CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter reviews the various methods used to establish the baseline models for landlord energy consumption. Simultaneously, the 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 help in creating a new successful method.

2.2.1 Regression-based models

Regression analysis is one of the important methods in which the 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 a similar form as equation 2.1, [52].

$$Z = f(X_1, X_2, X_3 \ldots\ldots X_k) + u \qquad (2.1)$$

where, Z is a variable whose movements and values may be described by the variables X_1 ,X_2 ,..., 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 use generalized least squares regression to determine the model coefficients.

2.2.2 Variable-based 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 variable, which can account for differing insulation levels, 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 loses through the envelope exactly to 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 ºF or 11.7 ºC 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.

These are basic functional forms of the VBDD models, [42].

1) For electricity use, electricity demand and water use (increase with outdoor temperature, T)

$$Y = \alpha + \beta c \cdot DD(\tau_c) \quad (2.2)$$

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

$$Y = \alpha + \beta h \cdot DD(\tau_h) \quad (2.3)$$

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

$$Y = \alpha + \beta_h \cdot DD(\tau_h) + \beta c \cdot DD(\tau_c) \quad (2.4)$$

Where α is the base energy use of the VBDD model, β_c is the slope for the VBDD cooling model, β_h is the slope for the VBDD heating model. $DD(\tau)$ are the degree-days to the base τ, 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, τ, 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 of energy consumption in buildings ranges from ± 4.5% for relatively high degree-day values (seasonally and yearly values) to ±10% for low degree-day values (monthly values).

2.2.3 Linear regression models:

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

2.2.4 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):

$$Y = Y_0 + RS.T_0 \quad (2.5)$$

2) For energy use that increases with decreasing outdoor temperature T_0 (eg: gas use)

$$Y = Y_0 - LS.T_0 \quad (2.6)$$

Where, Y_0 is the intercept that represents the value of energy use when T_0 =0 °F.RS means right-hand slope and LS means left-hand slope.

2.2.5 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 the derived equation is given as:

$$\hat{E} = a + bT_0 + cI + dI + eR_{dp} + + fq_{sol} + gq_i \quad (2.7)$$

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.

2.2.6 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.

2.2.7 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].

2.2.8 Calibrated simulation

Literature illustrates several calibration procedures were developed by different research groups to enhance the performance. Procedures for calibrating hourly simulation programs to create baseline models for building were developed in the 1990s .Following that developed a hybrid calibration procedure to take an hourly simulation model and apply it to data obtained from whole building electricity consumption and system level consumption. 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 situation 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 be easily 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 cost of options A or B is excessive. Option A is referred to partially measure retrofit isolation where, savings is 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 is 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 perform and how well they are calibrated to the actual performance.

2.2.10 Artificial neural networks

The Artificial Neural Networks or ANN is used in 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 and also by the 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 pre-processor 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 R^2 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 due to 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 result.

2.2.11 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.

2.2.12 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].

2.2.13 Support vector machines

Support vector machines (SVMs) is one of the important methods 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.

1) SVMs estimate the regression using kernel functions; a set of linear functions that is 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, to minimize 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 is located 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 is divided into two sets of inputs as given in equation (2.9). A2 is for training the data set while A3 is for testing &validation of empirical data set. The data arrangement for feeding to the inputs matrix (A1) is represented in equation (2.8) [29]:

A1=input Matrix = [timeDelayInput1, timeDelayInput11, timeDelayInput111, MaxTemperature_norm, Humidity_norm, SolarRadiation_norm] (2.8)

Where the time delay inputs represent electrical consumption values for the previous three steps; as a solution to the storage effect; the MaxTemperature_norm, Humidity_norm, SolarRadiation_norm are the indicator variables that visualize and

normalize the data of temperature, relative humidity & solar radiation respectively. The forecasting model for the training inputs is presented by equation (2.9).

A2 = ForecastingModel = 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),* it demonstrates the model file. Model_file is the file generated by svm-train & test file is the data that needs prediction. And after running the package, svm-predict will produce output in the output_file as per the equation (2.10), [30]

A3= [prediction] = svmpredict (TestOutput, TestInput, Forecasting Model) (2.10)
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: $(gamma * u' * v + coef0)$^$d$egree
2 -- radial basis function: $exp(-gamma * |u - v|$^2)
3 -- sigmoid: $tanh(gamma * u' * v + coef0)$
-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

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 $K(x_i,x_j) = (x_i * x_j + 1)^d$ and the radial-basis function (RBF) kernel, $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||)^2$;$\gamma > 0$, where d and γ are defined as kernel parameters.

The RBF kernel, which is based on Gaussian function, nonlinearly maps samples into a higher dimensional space and handles 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 $-\gamma = 1/k$ by Limsvm-2.6 where k is the number of attributes in the input data, $-\gamma$ is constantly set to 1/3 in the future modelling. Finally, all the training and test data sets are scaled to [0, 1].

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 regression 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, is performed. The stepwise search method is used to measure the performance as explained in Dong *et al.*, [27].

Selection of parameter c

According to equation 4.2 (chapter 4), parameter *C* determines the tradeoff between the model complexity and the degree to which deviations are larger than ε as tolerated in optimization formulation. Moreover, the regularization parameter *C* decides the range of values $0 \leq (a, a^*i) \leq C$, $i=1,...,l$ assumed by dual variables which are 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 undermine the training data because the weight placed on the training data is too small thus resulting in larger values of MSE on the test sets.

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

Number of support vectors increase slightly as *C* increases. When *C* gets larger, the optimization formula (4.2) emphasizes more on the empirical risk and makes the model a better fit for the training data, in comparison to 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 formula.

Selection of parameter ε

According to figure 4.3 (chapter 4), parameter ε is used to fit the training data and it controls the width of the ε -insensitive zone. The ε does not affect the performance of SVMs much; while the numbers of support vectors show a decreasing function of ε. Generally, the larger the ε, the fewer the number of support vectors, thereby resulting in the sparse representation of the solution (27). If the ε is too large, it will worsen the accuracy on the training data. The optimization for parameter γ (kernel parameter) also same as for the parameter C & ε, which is described in the chapter under 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 ε.

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

2) Stepwise Method: More efficient to quickly identify 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 adapted on parameter C (fixing the first result of ε) and ε (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

spotting the best *(C, ε)*, the whole training set is testified 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. In contrast, computer simulation methods (DOE-2, Energy Plus, BEST) are time consuming and need detailed information of the building operation characteristics, physical conditions and large inputs of energy consumption data. Other models such as ANN and Fourier series models are mostly simulated, based on hourly data and data inputs. The processing and results rely on certain software's and have quite high accuracy levels. Nevertheless, a very few models have been actually implemented in buildings as they need more improvement and computational efficiency to get wide spread applicability. The differences between several baseline-modelling methods used in this literature are summarized in table 2.1.

Table 2.1 Comparison of models discussed in literature

Models	Examples	Positive	Negative
Statistical Regression models	Linear, Multiple-linear, Change-point, degree-day	Easy to establish Most commonly used	Fair accuracy
Other models	Support Vector Machine, Neural Network, Fourier Series	High accuracy	Large pool data, however, SVM shows high accuracy with small pools of data

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 modelling.

Choosing a proper and practical methodology is important for baseline building energy consumption and energy savings estimation. And this is directly related to the time scales such as hourly, daily and monthly or on the levels such as system and facility. Regression based model is considered to be more practical for data of all time periods; while neural networks and Fourier Series are more suitable for modelling hourly building cooling and heating energy consumption.

Table 2.2 Summary for some baseline models.

Reference	Type of Models
Kusuda *et al*. (1981)	Variable-base degree-day
Fels *et al*. (1986)	Single-variable regression model
Pope (1987)	Modified Bin
Kissock *et al*.(1993)	Single-variable regression model
Katipamula *et al*. (1994)	Multiple linear regression
Kissock *et al*. (1998)	Linear and change-point models
Krarti *et al*. (1998)	Neural networks
Reddy *et al*. (1998)	Multiple linear regression
Dong *et al* (2005)	Support vector machine
Qiong li *et at* (2009)	Support vector machine
Milos *et at* (2010)	Support vector machine
Rishee *et al* (2014)	Support vector machine