option is RBF kernel which has less “numerical difficulties” than the polynomial kernel which results with less of a tendency to produce values approaching zero and infinity. However, when it comes to polynomial kernel, it creates a less restrained curve. Even though the sigmoid kernel behaves like RBF for certain parameters, Qiong li et al., [31]; the validity of the kernel is questionable under certain parameters. All above justifications pave the path to proof the best option as, RBF kernel for such kind of experiment.

4.5.3 Normalizing parameters

Before start the SVM regression it’s always beneficial to avoid the unnecessary calculations by improving the calculation efficiency. In order to avoid such kind of heavy calculations and perform a smooth process the input data and output data need to be normalized as follows Qiong li et al., [31].

\[ V'_i = \frac{V_i - V_{Min}}{V_{Max} - V_{Min}} \]
\[ q'_i = \frac{q_i - q_{Min}}{q_{Max} - q_{Min}} \] (4.10)

In the above equation (4.10) parameter \( v_i \) represent each input parameter, including the outdoor temperature, relative humidity with solar radiation. More over \( q_i \) represent, the building energy load. From the maximum & minimum terms of the above equation, they all represent corresponding minimum and maximum values the energy data, outdoor temperature, relative humidity, solar radiation data. From such methodology the equation (4.10) can be used to calculate the normalize data as an input factor to the SVM model. After normalizing the data as a preparation technique to select the appropriate kernel function, SVM model need to be established by the nonlinear relation between the building energy load and selected parameters. When the prediction is completed, the resultant output \( Y \) from SVM model, needs be converted into the actual prediction, Qiong li et al., [31]; value as shown in equation (4.11).

\[ \hat{q} = q_{min} + Y \cdot (q_{max} - q_{min}) \] (4.11)
4.6 Model Development for Monthly & Weekly Load Prediction

Figure 4.4 demonstrate the proposed model for load forecasting to both monthly & daily data. Thereby, firstly the data are separated in to two groups called training and testing data. After the separation the next step is to normalize the data by using pre-preparation techniques in order to avoid large numerical calculations. Then, these normalized training data used to model optimization which describe in section 4.5.3. After selecting the optimized parameters the next step is to create time lags in order to remove the storage effect. However, when it comes to daily load forecasting the storage effect significantly affect to forecasting model than the monthly load forecasting as stated in, section 4.6.5. After creating the time delays and parameter optimization the forecasting model is developed by using support vector regression on the training data set. Next, using this developed forecasting model with the testing data which obtained by splitting the data used to develop the prediction model. Finally, the prediction process conducted and the prediction results are obtained to monthly and daily basis as per the feeding data type (monthly or daily) and the performance of the models are measured by error measurement techniques as mentioned in the section 4.6.3.

4.6.1 Pre processing of collected data

As describe in section 4.4.3, the monthly energy demand data was collected from four peer office buildings, which were selected by analyzing the data availability among all the commercial buildings around Colombo, Sri Lanka. (The data collection period was monthly data from January 2010 to November 2013). For the simplicity, these buildings were assigned as A-D which named by the location wise as A-Town hall, B- Pettah, C- City, D- Bambalapitiya. January 2010 to December 2012 were chosen as the training years, while keeping year 2013 as the test year. In the Table 4.1 below, it summarizes the building physical details such as area and the average annual energy usage of these selected buildings.
4.6.1 Weather data

According to the previous studies T, RH and SR found to be the main parameters, that cause significant effects on the building energy consumption. Weather data was collected from the metrological department, (shows in figure 4.2), Colombo because of its accuracy & relativity to the building locations. Moreover to remove the year to year changes in air-conditioned areas are as affection and human capital changes of the buildings the research finding of the Fels and coworkers, [1]; are applied. Thereby, they evaluate a proportional relationship between annual daily energy usage and changes in air-conditioned area. Thus, for a particular year, the normalized area-changed energy usage is nearly resembles to the annual mean energy usage divided by the gross floor area as the air-conditioning facilitating not only the office areas but also most of the common areas in these four buildings. This study mainly focused on landlord energy consumption as it likely removes the affects coming from the population alteration. Thus, made an assumption that the normalized energy usage by these buildings are sufficiently enough for above mentioned changes.

4.6.1.2 Model identification

With the kernel parameter $\gamma$, the parameters $C$, and $\varepsilon$ were considered to be the performance parameters of the final prediction model. The prediction model was elaborated using a stepwise approach, in order to select the optimum parameters when the model is in its training phase.
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

Figure 4.4: Proposed model to energy forecasting.
4.6.2 Formatting Data:

Two sets of data need to Support Vector Machine Regression require to make its regression. They are namely the training data set and the test data set. As the first step to feed the data into the program a Mat lab code was written to format the data into two sets for use in SVM regression. The training set contains all of the data available. The training data set is used, along with the parameters modification & selecting the optimal, to build a multi-dimensional models and apply the kernel function. The first column of data in the training set, called the y values, contains a list of energy values beginning with the most recent and going back in time. The ultimate goal of the SVM regression is to output the future y values so the y values must be the same type of data that will be predicted (in this case, energy demand). When it comes to the testing data unlike the training set, the test set only contains the most recent data. It is used in conjunction with the model to predict future values. When Support Vector Machine regression is used, the y values predicted by the model are the values that would fit best as y values in the test set, [31].

4.6.3 Evaluation indices

The criterion used to select the most appropriate model is to maximize the goodness of-fit using the simplest model or combination of models, Draper and Smith, [52]. For the non-linear modeling, use the definitions in chapter 3, section 3.1 to evaluate the prediction performance namely, mean squared error (MSE), mean squared error of scaled value (S-MSE), % error, R-MSE and coefficient of variance based on root mean squared error (CV-RMSE).

The smaller the MSE and % error value, the closer are the predicted values to the actual values. CV-RMSE is a non-dimensional measure that is found by dividing RMSE by the mean value of Y. It is usually presented as a percentage. Hence, a CV-RMSE value of 5% would indicate that the mean variation in Y not explained by the regression model is only 5% of the mean value of Y, [41]. In addition, during the
model identification and processing period, S-MSEs are compared only and enough to decide the best model.

In order to measure the accuracy of the develop model there need to be some kind of evaluation method. In this paper to measure the performance of SVM models the following techniques were applied. They are root mean square error (RMSE), MSE and % error which are define in following equations.

\[
MRE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\% \quad (4.12)
\]

\(\hat{Y}_i\) is the prediction value of \(Y_i\)

The percentage error defined as follows.

\[
\text{% error} = \frac{Y - \hat{Y}}{Y} \times 100 \quad (4.13)
\]

Moreover, for further evaluations of these models following statistical metrics also used, namely, coefficient of variance based on root mean squared error (CV-RMSE). The CV-RMSE is defined below,[27]:

\[
CV - RMSE = \frac{RMSE}{\bar{Y}} \times 100
\]

Where, \(RMSE = \sqrt{[MSE]^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)^2} \times 100\% \quad (4.14)
\]

Here, \(N\) is number of observations.

As mentioned early in this paper, in order to apply the Support Vector Machine to the training data, number of parameters must be selected. The kernel function, \(C\) parameter \(\varepsilon\) and \(\gamma\) parameter specified in the training phase determine the performance of the model prediction.

Therefore, if wrong selection of the above mentioned parameters will leads to totally misleading models. In order to create the most effective model, the parameters need to select to create the best “goodness-of-fit”. When the experiment is going to
conduct to find out the best-fit need to make sure to not to make the model too computationally expensive to run on a personal computer.

4.6.4. SVM algorithm development.

When developing the SVM algorithm its need to find out the best platform to run the program. There are two different software packages, which can be used to implement the SVM algorithm in order to run SVM regressions on a personal computer. There are, Both LIBSVM and SVM\textsuperscript{light} the both software’s requirements were similar means the both need the same data format. Furthermore, these software divide the regression into a training function that trains the model and a testing function that is used to validate the model and predict into the future. When selecting the software package, the LIBSVM was chosen because the SVM\textsuperscript{light} training function took much longer to run when computationally expensive hyper parameters were specified. The average running time of LIBSVM was a few minutes, while SVM\textsuperscript{light} often took over an hour. Therefore, use on personal computers, LIBSVM was the clear choice software package, [28].

4.6.4.1 Variable selection

When it comes to model creation, the variable selection also plays a crucial role. Therefore to select the best combination of variable to the model, there need to have selection which combination of weather and energy variables would create the most efficient and accurate model. Therefore, to select the best models with variables, three different years of energy data and corresponding years of weather data were employed & conduct the experiment.

When selecting variables for such an experiment, to country like Sri Lanka, the most affecting parameters are temperature, relative humidity, dew point, wind speed, and solar radiation. Wind speed was eliminated as a possible weather variable from the model due to the low impact of these variables on energy use in heating, ventilation, and air conditioning which makes low impact on energy consumption of a building.
However, relative humidity and dew point temperature are two similar definitions which can be derived from each other. When employing similar variable to the model which can be derived from each other would introduce redundancy in the model by increasing the model complexity. Relative humidity can be defined as the ratio of water pressure in an air/water mixture. The main impact of relative humidity its changes with pressure and temperature, unlike the absolute humidity which only measurement of water content in air. Moreover, the dew point temperature, defined as “is the temperature to which a given parcel of humid air must be cooled, at constant barometric pressure, for water vapor to condense into water”. Therefore, from above two definitions to relative humidity & dew point temperature, its implies that, the both relative humidity and dew point take into account temperature, moisture content in air, and pressure. To this experiment, the relative humidity was chosen with the accuracy of data availability of the meteorological department.

Therefore, the 4 variables available to create the model were three separate years of energy, temperature, relative humidity and solar radiation. These variables were then ranked in terms of importance to the predictive model based on the variables. The variables were prioritized as 1st year of energy, 2nd year of energy, temperature, relative humidity and solar radiation. After selecting the variables it’s a need to select the best variable combinations to find out the best model. The LIBSVM-2.6, software was used to train and test data. Training data was used, starting January 2010 and ending December 2012. The following year was used as test data. Default parameters were used for each regression, since those values had not yet been specified. LIBSVM-2.6 produces an R-square and S-MSE value after running the testing function. These values can be used to validate the accuracy of the model if the class labels (the prediction variable values for SVMR) in the test set are the true values. The output is not a prediction but rather a comparison of the model’s performances to predict with the actual testing values. The methods section on the prediction cases outlines how replacing the class labels with random values creates the predictive capability of the model. However, in the case of validation of the model variables and parameters, the real values are used as class labels.
4.6.4.2 Parameter characteristics of SVM

When it comes to parameter modification of SVMR, the most important parameters are the penalty parameter $C$, the $\varepsilon$-insensitive loss function and $\gamma$ of Gaussian kernel. If chose a very small value of $C$, it will case to under-fit the training data. This is because weights placed on the training data are too small which would leads to result in large values of prediction error on the test sets. However, it’s not recommended to select too high values for penalty parameter $C$, which will leads to SVM to over-fit the training set. If placed a such a high value on the training set $C$ over fit with the data sets and the goes back to minimize the empirical risk which differs from the properties of structural risk minimization principle. Moreover, selecting larger $C$ causes to a larger range of the value of support vectors, by resulting need of more data points as support vectors in the optimization formulation, Dong et al., [27]. After selecting the penalty parameter, the other important parameter is $\varepsilon$. If the $\varepsilon$ is larger value according to the figure 4.8, the number of support vectors is fewer.

However, selection of too large $\varepsilon$ can leads to reduces the accuracy levels on the training data. Even though, $C$ & $\varepsilon$ plays a crucial role, there is no way to find the optimal values of $C$ and $\varepsilon$ are which will lead to optimal model design before conduct the experiments. In order to fund the best parameters which give the superior performances, some kind of parameter search must be done. There by can find an optimal values of $C$ and $\varepsilon$ so that the model can accurately predict the unknown data with high accuracy levels. Most of the past researches grid search method was conducted in order to find out the best $C$ & $\varepsilon$. However, the grid search method is a cost complex one with considerable reliability which tires all pairs of $(C, \varepsilon)$ to find out the best performances. In Bing et al., 2005,[27]; Qiong et al., 2009, [31]; its pointed out that use of exponential growing sequence ($C = 2^{-5}, 2^{-2}…2^{10}, \varepsilon = 2^{-10}…2^{-5}$) is a practical identify the best suit values for $C$ & $\varepsilon$. Dong et al., [27]; in his research developed a new searching method called the stepwise searching method to find out the performance of SVM and select the optimum values for $C$ & $\varepsilon$. In this research, in order to determine the best values of parameter and $\varepsilon$, the simulation processes of SVM load prediction model with various parameter settings were
analyzed. Firstly, in this process one time search method was used to find out the Mean Square Error (MSE)-1. Then, the experimental procedure was conducted on C (By fixing the first result of $\varepsilon$) and $\varepsilon$ (By fixing the second result of C) to get the second lowest MSE.

Likewise, the experimental procedure was conducted by fixing the $\varepsilon$ to be 0.1, and vary the value of C from $2^{-5}$ with an exponential growth till $2^5$ to train the SVM model using the training sample from year 2010 to 2012 data. The results of the prediction errors of MRE, and the number of support vectors are shown in figure 4.5.

From figure 4.5, its implies that the number of support vectors increases slightly with C. Furthermore, in figure 4.6, it shows for parameter C, there is a lowest MSE point for each and every building which pave the path of selecting the best value C. In figure 4.6, the S-MSE first decrease slightly when increasing the parameter C, and then increase after the reaching to a minimum value. The results, for the optimization process are listed as shown in table 4.1.

After obtaining minimum value of MSE to select the optimum value of C, the value of C is fixed. Then the $\varepsilon$ set at various values between $2^{-10}$ and $2^{-1}$, and trains the SVM model using the training data. The results are shown in figure 4.7. From figure 4.7, it can be seen that the MSE firstly almost remain constant, and then suddenly go up largely when $\varepsilon$ is between 0.03125 and 0.5.

However, figure 4.8, show that, the number of support vectors decreases largely when the $\varepsilon$ increases by reaching to zero. From this behavior it can conclude that $\varepsilon$ not plays a significant role when deciding the performance of the SVM model greatly. However, its affects to number of support vectors by showing a decreasing the number when the value of $\varepsilon$ increases largely.
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

Figure 4.5: Results of C value with MSE

Figure 4.6: Results of C value with NSV

Table 4.2: Results for C, ε, γ after SVMR–Buildings (A, B, C, D)

<table>
<thead>
<tr>
<th>Building</th>
<th>Type</th>
<th>C</th>
<th>ε</th>
<th>γ</th>
<th>NSV</th>
<th>% error</th>
<th>MSE</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Townhall</td>
<td>A</td>
<td>32</td>
<td>0.0625</td>
<td>0.125</td>
<td>30</td>
<td>-2.50</td>
<td>0.062</td>
<td>5.19</td>
</tr>
<tr>
<td>Pettah</td>
<td>B</td>
<td>0.5</td>
<td>0.125</td>
<td>0.125</td>
<td>23</td>
<td>-1.81</td>
<td>0.018</td>
<td>3.72</td>
</tr>
<tr>
<td>City</td>
<td>C</td>
<td>2</td>
<td>0.125</td>
<td>2</td>
<td>24</td>
<td>-1.49</td>
<td>0.025</td>
<td>4.07</td>
</tr>
<tr>
<td>Bambalapitiya</td>
<td>D</td>
<td>2</td>
<td>0.0625</td>
<td>0.125</td>
<td>28</td>
<td>2.09</td>
<td>0.033</td>
<td>3.67</td>
</tr>
</tbody>
</table>
Finally, to find out the best $\epsilon$ value the prediction error and the number of support vectors were considered. The choose parameters were shown in table 4.2.
4.6.4.3 Modification of the width parameter $\gamma$ of Gaussian kernel

When it comes to high dimensional feature space, the width parameter $\gamma$ of Gaussian kernel directly defines the structure of the high-dimensional feature space $\Phi(X)$. Furthermore, the Gaussian kernel can control the complexity of the final solutions by changing the complexity of the model. High value for $\gamma$ means the ultimate results will be a narrow kernel, by reducing the prediction error. If the value for $\gamma$ is too large it will result in increased prediction error by reducing the accuracy of the model. Therefore, the value of $\gamma$ plays a crucial role and it needs to be selected suitably according to the regression problems. To get an optimum value for $\gamma$ value, the stepwise search method is conducted again by setting the C and $\varepsilon$ as per the table no 2. The simulation was carried out for SVM load prediction model with various values of $\gamma$ between $2^{-1}$ and $2^{8}$. Due to difficulty of graphical representation of full scale of optimization, only critically affected ranges are included in this report. As done in parameter modification for C & $\gamma$ values prediction errors of MSE and % error and the number of support vectors are recorded. The results are shown in Figure 4.9. From figure 4.9, we can see that the best Results of $\gamma$ for this load prediction problem, [31].

Figure 4.9: Results of $\gamma$ value with MSE
4.6.5 Storage effects analysis to prediction

As mentioned earlier in this paper, when selecting the variables to model optimization, it’s also need to consider about the delay effects of outdoor temperature, solar radiation & relative humidity to the building energy consumption, especially on the building cooling load as it’s critically affects to the energy consumption. In order to find of the sensitivity effort (Storage effect) the model need to evaluate by using the data of the past 2 or 3 months in history as the input parameters, [31]. When it’s considering the sensitivity analysis the storage effects maybe go more than 3 month in history. In order to analyze the storage effects, the experiment was conducted for different five situations with different combinations of input parameters. The experimental combinations used for the analysis are shown in table 4.3.

![Figure 4.10: Results of γ value with NSV](image)

For these five cases, SVM model with parameters $C$, $\varepsilon$ and $\gamma$ determined above is established on the basis of the environmental factors and building energy load. And the established SVM model is used to predict the monthly building energy load. Table 4.3. Show the MSE of the training sample and testing sample under various cases. The table 4.4; shows that for case 1, only the outdoor temperature, relative
humidity & solar radiation of the past 1 in month is employed, the RMSE & % error is relatively larger than the other four cases. However, when the data of the past two or more months (cases 2–5,) in historical data considered to develop the model the RMSE and & % error are almost the same by resulting the minimum MSE in case 2.

Therefore, the input parameters were selected as like case 2, to gain a higher accuracy with simple model for prediction process.

Table 4.3: Historical monthly data for Input parameters setting.

<table>
<thead>
<tr>
<th>No.</th>
<th>Input Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>t1,RH,T,SR,O(t1,t2,t3)</td>
</tr>
<tr>
<td>Case 2</td>
<td>t1,t2,t3,RH,T,SR</td>
</tr>
<tr>
<td>Case 3</td>
<td>t1,t2,t3,RH,T,SR,O(t1)</td>
</tr>
<tr>
<td>Case 4</td>
<td>t1,t2,t3,RH,T,SR,O(t1,t2)</td>
</tr>
<tr>
<td>Case 5</td>
<td>t1,t2,t3,RH,T,SR,O(t1,t2,t3)</td>
</tr>
<tr>
<td>t1-3</td>
<td>Past Energy Data</td>
</tr>
<tr>
<td>O(t1--3)</td>
<td>Past environmental data</td>
</tr>
</tbody>
</table>

Table 4.4: The prediction errors of SVM models in different cases.

<table>
<thead>
<tr>
<th>No.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>0.0537</td>
</tr>
<tr>
<td>Case 2</td>
<td>0.0284</td>
</tr>
<tr>
<td>Case 3</td>
<td>0.0335</td>
</tr>
<tr>
<td>Case 4</td>
<td>0.0344</td>
</tr>
<tr>
<td>Case 5</td>
<td>0.0319</td>
</tr>
</tbody>
</table>

4.7 Results of SVM Prediction for Energy Load Forecasting

After parameters C, ε and γ are selected the selection of input parameters are finalized. There by, the SVM model is established with the environmental data and building energy load. The Mat lab R-2009a Toolbox is used to train and develop the SVM model to all four buildings energy load prediction.
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

Figure 4.11: Results of SVM prediction for Building A

Figure 4.12: Results of SVM prediction for Building B
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

Figure 4.13: SVM prediction for Building C

Figure 4.14: SVM prediction for Building D
4.7.1 Results of SVM for prediction of landlord energy consumption

The summary of results of SVM properties selections are shown in table 4.2. Table 4.2 show that Building A has the highest MSE of 0.062 & Building B has the lowest MSE of 0.018. When it comparing the number of support vectors, they are different to each building with the parameter C and γ & ε. When it’s comparing the CVs for all four buildings, representing the variances from the actual, are very small by resulting less than a value of 5%.

4.7.2 Validation of the results

Table 4.5: Comparison of results with standards

<table>
<thead>
<tr>
<th>Performance Measurement</th>
<th>ASHRAE Standard</th>
<th>Experimental Results</th>
<th>Literature Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV %</td>
<td>Less Than 10%</td>
<td>Less Than 5% for all 4 Buildings</td>
<td>3% Bing Dong et al-2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8% Qing Li et al-2009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5% Salomon et al-2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11% Rishie et al-2014</td>
</tr>
<tr>
<td>MSE</td>
<td>Minimum Value</td>
<td>Less Than 0.47 for 3 buildings</td>
<td>0.7 Bing Dong et all</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0 Qing Li et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Salomon et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rishie et al-2014</td>
</tr>
<tr>
<td>%Error</td>
<td>The Minimum Value</td>
<td>Within 2.5%</td>
<td>Within 4% Bing Dong et all</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Qing Li et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Salomon et al</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rishie et al-2014</td>
</tr>
</tbody>
</table>

From above results it can conclude that all the SVM models can be considered as accurate models for predicting the future energy consumption, Fels et al., 1986, [1]; Furthermore, compared with other research conducted using other techniques to predict the energy consumption, these models with SVM shows high prediction accuracy levels.

For example neural networks (NN)) Kreider and his co-workers, [47]; worked on chilled water shows much higher CVs more than 10%. However, in this prediction
Chapter 4: Baseline models of building monthly & daily landlord energy consumption

the results shows, the highest CV of 5.19% appears to be in the Building A, while the lowest CV of 3.67% appears in Building B which shows high level of accuracy on the prediction process.

4.7.3 Residual Values Calculation

The performances of residues of four buildings are shown in figure 4.15. In ordinary least square (OLS) regression, the undesirable performance of residues is always problem, [27]. In this study, it seems the same problem occurred again, which shows that the distributions of residues are not constant. However, since it appears randomly, the prediction model can be considered to be consistent.

![Residual Values](image)

**Figure 4.15: Residues of estimated landlord energy consumption**

4.7.4 Discussion and conclusion

SVM is used to predict landlord energy consumption in this study. With the gain popularity, SVMs is applied in the research of building load estimation. The performance of SVMs, in terms of MSE and % error, is proof better than the other established modeling tool used so far, including neural networks and genetic programming.
The results of comparison are shown in Table 4.5. It shows that in terms of performance measurements, the results from SVMs are much better than those from past researches carried out so far. Table 4.5 Results of comparison on baseline landlord energy consumption with three performance measurement techniques.

The results of MSE, % error, CV from SVMs in this research are much better than those from others, which indicate the strong tracking ability of SVMs. The reasons of outstanding results of SVMs can named as follows:

(1) Consideration of gamma, kernel parameter optimization.
(2) Developing time lags to both past energy consumption & environmental conditions to input vector can be pointed out.

4.8 Baseline Models of Building Weekly Landlord Energy Consumption

4.8.1 Data collection

Figure 4.16: Energy demand data profile of four years.

To develop an accurate weekly model for prediction process office building, E; is selected, among the Colombo city, Sri Lanka and figure 4.16 shows daily energy consumption values for the selected buildings for the modeling process. Weather
has a significant effect on the amount of energy required to functions like air-conditioning heating and cooling system of a building. Figure 4.17, shows, the selected weather parameters (temperature, relative humidity, solar radiation) in Colombo region, Sri Lanka for a period of 4 years.

### 4.8.2 Preprocessing of collected data

The energy demand data was collected from the office building, which were selected by analyzing the data availability among all the official buildings around Colombo, Sri Lanka. (The data period was monthly data from July 2011 to March 2014).

![Figure 4.17: Weather data profile of four years](image)

July 2011 to February 2014 are chosen as the training years, while keeping March 2014 as the test Month. In the table 4.6 below, it summarizes the building physical details such as area and the average annual energy usage of the building.

### 4.8.3 Weather Data

As per the monthly load forecasting T, RH and SR found to be the main parameters that cause significant effects on the building energy consumption. Weather data was
collected from the metrological department, Colombo because of its accuracy & relativity to the building locations. Same procedures are followed in order to remove the year to year changes in air-conditioned area & population changes. Thus, made same assumption that the normalized landlord energy usage by these buildings are sufficiently enough for above mentioned changes.

4.8.4 Model identification

With the kernel parameter $\gamma$, the parameters $C$, and $\varepsilon$ were considered to be the performance parameters of the final prediction model. The prediction model was elaborated using a stepwise approach, in order to select the optimum parameters when the model is in its training phase.

Table 4.6: Total energy consumption of four buildings

<table>
<thead>
<tr>
<th>Building</th>
<th>Training Year</th>
<th>Test Month</th>
<th>Gross Floor Area (m$^2$)</th>
<th>Landlord energy Consumption (kWh)/Yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>July 2011 to Feb 2014</td>
<td>March 2014</td>
<td>450.67</td>
<td>98,127</td>
</tr>
</tbody>
</table>

4.8.4.1 Formatting Data

The needed two sets of data to Support Vector Machine Regression namely the training data set and the test data set. As the first step to feed the data in to the program a Mat lab code was written to format the data into two sets for use in SVM regression. The training set contains all of the data available. The training data set is used, along with the parameters modification & selecting the optimal, to build a multi-dimensional models and apply the kernel function, [27].

4.8.4.2 Evaluation indices

The criterion used to evaluate the monthly load forecasting isselected in weekly forecasting too. For the non-linear modeling, the prediction performance is evaluated
using the following statistical metrics, namely, mean squared error (MSE), mean squared error of scaled value (S-MSE), percentage error (% error) and coefficient of variance based on root mean squared error (CV-RMSE).

Therefore, the 4 variables available to create the model were three separate years of energy, temperature, relative humidity and solar radiation. These variables were then ranked in terms of importance to the predictive model based on the variables. The variables were prioritized as 1st year of energy, 2nd year of energy, temperature, relative humidity and solar radiation. After selecting the variables it’s a need to select the best variable combinations to find out the best model. The LIBSVM-2.6 software was used to train and test data. Training data was used, starting July 2011 and ending February 2014. The following month was used as test data. Default parameters were used for each regression, since those values had not yet been specified. LIBSVM-2.6 produces an R-square and MSE value after running the testing function.

4.8.4.3 Parameter characteristics of SVM

As mentioned in section 4.8.4, the most important parameters are the penalty parameter C, the ε-insensitive loss function and γ of Gaussian kernel. In the short term load forecasting also, in order to determine the best values of parameter C, γ & ε the simulation processes of SVM load prediction model with various parameter settings were analyzed. Firstly, in this process one time search method was used to find out the Mean Square Error (MSE)-1. Then, the experimental procedure was conducted on C (By fixing the first result of ε) and ε (By fixing the second result of C) to get the second lowest MSE2. Likewise, the experimental procedures are conducted by fixing the ε and vary the value of C to train the SVM model using the training sample from year 2011 to 2014 data. The results of the prediction errors of MSE, and the C values are shown in figure 4.18.
Figure 4.18: Results of SVM prediction for Building E (Borella)

Figure 4.19: Results of SVM prediction for Building E (Borella)

Again it implies from figure 4.18, that the number of support vectors increases slightly with C. Furthermore, in figure 4.18, it shows for parameter C, there is a lowest MSE point for each and every building which pave the path of selecting the best value C. In figure 4.18, the MSE first decrease slightly when increasing the parameter C, and then increase after the reaching to a minimum value. The results, for the optimization process are listed as shown in table no 4.7.
After obtaining minimum value of MSE to select the optimum value of C, the value of C is fixed. Then the ε set at various and trains the SVM model using the training data. The results are shown in figure 4.20 & 4.21.

From figure 4.20, it can be seen that the MSE firstly almost remain constant, and then suddenly go up largely when ε is between 0.03125 and 0.5.
However, figure 4.21 shows that, the number of support vectors decreases largely when the $\varepsilon$ increases by reaching to zero. Again its proves that the $\varepsilon$ not plays a significant role when deciding the performance of the SVM model greatly. However, its affects to number of support vectors by showing a decreasing the number when the value of $\varepsilon$ increases largely.

4.8.4.4 Modification of the width parameter $\gamma$ of Gaussian kernel

The same procedures were followed as mentioned in section 4.8.4 to get an optimum value to $\gamma$ value. Here also followed the stepwise search method by setting the $C$ and $\varepsilon$ as per the table no 4.7.

![Figure 4.22: Results of SVM prediction for Building E (Borella)](image)

Table 4.7: Results for $C$, $\varepsilon$, $\gamma$ after SVMR – Building (E)

<table>
<thead>
<tr>
<th>Building</th>
<th>Type</th>
<th>$C$</th>
<th>$\varepsilon$</th>
<th>$\gamma$</th>
<th>nSV</th>
<th>% error</th>
<th>MSE</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borella</td>
<td>E</td>
<td>1</td>
<td>0.125</td>
<td>0.25</td>
<td>113</td>
<td>1.44</td>
<td>0.0036</td>
<td>8.16</td>
</tr>
</tbody>
</table>


4.8.4.5 Storage effects analysis to prediction

To overcome the sensitivity effort (storage effect) the model need to evaluate by using the data of the past 2 or 3 months in history as the input parameters. When it’s considering the sensitivity analysis the storage effects maybe go more than 3 month in history, [27].

In order to analyze the storage effects, the experiment was conducted for different four situations with different combinations of input parameters. The experimental combinations used for the analysis are shown in Table 4.8 & 4.9.

Table 4.8: Results for Storage effect to Building (E)

<table>
<thead>
<tr>
<th>No.</th>
<th>Input Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>t1,RH,T,SR,O(t1,t2,t3)</td>
</tr>
<tr>
<td>Case 2</td>
<td>t1,t2,t3,RH,T,SR</td>
</tr>
<tr>
<td>Case 3</td>
<td>t1,t2,t3,RH,T,SR,O(t1)</td>
</tr>
<tr>
<td>Case 4</td>
<td>t1,t2,t3,RH,T,SR,O(t1,t2)</td>
</tr>
</tbody>
</table>

| t1-3 | Past Energy Data       |
| O(t1--3) | Past environmental   |
Table 4.9: Results for storage effect to Building (E)

<table>
<thead>
<tr>
<th>No.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>0.0642</td>
</tr>
<tr>
<td>Case 2</td>
<td>0.0241</td>
</tr>
<tr>
<td>Case 3</td>
<td>0.0293</td>
</tr>
<tr>
<td>Case 4</td>
<td>0.0254</td>
</tr>
</tbody>
</table>

4.8.5 Results of SVM for prediction of landlord energy consumption

The summary of results of SVM parameter properties selections are shown in table 4.7. Table 4.7; shows that Building E has the MSE of 0.074. When it’s comparing the % error for the building, representing the variances from the actual, are very small by resulting less than a value of 2%.

Figure 4.24: Results of SVM prediction for Building E (Borella)

From above results it can conclude that all the SVM models can be considered as accurate models for predicting the future energy consumption, Greg et al., [15]. Furthermore, compared with other research conducted using other techniques to predict the energy consumption, these models with SVM shows high prediction accuracy levels. When compare to monthly load prediction, in this prediction the
results shows, the highest CV of with high level of accuracy on the prediction process. The main reason can be the data pool used to training stage is significantly larger & the weather parameters effects also considerable amount compared to monthly load forecasting.

4.8.6 Residual values calculation

In this study, it seems the same problem occurred again, which shows that the distributions of residues are not constant. However, since it appears randomly, the prediction model can be considered to be consistent.

Figure 4.25: Residues of estimated landlord energy consumption, (E)

4.9 Observations of the Experiment on Both Daily & Monthly Load Forecasting

SVM is used to predict landlord energy consumption in this study. With the gain popularity, SVMs is applied in the research of building load estimation. The performance of SVMs, in terms of CV and % error, is proof better than the other established modeling tool used so far, including neural networks and genetic programming.
The results of comparison are shown in table 4.5. It shows that in terms of performance measurements, the results from SVMs are much better than those from past researches carried out so far. Table 4.5 Results of comparison on baseline landlord energy consumption with three performance measurement techniques. The results of CV from SVMs in this research are much better than those from others, which indicate the strong tracking ability of SVMs. The reasons of outstanding results of SVMs can named as follows:

(1) Consideration of gamma, kernel parameter optimization.
(2) When creating time lags both past energy consumption & environmental conditions to input vector can be pointed out.
Furthermore, with above observation the below outstanding features of SVM also caused to such an improvement of results.

(1) Structural risk minimization (SRM) principle, which is the most outstanding feature of SVM, is implemented to minimize the upper bound of the generalization error rather than the training error, which is applied in other learning methods such as NN.

(2) There are fewer free parameters to optimize compared to neural network and genetic programming. As investigated in this study, only parameter C, γ and ε are important parameters to the prediction models. However, for the neural networks, there are lots of free parameters needed to adjust such as number of neurons in the hidden layers, the learning rate, number of epochs, the stop criteria and the transfer functions. Furthermore, NN can never reach a global solution. However, the solution of SVM is unique and optimal because SVM is like solving a linearly constrained quadratic programming.

Furthermore, a stepwise search is employed in this study, which is more reliable yet simple. The final results demonstrated that SVMs is feasible and applicable in prediction of monthly landlord utility bills in the tropical region. Moreover, the application of this methodology is not limited to only the tropical region based on its
strongly theoretical background and regression characters. Since SVMs presents many advantages in prediction, short-term (weekly) load data also explored in this research and found the prediction in commercial building also gives high accuracy results.

Furthermore, when it compared to monthly load forecasting in section 4.4, again it proves that the short term load forecasting gives high performances with low error results that the monthly model. The reason for such prediction can be pointed out as the large daily data input to the model allows high training facility to model rather than the monthly one. Further the effect of environmental factors to daily load forecasting may be much higher than the monthly period as in monthly load forecasting it takes the mean values to all predictions inputs. Table 4.10 shows the accuracy levels of short term and medium term load forecasting results.

<table>
<thead>
<tr>
<th>Performance measurement</th>
<th>Medium term</th>
<th>Short term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.018</td>
<td>0.0036</td>
</tr>
<tr>
<td>% error</td>
<td>within 2%</td>
<td>within 2%</td>
</tr>
</tbody>
</table>

4.10 Implication

Two main analysis, support vectors machines, for baselining building (monthly & weekly) landlord energy consumption are presented and discussed in this chapter. The applications of these two systems are feasible and applicable. Furthermore, SVMs is verified to be better method for regression estimation of non-linear performance data pool. The results show that the development of baseline models of building landlord energy consumption is successful and the errors are in secure. However, more buildings are needed to constitute a larger data pool and consequently, more reliable and confident for the development of a standard baseline method of landlord energy consumption.
CHAPTER 5
CONCLUSION

Five models (four with monthly electricity data & one with daily electricity data) are innovated, to predict the energy demand for commercial buildings in Colombo city, Sri Lanka. The CVs, MSE and % error values for the final regression indicate the accuracy of the model to predict energy demand [27]. Moreover, lower MSE value will provide a good estimate of how the building will behave in the future and accordingly can suggest optimistic recommendations for improvements.

5.1 Achievements of research objectives

The objectives of the thesis are:
1) To analyze the existing methods for building energy consumption applied in the tropical region

2) To explore and establish Support Vector Machine in building energy consumption in tropical region.

The results show that all the predictions for monthly load forecasting, CVs are less than 5% & for daily load forecasting nearly 8%. Low CV means, Qiong et al., [31] a good tracking ability along the monthly and daily bills, which is what building owners want to know. This method is verified to have good prediction ability, and is sufficient to establish the baseline models for landlord energy consumption.

Moreover, when this is compared to monthly load forecasting in chapter 4, section 4.9; it proves that the short term load forecasting gives high performances with low MSE results than the monthly model. The reason for such prediction can be pointed out as the large daily data input to the model allowing it for high training facility to model rather than the monthly one. Further, the effect of environmental factors to daily load forecasting are much significant, Dong et al.,[27], than the monthly period.
Chapter 5 Conclusion

as in monthly load forecasting since it takes the mean value of all data as predictions inputs. Table 4.10 in chapter 4, shows the accuracy level of short term and medium term load forecasting results. Furthermore, for all five buildings the % errors are within 2% which shows the accuracy of all five models, [27].

Furthermore, when doing such kind of research the seasonal variation & the fluctuation of the energy usage need to study and the prediction need to carried out for high fluctuation of energy consumption in buildings rather than choosing small variation (section 3.7.1) of energy consumption in buildings. The main reason for that is, it cannot be model a system with high accuracy, if the energy consumption variation is a small value and if there is a seasonal pattern of the energy consumption the prediction will follow the pattern of energy consumption rather than predicting the future consumption.

5.2 Contributions of the study

1) A comprehensive review of the literature in the baseline model methods for topics with direction in future research.

2) A modelling method for building electricity consumption for commercial building in tropical region for both short term & medium term forecasting.

5.3 Recommendations for further research

The development of baseline model is very important. It is not limited to utility bill tracking or prediction although it is the easiest data to obtain. The baseline model plays a significant role in the development of measurement & verification protocol. Hence, there are few recommendations for future research.

By examining the feasibility of more independent variables such as occupancy rate, extension of office area, it can be used to increase the performance of the model. Although most recent studies on the development of baseline models are focusing on
the weather data variables, the variable such as occupancy, daily extension too have impact on load prediction. However, modeling by using such input is difficult due to the difficulty of quantifying the relationship of these variables with electricity consumption. Moreover, than occupancy, there are other potential variables such as equipment operation hours and frequent increment of conditioned room area. A reliable and robust baseline model should include all important factors that affect the building electrical consumption in order to improve the accuracy of the model. Therefore, future research should emphasize on performance driven by considering above facts to develop an improved electricity forecasting models.
APPENDIX A: AN OVERVIEW OF STATISTICAL LEARNING THEORY
APPENDIX A: AN OVERVIEW OF STATISTICAL LEARNING THEORY

Vapnik (1995) presents all the machine learning problems as the following:

Given a set of data points \(((x_i, y_i), (x_2, y_2), \ldots, (x_l, y_l))\) \((x_i \in \mathcal{X}, y_i \in \mathcal{Y}, l \text{ is the number of data points, for regression estimation and density estimation and } y_i \in \mathcal{Y} \subseteq \mathbb{R} \text{ for pattern recognition})\) randomly and independently generated from an unknown probability distribution \(p(x, y)\), find a function \(f(x, a)\) that has the minimal risk function \((A.1)\).

\[
R(f) = \int_{x,y} L(y, f(x, a)) p(x, y) dx dy
\]  

\((A.1)\)

Where \(a\) is the parameter of \(f(x, a)\). \(R(f)\) is called the generalization error or the expected test error. It is a measure of the generalization performance of \(f(x, a)\). \(L(y, f(x, a))\) is called the loss function. It is a measure of the deviations between the actual values and the estimated values on the data points generated from \(p(x, y)\).

As \(p(x, y)\) is unknown, traditional methods attempt to estimate \(f(x, a)\) by minimizing the empirical risk function:

\[
R_{emp}(f) = \frac{1}{l} \sum_{i=1}^{l} L(x, a))
\]  

\((A.2)\)

\(R_{emp}(f)\) is called the empirical error. That is, is estimated by training samples. Empirical Risk Minimization (ERM) principle is to minimize the generalization error by minimizing the empirical error \(R_{emp}(f)\). Traditional neural networks utilize this principle.

However, because of the limited number of \(l\), sometimes \(R_{emp}(f)\) cannot estimate \(R(f)\) well. As described in the statistical learning theory, and have the following relationship:

\[
R(f) \leq R_{emp}(f) + \Omega\left(\frac{l}{n}\right)
\]  

\((A.3)\)
Where $h$ is called the Vapnik-Chervonenkis (VC) dimension. It is a measure of the capacity of $f(x,a)$, which means that the ability of $f(x,a)$ to learn any training data point without error. $\Omega(\frac{1}{h})$ is called the confidence interval, a decreasing function of $\frac{l}{h}$, which is the ratio of the number of training samples into the VC dimension of the estimator. Equation (A.3) shows that the value of $R(f)$ depends both on $R_{emp}(f)$ and $\Omega(\frac{1}{l})$. Hence, $R_{emp}(f)$ can accurately estimate $R(f)$ only when $\Omega(\frac{1}{h})$ is small enough. Because of this, Vapnik developed the Structural Risk Minimization (SRM) principle. The SRM principle is: one defines a nested structure, $S_1 \subset S_2 \subset \ldots \subset S_m \ldots$, as shown in Appendix B.1, on the set of functions $S = \{f(\cdot, a), a \in \Lambda\}$ with their VC-dimensions satisfying $h_1 \subset h_2 \subset \ldots \subset h_m \ldots$, and then chooses the structure element $S_k$ with the minimal upper bound of the generalization error $R(f)$.

Appendix A.1: A structure on the set of functions is determined by the nested subsets of functions.

The objective of SRM principle is to estimate $f(x,a)$ by minimizing both the empirical error $R_{emp}(f)$ and the confidence interval $\Omega(\frac{1}{h})$, as shown in Appendix A.2. It defines a trade-off between the quality of the approximation of the given data and the complexity of the approximating function.
Appendix A.2: The bound on the risk is the sum of the empirical risk and of the confidence interval. The smallest bound of the risk is achieved on some appropriate element of the structure (Source: Vapnik, 1995)
APPENDIX B: AN INTRODUCTION TO LIBSVM 2.6 PROGRAM
APPENDIX B: AN INTRODUCTION TO LIBSVM 2.6 PROGRAM

Libsvm is a simple, easy-to-use, and efficient software for SVM classification and regression. Libsvm 2.6 was developed by Chih-Chung Chang and Chih-Jen Lin in 2001. It can solve C-SVM classification, nu-SVM classification, one-class-SVM, epsilon-SVM regression, and nu-SVM regression. It also provides an automatic model selection tool for C-SVM classification.

Libsvm 2.6 is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

The format of training and testing data file is:
<label><index1>:<value1><index2>:<value2> ...

<label> is the target value of the training data. For classification, it should be an integer which identifies a class (multi-class classification is supported). For regression, it's any real number. For one-class SVM, it's not used so can be any number. <index> is an integer starting from 1, <value> is a real number. The labels in the testing data file are only used to calculate accuracy or error. If they are unknown, just fill this column with a number.

‘svm-train’ Usage

options:
-s svm_type : set type of SVM (default 0)
 0 -- C-SVC
 1 -- nu-SVC
 2 -- one-class SVM
 3 -- epsilon-SVR
 4 -- nu-SVR
-t kernel_type : set type of kernel function (default 2)
Appendix

0 -- linear: u'*v
1 -- polynomial: (gamma*u'*v + coef0)^degree
2 -- radial basis function: exp(-gamma*|u-v|^2)
3 -- sigmoid: tanh(gamma*u'*v + coef0)

-d degree : set degree in kernel function (default 3)
-g gamma : set gamma in kernel function (default 1/k)
-r coef0 : set coef0 in kernel function (default 0)
-c cost : set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1)
-n nu : set the parameter nu of nu-SVC, one-class SVM, and nu-SVR (default 0.5)
-p epsilon : set the epsilon in loss function of epsilon-SVR (default 0.1)
-m cachesize : set cache memory size in MB (default 40)
-e epsilon : set tolerance of termination criterion (default 0.001)
-h shrinking: whether to use the shrinking heuristics, 0 or 1 (default 1)
-b probability_estimates: whether to train an SVC or SVR model for probability estimates, 0 or 1 (default 0)
-wi weight: set the parameter C of class i to weight*C in C-SVC (default 1)
-v n: n-fold cross validation mode

The k in the -g option means the number of attributes in the input data.
option -v randomly splits the data into n parts and calculates cross validation accuracy/mean squared error on them.

‘svm-predict’ Usage

Usage: svm-predict [options] test_file model_file output_file
-b probability estimates: whether to predict probability estimates, 0 or 1 (default 0); one-class SVM not supported yet
model_file is the model file generated by svm-train.
test_file is the test data you want to predict.
svm-predict will produce output in the output_file.
(Source: Chang and Lin, 2001)
APPENDIX C: MAT LAB CODES FOR PROGRAM
APPENDIX C: MAT LAB CODES FOR PROGRAM

devidePoint=600;
shift=10;

% output data vection creation for svmtrain
% normalize the output
output = (ElectricalConsumption(:,3)
min(ElectricalConsumption(:,3)))/(max(ElectricalConsumption(:,3))
min(ElectricalConsumption(:,3)));

TrainOutput = output(shift+1:devidePoint+1,:); TestOutput =
output(devidePoint+2:end,:);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% this section belongs to input data matrix creation
% normalize inputs
Humidity_norm = (Humidity(:,4)
min(Humidity(:,4)))/(max(Humidity(:,4))
min(Humidity(:,4)));
MaxTemperature_norm = (MaxTemperature(:,4)
min(MaxTemperature(:,4)))/(max(MaxTemperature(:,4))
min(MaxTemperature(:,4)));
SolarRadiation_norm = (SolarRadiation(:,4)
min(SolarRadiation(:,4)))/(max(SolarRadiation(:,4))
min(SolarRadiation(:,4)));

% moving average input & Outputs
%outputMa1 = transpose(tsmovavg(transpose(output), 'e', shift));
%outputMa2 = transpose(tsmovavg(transpose(MaxTemperature_norm ), 'e',
shift));
%outputMa3 = transpose(tsmovavg(transpose( Humidity_norm), 'e', shift));
%outputMa4 = transpose(tsmovavg(transpose(SolarRadiation_norm ), 'e', shift));
%create time delay input
timeDelayInput1 = [output(1:end)];
timeDelayInput11 = [output(1);output(1:end-1)];

timeDelayInput111 = [output(1:2);output(1:end-2)];

%timeDelayInput2  = [output(1:end)];
%timeDelayInput22  = [output(1);output(1:end-1)];
%timeDelayInput222 = [output(1:2);output(1:end-2)];

%timeDelayInput3  = [outputMa3(1:end)];
%timeDelayInput33  = [outputMa3(1);outputMa3(1:end-1)];
%timeDelayInput333 = [outputMa3(1:2);outputMa3(1:end-2)];

%timeDelayInput4  = [outputMa4(1:end)];
%timeDelayInput44  = [outputMa4(1);outputMa4(1:end-1)];
%timeDelayInput444 = [outputMa4(1:2);outputMa4(1:end-2)];

input =
[timeDelayInput1,timeDelayInput11,timeDelayInput111,MaxTemperature_norm,Humidity_norm,SolarRadiation_norm];%,outputMa2,outputMa3,outputMa4];%

devide input vector for training and testing
TrainInput = input(shift:devidePoint,:);
TestInput = input(devidePoint+1:end-1,:);

% train svm model and set parameters
% Usage: model = svmtrain(training_label_vector, training_instance_matrix, 'libsvm_options');

% libsvm_options:
% -s svm_type : set type of SVM (default 0)
% 0 -- C-SVC
% 1 -- nu-SVC
% 2 -- one-class SVM
% 3 -- epsilon-SVR
Appendix

% 4 -- nu-SVR
% -t kernel_type : set type of kernel function (default 2)
%  0 -- linear: u*v
%  1 -- polynomial: (gamma*u*v + coef0)^degree
%  2 -- radial basis function: exp(-gamma|u-v|^2)
%  3 -- sigmoid: tanh(gamma*u*v + coef0)
%  4 -- precomputed kernel (kernel values in training_instance_matrix)

% -d degree : set degree in kernel function (default 3)
% -g gamma : set gamma in kernel function (default 1/num_features)
% -r coef0 : set coef0 in kernel function (default 0)
% -c cost : set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1)
% -n nu : set the parameter nu of nu-SVC, one-class SVM, and nu-SVR (default 0.5)
% -p epsilon : set the epsilon in loss function of epsilon-SVR (default 0.1)
% -m cachesize : set cache memory size in MB (default 100)
% -e epsilon : set tolerance of termination criterion (default 0.001)
% -h shrinking : whether to use the shrinking heuristics, 0 or 1 (default 1)
% -b probability_estimates : whether to train a SVC or SVR model for probability estimates, 0 or 1 (default 0)
% -wi weight : set the parameter C of class i to weight*C, for C-SVC (default 1)
% -v n : n-fold cross validation mode
% -q : quiet mode (no outputs)

%crossvalidation = svmtrain(TrainOutput,TrainInput,'-m 800 -v 5')
%K1 = [(1:630), 'K']; % include sample serial number as first column
%model = svmtrain(label_vector, K1, '-t 4');
%matlab> [predict_label, accuracy, dec_values] = svmpredict(label_vector, K1, model); % test the training data
ForecastingModel = svmtrain(TrainOutput,TrainInput,'-s 3 -t 2 -g 0.25 -c 1 -p 0.125');

%svm-scale -l -1 -u 1 -s range train > train.scale
%svm-scale -r range test > test.scale
%Scale each feature of the training data to be in [-1,1]. Scaling factors are stored in the file range and then used for scaling the test data.

% svm-train -s 0 -c 5 -t 2 -g 0.5 -e 0.1 data_file
%Train a classifier with RBF kernel exp(-0.5|u-v|^2), C=10, and stopping tolerance 0.1.
%svm-train -s 3 -p 0.1 -t 0 data_file
%Solve SVM regression with linear kernel u'v and epsilon=0.1 in the loss function.
%svm-train -c 10 -w1 1 -w2 5 -w4 2 data_file
%Train a classifier with penalty 10 = 1 * 10 for class 1, penalty 50 = 5 * 10 for class 2, and penalty 20 = 2 * 10 for class 4.
%svm-train -s 0 -c 100 -g 0.1 -v 5 data_file
%Do five-fold cross validation for the classifier using the parameters C = 100 and gamma = 0.1
%svm-train -s 0 -b 1 data_file
%svm-predict -b 1 test_file data_file.model output_file
%Obtain a model with probability information and predict test data with probability estimates

%svm_parameter(svm_type=0, kernel_type=2, gamma=1, cache_size=40, eps=0.001, C =1, nr_weight=0, shrinking=1)
%crossvalidation = svmtrain(TrainOutput,TrainInput,'-m 800 -v 5')
%svm-train -c 2 -g 2 svmguide1.scale
%svm-predict svmguide1.t.scale svmguide1.scale.model svmguide1.t.predict
%. /svm-train -s 4 -t 2 -g .1 -c 120 TrainFile.txt ModelFile.txt
%. /svm-predict TestFile.txt ModelFile.txt OutputFile.txt.
%./svm-train svmguide1
%./svm-predict svmguide1.t svmguide1.model svmguide1.t.predict

% test svm model

[predictionP, accuracyP, decvalueP] = svmpredict(TestOutput, TestInput, ForecastingModel, '-b 0');

predictionP;
accuracyP;
decvalueP;
TestOutput;

p1=plot(predictionP);
set(p1,'Color','red')
hold on
p2=plot(TestOutput);
set(p2,'Color','black')

% Whether data
%p2=plot(Humidity_norm)
%set(p2,'Color','red')
%hold on

%p3=plot(MaxTemperature_norm)
%set(p3,'Color','black')
%hold on

%p5=plot(SolarRadiation_norm)
%set(p5,'Color','blue')
%hold on

%MSE & R^2, Nsv data

For constant gamma = 0.1 and varying C values

%p2=plot(X,Y)
%set(p2,'Color','red')
%p3=plot(X,Z)
%set(p3,'Color','black')

%p5=plot(X,xxa)
%set(p5,'Color','blue')
%hold on
REFERENCES


[29] LibSVM “README” tutorial
[30] Hsu., Chang, C-W., Lin, C-C., A practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University.


