T.C.
SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

ASSESSMENT OF SEASONAL EFFECTS ON CITY BASED DAILY ELECTRICITY LOAD FORECASTING USING LINEAR REGRESSION

M.Sc. THESIS

Shanga Othman KAREEM

Department : COMPUTER AND INFORMATION ENGINEERING

Supervisor : Assist. Prof. Dr. Mustafa AKPINAR

July 2021

T.C.
SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

ASSESSMENT OF SEASONAL EFFECTS ON CITY BASED DAILY ELECTRICITY LOAD FORECASTING USING LINEAR REGRESSION

M.Sc. THESIS

Shanga Othman KAREEM

Department : COMPUTER AND INFORMATION ENGINEERING

Supervisor : Assist. Prof. Dr. Mustafa AKPINAR

This thesis has been accepted unanimously / with majority of votes by the examination committee on 29.07.2021.

Head of Jury **Jury Member** **Jury Member**

DECLERATION

I hereby declare that the work in this thesis is my original work in the Department of Computer Engineering at Sakarya University and has not previously been sent to any agency for evaluation purposes. All visual and written information and results were presented by academic and ethical rules. In addition, I have noted and listed all references included.

Shanga KAREEM

29.07.2021

ACKNOWLEDGEMENT

I would like to express my gratitude to my supervisor, Dr. Mustafa AKPINAR, for his valuable guidance, flexibility, constant advice, support, and assistance during the time. The majority of his invaluable comments and suggestions will be greatly appreciated.

I am especially grateful to my beloved father: Othman KAREEM, for his consistent encouragement and support. Sincere gratitude to my friend Rebaz Sabir Salih for his support and patience.

I would like to thank Mr. Daban Abdulkarim Hasan, head of Lightning Department from Distribution of Electricity Management of Sulaimanyah Electricity Directorate, who is helping me to get the dataset from the control center of Sulaimaniyah/Iraq with providing and explaining the distribution plan, process, issues, and challenges.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

◦C	: Celsius degree.
CU	: Control Unite
ECMWF	: European Center for Mid-Range Weather Forecast
GDP	: Gross Domestic Product
GHG	: Green House Gas
HVAC	: Heating, Ventilation and Air Conditioning
MAPE	: Mean Absolut Percentage Error
MLR	: Multiple Linear Regression
R^2	: Coefficient of Determination
Amp	: Ampere
STLF	: Short-Term Load Forecasting
GP	: Gaussian Process
α	: Level of Significance and Type I Errors
β	: Regression Coefficient
ε	: “Error Term” in regression/statistics residual
σ	: Variance of population values.

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Keywords: Multiple Linear Regression, STLF, Load Forecasting, Control Center Component.

Due to the importance of electricity and its impacts on the human living environment, several studies have been conducted to forecast and possibly reduce forecast errors of the electricity load.

In this thesis, the load data of Iraq Sulaymaniyah city are used. Forward selection, backward elimination, and stepwise approaches are used to determine the variables in the multiple regression equation. The 6-year data of 2014-2019 was used to develop the day-ahead forecasting models, while the 2019 year was used as a test dataset to validate the model. Each year was divided into two-time intervals according to the change in load behaviour. The long-term seasonality effect was tried to be determined.

The results show that instead of all data, the series divided into two according to long-term seasonality could estimate the load with lower errors. Divided series will help the control center to have a better estimation of electricity demand and energy purchase. Using our model, they will be capable of forecasting electricity load for upcoming months and years to replace the traditional way of calculating and reporting load.

LİNEER REGRESYON İLE ŞEHİR BAZLI GÜNLÜK ELEKTRİK YÜK TAHMİNİ ÜZERİNDE MEVSİMSEL ETKİLERİN DEĞERLENDİRİLMESİ

ÖZET

Anahtar Kelimeler: Çoklu Doğrusal Regresyon, STLF, Yük Tahmini, Kontrol Merkezi Bileşeni.

Elektriğin önemi ve insan yaşam ortamı üzerindeki etkileri nedeniyle, elektrik yükünün tahmini ve elektrik yük tahmin hatalarını düşürmek için çeşitli çalışmalar yapılmıştır.

Bu çalışmada Irak Süleymaniye şehrinin yük verileri kullanılmıştır. Çoklu regresyon denklemindeki değişkenleri belirlemek için ileriye doğru seçim, geriye doğru eleme ve kademeli yaklaşımlar kullanılmaktadır. Gün öncesi tahmin modellerini geliştirmek için 2014-2019 yılları arasındaki 5 yıllık veri, modeli doğrulamak için ise 2019 yıla ait test veri seti kullanılmıştır. Her yıl, yük davranışındaki değişime göre iki zaman aralığına bölünmüştür. Böylece, uzun dönemli mevsimsellik etkisi belirlenmeye çalışılmıştır.

Sonuçlar, tüm veriler yerine uzun dönemli mevsimselliğe göre ikiye ayrılan serinin daha düşük hata ile yükü tahmin edebileceğini göstermektedir. Bölünmüş seriler, kontrol merkezinin elektrik talebini ve enerji alımını daha iyi tahmin etmesine yardımcı olacaktır. Modelimizi kullanarak, geleneksel yük hesaplama ve raporlama yönteminin yerine gelecek aylar ve yıllar için elektrik yükünü tahmin edebilecekler.

CHAPTER 1. INTRODUCTION

Electricity is produced from two different types of sources which are called renewable and non-renewable sources. Renewable includes solar, wind, radiation, etc. Non-renewables are coal, natural gas, and oil. Unfortunately, our region entirely depends on non-renewable energy.

Most of the energy is produced by electrical generators as well as water turbines. Due to the lack of enough water sources, the majority is produced using electrical generators. Once produced, they are transferred to the substations in the cities through overhead power lines suspended by towers.

Supplying power is one of the significant challenges in Iraq. The Government can barely provide 24 hours of electricity in a day. A portion of the energy is supplied by generators that belong to the private sector. Government or private, in both scenarios where electricity is generated, the source of the power is fuel (examples; gasoline, petroleum). Therefore, it has to consider the amount of fuel required to provide the necessary power. Besides, for the rest of the electrical energy generated, Government has to dedicate a particular portion of the annual budget for the companies that work in the sector to overcome the shortage of power. Nevertheless, Government could not fully provide the 24 hours of electricity.

There are several control centers taking care of power distribution in each city. Electricity is supplied based on a specific schedule derived from recent days, months, and years. For instance, the Government provides power for 14 hours in spring and 8 hours in summer.

Due to high energy demand and low supply in Sulaimaniyah city, there is insufficient energy to be stored. Once produced, it is either distributed or sent to other parts of Iraq. But in general, electricity can not be stored; it needs to be transformed to other forms of energy and reuse once required. One of the most critical storage media is the battery, then comes flywheel, etc.

Lack of fuel, shutting down the feeders for maintenance are examples of rapid declines that cause the schedule changes. Forecasting load consumption helps government/companies make a tactical decision about the power supply and distribution for upcoming days, months, and even years.

This traditional power distribution approach has been working for the past years. However, there is no formal way of predicting how much power will be needed next year(s). They should formulate a strategic plan to balance the load, minimize the immediate power shortage, and even develop an annual report that will help reduce the chance of errors, estimate the required power, and maximize hours of dedicating the power supply to people. The governors should do accurate demand forecasting to achieve balanced electricity with minimum loss and power shortage. Demand forecasting of electricity load is one of the aims of this thesis.

Forecasting is a method of predicting what the future will bring. Every function inside an organization requires an estimate of what the future will look like to create its current plans. Today, all firms work in an unpredictable environment. Organizations should analyze the environment using various forecasting tools, determine their strengths and weaknesses, and develop plans. Predicting is making plans based on a set of assumptions or forecasts that may or may not be accurate. Demand forecasting is critical for businesses to receive the most precise estimate of the changes feasible to survive, strive for operational excellence, and maintain a competitive advantage.

According to William J. Stevenson of Operations Management, a good forecast should have the following characteristics: Accuracy, reliability, timely, simple to use and understand, expense [1].

The projections are based on historical data. The current behavior of these components and the likelihood of their occurrence in the future is, to some extent, an extension of how they have occurred in the past and present. However, unanticipated alterations can always occur. Forecasting is inextricably linked to planning.

According to the businesses, annual, monthly, weekly, daily forecasts can be done. For example, while annual population growth is more likely to forecast, monthly forecasting is more important in estimating interest rates. A household appliances store is more likely to provide weekly sales forecasts, while a gas supply company may need to forecast daily. The concept of time in the estimation changes according to each business and condition. For a person who trades in the stock market, minute forecasts can be "short-term forecasts," hourly or daily forecasts can be "medium-term forecasts", weekly or monthly forecasts can be "long-term forecasts". Daily forecasts in this thesis are referred to as "short-term forecasts" in the electricity market, monthly forecasts are referred to as "medium-term forecasts", annual and higher forecasts are referred to as "long-term forecasts".

This thesis examines multiple linear regression (MLR) to predict electricity load for the upcoming year based. The data mainly composed 24 hours load on each feeder within the regions that belong to the Sulaimaniyah city in Iraq collected from 2014 to 2019. Since data is hourly-based, it is converted to a daily-based for daily forecasting.

1.1. Literature Review

Different models were focused on and discussed on applying these models in forecasting electricity demands in the literature.

Saber et al. conducted short-term load forecasting using multivariable linear regression with stochastic and dependable big data. The multicore parallel processing was used in all matrix operations; the mean absolute percent error is 3.99% of actual recorded data [2].

Furthermore, Kolasa-Wiecek and Alicja analyzed greenhouse gas (GHG) emissions generated by Poland's energy sector. In their study, a multiple stepwise regression model was applied to the data from the 1989-2011 period, and it was shown that the obtained regression model could explain 90% of the variability [3].

Fumo and RafeBiswas analyzed the hourly and daily energy consumption of an HVAC (Heating, Ventilation, and Air Conditioning) system in a research house using simple linear regression, multiple linear regression, and quadratic regression. Their results proved that the accuracy of the models increased with the time of the observed data [4].

In addition, Amber et al. used multiple regression to create a mathematical equation to estimate the daily energy demand in university buildings on London's South Bank University's Southwark Campus. Their study determined that temperature, weekday-weekend situations, and building type excessively affect using five-year data [5].

Also, Akpinar and Yumusak used MLR in the day-ahead natural gas forecast for 2012. In their work, expanding data and sliding window approaches were compared. The results showed sliding window approach could predict lower errors. The rest of the study tested models with different window sizes and showed that the 4-week window size had the lowest MAPE value [6].

Moreover, Kim et al. conducted a study forecasting peak load demand for an institutional building in Seoul; they used ARIMA models, ARIMA-GARCH models, multiple seasonal exponential smoothing, and ANN models. The data were collected from 23 facilities in the campus area. The best model was found with moving window simulations and step-ahead forecasts. In addition, they used weather and holiday variables, which were critical for load estimation. The ANN model with external variables (NARX) was the best for 1-hour to 1-day ahead forecasting [7].

Yan et al. Prerared a factor-based bottom-up forecasting model to estimate the electricity consumption and carbon emission during 2015-2040 periods for the

Japanese residential sector. The models consist of nine scenarios that combine three levels of household size, and three levels of per capita gross domestic product growth are taken into account to estimate the electricity consumption for space heating and cooling, water heating, cooking, and appliances. They performed that the total residential electricity consumption will reach a peak during the 2020s. And the total carbon will keep decreasing by 51.14-72.16 Mt between 2015 and 2040 [8].

Also, Larsen et al. introduced a new estimation methodology for electricity usage with the daylight and occupancy-controlled artificial lighting in an office, which is accurate and rapid. The technique is validated for an office building in Oslo, Norway, utilizing data from the Building Management System and experimentally generated data. They applied on a case-study and cell-office, during the 6-day measuring period, used measured external irradiance and the actual occupancy profile for the office. During the measurement period, the calculated electricity use is predicted to be 0.3 kWh/m^2 (6 days). For the case-study office, a rough assessment of electricity use can be performed using the same methodology, but with the Norwegian reference year weather file instead of measured irradiance as a background. In this case, annual electricity use in the office corresponds to 10.5 kWh/m^2 year, which is approximately 0.32 kWh/m^2 (6 days) for the period of measurements [9].

In addition, Seyedzadeh et al. used ML models for building energy estimation and benchmarking, as well as the benefits and downsides of each model. Beside ML techniques and other black box methods, only Gaussian Process (GP) was used for model training with uncertainty estimations. ANN produced a fast and precise short-term load forecasting for Energy Management System (EMS)s where temerature and humidity data is collected using sensors, while GP is more beneficial for long-term energy estimation when there is uncertainty in input variables [10].

Also, Motlagh et al. produced a joint probability model of electricity demand based on occupant’s age grades and household income levels. They designed the bottom-up technique by using a micro-level database for 70 houses in Australia. A neural

regression generalization technique using back-propagation and cognitive mapping is developed to estimate electricity consumption.

The aggregated result is then confirmed against the Australian national census data from 2012. As a result, the model is improved through a top-down review. The findings also suggest a higher percentage of per capita demand for adults in the high and medium-income classes and a lower percentage for individuals in the low-income category. The ratio of child demand to adult demand is highest in low-income households and lowest in middle-income households, with high-income households having the best balance of adult and child per capita demand [11].

Cao, et al. used data from the Chinese Urban Household Survey over 2009–2025 using alternative linear and nonlinear autonomous trends. They develop a preferred forecast range of 85–143 percent growth in residential per capita power demand.

According to their analysis, per capita, income growth accounts for 43% percent of the rise, with the remaining due to unexplained historical trends. Increases in the stock of specific essential appliances, particularly air conditioners, account for around one-third of the income-driven demand. The remaining two-thirds are derived from non-specific sources of income-driven growth and are based on an estimated income elasticity that decreases from 0.28 to 0.11 as income increases.

While the supply of refrigerators is not expected to grow, it discovered that they account for approximately 20% of household electricity demand. The extensive range of 85–143 percent is due to alternative credible temporal trend assumptions. However, the estimation price of electricity result was -0.7. These estimations indicate that carbon price and appliance efficiency policies might significantly lower demand [12].

In addition, Al-Mosawy et al. analyzed household electricity consumption of residential areas in Baghdad by studying a set of factors, which are the average daily outdoor air temperature, the plot area of residential. It has been discovered that the

annual electricity consumption is directly proportional to the plot coverage ratio for residential units, the number of housing unit residents, and the household income.

While the increase in the number of members of the housing unit and the increase in monthly household income are inversely proportional to the coverage ratio, the lowest electricity consumption achieved is (45 kWh) when the average daily temperature is (23°) Celsius in April. It reaches its highest value of (169 kWh) when the average daily temperature is (39°) Celsius in July.

The results revealed that small plot areas and high coverage residential units consume less electricity than large plot units. Six quantitative models were developed to describe electricity consumption behavior regarding the variables studied [13].

In another study, Abuella et al. prepared an analysis model for the European Center for Mid-Range Weather Forecast (ECMWF) using MLR to produce probabilistic estimates of solar energy, and they performed a short-term load prediction [14].

Also, Hong et al. prepared MLR models using the 3-year hourly energy demand of the U.S. utility, and they used a year as a hold-out sample. They also showed the relationship between system load and temperature in graphical form monthly [15].

Yildiz et al. first modeled the monthly delivered natural gas demand estimation by decomposition of time series. Next, the residuals were examined by various independent variables regression. As a result of the regression model, it is seen that the variables of standard precipitation index of 24 months, natural gas sold to commercial consumers, and total natural gas underground storage capacity are determined as significant. The hybrid approach has yielded lower percentage errors [16].

In addition, Zhang, et al. applied three forecast models: multiple linear regression, random forest, and gradient boosting. They were merged solar capacity to estimate

hourly load in southern California, and the result shows that the models were more accurate where lower loads [17].

Amiri et al. used multiple regression to develop a cooling and heating load model to estimate energy consumption and performance for commercial buildings in the U.S. The difference between heating and cooling loads compared to the result demonstrated from energy simulations; their result showed that outlined the benefits and prospects of this method for determining the energy efficiency of commercial buildings [18].

Braun et al. demonstrated an examination of a supermarket's electricity use in northern England. A multiple regression model has been used data from the 1961–1990 interval in their study, and the model was obtained forecasts the climatic period 2030–2059. The estimated outcome is then compared between these two times. As a result, power usage is expected to grow by 5.5%, with 2.1% being the most conservative estimate. Gas usage is expected to decline by up to 28%. (13% central estimate) [19].

Vu et al. applied multicollinearity and backward elimination methods to identify the most influential variables and generate a multiple regression model for monthly forecasting of energy consumption in the Australian state of New South Wales. The outcome demonstrated that the suggested model had a reduced prediction error [20].

Wu et al. constructed a linear regression model to forecast energy usage based on online monitoring data from 30 single functional and 20 multipurpose buildings. On-site studies revealed that the sub-item energy consumption index obtained from multifunctional structures had lower inaccuracy than single functional buildings. The inaccuracy was the lowest in the hotel industry, totaling about 1.1%. In the office sector, the variance was 3.9 % [21].

Vasquez et al. prepared an analysis model for the annual energy consumption of the Puerto Princesa Distribution System for 2019-2028, using MLR to provide a forecasting model. The variables considered for the regression analysis were peak

demand and the number of customers. The result showed the MLR was a decent match, and the error performance test proved that the mean percent error was 0.74% [22].

Bianco et al. used simple and multiple linear regression to estimate energy consumption for the historical data are from 1970–2007 in Italy, with the variables gross domestic product (GDP), GDP per capita, and population. The result showed that R2 is 0.981 for the total residential and non-residential consumption. Then the results of the models are compared with the national forecasts available in Italy, which showed excellent accuracy [23].

Aranda et al. analyzed the energy consumption of annual Spanish banking sectors, and a multiple linear regression model was applied. They obtained three models; the first one used to predict the energy consumption for the whole banking sector, second and third is to estimate the energy consumption for the branches with low winter climate severity (Model 2), with high winter climate severity (Model 3). Results showed that the verification of the first model had the lowest determination coefficient that allows for the detection of weak bank branches [24].

Asadi et al. developed a novel model that applied regression equations to forecast and figure out energy consumption in commercial buildings during the early phases of building design based on construction features, form, and occupancy schedule in the United States. Their findings were included in a set of regression equations to estimate energy consumption in each design scenario. The best agreement was found between the projected data and the DOE simulation based on the constructed regression model. The highest error rate was less than 5% [25].

Siyu Zhou and Neng Zhu; developed an analytical model for four Chinese climates: hot summer and warm winter, hot summer and cold winter, cold, and extremely cold They created several regression models to predict the energy consumption of office buildings in various climates when diverse building envelope designs are factored into the equation. Simulation evaluations and actual case evaluations were performed to evaluate the feasibility and correctness of the regression models during the regression

model assessment. The simulated assessments had a ±5% mistake rate, whereas the actual case evaluation had a ±15% error rate [26].

Mohammed et al. used a regression model to estimate the energy consumption of one of Saudi Arabia's most critical energy-consuming categories of facilities: schools. The model was created using 350 actual data points of energy usage collected from Saudi Arabia's eastern region schools. According to the results, the model correctly forecasted the energy consumption of school buildings with a greater than 90% accuracy [27].

Dhaval and Deshpande; applied For day-ahead load forecasting, multiple linear regression (MLR) is used. The load in an electrical power system is affected by temperature, the due point with seasons, and the load corresponds with past load consumption (Historical data). The findings revealed that the model predicted with 95% accuracy [28].

Supapo et al. multiple linear regression models were developed for predicting load consumption on Palwan's Aborlan-Narra-Quezon supply system from 2016 to 2025. The results demonstrated that the proposed approach is adequate. For prediction accuracy, the mean average percentage error (MAPE) for each year between historical and anticipated load data was calculated to be 2.26 % [29].

Tuaimah et al. conducted short-term (up to 24 hours) load forecasting for historical data of Iraqi Power System, using multiple linear regression model (MLR) method, produced two models: (winter season, summer season) models. They got the compassionate model to the fluctuation of temperature. It needed a very accurate temperature forecast, as a slight change of temperature is causes a significant difference in load prediction [30].

Moreover, in recent years, there are several studies with the other techniques used, such as computational intelligence (artificial intelligence), learning-based techniques,

including Auto-Regressive Integrated Moving Average (ARIMA), that have been used in [31], [32].

In studies on multiple regression techniques, electrical consumption estimation, and energy planning in the literature, it has been observed that seasonal effects, temperature, and calendar events are effective in load consumptions. With this motivation, it was thought that evaluating the seasonal impact with separate models would reduce the error by making more accurate predictions. City electricity consumption data was obtained to carry out this thesis. The data mainly composed 24 hours load on each feeder within the regions that belong to the Sulaymaniyah City in Iraq. Since it is daily data, use Short-Term Load Forecasting (STLF) to create the model. This thesis examines multiple linear regression (MLR) applications to predict electricity load for the upcoming year based on the data collected from 2014 to 2019.

In Section 2, the multiple regression technique is explained. The data and seasonality are mentioned in Section 3. In addition, the hypothesis question of the thesis is given in Section 3. In Section 4, the results are shown, and a general evaluation is done. In the last section, the results are summarized as a conclusion.

CHAPTER 2. METHODOLOGY

This chapter goes through types of regression methods and how they are applied in the thesis. Forward selection outperforms the others in terms of performance. The details will be further discussed in this chapter.

2.1. Multiple Linear Regression

Over the last few decades, there has been a great deal of study into electric load forecasting. Most research in this field aims to design models that can predict the energy load profile with greater precision.

Predictive models are used to estimate events at any period. It is widely used in sports, weather and healthcare, auto insurance [33], [34]. One of the most common methods used in this area is linear [35], [36]. In the behavioral sciences, it is one of the most commonly used predictive analyses. It is essentially the relationship between one or more explanatory variables and a dependent variable. Multiple Linear Regression (MLR), frequently known as multiple regression, is a statistical method that predicts the outcomes of a response variable and use many explanatory factors [5]. It is a tried and true approach. They are widely applied in forecasting in sectors. It is accurate and robust [37].

Many analysts have addressed the issue of using both qualitative and quantitative factors in regression or multivariate analysis [38],[39].

The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables; the independent variables can be continuous or categorical [40], [19]. The goal of

reducing the gap between the observed and estimated values is often accomplished by curve fitting using the regression approach. The mathematical expression of the MLR is given in Equation (2.1).

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad (2.1)$$

where:

i = the number of days,

y_i = dependent variable, the electrical charge,

x_i = explanatory variables,

β_0 = the average coefficient of the model,

β_p = slope coefficients for each explanatory variable,

β_0 ..... β_p = the linear coefficient of each explanatory variable.

ε: The residual (fitted error) is used to assess the overall significance (F-test) of the equation as well as the significance of each regression coefficient (t-test). To achieve accurate findings from these analyses, the residual must be average and stable, with a mean of zero and a constant variance of σ^2 [19], [41].

MLR model works better when the relationship between dependent and explanatory variables is linear. Multiple linear regression analysis predicts trends and future values, either points or ranges. Until now, MLR has been used in numerous load forecasting studies. Different methodologies, prediction periods, mathematical models, and datasets results were discussed in the literature review of the thesis.

Short Term Load Forecasting (STLF) is a reliable technique for estimating system loads from hours to days in advance. A strong forecasting strategy is crucial for generating economic output, securing systems, managing them, and planning. Linear regression analysis is a strong approach for predicting unknown values of a variable based on the actual value of another variable (factors) [2].

2.2. Forward Selection

The forward selection approach continuously applies variables to the model. The first variable in the model has the most significant similarity to the independent variable *y*. Since adjusting for the influence of the first variable, the variable that entering the model as the second independent variable has the strongest relationship with *y*. When the last variable entering the sample has an irrelevant regression coefficient, or when all variables are used in the model, the mechanism ends as shown in Figure 2.1.

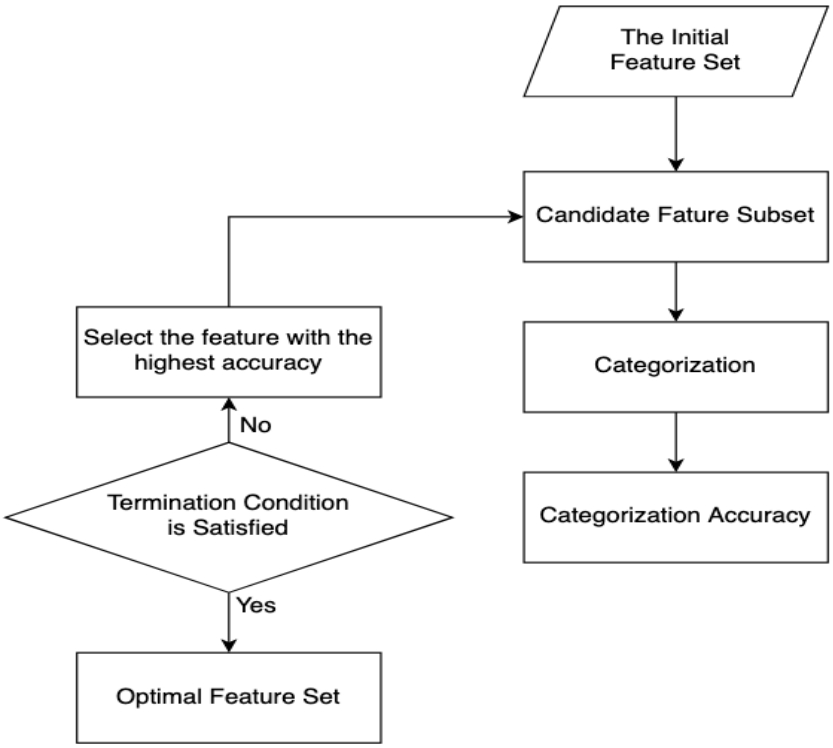
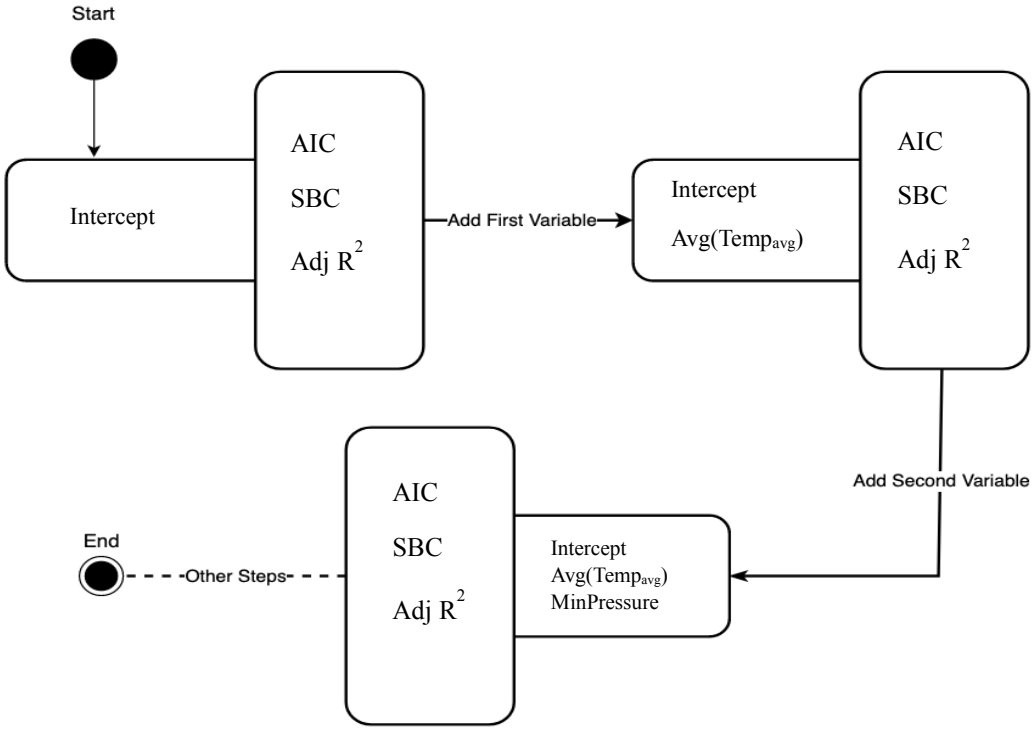


Figure 2.1. Forward selection diagram

Figure 2.2. shows the first three steps of the forward selection method for the Winter to Summer Train dataset (details are given in Section 3), which starts with no variable called the Null Model. In the following steps, the model selects the variables by comparing the p-value or R^2 or other predicators involved in the selection process, such as SBC and AIC (Figure 2.2.).



The model starts with intercept, which means that it has not selected an attribute for comparison yet. In this step, the parameter values are given in Table 2.1. (Adj $R^2 = 0$, AIC = 21,717, SBC = 20,872). As a result, the estimated intercept, t-value, and standard error are 1,346,456, 175.2, 7,685.300005, respectively. It is ready to determine the most significant variables to inter the model (Table 2.2.).

Table 2.1. Firs-step result of Forward selection method (Null model)

Step 0. Effect	Errors	Adj R^2 AIC SBC	0 21,717 20,872	
	Parameter Estimates	Estimates	t-Value	Standard Error
Entered: Intercept	Intercept	1,346,456	175.2	7,685.300005

The first variable is $Avg(Temp_{avg})$ selected then added to the model as a first step, it is one of the significant entry candidates variables by pre-determined criteria Table 2.2.

Table 2.2. Most significant candidate variables for Step1

Best 10 Entry Candidates for Step 1			
Rank	Effect	Log p-Value	Pr > F
1	Avg($Temp_{avg}$)	-532.387	<.0001
2	Avg($Temp_{min}$)	-523.308	<.0001
3	Avg($Temp_{max}$)	-512.609	<.0001
4	Max($Temp_{max}$)	-428.8936	<.0001
5	Max($Temp_{avg}$)	-413.5074	<.0001
6	Max($Temp_{min}$)	-400.5002	<.0001
7	Min($Temp_{avg}$)	-369.4979	<.0001
8	Min($Temp_{min}$)	-364.7489	<.0001
9	Min($Temp_{max}$)	-363.7383	<.0001
10	AvgPressure	-270.3600	<.0001

Avg($Temp_{avg}$) added to the model, which has the highest Adj R^2, lowest AIC, SBC, and BIC, among other candidate variables.

Table 2.3. Result of Forward selection model with the first variable

Step 1.	Errors	Adj R^2	0.7137	
		AIC	20,658	
		SBC	19,819	
Effect Entered: Avg($Temp_{avg}$)	Parameter Estimates	Estimates	t -Value	Standard Error
	Intercept	277,153	11.72	23,638
	Avg($Temp_{avg}$)	37,969	45.94	826.557822

The next step is to add another significant variable to the model, which is (MinPressure) in the candidate variables (Table 2.4.); all error types and parameter estimates results are shown in (Table 2.5.).

Table 2.4. Most significant candidate variables Step2

Best 10 Entry Candidates for Step 2			
Rank	Effect	Log p-Value	Pr > F
1	Min. Pressure	-35.4757	<.0001
2	Avg. Pressure	-29.9936	<.0001
3	Max. Pressure	-24.415	<.0001
4	Avg. Humidity	-22.5411	<.0001
5	Max. Humidity	-18.7974	<.0001
6	Max($Temp_{min}$)	-12.4616	<.0001
7	Avg. Type of Weather	-11.1241	<.0001
8	Max($Temp_{avg}$)	-10.371	<.0001
9	Min. Humidity	-9.2327	<.0001
10	Avg($Temp_{max}$)	-7.4911	0.0006

Table 2.5. Forward selection model with 2 variables

Step 2.	Errors	Adj R^2 AIC SBC	0.735 0.735 19,759	
	Parameter Estimates	Estimates	t -Valsue	Standard Error
Effect Entered:	Intercept	7,871,199	8.61	914,559
	Avg($Temp_{max}$)	32,778	32.41	1,011.377055
Min. Pressure	Min. Pressure	-7,399.977316	-8.31	890.911445

Moreover, As the model improves per the same criteria (such as fixed value (p-value 0.05), AIC, SBC), repeats the procedure. When all remaining variables include a p-value more significant than a certain level when added to the model, the stopping criteria are fulfilled. As mentioned before, it reaches this point, forward selection will stop and will be left with a model that only contains variables with p-values greater than a certain threshold.

One of the estimates used by the regression model is a standardized coefficient. It is essentially the measure of the standard deviation of the variable as it progresses. Its primary advantage is unitless. For instance, the dataset includes temperature, wind, and holidays, which might affect the output while having different units. Temperature is measured using Celsius while the wind is measured in km/hour and holidays are essentially a boolean value.

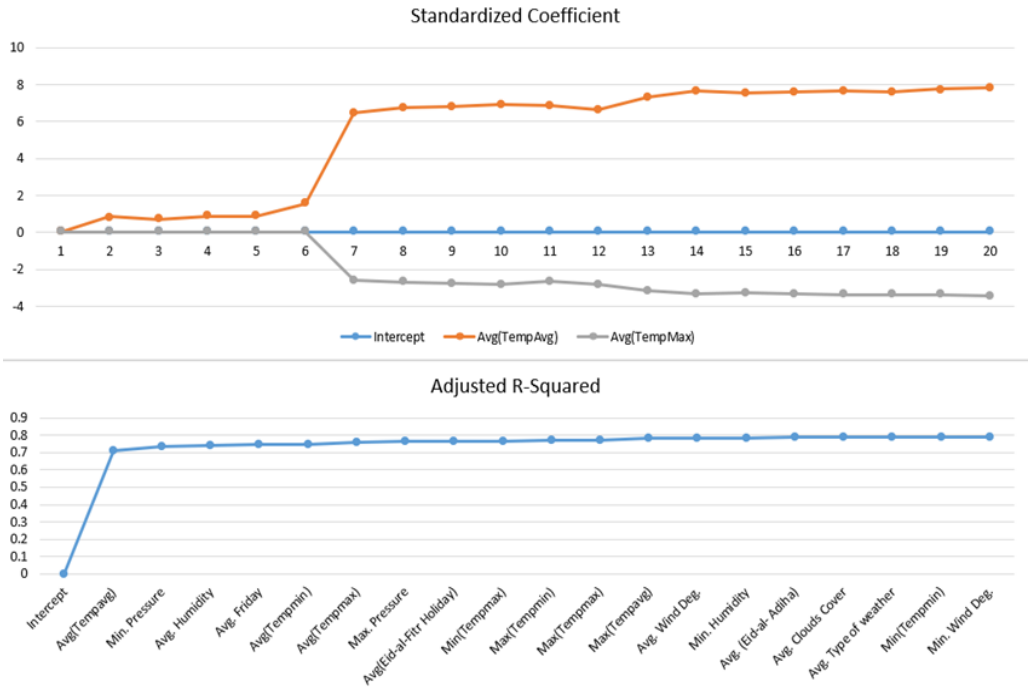


Figure 2.3. Coefficient Progression for Load

The coefficient progression for load falls between 10 and -4, where the majority of the variables are between -1 and +1. The max coefficient progression for load demand belongs to Avg($Temp_{avg}$), and Avg($Temp_{max}$) has the lowest coefficient progression for load value as it progresses.

As it is mentioned before, one of the measures that the model uses to measure the significance of a predictor is adjusted R^2. The adjusted R^2 is a modified version of R^2. The value of R^2 determines the importance of the given predictor.

In this step, most of the predictors are quite close to each other, which indicates that they have a similar impact on load consumption. It can be clearly seen after moving from step-0, which is a null model (intercept), to the following steps, which start incorporating the predictors into the steps, the values remain almost the same for all predictors as is shown in Figure 2.3.

The algorithm uses several criteria to select a model from a finite set of models. BIC, SBC, AIC, Adj R^2 are among the criteria used by the regression model. Due to the fact,

the predictors have a relatively similar impact ratio on the data, and these criteria produce an identical flow of value as they move away from the intercept, as shown in Figure 2.4. The graph shows that the model has selected the last value as the best criteria value. However, it is different for SBC as it established the 13th step instead of the last step. Nevertheless, in terms of the output, it remains as is.

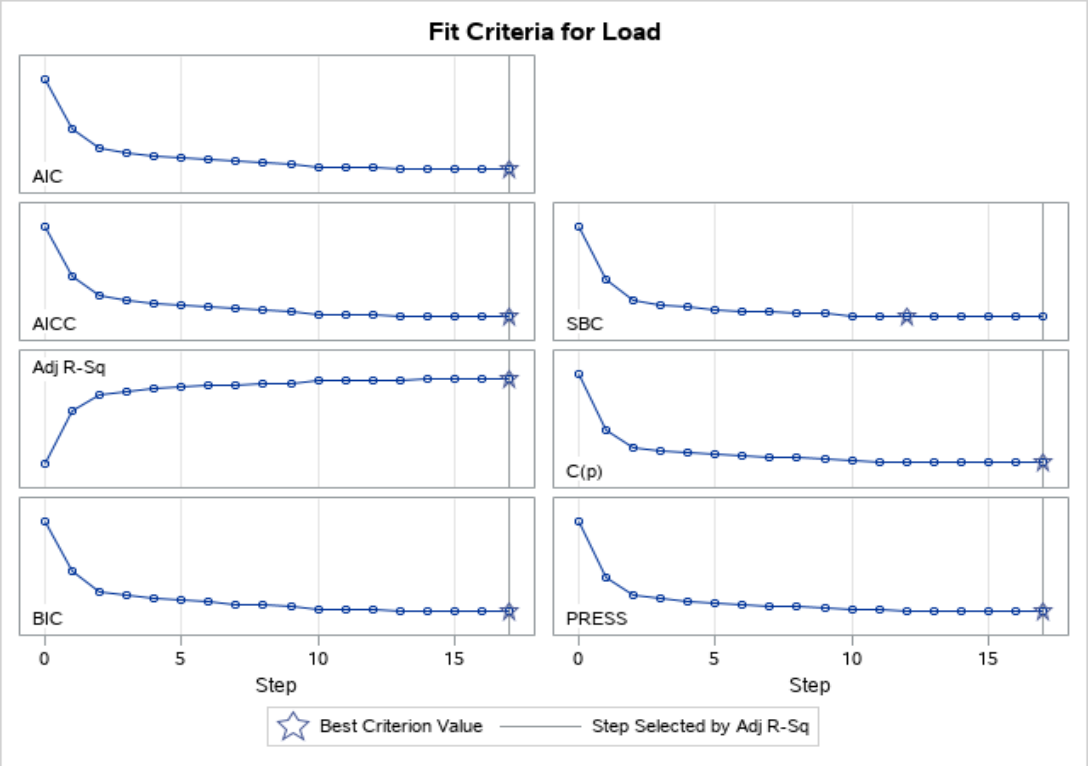


Figure 2.4. Fit criteria for load

2.3. Backward Elimination

In contrast to forward selection, backward selection starts with the entire set of attributes instead of one single attribute. Then removes the least significant attributes gradually according to an objective function until it satisfies the termination condition.

The first variable to be removed is the one that relates the minimum to the reduction of predictive error sum of squares (PRESS). Assuming that there are many insignificant variables, the procedure begins by removing the following most

insignificant variable. When all variables are essential or all, but one has been discarded, the process is terminated (Figure 2.5.).

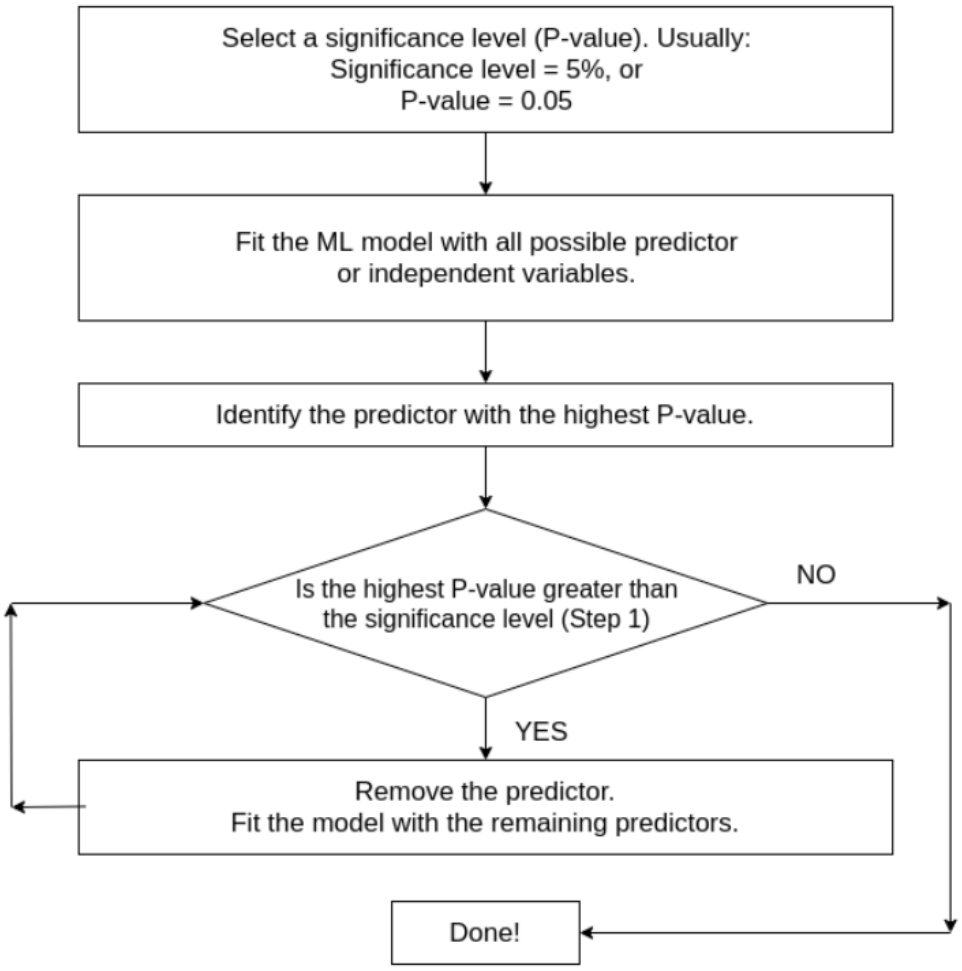


Figure 2.5. Backward elimination diagram

The first step in backward elimination is to choose a confidence interval or p-value. In most situations, a 5% significance threshold is selected, which results in a p-value of 0.05. then fit the model with all of the attributes provided. After that, obtain the characteristic or predictor with the greatest p-value. Moreover, if the p-value is more significant than the significance level, the first step removes the variable from the dataset. If the p-value is the highest in the set, less than the significance level, it is the last step, which means that it is done, trying to delete the component from the dataset and then re-fitting the model using the new dataset. Return to step 3 after fitting the model to the new dataset. Once if it is in step 6, It signifies that the feature selection procedure is complete. Furthermore, it performed backward elimination to

successfully filter out elements that were not important enough for the model. (Figure 2.5.) [20].

Figure 2.6. depicts the first three steps of the backward elimination procedure for one of the models in this study. All variables are introduced to the model and then eliminated one by one, with the variable with the highest probability of (p-value) removed. For instance, in the first step, the variable (Max. Type of Weather) with the highest p-value among the other variables were excluded from the dataset.

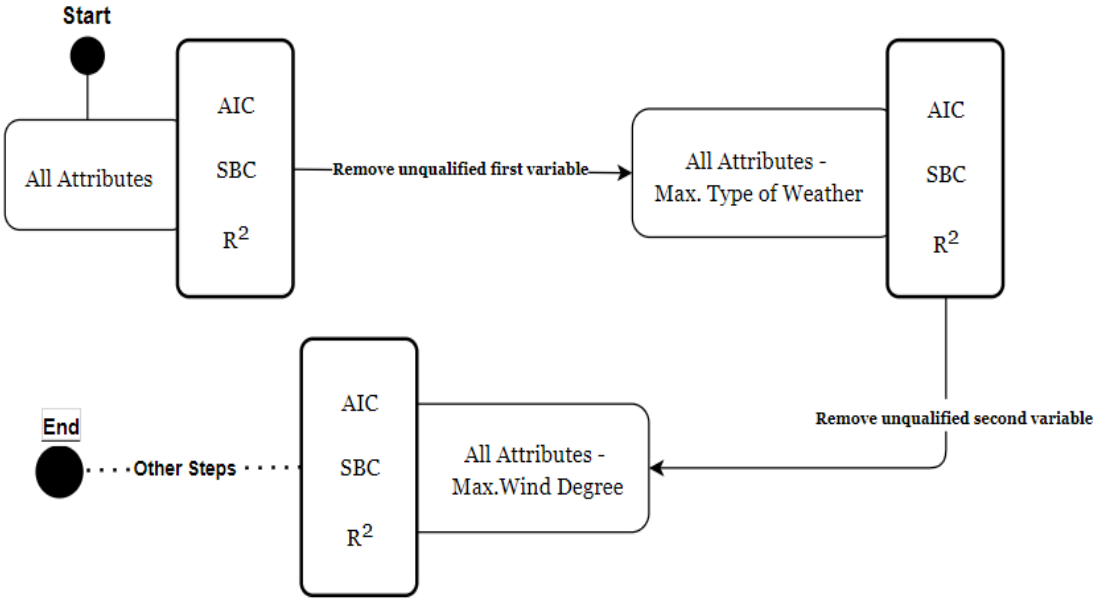


Figure 2.6. Three steps of Backward elimination diagram

The same technique is repeated for the subsequent steps, whereas the final step is if one of the variables has the highest p-value in the set and is less than the significance level.

Table 2.6. First step of Backward elemination model

Step 0.	Errors	Adj R^2 AIC SBC	0.7892 20,430 19,742	
Effect Entered: Intercept	Parameter Estimates	Estimates	t-Value	Standard Error
	Intercept + All variables	6,722,182	4.89	1,375,210

The results show that the first model starts with all variables in the dataset. For the next step, the eliminator removes one of the variables (Max. Type of Weather) that has the highest p-value compared to all other candidate variables shown in Table 2.7.

Table 2.7. Backward elimination in the second step (removing one variable)

Step 0.	Errors	Adj R^2 AIC SBC	0.7895 20,428 19,736	
	Parameter Estimates	Estimates	t-Value	Standard Error
Effect Entered: Intercept	Intercept	6,722,182	4.89	1,375,210
	+ All variables – Max.Type of Weather			

Also, for the next step, the model has removed the variable (Max. Wind Deg.), which has the highest p-value between all other variables, shown in Table 2.8. Table 2.9.

Table 2.8. Best 10 Removal Candidates variables

Best 10 Removal Candidates			
Rank	Effect	Log p-Value	Pr > F
1	Max. Type of Weather	-0.1732	0.8410
2	Max. Wind Deg.	-0.2907	0.7478
3	AvgSunday	-0.3656	0.6938
4	AvgHolidayOther	-0.3704	0.6905
5	Max. Wind Speed	-0.5609	0.5707
6	Avg. Wind Speed	-0.5829	0.5583
7	Avg. Pressure	-0.6041	0.5466
8	$Min(Temp_{avg})$	-0.6156	0.5403
9	Min. Type of Weather	-0.7203	0.4866

Table 2.9. shows the result of the third step, which removed (Max. Wind Deg.) from the dataset.

Table 2.9. Backward elimination in the third step removing (MaxWind_Deg.)

Step 0.	Errors	Adj R^2 AIC SBC	0.7897 20,426 19,729	
Effect Entered: Intercept	Parameter Estimates	Estimates	t-Value	Standard Error
	Intercept	6,722,714	4.90	1,372,637
	+ All variables – Max.Type of Weather – Max. Wind Deg.			

In addition, the standardized coefficient is essentially the measure of the standard deviation of the variable as it progresses. Its primary advantage is that it is unitless, as mentioned before.

The coefficient progression for load falls between 14 and -8, where the majority of the variables are between -1 and +1. The max coefficient progression for load demand belongs to Avg($Temp_{avg}$), and Max($Temp_{avg}$) has the lowest coefficient progression for load value as it progresses.

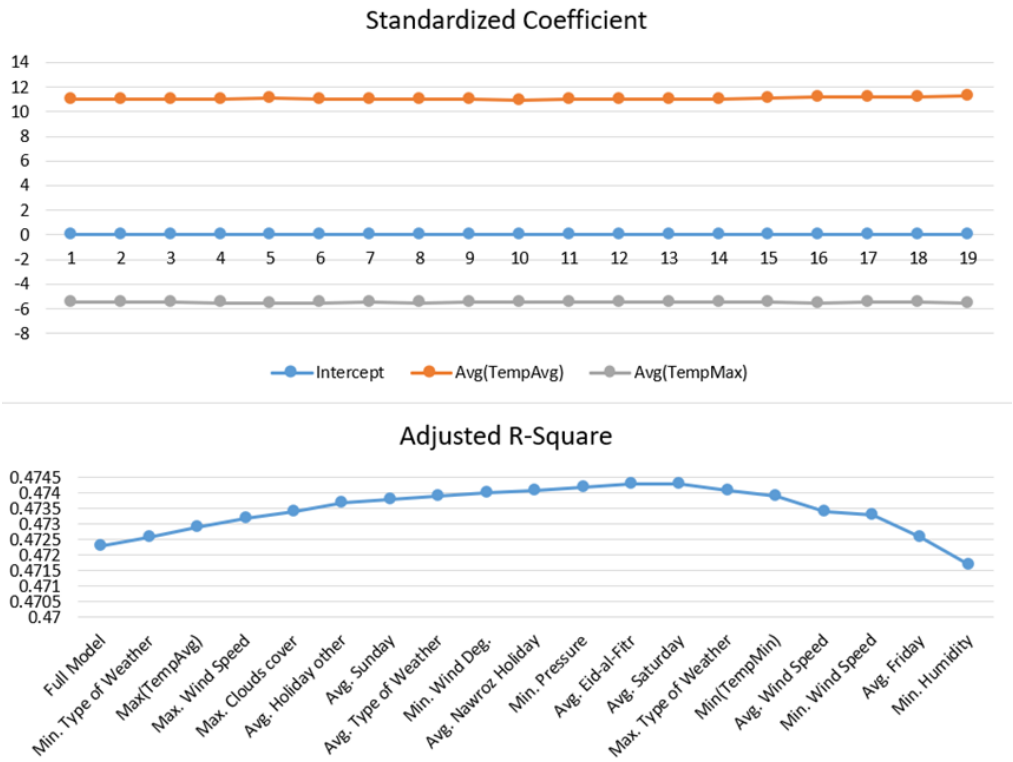


Figure 2.7. Coefficient Progression for Load

Figure 2.7. shows the changes in standard coefficient through each step. One of the most important factors affecting the method decision of eliminating the attributes is the standard coefficient. But as it can be seen in the figure, there is not any change in value. They stay the same for both Avg($Temp_{avg}$) and Avg($Temp_{max}$) throughout the process.

The same criteria are used by the backward elimination method as forward method: BIC, SBC, AIC, Adj R^2. for SBC. It can be clearly seen that all criteria have the same flow except Adj R^2. The algorithm's peak, top, or most preferred value lies in the 10th step of the elimination. The flow of others is pretty similar, although the preferred values differ from one method to another, as shown in Figure 2.8.

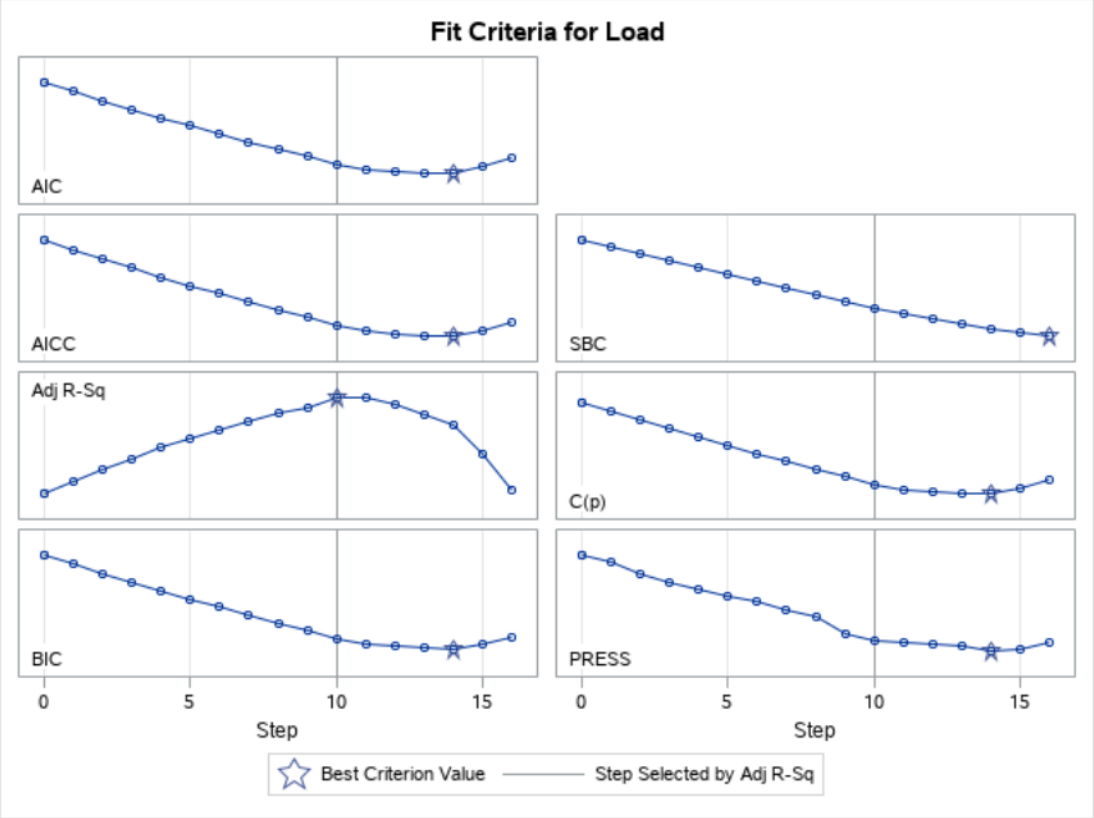


Figure 2.8. Fit Criteria for Load of Backward elimination method

2.4. Stepwise Method

A variable that entered the model in the earlier stages of selection may be discarded in the later stages in a stepwise process. The computations used for variable similarities and differences are the same as those used for forward and backward selection. The stepwise approach is a forward selection process, but the probability of eliminating a variable is contemplated at each point, as in backward elimination. The number of variables maintained in the model is determined by the degree of importance assumed for variable inclusion and exclusion. It depicted one of the sample's phases in (Figure 2.9.).

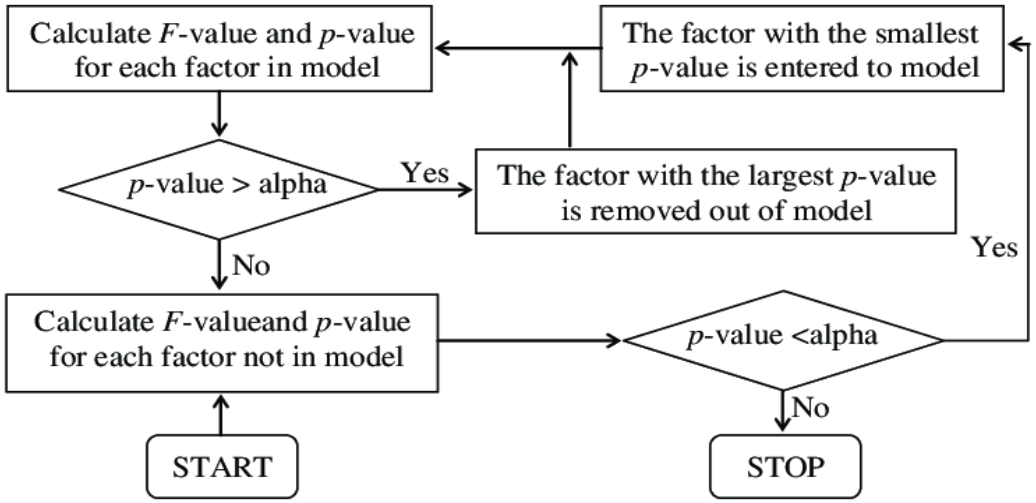


Figure 2.9. Stepwise method diagram

The first step is to establish a significance threshold for determining when to include a predictor in the stepwise model. This threshold is known as the Alpha-to-Enter significance level, and it is denoted by α_E. It should also provide a significance level for determining when a predictor should be removed from the stepwise model. That is, first:

a. Indicate the Alpha-to-Enter significance level that is usually larger than the standard 0.05 threshold so that entering predictors into the model is not too difficult.

b. Specify an Alpha-to-Remove significance level that will typically be greater than the usual 0.05 level so that it is not too easy to remove predictors from the model [30].

At each step, the process applies an important independent variable to the model (if any), selecting the variable that minimizes the Akaike information criterion (AIC), which calculates the relative consistency of a predictive model.

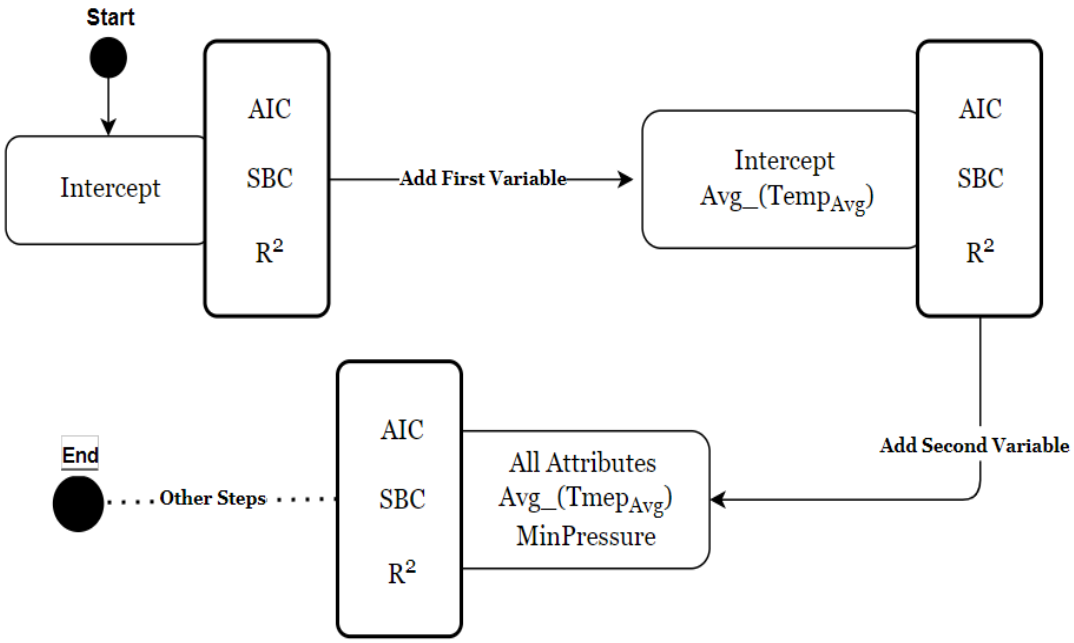


Figure 2.10. Three steps of Stepwise selection method diagram

Since stepwise regression is the combination of both backward and forward regression, there is a high probability of getting similar results, especially in the early steps of the process. It starts with intercept then applies backward regression to the model. The intercept has a t-Value of 175.2 with Adj R^2 of 0, shown in Table 2.10. The attributes were candidate of ten significance variables for the next step shown in Table 2.11.

Table 2.10. First-step result of Stepwise selection method

Step 0.	Errors	Adj R^2	0	
		AIC	21,717	
		SBC	20,872	
Effect Entered: Intercept	Parameter Estimates	Estimates	t-Value	Standard Error
	Intercept	1,346,456	175.2	7,685.300005

Table 2.11. Best 10 entry candidates variables for Step 1

Best 10 Entry Candidates for Step 1			
Rank	Effect	Log p-Value	Pr > F
1	Avg($Temp_{avg}$)	-532.387	<.0001
2	Avg($Temp_{min}$)	-523.308	<.0001
3	Avg($Temp_{max}$)	-512.609	<.0001
4	Max($Temp_{max}$)	-428.8936	<.0001
5	Max($Temp_{avg}$)	-413.5074	<.0001
6	Max($Temp_{min}$)	-400.5002	<.0001
7	Min($Temp_{avg}$)	-369.4979	<.0001
8	Min($Temp_{min}$)	-364.7489	<.0001
9	Min($Temp_{max}$)	-363.7383	<.0001
10	AvgPressure	-270.3600	<.0001

Avg($Temp_{avg}$) has the highest rank based on the p-values, while AvgPressure has the lowest rank. For this reason, Avg($Temp_{avg}$) has a higher chance of getting passed to the next step, as it is shown in Table 2.11.

Table 2.12. Second-step result of Stepwise selection method

Step 1.	Errors	Adj R^2	0.7137	
		AIC	20,658	
		SBC	19,819	
Effect Entered: Avg($Temp_{Avg}$)	Parameter Estimates	Estimates	t -Value	Standard Error
	Intercept	277,153	11.72	23,638
	Avg($Temp_{avg}$)	37,969	45.94	826.557822

As the model proceeds and passes the next step, it has to recalculate values once again for each attribute. According to Table 2.12, Avg($Temp_{avg}$) has been added to the selected candidates in the second step. The t-value is 45.94, with a standard error of roughly 826. The Adj R^2 slightly increases from 0 to 0.7 as it moves away from intercept in the next step.

After the recalculation process, they are again ranked in the selection pool, as it is shown in Table 2.13. The Min. Pressure has the highest chance as it is ranked number one in the collection. It can be clear that there is a quite difference between step one and step two entry candidate p-values as the previously selected candidates play an important role in determining the t-Value and p-Value.

Table 2.13. Best 10 entry candidates variables for Step 2

Best 10 Entry Candidates for Step 2			
Rank	Effect	Log p-Value	Pr > F
1	Min. Pressure	-35.4757	<.0001
2	Avg. Pressure	-29.9936	<.0001
3	Max. Pressure	-24.415	<.0001
4	Avg. Humidity	-22.5411	<.0001
5	Max. Humidity	-18.7974	<.0001
6	$Max(Temp_{min})$	-12.4616	<.0001
7	Avg. Type of Weather	-11.1241	<.0001
8	$Max(Temp_{avg})$	-10.371	<.0001
9	Min. Humidity	-9.2327	<.0001
10	$Avg(Temp_{max})$	-7.4911	0.0006

For the third step, Min. Pressure is the selected candidate with a t-Value of -8.31 and Adj R^2 of 0.735, as is shown in Table 2.14. This way, the process continues until it reaches the final step and decides which candidates to choose and which has the most impact on the predicator. It helps to eliminate in-significant predicators on the resulting predicted value.

Table 2.14. Third-step rsult of Stepwise sselection method

Step 2.	Errors	Adj R^2	0.735	
		AIC	0.735	
		SBC	19,759	
	Parameter Estimates	Estimates	t -Value	Standard Error
Effect Entered: Min. Pressure	Intercept	7,871,199	8.61	914,559
	$Avg(Temp_{max})$	32,778	32.41	1,011.377055
	Min. Pressure	-7,399.977316	-8.31	890.911445

The standard coefficient goes through quite a change as it moves from the first step to the latest step. $Avg(Temp_{avg})$ gets to the highest value in 13 and stays the same throughout the process. In terms of intercept, it is always the same and remains the

same throughout the process. However, Min($Temp_{avg}$) undergoes a significant decline in value as it passes through step 12, as shown in Figure 2.11.

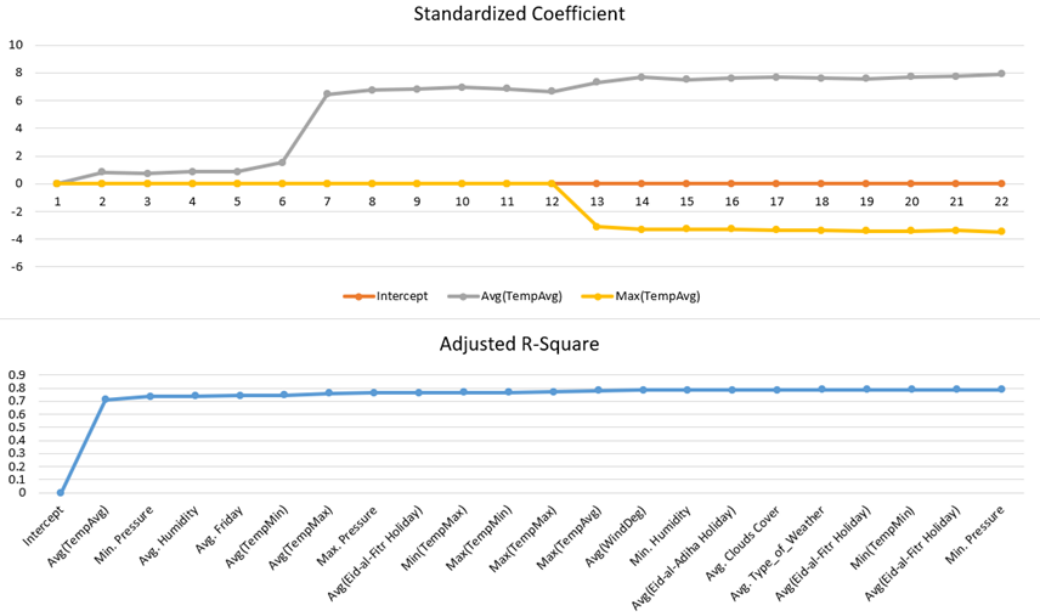


Figure 2.11. Fit Criteria for Load in Stepwise method

According to (Figure 2.12.), all of the fit criteria reach the peak in the last step of the process. Although they reach that value quite before, the model has to recalculate for each step. They all decline except Adj-R^2, which gradually rises as it reaches step-4. Adj-R^2 is among the most significant factors which contribute to candidate selection.

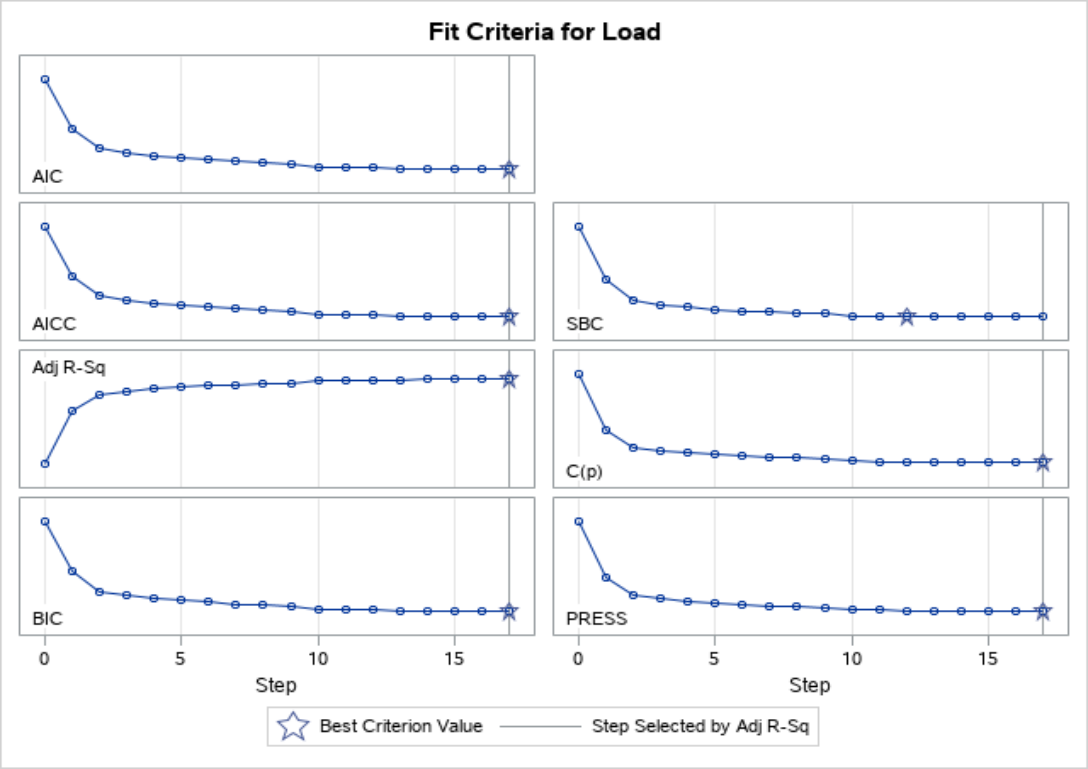


Figure 2.12. Fit Criteria for Load in the Stepwise method

Defining the forecasting method (Multiple linear regression) and using it for Short term Load forecasting such a statistical technique, after that defining each method of Multiple linear regression like Forward selection, Backward elimination and Stepwise method, with showing a sample of 3 steps to one of the models in the study, with the working technique diagram for each method.

CHAPTER 3. MODELING

Dataset is one of the major elements of prediction. The results highly depend on the provided dataset. In addition, preparing and processing the dataset also highly impacts the performance of the algorithm. We explain the details of the dataset, algorithm, and modeling in this chapter.

3.1. The Dataset

Three datasets are used in this thesis. The first one is the load demand dataset provided by the control distribution center of Sulaimanyah city, which contains (Location name, Feeder Name, 24 hours of Load consumption) for 6-years (2014-2019). The data for each feeder in the city is recorded once every hour. Therefore, it is concluded that there are 24 records for each feeder. The load consumption unit of the dataset is Ampere.

The second dataset contains the weather data as described below. This dataset used in this study is to find the effect of weather on the amount of power distribution in the city. The region has four complete seasons in terms of weather. However, summer constitutes most of the days of the year. Then comes winter, fall, and spring consecutively. It is worth mentioning that power distribution during the season days is almost the same unless there is a hardware failure or any other faults that reduces power generation.

The plan for distributing power generally changes with variations of the weather. For example, people consume more power in the summertime because they have to use a cooling system such as air conditioning due to high temperatures. For this reason, the Control Unit (CU) will have to decrease the supplying period to make sure they can even provide the power for the season. While weather changes significantly impact power distribution, CUs do not keep track of the temperature.

The last dataset used in the study is the holidays. During the holidays, the control center usually provides more hours than any other regular day. While weekends are off, but it does not affect the hours. The most effective holidays are Eid al Fitr, Eid al Adha, Newroz, etc.

3.2. Data Preparation

As mentioned before, the control distribution centers in the cities record the hourly load for each feeder in each area. The attributes placed in the dataset are region, feeder name, and 24-hours Table (Table 3.1.).

It can be clearly seen that the data is recorded in such a way that the attributes are hours from 00:00 to 23:00, and the feeders are listed as a row. There is a separate sheet for each day of the month in the year. Therefore it is required to integrate the data and collect all hours and days in one large dataset to analyze the data further. It had to be transposed the data to accomplish. After, the columns become the rows, and the rows are moved to columns. This way, it is much easier to manage and further process the data. The outcome of the process and the transition of data is shown in (Table 3.3.).

The input data is converted from hourly form to daily form to organize the data further as it is more interested in the sum of the power load based on daily demand rather than hours. In daily conversions, the number of independent variables has increased. For example, there are 24 highest temperatures in a day. The highest, lowest and average values of the 24 hourly highest temperatures can be found as shown in a subset of the data in Table 3.2.

In other words, when a variable is converted into a daily variable, it is shown as three variables. (Figure 3.1.) shows that temperature is the essential variable. For this reason, after the daily conversion of temperatures, nine independent variables were included in the model.

The city's data was fetched from Open Weather website to integrate and formalize the impact of weather. The data is based on altitude and latitude. The collected variables were latitude, longitude, minimum, average, and maximum temperature, wind speed, wind degree, humidity, precipitation, in 1, 3, 6, 12, and 24 hours, cloud cover, pressure, weather condition. Finally, the weather data is available in altitude and latitude. Therefore, it had to determine the coordinates for each area to match the weather data with the correct locations by date. At this stage, the weather attributes were combined with the original load dataset [42].

In addition, many fields, such as religious holidays (Ramadan, Adha, Nawroz, and weekends (Friday, Saturday, Sunday)), that would possibly affect the electricity demand are added. Ramadan, Adha, Newroz holidays are 3,4,5 days, respectively.

Table 3.1. A subset of raw hourly load demand data according to the location

Hour	Location / Feeder			
	Rzgary / 1 (Amp)	Shaheed / 20 (Amp)	Malkandi / 39 (Amp)	Azmer / 43 (Amp)
01:00	70	55	55	100
02:00	70	50	35	80
03:00	70	50	50	70
04:00	50	45	45	75
05:00	50	40	50	75
06:00	60	55	60	80
07:00	90	55	60	90
08:00	135	45	60	90
09:00	225	45	60	95
10:00	265	50	70	100
11:00	280	50	75	110
12:00	285	55	80	115
13:00	290	50	75	115
14:00	275	50	75	90
15:00	210	55	60	90
16:00	200	55	75	110
17:00	100	50	85	120
18:00	55	50	75	115
19:00	85	60	75	115
20:00	90	60	80	110
21:00	80	55	75	115
22:00	90	55	75	110
23:00	80	55	70	110
00:00	75	55	65	105

Table 3.2. Subset of the dataset after migrating data from hour to daily load

Day	Month	Year	Date	Avg_TempAvg	Max_TempAvg	Min_TempAvg
1	2	2014	2/1/2014	5.73875	10.33	1
2	2	2014	2/2/2014	1.65	5	-1
3	2	2014	2/3/2014	-1.12	0	-2
4	2	2014	2/4/2014	0.042083	1.32	-1
5	2	2014	2/5/2014	1.813333	5.95	-2.4
6	2	2014	2/6/2014	0.687916	9	-7
7	2	2014	2/7/2014	0.664583	10	-7
8	2	2014	2/8/2014	2.395	11.99	-7
9	2	2014	2/9/2014	3.98375	13.67	-5
10	2	2014	2/10/2014	4.773333	14.12	-6
11	2	2014	2/11/2014	6.975416	14	-4
12	2	2014	2/12/2014	10.96875	16.28	6.53

3.3. Seasonal Data

Seasons have a direct effect on how much energy is used. The summer season is scorching sand dry (T_{max} = 46 ◦C), and the winter season is icy (T_{min} = -8 ◦C) in Sulaymaniyah. In the preliminary analysis phase of the study, the distributions of all independent variables with electrical load were prepared. As a result of the scatter plots, a high nonlinear relationship with temperature and load has been observed (Figure 3.1.).

When the coefficient of determination, which examines the linear relationship with electric load, was considered, it was seen that the highest relationship was again temperature (R^2_{Tmean} = 0.35). However, as seen in (Figure 3.1.), the coefficient of determination could preferably represent the electric load in the nonlinear state (R^2_{Tmax} = 0.65). While a linear relationship is expected between temperature and electric load, the nonlinear relationship is an essential part of this study.

It can be clearly seen that load is more affected by the mean temperature in (Figure 3.1.) As can be seen in the same way, the variable in which the distribution of temperature in electricity load is the most robust and the little spread is the mean temperature ($R^2_{Tmean} \approx 0.715$). It has been observed that if the mean temperature is nearly 20°C, load demand is the lowest. There is an increase in consumption both to the right and to the left of the temperature. This situation causes nonlinearity.

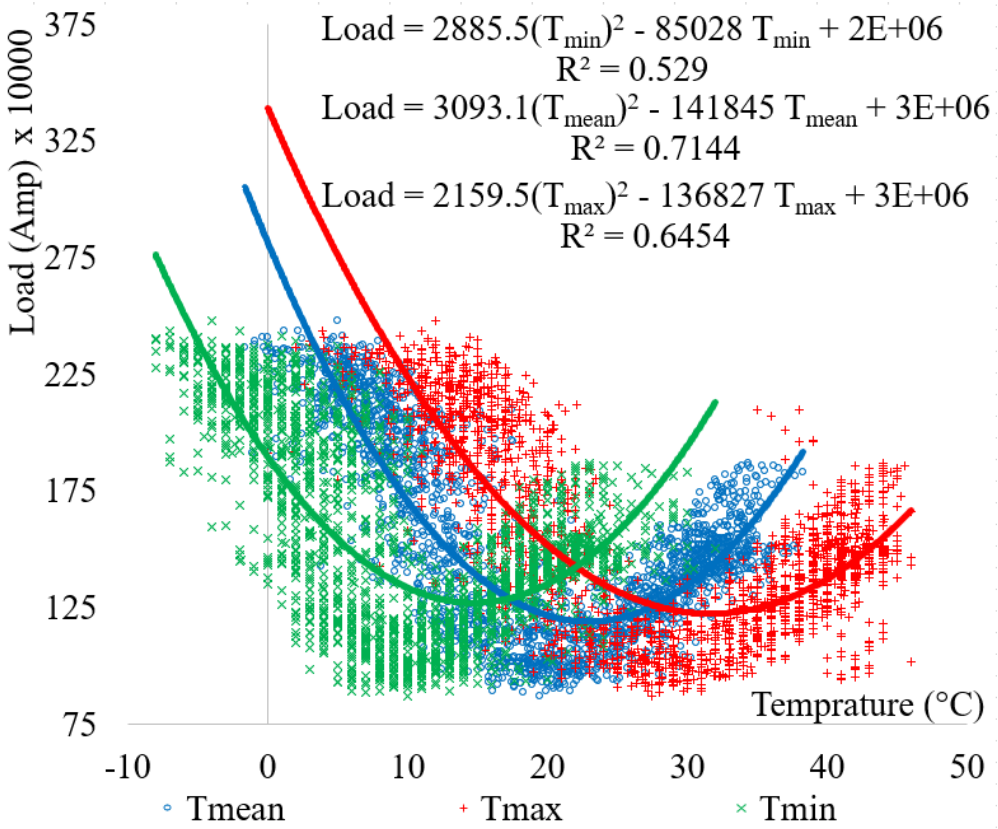


Figure 3.1. Load – Temperature plot of the city

After determining the nonlinear behavior, the demand for the electricity load over time series is investigated. In the time series, sudden increases and decreases were observed in the transition from winter to summer and from summer to winter. When the load is analyzed as a time series, the dates when load demand behavior changed are shown in (Table 3.3.), and seasonal models in the study are divided according to these dates.

Table 3.3. Split dates of the models

Winter to Summer (WS Model)		Summer to Winter (SW Model)	
First date	Last date	First date	Last date
-	-	01/02/2014	27/04/2014
28/04/2014	19/10/2014	20/10/2014	03/05/2015
04/05/2015	18/10/2015	19/10/2015	01/05/2016
02/05/2016	16/10/2016	17/10/2016	30/04/2017
01/05/2017	15/10/2017	16/10/2017	29/04/2018
30/04/2018	14/10/2018	15/10/2018	05/05/2019
06/05/2019	19/10/2019	20/10/2019	30/11/2019

Electric load estimation has been carried out over three different models. In this study, it is foreseen to make improvements by trying to be seasonally adjusted. In the first model, all historical data were used between 01/02/2014 - 31/11/2018 (Model 1). In the second approach, the data divided according to (Table 3.3.) was used in winter to summer (WS Model), and summer to winter (SW Model) models divided by seasonality. The general estimation model made using WS and SW models is expressed as Model 2.

3.3. Goodness-of-fit statistics

Mean absolute percent error (MAPE) and coefficient of determination (R^2) values were calculated for the errors and fit statistics of the models obtained in this study. MAPE and R^2 equations are given in (Equation 3.1) and (Equation 3.2), [21], [25], [26].

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Y}Fi - yi}{yi}\right| 100\% \tag{3.1}$$

$$R^2 = \left[\frac{\sum_i (yi - \hat{y})^2}{\sum_i (yi - \bar{y})^2}\right], 0 \leq R^2 \leq 1 \tag{3.2}$$

In the equations, the value of Y_i indicates the actual electric load demand, $\hat{y}_{Fi}$ shows forecasts, $\bar{y}$ is the average of the load series, and n is the number of days.

Here, the choice of MAPE and R^2 as error terms is important. R^2 shows the behavior of the entire series, while MAPE shows the error between realization and forecast. Suppose there is always the same difference between the forecast and the realization. For example, if the realization is 80% more than the forecast throughout the whole series, the MAPE will be 80%, even though the R^2 value is close to one. R^2 is usually expected to be high when MAPE is low. Otherwise, the data series may be random and away from seasonality.

This section provides all modeling steps, such as preparing the dataset as a required format for the models, collecting the data, and integrating with the weather attribute data after cleaning and organizing the data format. The religious holidays are added data. Dividing the dataset according to people's consumption of electric power at the seasons (Summer season and Winter season) seriously influences the models. Following the determination of the nonlinear behavior, the demand for the power load over time is examined in three models (All data, Winter to Summer dataset, and Summer to Winter dataset). The errors were measured by MAPE and R^2, which were defined in detail before.

CHAPTER 4. RESULTS

In this thesis, two approaches are compared. The first of these approaches is load estimation using all data. The second approach is to determine whether dividing the seasonality effect dataset shown in (Figure 4.1.) will improve the load estimate or not.

The lowest, highest, and average of the pressure, humidity, wind speed, wind direction, cloud cover are also included in the models. Weather events are listed ordinally and converted from hourly data to daily data accordingly, and the minimum, maximum and average value has been obtained. Weekend holidays in Iraq are Fridays and Saturdays. For compatibility with other countries, Fridays, Saturdays, and Sundays were included as dummy variables in the model. Similarly, it has different effects on religious and cultural holidays. For this reason, Eid al-Fitr, Eid al-Adha, and Nowruz holidays were included in the model as separate dummy variables. Other public holidays are kept under a single dummy variable.

In the next step, separate regression equations were obtained for the model covering the whole time (Model 1), WS model, and SW models. Forward selection, backward elimination, and stepwise selection were used to eliminate the independent variables (Table 4.1.). In the elimination process, Akaike information criterion (AIC), Schwarz Bayesian information criterion (SBC), adjusted R^2 (Adj R^2) were used as criteria in determining the best model. Another method is to look at the statistical significance of variables in the elimination process. Models in statistical significance where p-value is below 0.1 ($p<0.1$) and 0.05 ($p<0.05$) were also determined. Thus, the equations of Model 1, WS Model, and SW Models; were determined with three different elimination methods and five different elimination criteria (15 different approaches). The numbers of variables that are useful in these models are shown in (Table 4.1.).

Table 4.1. Number of independent variables in the proposed models

	Forward			Backward			Stepwise		
Models	All	WS	SW	All	WS	SW	All	WS	SW
AIC	21	20	16	21	20	15	18	20	15
SBC	13	8	8	13	15	9	13	8	5
Adj R^2	23	24	17	23	24	17	21	24	17
$p<0.1$	19	20	14	23	24	17	18	20	12
$p<0.05$	18	19	10	23	24	17	17	19	7

It was seen that all three models have common independent variables as a result of 15 different approaches (Table 4.2.) Model 1 has the average of hourly maximum, minimum and average temperatures, maximum of hourly minimum temperatures, average and maximum humidity, average and minimum cloud cover in all 15 approaches. While the Eid al-Adha variable was found in 14 models, the minimum hourly minimum temperatures, average pressure, maximum pressure, lowest humidity, average wind direction, maximum wind direction, and Friday variables were observed in at least ten models.

The maximum of the hourly minimum temperatures, maximum pressure, and minimum cloud cover is also found in 15 approaches in the SW Model. The average and minimum hourly average, maximum and minimum temperatures, maximum and minimum wind speeds, and holiday dummy variables existed in at least ten models.

In the WS Model, the average of hourly maximum, minimum and average temperatures, maximum pressure, average humidity, and Friday variables are in 15 approaches; The minimum pressure is in 14 approaches. The variables of the maximum of hourly average, minimum and maximum temperatures, minimum of hourly minimum and maximum temperatures, minimum humidity, average and minimum wind direction, average weather events, average cloud cover, Eid al-Fitr, and Eid al-Adha holidays were observed in at least ten models.

Table 4.2. Number of usage independent variables in proposed models

Variable	Model 1	SW Model	WS Model
Intercept	15	15	15
Avg($Temp_{avg}$)	15	13	15
Max($Temp_{avg}$)	8	0	13
Min($Temp_{avg}$)	5	11	0
Avg($Temp_{min}$)	15	13	15
Max($Temp_{min}$)	15	15	13
Min($Temp_{min}$)	13	12	12
Avg($Temp_{max}$)	15	13	15
Max($Temp_{max}$)	9	1	13
Min($Temp_{max}$)	4	10	13
Avg. Pressure	13	5	0
Max. Pressure	13	15	15
Min. Pressure	6	0	14
Avg. Humidity	15	0	15
Max. Humidity	15	10	5
Min. Humidity	12	0	13
Avg. Wind Speed	6	0	0
Max. Wind Speed	0	11	0
Min. Wind Speed	6	10	5
Avg. Wind Degree	12	0	13
Max. Wind Degree	12	0	0
Min. Wind Degree	0	0	10
Avg. Clouds Cover	15	0	13
Max. Clouds Cover	0	0	5
Min. Clouds Cover	15	15	0
Avg. Type of Weather	1	8	13
Max. Type of Weather	5	0	0
Min. Type of Weather	0	0	0
Friday	10	8	15
Saturday	0	0	5
Sunday	0	0	0
Eid al-Fitr Holiday	0	0	12
Eid al-Adha Holiday	14	0	12
Nowruz Holiday	0	0	0
Oher Holidays	0	11	0

4.1. Forecasting Results

The results of three models and 15 approaches obtained were evaluated on MAPE and R^2. When Model 1 was examined, it was seen that the lowest MAPE and highest R^2 values were found in the backward elimination approach and the highest adj R^2, $p<0.1$, and $p<0.05$ elimination in the training dataset (Table 4.3.). In all three screening methods, the same variables are included in the model. In the training dataset, MAPE

and R^2 values were 16.48% and 0.48, respectively. The test dataset MAPE was determined to be 23.95% (Table 4.3.), that the models that gave the best results in training had the lowest MAPE value. Again, in the test dataset, the forward selection had the highest R^2 value (0.3484) in the AIC as a selection criterion.

In the SW Model, the lowest MAPE and highest R^2 were found in four approaches in the training dataset and were determined as 11.01% and 0.7702, respectively. The same independent variables in these four approaches are Adj R^2 criteria in the forward selection, stepwise selection approaches, Adj R^2, and $p<0.05$ criteria in the backward elimination approach. In the test dataset, the lowest MAPE and R^2 values were 12.14% and 0.7673 in the stepwise selection approach SBC criterion and forward selection approach $p<0.05$ criteria, respectively.

In the WS Model, the lowest MAPE and highest R^2 values in the five approaches in the training dataset were 5.93% and 0.7965, respectively (Table 4.3.). The lowest MAPE was found in the test dataset as 12.70% according to SBC criteria in the backward elimination approach. The highest R^2 was determined as 0.6785 according to SBC criteria in the forward and stepwise selection approach.

The four approaches with the lowest MAPE and the highest R^2 values in SW and WS models are common in training (Table 4.3.). In these approaches (Forward Selection - Adj R^2, Backward Elimination - Adj R^2, Backward Elimination - $p<0.05$, Stepwise Selection - Adj R^2), the same independent variables exist in the model. The results of this approach are expressed as Model 2 in the graphs. In SW and WS models, the combination of approaches with the lowest MAPE value in the test dataset (Stepwise Selection - SBC for SW Model, Backward Elimination - SBC WS Model) is expressed as Model 3. The combination of approaches in the test dataset with the highest R^2 value (Forward Selection - $p<0.05$ for SW Model, Forward Selection - SBC for WS Model) is expressed as Model 4. While Model 3 is expected to give the best result as a prediction model, Model 2 can be stated as the best seasonally adjusted training model.

Table 4.3. Goodness-of-fit statistics of the proposed approaches

Approach-Criteria	Model 1				SW Model				WS Model			
Dataset	Train		Test		Train		Test		Train		Test	
Goodness-of-fit statistics	*MAPE*	*Adj R^2*	*MAPE*	*Adj R^2*	*MAPE*	*Adj R^2*	*MAPE*	*Adj R^2*	*MAPE*	*Adj R^2*	*MAPE*	*Adj R^2*
Forward Selection AIC	16.51%	0.4788	25.49%	0.3484	11.03%	0.7698	22.36%	0.1895	5.96%	0.7950	12.90%	0.6261
Forward Selection SBC	16.67%	0.4687	26.35%	0.3481	11.18%	0.7625	19.13%	0.7633	6.32%	0.7656	14.31%	0.6785
Forward Selection Aj R^2	16.50%	0.4798	25.49%	0.3483	11.01%	0.7702	22.55%	0.1758	5.93%	0.7965	12.87%	0.6213
Forward Selection p<0.1	16.53%	0.4775	26.60%	0.3361	11.06%	0.7687	22.26%	0.1881	5.96%	0.7950	12.90%	0.6261
Forward Selection p<0.05	16.54%	0.4765	26.67%	0.3343	11.15%	0.7648	18.85%	0.7673	5.97%	0.7942	12.74%	0.6207
Backward Elimination – AIC	16.49%	0.4796	25.62%	0.2221	11.03%	0.7695	24.77%	0.1060	5.96%	0.7950	12.90%	0.6261
Backward Elimination – SBC	16.67%	0.4693	26.32%	0.3410	11.13%	0.7649	24.15%	0.1252	5.98%	0.7900	12.70%	0.5964
Backward Elimination - Adj R^2	16.48%	0.4805	23.95%	0.2707	11.01%	0.7702	22.55%	0.1758	5.93%	0.7965	12.87%	0.6213
Backward Elimination - p<0.1	16.48%	0.4805	23.95%	0.2707	122.2%	0.0652	105.15%	0.0346	5.93%	0.7965	12.87%	0.6213
Backward Elimination - p<0.05	16.48%	0.4805	23.95%	0.2707	11.01%	0.7702	22.55%	0.1758	5.93%	0.7965	12.87%	0.6213
Stepwise Selection - AIC	16.51%	0.4781	25.71%	0.3382	11.03%	0.7695	24.77%	0.1060	5.96%	0.7950	12.90%	0.6261
Stepwise Selection - SBC	16.67%	0.4687	26.35%	0.3481	11.56%	0.7539	12.14%	0.7434	6.32%	0.7656	14.31%	0.6785
Stepwise Selection - Adj R^2	16.50%	0.4796	25.65%	0.3391	11.01%	0.7702	22.55%	0.1758	5.93%	0.7965	12.87%	0.6213
Stepwise Selection - p<0.1	16.52%	0.4774	26.73%	0.3325	11.09%	0.7672	21.82%	0.2093	5.96%	0.7950	12.90%	0.6261
Stepwise Selection - p<0.05	16.53%	0.4763	26.81%	0.3305	11.49%	0.7564	12.19%	0.7465	5.97%	0.7942	12.74%	0.6207

Four model training dataset results are shown in (Figure 4.1.). The positive results of seasonal decomposition can be easily seen in models other than Model 1 that shows the results with no seasonal split. While Model 1 more consistently predicted winter load results, it was not able to predict summer results. In general, it was observed that the results were within acceptable tests in the training dataset.

In the test dataset, it is easily seen that the results of Model 1 and Model 2 are insufficient, as is shown in Figure 4.2., Figure 4.3., Figure 4.4., Figure 4.5. Model 1 forecasted demands inaccurately in the summer season, just like in the training phase.

Model 2, on the other hand, made very high inaccurate forecasts in the transition from winter-summer and summer-winter. Model 3 and Model 4 made close and acceptable forecasts. It is seen that Model 3 provides better predictions in summer-winter and winter-summer transition compared to Model 4. Better projections in Model 3 show that the MAPE variables are more effective in predictions than the R^2 variables.

Model results are shown in (Table 4.4.) Model 1 had the highest MAPE and lowest R^2 value in both the training and the test dataset. Model 2, the best approach considering only the training dataset according to the seasonal decomposition approach, reduced the MAPE value by 25% in the test dataset and obtained more accurate and compatible results by doubling the R^2. In other words, such an approach will reduce model estimation errors.

Model 3 and Model 4 show the best results obtained according to the test data. Model 3 had the lowest MAPE and highest R^2. This is expected. Another dramatic situation in the results is Model 4, which was obtained according to the highest R^2, has a lower R^2 than Model 3. It is because the errors of the results in the WS Model and the SW Model are high, eliminating the continuity in the prediction. This situation decreased the R^2.

Table 4.4. Goodness-of-Fit Statistics of The Models

Dataset	**Train**		**Test**	
Errors	MAPE	Adj R^2	MAPE	Adj R^2
Model 1	16.48%	0.4805	23.95%	0.2707
Model 2	8.57%	0.8188	18.12%	0.4191
Model 3	8.88%	0.8071	12.40%	0.7133
Model 4	8.83%	0.8105	16.78%	0.5581

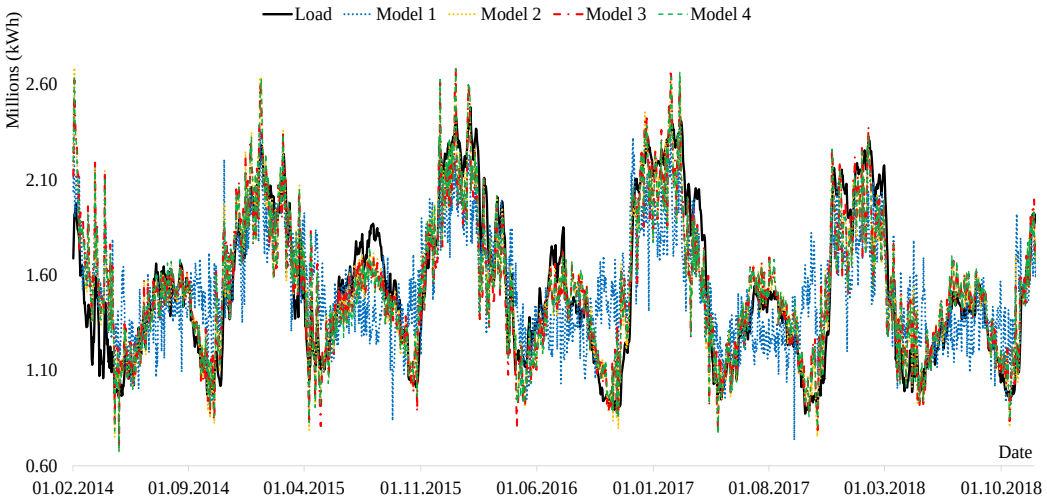


Figure 4.1. Training dataset load estimations

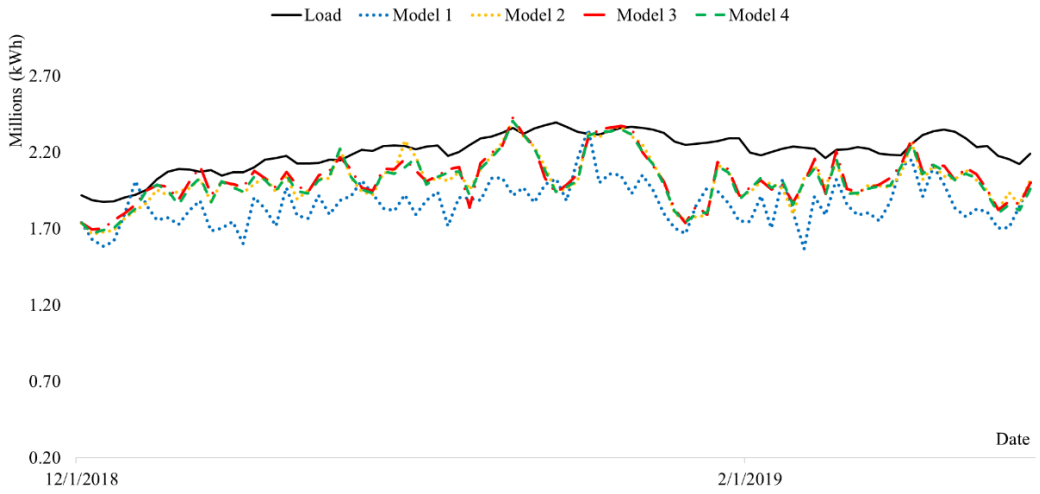


Figure 4.2. Test dataset load forecasts (part-1)

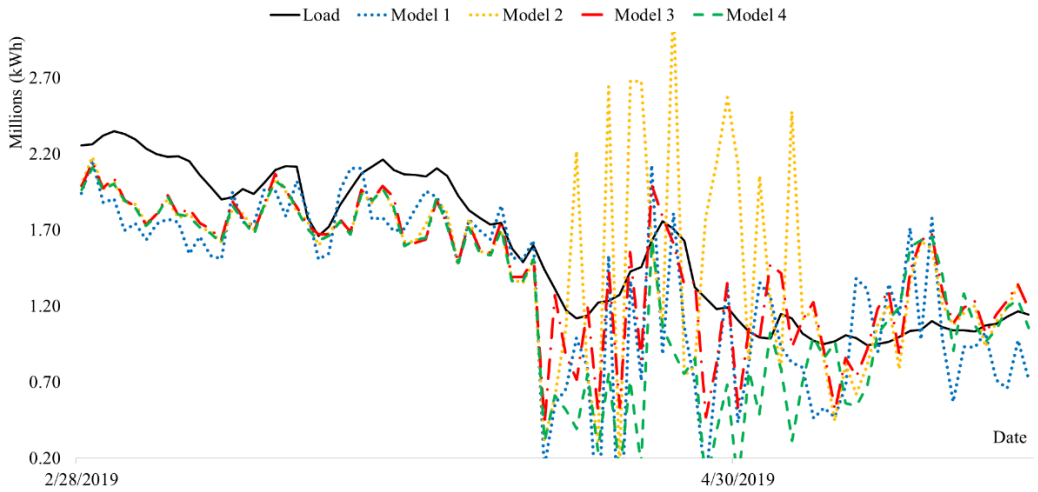


Figure 4.3. Test dataset load forecasts (part-2)

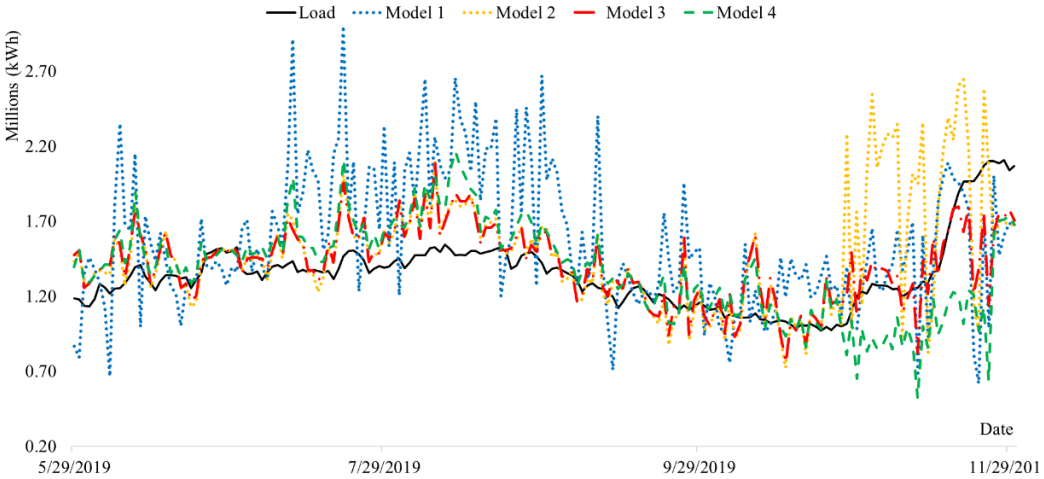


Figure 4.4. Test dataset load forecasts (part-3)

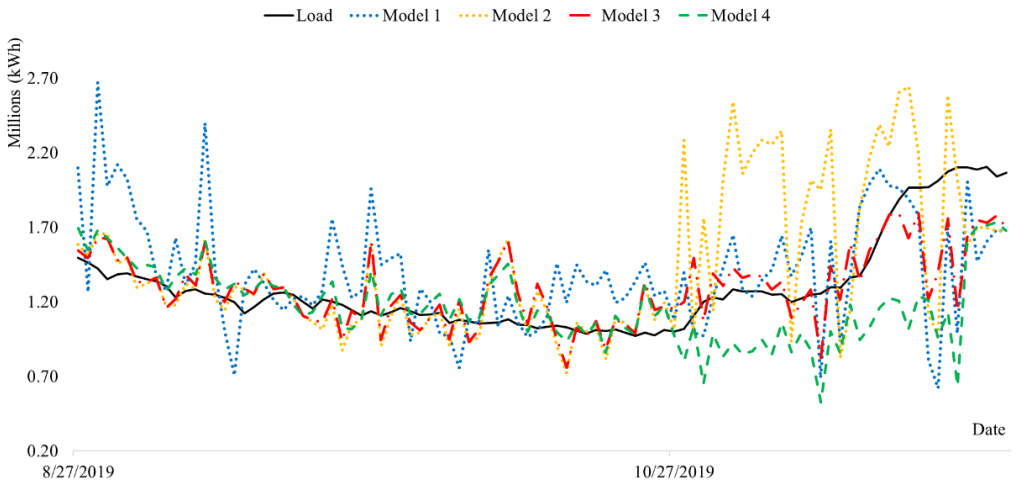


Figure 4.5. Test dataset load forecasts (part-4)

In this section, the results of the thesis are presented. The performance of the models is compared based on the data for the whole year with respect to two time periods from winter to summer and summer to winter. Estimates can be more accurate if they are based on consumption habits. Another significant finding in the thesis is that in the training dataset, backward elimination produces more accurate results, which is lowest MAPE is (16.48 %), and highest R^2 is (0.4805), compared with forward and stepwise selection. Similar to Tuaimah et al. used MLR to create the estimation models for historical data of Iraqi power system which are; Winter season model and Summer season models. The instability of the temperature in Iraq has a considerable influence on the accuracy of the models; that is why they got the critical temperature model [30].

Also, Vasquez et al. used MLR, for the annual energy consumption. The performance of their model result was measured by MAPE was (0.74) [22]. Otherwise, Amber et al. used the same model to estimate the daily energy demand, but only for university buildings, their results affected by temperature, weekday-weekend situations [37].
In contrast, our study is quite different from the other studies because of the sensitivity of the load consumption at the seasons.

CHAPTER 5. CONCLUSION

The power distribution centers in Iraq use a traditional approach to predict power consumption for upcoming weeks, months, and years. The major issue with this approach is the lack of accuracy and the time required to foresee the power load, consumption, and demand.

This thesis was conducted to determine whether machine learning algorithms can be used to mitigate this process or not. MLR, which is one of the most common algorithms used in the area.

According to the study results, load pre-analysis is needed for effective MLR forecasting. This is since the load in Sulaimaniyan city is affected by both temperature fluctuations and weather conditions. The demand for power changes with changes in weather conditions. They need a lot more energy during summer than in winter.

This study has investigated whether the forecast error will decrease or not by dividing the data according to the season in the series with seasonal effects. The essential factor in the expectation of a reduction in prediction error is nonlinear electrical load consumption behavior concerning temperature. In the study, it has been observed that using models to reduce seasonal effects reduces prediction errors. In other words, dividing data according to consumption behavior enables more accurate estimates. Another important finding in the study is that backward elimination finds more accurate results than forward and stepwise selection in the training dataset.

The following stages of the study will be carried out to obtain separate MLR equations for the sudden increase and decrease regions of the insufficient models in winter-summer and summer-winter transitions. Thus, it is thought that it will reduce the error of the prediction. In addition, the bottom-up consumption of sub-regions in the city

could be calculated according to the approach, and the results will be evaluated as future research.

REFERENCES

[1] G. Sites, “The purpose and need for forecasting - forecasting project,” 19 aug 2021.

[2] Saber, Ahmed Yousuf and Alam, AKM Rezaul, “Short term load forecasting using multiple linear regression for big data,” 2017 IEEE symposium series on computational intelligence (SSCI), pp. {1--6}, 2017.

[3] A. Kolasa-Wiecek, “Stepwise multiple regression method of greenhouse gas emission modeling in the energy sector in Poland,” Journal of Environmental Sciences, vol. 30, pp. 47--54, 2015.

[4] Fumo, Nelson and Biswas, MA Rafe, “Regression analysis for prediction of residential energy consumption,” Renewable and sustainable energy reviews, vol. 47, pp. 332--343, 2015.

[5] K.P., Amber; R., Ahmed; M.W., Aslam; A., Kousar; M., Usman; M.S., Khan, “Intelligent techniques for forecasting electricity consumption of buildings,” Energy, vol. 157, pp. 886--893, 2018.

[6] Akpinar, Mustafa and Yumusak, Nejat, “Naive forecasting of household natural gas consumption with sliding window approach,” Turkish Journal of Electrical Engineering & Computer Sciences, vol. 25, pp. 30--45, 2017.

[7] Yunsun Kim, Heung-gu Son, Sahm Kim, “Short term electricity load forecasting for institutional buildings,” Energy Reports, vol. 5, pp. 1270-1280, 2019.

[8] Yan, Yamin and Zhang, Haoran and Long, Yin and Zhou, Xingyuan and Liao, Qi and Xu, Ning and Liang, Yongtu, “A factor-based bottom-up approach for the long-term electricity consumption estimation in the Japanese residential sector,” Journal of Environmental Management, vol. 270, p. 110750, 2020.

[9] Larsen, Olena Kalyanova and Jensen, Rasmus Lund and Antonsen, Therese and Str{\o}mberg, Ida, “Estimation methodology for the electricity consumption with daylight-and occupancy-controlled artificial lighting,” Energy Procedia, vol. 122, pp. 733--738, 2017.

[10] Seyedzadeh, S., Rahimian, F., Glesk, I. et al., “Machine learning for estimation of building energy consumption and performance: a review.,” vol. 6, no. 1, p. 5, 2 October 2018.

[11] Omid Motlagh, George Grozev, Chi-Hsiang Wang & Melissa James, “A neural approach for estimation of per capita electricity consumption due to age and income,” Neural Computing and Applications, vol. 28, no. 7, pp. 1747-1761, 1 July 2017.

[12] Jing Cao, Mun Sing Ho, Yating Li, Richard G. Newell, William A. Pizer, “Chinese residential electricity consumption: Estimation and forecast using micro-data,” Resource and Energy Economics, vol. 56, pp. 6-27, 2019.

[13] Al-Mosawy, SK and Al-Jawari, SM and Al-Yassri, IJ, “Estimation of domestic urban electricity consumption: A case study of Baghdad, Iraq,” Periodicals of Engineering and Natural Sciences (PEN), vol. 9, pp. 678--683, 2021.

[14] Abuella, Mohamed and Chowdhury, Badrul, “Solar power probabilistic forecasting by using multiple linear regression analysis,” SoutheastCon 2015, pp. 1--5, 2015.

[15] Hong, Tao and Gui, Min and Baran, Mesut E and Willis, H Lee, “Modeling and forecasting hourly electric load by multiple linear regression with interactions,” IEEE PES General Meeting, pp. 1--8, 2010.

[16] Yildiz, Enes Mesut and Akpinar, Mustafa and Yumusak, Nejat, “Demand Forecasting Using Decomposition and Regressors of Natural Gas Delivered to Consumers in the US,” 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1--5, 2018.

[17] Zhang, Ning and Li, Zhiying and Zou, Xun and Quiring, Steven M, “Comparison of three short-term load forecast models in Southern California,” Energy, vol. 189, p. 116358, 2019.

[18] Shideh Shams Amiri, Mohammad Mottahedi and Somayeh Asadi, “Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S.,” Energy and Buildings, vol. 109, pp. 209--216, 2015.

[19] M.R. Braun, H. Altan, S.B.M. Beck, “Using regression analysis to predict the future energy consumption of a supermarket in the UK,” Applied Energy, vol. 130, pp. 305--313, 2014.

[20] D.H. Vu, K.M. Muttaqi, A.P. Agalgaonkar, “A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables,” Applied Energy, vol. 140, pp. 385--394, 2015.

[21] Zhang, Jialin Wu and Zhiwei Lian and Zhuling Zheng and Huibo, “A method to evaluate building energy consumption based on energy use index of different functional sectors,” Sustainable Cities and Society, vol. 53, pp. 385--394, 2020.

[22] Vasquez, Alfred Rey G and Rodriguez, Michael Ernie F and Dayupay, Roy C, “Energy Consumption Forecasting Model for Puerto Princesa Distribution System Using Multiple Linear Regression,” Technology, vol. 5, 2020.

[23] Nardini, V. Bianco and O. Manca and S., “Linear Regression Models to Forecast Electricity Consumption in Italy,” Energy Sources, Part B: Economics, Planning, and Policy, vol. 8, pp. 86-93, 2013.

[24] Alfonso Aranda and Germán Ferreira, M.D. and Mainar-Toledo and Sabina Scarpellini and Eva Llera Sastres, “Multiple regression models to predict the annual energy consumption in the Spanish banking sector,” Energy and Buildings, vol. 49, pp. 380--387, 2012.

[25] Somayeh Asadi, Shideh Shams Amiri, Mohammad Mottahedi, “On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design,” Energy and Buildings, vol. 85, pp. 246--225, 2014.

[26] N. Z. Siyu ZHOU, “Multiple regression models for energy consumption of office buildings in different climates in China,” Frontiers in Energy, vol. 7, pp. 103–-110, 2013.

[27] Mohammed, Awsan and Alshibani, Adel and Alshamrani, Othman and Hassanain, Mohammad, “A regression-based model for estimating the energy consumption of school facilities in Saudi Arabia,” Energy and Buildings, vol. 237, p. 110809, 2021.

[28] Dhaval, Bhatti and Deshpande, Anuradha, “Short-term load forecasting with using multiple linear regression,” International Journal of Electrical and Computer Engineering, vol. 10, p. 3911, 2020.

[29] M. C. Supapo and K. R. M. and Santiago and R. V. M. and Pacis, “Electric load demand forecasting for Aborlan-Narra-Quezon distribution grid in Palawan using multiple linear regression,” 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1--6, 2017.

[30] Abdul, Tuaimah and Firas M and Abass and Huda M, “Short-term electrical load forecasting for Iraqi power system based on Multiple Linear Regression method,” International Journal of Computer Applications, vol. 100, 2014.

[31] B. K. Nelson, “Time series analysis using autoregressive integrated moving average (ARIMA) models,” Academic emergency medicine, vol. 5, pp. 739--744, 1998.

[32] Yi-Shian Lee, Lee-Ing Tong, “Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming,” Knowledge-Based Systems, vol. 24, pp. 66--72, 2011.

[33] Wikipedia, “Population Republic of Iraq,” 2021.

[34] Jolliffe, Ian T and Stephenson, David B, “Forecast verification: a practitioner’s guide in atmospheric science,” 2012.

[35] Yan, Xin and Su, Xiaogang, “Linear regression analysis: theory and computing}„” 2009.

[36] H. T. Syarifah Diana Permai, “Linear regression model using bayesian approach for energy performance of residential building,” Procedia Computer Science, vol. 135, pp. 671--677, 2018.

[37] Amber, Khuram Pervez and Aslam, Muhammad Waqar and Mahmood, Anzar and Kousar, Anila and Younis, Muhammad Yamin and Akbar, Bilal and Chaudhary, Ghulam Qadar and Hussain, Syed Kashif, “Energy consumption forecasting for university sector buildings,” Energies, vol. 10, p. 1579, 2017.

[38] Cuadras, CM and Arenas, C, “A distance based regression model for prediction with mixed data,” Communications in Statistics-Theory and Methods, vol. 19, pp. 2261--2279, 1990.

[39] María, Cuadras and Carlos, “A distance based approach to discriminant analysis and its properties,” 1991.

[40] A. Hayes, “How Multiple Linear Regression Works,” 2021.

[41] Srinidhi and S. (n.d.), “Backward Elimination for Feature Selection in Machine Learning. Medium.”.

[42] W. Krzanowski, “Principles of multivariate analysis,” vol. 23, 2000.

[43] C. K. T. H. Kwangbok Jeong, “An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network),” Energy, vol. 71, pp. 71--79, 2014.

[44] Khan, Abdullah, Haruna Chiroma, Muhammad Imran, Javed Iqbal Bangash, Muhammad Asim, Mukhtar F Hamza, Hanan Aljuaid et al., "Forecasting electricity consumption based on machine learning to improve performance: A case study for the organization of petroleum exporting countries (OPEC)," Computers \& Electrical Engineering, vol. 86, p. 106737, 2020.

[45] N. A. Mohammed, "Modelling of unsuppressed electrical demand forecasting in Iraq for long term," Energy, vol. 162, pp. 354 -- 363, 2018.

[46] Mordjaoui, Mourad, Salim Haddad, Ammar Medoued and Abderrezak Laouafi, "Electric load forecasting by using dynamic neural networ," International journal of hydrogen energy, vol. 42, pp. 17655--17663, 2017.

[47] Le Cam, M, A Daoud and R Zmeureanu, "Forecasting electric demand of supply fan using data mining techniques," Energy, vol. 101, pp. 541--557, 2016.

[48] Xu, Lei, Shengwei Wang and Rui Tang, "Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load," Applied Energy, vol. 237, pp. 180--195, 2019.

[49] JianzhouWang, Xiaobo Zhang amd and Kequan Zhang, "Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by Cuckoo search algorithm," Electric Power Systems Research, vol. 146, pp. 270--285, 2017.

[50] Proti´c, Milan, Shahaboddin Shamshirband, Dalibor Petkovi´c, Almas Abbasi, Miss Laiha Mat Kiah, Jawed Akhtar Unar, Ljiljana ˇ Zivkovi´c and Miomir Raos, "Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm," Energy, vol. 86, pp. 343--351, 2015.

[51] J. C. Changhao Xia and Mi Zhang, "A hybrid application of soft computing methods with wavelet SVM and neural network to electric power load forecastin," Electrical Systems and Information Technology, vol. 5, pp. 681--696, 2018.

[52] Sasan Barak, S. Saeedeh Sadegh, "Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm," International Journal of Electrical Power & Energy Systems, vol. 82, pp. 92--104, 2016.

[53] R. S. Anurag, "Load Forecasting by using ANFIS," International Journal of Research and Development in Applied Science and Engineering, vol. 20, 2020.

[54] F. a. K. M. Chahkoutahi, "A seasonal direct optimal hybrid model of computational intelligence and soft computing techniques for electricity load forecasting," Energy, vol. 140, pp. 988--1004, 2017.

[55] Panapakidis, Ioannis P and Athanasios S Dagoumas, “Day-ahead natural gas demand forecasting based on the combination of wavelet transform and ANFIS/genetic algorithm/neural network model,” Energy, vol. 118, pp. 231--245, 2017.

RESUME

Name and Surname **:** Shanga Othman KAREEM

EDUCATION STATUS

Degree	Education Unit	Graduation Year
Bachelor	Sulaimaniyah University / College of Commerce / Statistical and Computer Department	2008
High Scholl	Zheen High School/ Sulaimaniyah	2004

WORK EXPERIENCE

Year	Location	Task
2011-present	Shar Hospital/ Sulaimaniyah	Programmer
2010-2011	Asiacell For telecommunication	Administrative Assistant

FOREIGN LANGUAGE

English

Arabic