# Relative Evaluation of Regression Tools for Urban Area Electrical Energy Demand Forecasting

**Nils Jakob Johannesen and Mohan Kolhe and Morten Goodwin**

Faculty of Engineering and Science, University of Agder, PO Box 422, NO 4604 Kristiansand, Norway.

***Abstract* - Load forecasting is the most fundamental application in SmartGrid, which provides essential input to Demand Response, Topology Optimization and Abnormally Detection, facilitating the integration of intermittent clean energy sources. In this work, several regression tools are analyzed using larger datasets for urban area electrical load forecasting. The regression tools which are used are Random Forest Regressor, k-Nearest Neighbor Regressor and Linear Regressor. This work explores the use of regression tool for regional electric load forecasting by correlating lower distinctive categorical level (season, day of the week) and weather parameters. The regression analysis has been done on continuous time basis as well as vertical time axis approach. The vertical time approach is considering a sample time period (e.g seasonally and weekly) of data for four years and has been tested for the same time period for the consecutive year. This work has uniqueness in electrical demand forecasting using regression tools through vertical approach and it also considers the impact of meteorological parameters. This vertical approach uses less amount of data compare to continuous time-series as well as neural network techniques. A correlation study, where both the Pearson method and visual inspection, of the vertical approach depicts meaningful relation between pre-processing of data, test methods and results, for the regressors examined through Mean Absolute Percentage Error (MAPE). By examining the structure of various regressors they are compared for the**

**lowest MAPE. Random Forest Regressor provides better short-term load prediction (30 min) and kNN offers relatively better long-term load prediction (24 hours).**

***Keywords*** **-** *Electrical Energy Demand Forecasting, Impact of meteorological parameters on demand forecasting, Smart-Grid management, Machine Learning, Regression Tools, Random Forest Regressor, k-Nearest Neighbor Regressor, Linear Regressor.*

## A.1 Introduction

Urban area electrical energy demand forecasting is necessary for optimizing the electrical power generation scheduling in coordination with distributed generators including intermittent renewable energy sources. It will also be beneficial for demand side management considering grid constraints. In the literature, most of the electrical energy prediction studies are using shallow neural networks and support vector [2, 3]. Popular stochastic models, such as hidden Markov models, are also used for energy prediction [4, 5]. In the EU FP7 SEMIAH (Scalable Energy Management Infrastructure for Aggregation of Households) project [5], the domestic demand has been predicted using a two-stage linear stochastic optimization for managing operation of non-critical power intensive loads (for example, thermal load).

Recent research from 2018 on Computational Intelligence Approaches for Energy Load Forecasting [6], that reviewed more then 50 research papers related to the subject outlines the complexity of demand patterns as potentially influenced by factors like climate, time periods, holiday or working days and other factors such as social activities, economic factors including power market policies. Electrical energy demand is influenced by meteorological weather conditions, therefore it is necessary to include the impact of meteorological weather parameters in electrical energy demand forecasting also renewable electrical energy production is nature dependent. The future electrified grid will increasingly depend on renewable intermittent energy sources (solar, wind), and the individual load profiles of such a system will change radically as home appliances includes new energy demanding appliances (e.g. heat pump, electric vehicles and induction stove) [7]. The new electrified grid is Smart Grid System, as it is a complex whole of two-way communication aided by intelligent agents. The information will be used to provide demand side management such as peak shaving, where non-critical load demands are shifted to other periods where the stress on the grid is less intense. Electric load forecasting by machine learning will be useful in the operation of load shifting, with an accurate prediction of the load demand. Machine learning falls into two categories of Supervised, where the data points have a known value, and Unsupervised, where data points have unknown outcome. The types of supervised learning is divided into regression and classification. The first where the outcome is continuous (numerical), the latter categorical. Regression models considers the relationship to independent variables, predictors, and a dependent variable, known as target.

The regression models k-Nearest Neighbor (kNN), Linear Regression (LR), and Random Forests (RF) are supervised machine learning algorithms with a numerical outcome. The model is trained to find rules for pattern recognition in the input to output relation. The input to the model are known as features. Neural Networks is the preferred machine learning tool and are known as both feedforward and backpropagating networks, where a number of inputs are weighted in order to provide a predictive outcome. Neural networks are good for detecting non-linearities and therefore preferred as a predictive tool in electrical load forecasting, yet also often criticized for low transparency and lack of interpretability because of the black box approach, and using large amount of data. Overfitting is still a challenging issue when applying Neural Networks to electrical demand prediction [8]. The literature distinguishes between short term prediction and long term prediction time. In this article short term is defined as the 30 minute prediction time interval, and long term prediction is defined as 24 hour time prediction interval.

Urban area load is influenced by meteorological conditions therefore it is important to include impact of weather parameters on load prediction, yet this impact is governed by the prediction time, greater for long term and decreases as the prediction time is narrowed. The electrical energy demand is influenced by the user behavior as well as weather conditions. Individual human behavior and weather are so random that a complex neural network would not predict the outcome better then a coin toss. Hence, if one has to analyze the load demand of larger area like the urban area, systematic load behavior with correlation to weather parameters and continuous load profile, should be investigated.

This work has uniqueness in electrical demand forecasting using regression tools through vertical approach and it also considers the impact of meteorological parameters. This vertical approach uses less amount of data compare to continuous time-series as well as neural network techniques.

The objectives of this work are to explore the use of regression tools for regional electrical load forecasting by correlating lower distinctive categorical levels (season, day of the week) and weather parameters. The vertical time approach is considering a sample time period (e.g seasonally and weekly) of data for four years and has been tested for the same time period for the consecutive year. A vertical axis approach, showed to be competitive to Artificial Neural Networks (ANN), with a low amount of data.

The paper is organized as follows: Review of electrical load forecasting is presented in Section 2. In Section 3, various parameters (e.g. weather parameters, seasonal impact and time as well as random effects) are discussed for urban area electrical energy demand forecasting. Section 4 shows analysis both by Pearson correlation method and visual inspection to find correlation of meteorological parameters and previous load patterns on Urban Area Load Forecasting and shows the Regressor Model and gives regression model analysis. Results and Discussions are provided in Section 5. Finally, in Section 6, the conclusions are presented.

## A.2 Review of electrical load forecasting

In most of the work, hourly electrical energy predictions are considered. It is important to have precise prediction for short-term (e.g. 30 min) using less amount of data as well as for long-term (e.g. 24 hours) for urban area electrical demand for electrical power generation coordination. The small area of Tunis (with only installed capacity of 4425 MW) is considered for analysis of load prediction with seasonal variations [22]. A variation in load due to season is only once a year during heat wave in the summer. For training set they have used horizontal time-series approach, where almost 10 months (more than 14400 datapoints) of training was used for testing on one week. According to Lahouar and Slama (2015) [22], who used random forest for day-ahead load forecast for the Tunisian market with historical data from 2009-2014, they obtained an average MAPE of 2.24% when crediting for the next 24 hours. Presented method of [22] does not improve, when predicting for the heat wave season, as the average MAPE for heat wave period (7-13 July) has increased to 2.6899%. During the Arabic spring in Tunis 2011, Tunis experienced a Random effect caused by a much lower power demand during the Tunisian Revolution, the MAPE for some 24 hour intervals of prediction as high as 19.61%. It was even worse during the Blackout of August 2014 where the MAPE rose to 398.09%. This show the machine learning algorithms inabilities in forecasting rare events. [22] also makes a comparison with ANN, and for the testing period of 7-13 of July it scores 2.9140 MAPE. They state that the main advantage with Random Forest over other methods is that there are few hyper-parameters to set and generalize by saying default settings is normally enough to compete with ANN and Support Vector Machine (SVM)/Support Vector Regressor (SVR), which accuracy depends on the tuning of their hyper-parameters. In our work we have used the experiences from Tunis to understand the random effects and their input on electrical energy demand forecasting as well as the understanding of hyper-parameters.

Jinkyu and Sup (2015) [23] recognizes artificial intelligence techniques like ANN or Kalman filter, to show promising results in the load forecasting predictions, although the hidden structures in AI might limit the understanding of the complex spatiotemporal developments in correlation between meteorological conditions and electricity demand. Electrical load demand and the temperature effects have been studied and short term load forecast needs to take temperature effects into consideration for day-ahead predictions. In the very short time load prediction the time scale is to short for the temperature to have any effect, and in the long run the effect tend to even out [24, 25]. On the load forecasting for the UK electricity demand Al-Qahtani and Crone (2013), proposes a multivariate k-NN approach that, opposed to the univariate model that does not take into account the underlying sub-categories of the calender, create a binary dummy variable where dt = 1 for all nonworking days and dt = 0 for working days. The load forecast MAPE of both univariate and multivariate show improved results by the use of dummy variables. A MAPE of 2.3284 was found using the univariate model, and a 1.8133 was found using the multivariate model. The dataset contained data for more then 7 years (2001-2008). The complete year of 2004 was used for training and 2005 used for validation [26]. Based on their research we developed the relevance of doing multiple correlation analysis with

different time factors, where we can observe that meteorological parameters increase their importance on the prediction output as time window increases. In this context we regarded the work of Afkhami and Yazdi who proposes a way to quantisize the day into 3 periods of 8 hours for neural networks to enhance their performance [27].

Local Interpretable Model-Agnostic Explanations (LIME) aims to reflect the behavior in proximity to the predicted outcome, and does so by offering a interpretation that can explain doubts about the model. By explaining here means to provide some mean of qualitative understanding in the relation between a decision making and the predictive outcome. In medical diagnosis LIME highlights what features in the dataset that led to the prediction, and what was evidential against it [28].

ANN studies have shown an MAPE of 1.9, resulting in a Mean Absolute Error (MAE) of 167.91 MW, based on training data for a whole year. The research includes studies of temperature effects and introduces two threshold values where the load and temperature exhibits close correlation, at below 10 degrees Celsius due to heating, and above 23 degrees because of cooling [29].

The focus of this work is to verify the regression tools for electrical energy demand forecasting and we have not considered the prediction from the supply side. We considered regional area electrical energy demand forecasting with impact of weather parameters. And we have the availability of the required data for mentioned period.

## A.3 Urban Area Electrical Demand Forecasting

The purpose has been to test the regression tools on the available real data. Urban area electrical energy demand forecasting is very important for generation scheduling, as well as effectively taking contribution from renewable energy sources and demand side management. Urban area electrical energy demand predictions for short term (30 min) and long term (24 hours) are necessary for scheduling power generation units as well as participating them in short term and day ahead energy market. When predicting the electrical load demand for a particular time window, in this case the next 30 minutes or 24 hours, the machine learning algorithms search for patterns and rules for the predictive outcome in the Supervised category with a continuous numerical output.

The following three parameters are important for system electrical energy demand:
(i) Time
(ii) Weather
(iii) Random effects
The seasonal patterns are repeating with the same upper and lower limits (e.g repeating on annual basis) and therefore considered as no economic effects are influencing the load behavior in the urban area of Sydney during the years 2006-2010. When investigating the Sydney dataset, see Figure A.1, we find that the load curves, yet containing cyclic

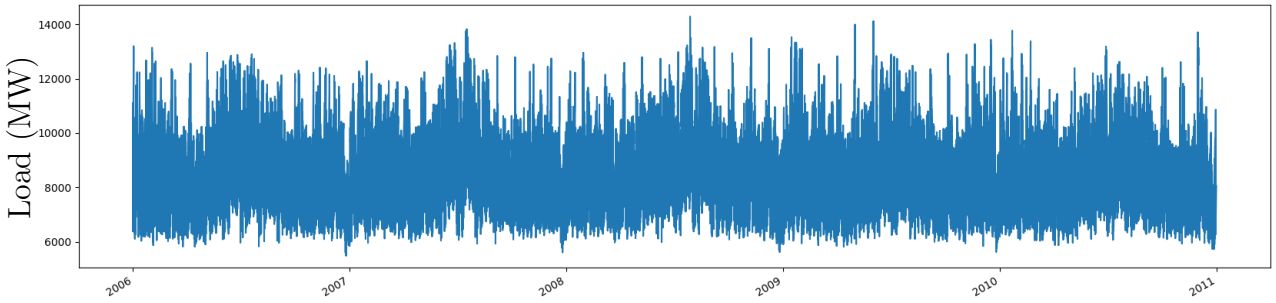



Figure A.1: Load curve of Sydney dataset containing five years of half hour values.

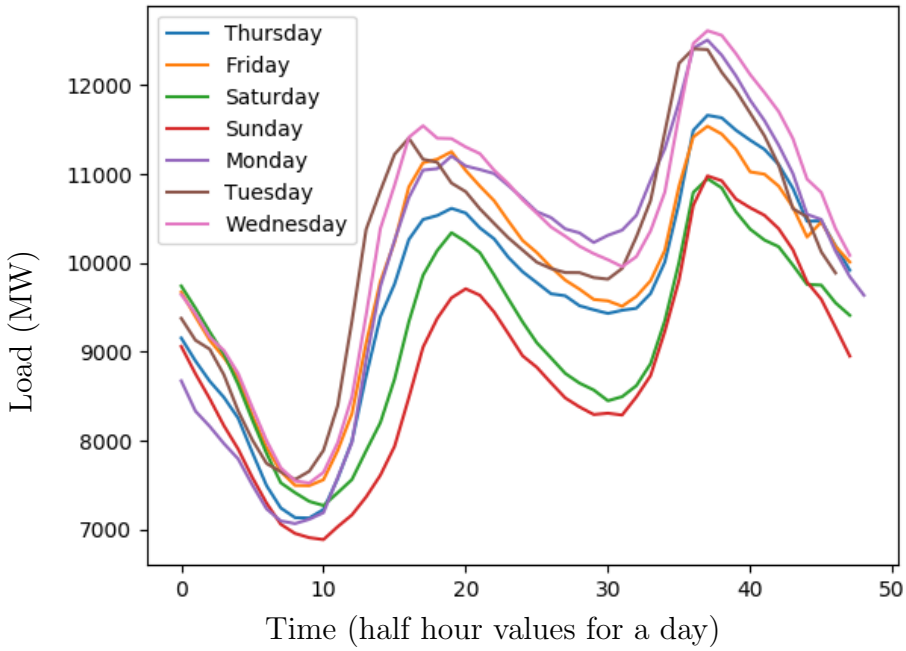
and seasonal differences, do not contain significant changes on the system load due to changing economic trends [57].

### A.3.1 Time

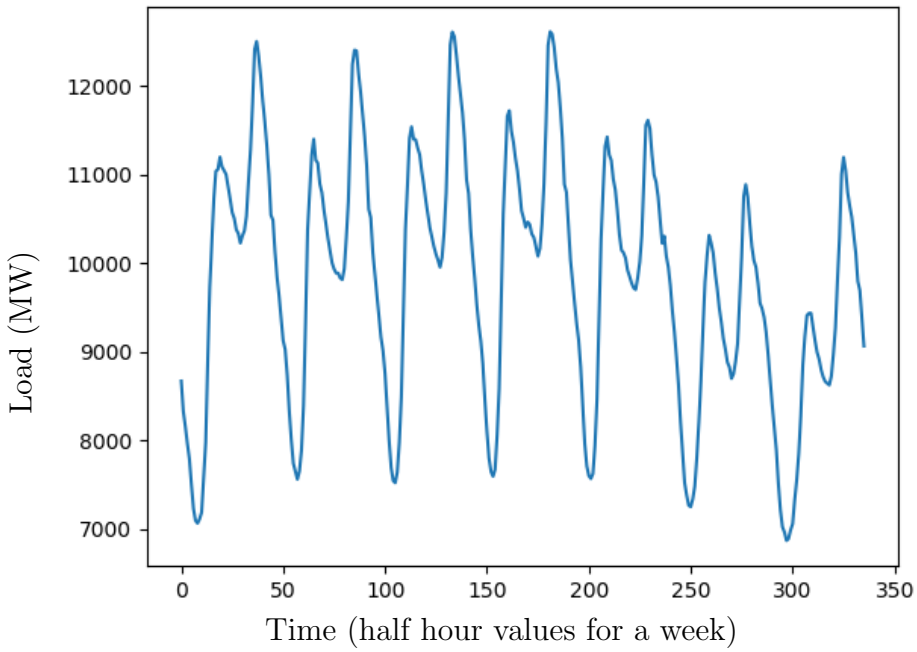
Apart from the seasonal effects, shown in Figure A.1, underlying patterns emerges in the system load demand. There are different peaks throughout the seasons, whether it is a winter peak or a summer peak. Emerging under this seasonal patterns are daily- and a weekly-cycles. The daily routines of human behavior are manifested in systematic load patterns on a daily basis. Day of the week is also significant. Public buildings and offices demands large amount of electrical load and whether it is a working day or not, influences load patterns.

When inspecting the daily- and weekly-cycle in Figure A.2, we can clearly see a load pattern emerging from a very low activity during the early hours of the day, into one peak at morning (between 8-10 hours), and another peak in the evening (between 19-21 hours) in Subfigure A.2a. The same daily repeating pattern, with a low activity followed by two peaks, are also evident in the weekly cycle, seen in Subfigure A.2b, except for that the two last days of the week (Saturday and Sunday) the peaks and general load is lower. It can be seen that urban area load predominantly reflects the domestic load and it can be correlated to human behavior. The periodicity in the load patterns reveal a load demand that reflects consumer-lifestyle.

The periodicity reflected in the daily load curve is significant in weekly cycles as well as monthly, seasonal and yearly load curves, as seen in Figure A.1 and A.2. Sub-categorical levels like working/non-working days are referred to in the literature as an indicator variable. In this work the time has been used as a variable which can be categorized as day of the week or working/non-working days or time of the day. To give this properties to our algorithms are very important as it makes prediction of forecast load more efficient [48]. The use of such type of variables has been successfully employed in electric market forecasting in the Tunis as well as the UK [22, 26].




(a) Daily cycle




(b) Weekly cycle

Figure A.2: Load patterns in daily- and weekly- cycles

### A.3.2 Weather

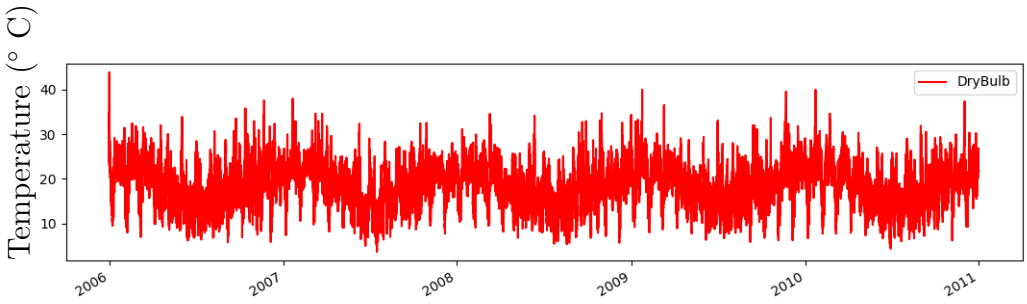



Figure A.3: DryBulb temperature curve of Sydney dataset containing five years of half hour values.

The features enlisted in the Sydney dataset, has two time indicators 'Date' and 'Hour', four weather parameters, information about the electricity price, 'ElecPrice' and information about the electricity load consumption, 'SYSLoad', these features have been developed in the pre-processing to match the requirements of the prediction tool, see Figure A.4.

The four weather parameters enlisted are DryBulb, DewPnt, WetBulb and Humidity. Dry Bulb Temperature (DBT) is temperature measured from the air, yet not exposed to solar radiation or moisture. Wet Bulb Temperature (WBT) is measured from a thermometer where the bulb of the measurement device is soaked by a wet cloth. As long as the air is not saturated, evaporation from the moist cloth keeps the WBT lower then the DBT. From the DBT and WBT one can then derive the relative humidity of the air and the dew point from a Mollier Chart by psychometrics.

Many electrical utilities are weather-sensitive, such as heating and air conditioning. Temperature, as well as past temperature effect on the load are important effect on the electrical demand, the temperature on a hot summer day may reach its peak after sunset due to heat buildup in the construction materials of buildings. In addition to the daily heat buildup, will a sequence of days with high temperature create new system peak.

The complexity in the control system engineering of maintaining thermal comfort as well as optimizing for energy is important to know. At the same time it is important to acknowledge that most houses are designed to resist the worst meteorological conditions[50]. There are also limitations in the heating system itself that might cause load peaks, like the inertia in the floor heating system, known as thermal lag [51].

In humid and hot places it is likely that humidity will effect the load pattern in similar ways as temperature. Humidity explains the complex relation between temperature and load, and therefore mathematical models is not enough in a thorough analysis. Humidity is the amount of water vapor in the air and might increase the gap between actual and apparent or felt temperature. When regulating temperature the body utilizes evaporate cooling, and the rate of evaporation through the skin is correlated to humidity, and because of the conductive properties of water, we feel warmer at high humid conditions.

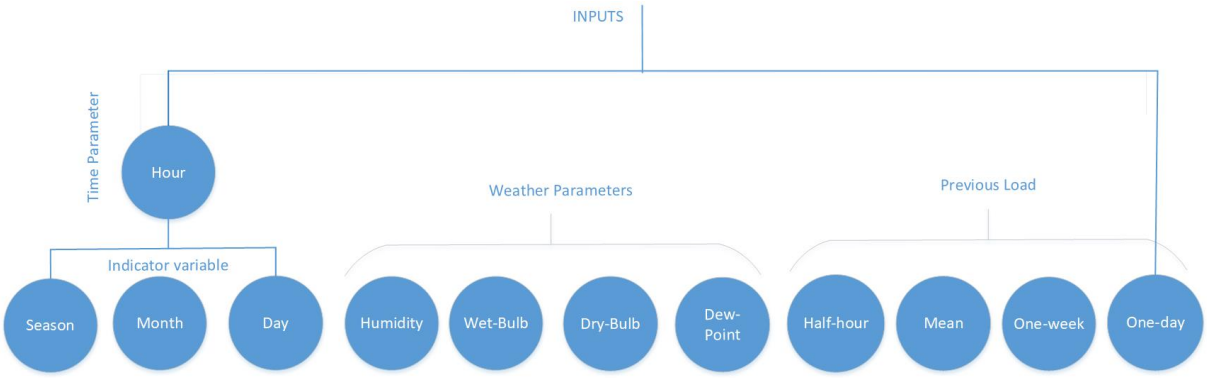



Figure A.4: Model input parameters: indicator variables, weather parameters and previous load consumption.

Also, due to the seasonal changes of weather data, the correlation to the electrical load will vary during the year.

### A.3.3 Random effects

Infrastructural changes in the urban area and maintenance work are random effects that will not be detected by pattern recognition. When examining the Sydney dataset load curve as shown in Fig. A.1, there is consisting seasonal patterns. Load pattern are consistent from year to year, and show reoccurring seasonal pattern. When the yearly load curves do not vary from year to year show that there are no economic trends.

### A.3.4 Relevance

It is important to investigate the main effects on the system load pattern as these are the main predictors in load forecasting.

To look for causalities in load and effect has been the topic of previous studies in load forecasting. Knowledge about the cause and effect about external parameters and system load is needed for accurate prediction. In the literature concerns have been voiced about more complicated forecast scenarios based on deregulated markets [22] and demand side management.

When energy consumers are free to choose suppliers the varying energy prices are incentives to attempt to shift non-critical load demands to periods where the stress on the grid is less intense, otherwise known as peak shaving. The other aspect is the integration of the district level environment friendly power plant, relying on intermittent renewable energy sources. In Figure 2.4, the load and temperature are plotted in the same plot. The plot will help searching for linearity among the features. The upper side of the plot forms a v-shape, separating the plot into two linear relationships at around 21°C. While the lower end has a more round u-shape.

## A.4 Correlation Analysis of Electrical Load with Meteorological Parameters

Correlation is a measurement to how two ranges of data move together, and will give us an indication of how to assess feature engineering. Other means to measure the relevance between variables is Shannons concept of Mutual Information (MI), a method based in the entropy function that gives the certainty of a variable [21]. Correlation is widely used in contemporary research, where regression tools and other machine learning methods are applied to various engineering features (e.g. power transformers health index [22], emission prediction of Combined Cycle Gas Turbine [23], wind power prediction based on weather data and local terrain [24]). The Pearson Correlation Coefficient (r) computes the linear relationship between two datasets, in a range from -1 to +1. [36]. If the relationship is in the proximity of 1, it means that when x increases so does y and at exact linearity, the opposite is true for -1, it means that when a dataset is increasing the other dataset is decreasing.

$$r = r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{A.1}$$

One of the means to improve prediction accuracy in spite the seasonal differences, is to create a dummy variable that increases the precision of the algorithm while differentiating the seasonal changes.

A dummy variable or Indicator variable is an variable created to represent more distinct categorical level. In this paper one was made to categorize on day of week:

$$df['season'] = (df['month']\%12 + 3)// \tag{A.2}$$

The use of dummy variables has been successfully employed by forecasting on the UK electricity market, to categorize days into working and non-working days [26].

Other papers conclude what this research also experienced that the most accurate prediction comes from either from predicting on the same hour or for 24 hours, probably due to the habitual individual behaviour like showering and putting on the coffee at the same time every day [25].

### A.4.1 Regressor Model

The input for the model are based on tree parameters, time, weather and previous load consumption, see Figure A.4. The time parameters are divided into sub-categories in lower categorical level as the indicating variables day of the week, working-/non-working days and season. Included are also the previous load consumption are organized by the lag method and weather parameters.

The preprocessed inputs are then computed using regression tools, in Figure A.5, represented by the k-Nearest Neighbour regressor. Figure A.5, is showing the k-Nearest

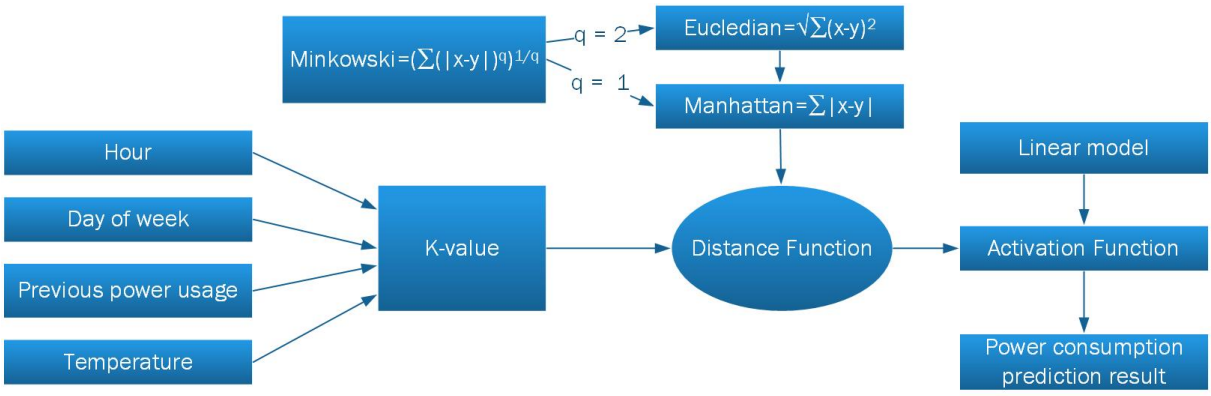



Figure A.5: The regressor model

Neighbour algorithm, where the model shows the algorithm consider a k-value and distance function based on the inputs. The regressors are taken from the scikit learn library [29], and further the hyper-parameters are tuned for optimal performance.

### A.4.2 Regression Models Analysis

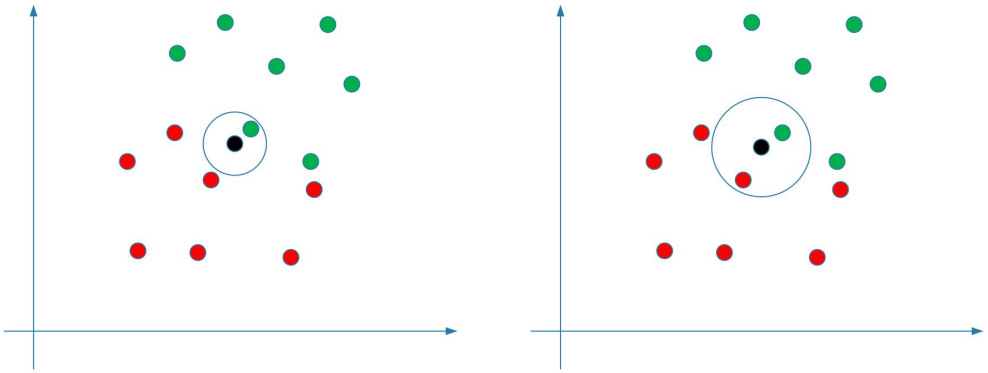
The regression-tools considered in this article are kNearestNeighborRegressor, LinearRegressor and RandomForestRegressor. To elaborate further on the model used in this research in the following the kNearestNeighborRegressor is explained: The k-Nearest Neighbour is an algorithm that computes the numerical value of the distance between given features or data points and a query point in an multi-dimensional array, and then find the point in vicinity to the query point [26].

In Figure A.5, the model takes a set of inputs, based on time, date, previous power consumption and weather parameters, based on the features of the Sydney dataset and further created.

#### A.4.2.1 kNN Regression Tool

The kNN-classifier is illustrated in Figure A.6, where Subfigure A.6a, depicts a nearest neighbour of k=1, where simply the nearest neighbour decides the class of prediction, whilst in Subfigure A.6b, the number of k is increased to more then one [70].

Using k=1 can lead to false prediction, and a set of k-Nearest Neighbours are often used. When classifying the dependent variable is categorical can easily been made numerical by regression. The k-NN regressor makes a regression based on the number of k-Nearest Neighbours to minimize the false predictions. The model considers a range of different k-values to find the optimal value.



(a) k=1 (b) k=2

Figure A.6: kNN-classifier

#### A.4.2.2 Distance

A variety of distances is used in the algorithm. As seen in Equation C.10, C.11, C.12, and C.13, they are most used since it is easy to intersect by changing the variable q. The variable q is also considered to find the optimal value.

#### A.4.2.3 Manhattan/City Block Distance

$$d(x, y) = \sum_{i=1}^{k} |x_i - y_i| \tag{A.3}$$

#### A.4.2.4 Euclidean distance

$$d(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{A.4}$$

#### A.4.2.5 Minkowski Distance

$$d(x, y) = \left( \sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{\frac{1}{q}} \tag{A.5}$$

#### A.4.2.6 Chebychev Distance

$$d(x, y) = \lim_{q \to \infty} \left( \sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{\frac{1}{q}} \tag{A.6}$$

Similarly the all the regression-tools have parameters viable for optimisation. This research employs a systematic grid-search on selected parameters.

## A.5 Results and Discussions

In k-fold cross validation the dataset D is divided into an equally adjusted amount of k's. For the Sydney dataset the subsets are $D_{2006}$, $D_{2007}$, $D_{2008}$, $D_{2009}$ and $D_{2010}$. One subset when is taken apart for testing $D_i$, and the remaining four is used for training. The method is repeated until all the subsets are tested on an equal shifted amount of training data [28]. In the case of the Sydney dataset containing 87648 datapoints, each k-subset will contain approximately 17530 datapoints depending if there is a leap year or not.

The cross validation was done on various regressors from the Python library scikit-learn [29]. All regressors were set to the default values. In Table A.1, the validation is done for short-term (30 min) time prediction window, denoted *t-1*, and long-term (24 hour) time prediction window, denoted *t-48*. The MAPE of the cross validation show little variation between the subsets. In this work the weather parameters and the load data from the urban area of Sydney city is used. The results are analyzed for correlation among the dataset variables, graphical inspection for understanding some patterns between load and temperature, impact analyses of q-values on load prediction, and analyses of results for load and indicator variables.

### A.5.1 Correlation Analysis

Correlation analysis between the variables enlisted in the Sydney dataset (Date and Hour, four weather parameters; DryBulb, DewPnt, WetBulb and Humidity, information about the electricity load consumption, 'SYSLoad') are presented in Table A.2.

### A.5.2 Graphical Inspection between Load and Temperature

In section 8.1 it is observed that there is significant impact of temperature on the load. Therefore it is also investigated through graphical depiction the complex relation between DryBulb Temperature and the load patterns emerging from human lifestyle behavior, influenced by the weather conditions. The correlation of System Load to Last half hour value correlates highest at 0.98, and is also the most effective variable for short-term load forecast. Preceding the last half hour value is the variable Hour at 0.48, giving high impact on the periodicity. It has been observed that among the weather parameters, DryBulb has a better correlation with the load. The correlation for DryBulb to the load improves further when it is correlated to the previous 24 hour load data. This might explain why the 24 hour prediction results improves when impact of the weather parameters are included. When investigating the correlation between load and temperature from the graphical depiction, as seen in Figure A.7, where seasonal effects influences the load patterns we find complex patterns, but also periodicity. From these observations it can be seen that the vertical approach (considering a sample time period - e.g seasonally and weekly - of data for four years, and tested for the same time period of the consecutive year) enables the algorithm to reveal the complexity of load and temperature for better prediction results [30].

Table A.1: k-fold validation result in MAPE

| Regr. | *2006 t-1* | *2006 t-48* | *2007 t-1* | *2007 t-48* | *2008 t-1* | *2008 t-48* | *2009 t-1* | *2009 t-48* | *2010 t-1* | *2010 t-48* |
|---|---|---|---|---|---|---|---|---|---|---|
| Rand. | 1.02 | 4.47 | 1.00 | 4.07 | 0.98 | 3.96 | 1.02 | 4.35 | 1.05 | 4.35 |
| k-NN | 1.83 | 4.93 | 1.73 | 4.65 | 1.64 | 4.43 | 1.79 | 4.92 | 1.82 | 4.88 |
| Linear | 2.22 | 5.49 | 2.12 | 5.07 | 2.13 | 4.95 | 2.17 | 5.24 | 2.11 | 5.11 |
| Bayes | 2.22 | 5.49 | 2.12 | 5.07 | 2.13 | 4.95 | 2.17 | 5.24 | 2.11 | 5.11 |

Table A.2: Correlation of Dataset

| | Hour | DryBulb | WetBulb | Humidity | SYSLoad | weekday | LastHalfHr |
|---|---|---|---|---|---|---|---|
| Hour | 1.00 | 0.20 | 0.11 | -0.23 | 0.48 | 0.00 | 0.51 |
| DryBulb | 0.20 | 1.00 | 0.90 | -0.22 | 0.09 | 0.01 | 0.09 |
| WetBulb | 0.11 | 0.90 | 1.00 | 0.20 | -0.02 | 0.00 | -0.02 |
| Humidity | -0.23 | -0.21 | 0.20 | 1.00 | -0.30 | -0.02 | -0.30 |
| SYSLoad | 0.48 | 0.09 | -0.02 | -0.30 | 1.00 | -0.14 | 0.98 |
| weekday | 0.00 | 0.01 | 0.00 | -0.02 | -0.14 | 1.00 | -0.14 |
| LastHalfHr | 0.51 | 0.10 | -0.02 | -0.30 | 0.98 | -0.14 | 1.00 |

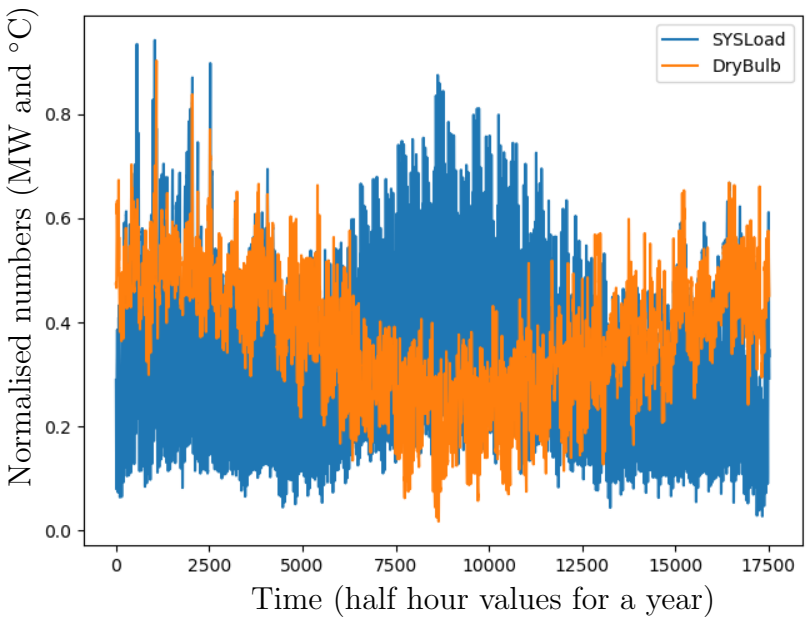


Figure A.7: Correlation of DryBulb Temperature and Electric Load consumption.

Table A.3: Seasons

| Season | Months | | |
|---|---|---|---|
| Season 1 | December | January | February |
| Season 2 | March | April | May |
| Season 3 | June | July | August |
| Season 4 | September | October | November |

### A.5.3 Impact Analysis of q-value on MAPE

In this work the annual load profile has been divided in four seasons and time frames are given in Table A.3. Observing the results of the impact of q-values on prediction, the preferred value is 1, which is the absolute value. Only occasionly are other q-values the preferred output, meaning the one with the lowest MAPE. On these occasions the highest q-value was 4. Load prediction has been analyzed for all seasons for different regressors and the MAPE for short-term (30 minutes) and long term (24 hours) are presented in Table A.4. In this analysis only the previous load pattern were taken into account. MAPE analysis has been carried out for horizontal (continous time series) as well as vertical approach. It has been observed that Random Forest Regressor provides better results for 30 minutes prediction in horizontal as well as vertical approach for all seasons. For 24 hours prediction it has been observed that in most of the season k-Nearest Neighbour Regressor performs well compared to other regression tools. But in season one for vertical approach Linear Regression has given better result. In season three k-Nearest Neighbour regressor performs well especially considering the vertical approach.

The load prediction using Random Forest Regressor, k-Neareast Neighbor Regressor and Linear Regression has been presented in Figure A.8. These regression results for 24 hour load prediction in season three using vertical approach. Tests conducted by including previous load consumption, weather parameters and indicator variables.

### A.5.4 Lowest MAPE for short term and long term prediction

The relative comparison of the MAPE for different regression tools for 30 minutes and 24 hours have been done using both horizontal and vertical approach for all seasons, as shown in Table C.4. It has been found that the the lowest MAPE was achieved with the use of previous load patterns together with indicator variables, and noticeably disregarding weather variables. This goes well with the previous analysis of correlation, which confirms that previous load patterns and indicator variables have higher correlation to the actual load, then the weather parameters.

It has been observed from the test results the lowest MAPE is found through Random Forest Regressor for 30 minutes prediction using the vertical approach. For the 24 hour time period k-Nearest Neighbor is providing lowest MAPE, again through the vertical approach. The lowest MAPE for 30 minutes prediction in season three using vertical ap-

Table A.4: q-Value Results

| Time | Regressor: Random Forest | Regressor: k-Nearest Neighbour | Regressor: Linear Regression |
|---|---|---|---|
| **Season One Horizontal Approach** | | | |
| 30 minutes | 1.12(16*) | 1.29(5**,1***) | 2.02 |
| 24 hours | 5.21(13*) | 4.30(16**,4***) | 5.55 |
| **Season One Vertical Approach** | | | |
| 30 minutes | 1.01(16*) | 1.44(11**,1***) | 1.78 |
| 24 hours | 6.75(13*) | 6.63(19**,1***) | 6.29 |
| **Season Three Horizontal Approach** | | | |
| 30 minutes | 1.13(15*) | 1.43(7**,1***) | 2.29 |
| 24 hours | 4.00(15*) | 3.60(19**,3***) | 5.03 |
| **Season Three Vertical Approach** | | | |
| 30 minutes | 0.93(18*) | 1.17(7**,1***) | 2.22 |
| 24 hours | 3.73(13*) | *3.58(7**,1***)* | 5.09 |

* n-estimator
** k-value
***q-value

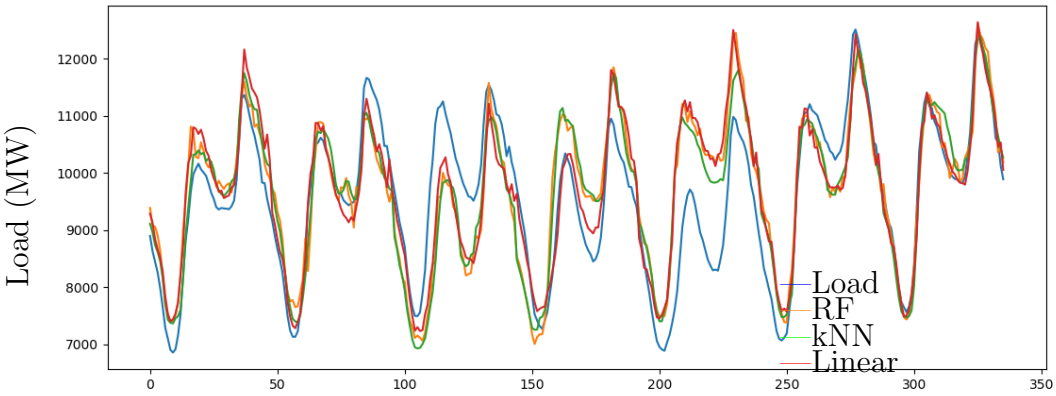



Figure A.8: Regression results on 24 hour prediction for season tree using vertical approach. Tests conducted by including previous load consumption, weather parameters and indicator variables.

proach is shown in Figure A.9, and similarly for 24 hours in Figure A.10. The MAPE for 30 min prediction results using 'random forest regressor' is varying between 1-2%, as seen in Figure A.9, and providing very good results compare to other regressions techniques, which have been used in this work. The 24 hours predictions results using 'k-Nearest Neighbor Regressor' technique has MAP of 2.61%, as seen in Figure A.10, which is much

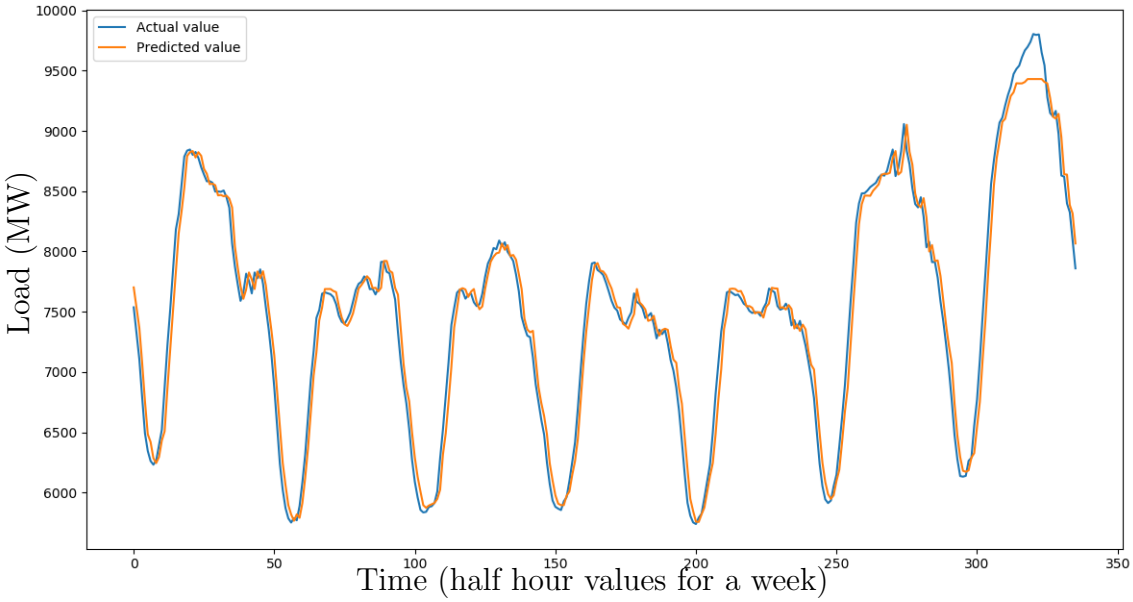


Figure A.9: Best performance for 30 min prediction by Random Forest Regressor

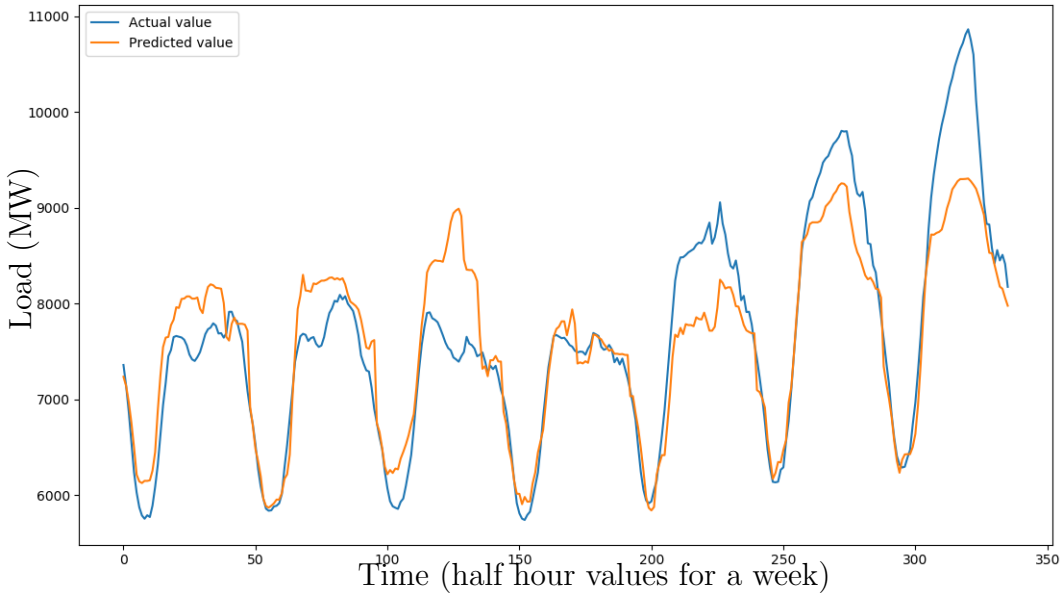


Figure A.10: Best performance for 24 hour prediction by kNN regressor

better compare to other regressors, which have been studied in his work. From the results, it has been observed that for short-term predictions (30 min) the 'random forest regressor' should be used; and for long-term predictions (24 hours) the 'k-Nearest Neighbor Regressor' should be considered.

Table A.5: BEST RESULTS (MAPE Load and Indicator aggregated version test results)

| | Regressor | | |
|---|---|---|---|
| Time | Random Forest | k-Nearest Neighbour | Linear Regression |
| Season One Horizontal Approach | | | |
| 30 minutes | 1.11(9*) | 1.98(7**,1***) | 2.04 |
| 24 hours | 5.32(13*) | 6.53(4**,1***) | 5.15 |
| Season One Vertical Approach | | | |
| 30 minutes | 0.94(16*) | 1.85(8**,1***) | 1.76 |
| 24 hours | 5.88(13*) | 5.49(5**,2***) | 5.83 |
| Season Three Horizontal Approach | | | |
| 30 minutes | 1.12(17*) | 2.36(5**,1***) | 2.29 |
| 24 hours | 4.76(9*) | 5.41(19**,1***) | 5.27 |
| Season Three Vertical Approach | | | |
| 30 minutes | *0.86(17*)* | 1.19(6**,1***) | 2.15 |
| 24 hours | 2.71(17*) | *2.61(17**,1***)* | 4.26 |

* n-estimator
** k-value
***q-value

## A.6 Conclusion

In this paper the regression tools, Random Forest Regressor, k-Nearest Neighbor Regressor and Linear Regression are used for analyzing the urban area electrical energy demand forecasting. Using larger dataset of Sydney region. This work has explored the use of regression tools for electrical energy load forecasting through correlating weather parameters as well as the time period. Load prediction analysis using regression tools have been done continuous time basis (horizontal) as well as vertical time approach.

A correlation study, where both the Pearson method and visual inspection, of the vertical approach depicts meaningful relation between pre-processing of data, test methods and results, for the regressors examined. Data correlation over seasonal changes have been argued by means of improving Mean Absolute Percentage Error (MAPE). By examining the structure of various regressors they are compared for the lowest MAPE. The regressors show good MAPE for short term (30 min) prediction and Random Forest Regressor scores best in the range of 1-2 % MAPE. kNN show best results for 24 hour prediction, with a MAPE of 2.61%.

Results of this work is going to be useful for predicting the short term 30 minutes electrical energy using vertical approach and considering Random Forest Regression Tool. For long term prediction of 24 hours kNN Regression Tool can provide better results using vertical approach. It is also important to consider further investigations of the impact of various weather parameters on load prediction.

The presented regression techniques can forecast electrical energy demand for short-term (30 min) and long-term (24 hours) using limited datasets. Vertical axis approach has shown competitiveness to ANN due to use of low amount of data and considering the impact of meteorological parameters. Load forecasting is the most fundamental application in Smart-Grid, which provides essential input to other applications such as Demand Response, Topology Optimization and Abnormally Detection, facilitating the integration of intermittent clean energy sources. Presented regression techniques can also be used for predicting energy output (short- and long-term) from the intermittent renewable energy sources.

## Acknowledgement:

Authors are very grateful to the scientific committee of IEEE SpliTech 2018 for recommending this article for publication in a Special Issue of Journal of Cleaner Production. Authors appreciatively acknowledge the anonymous reviewers for providing suggestions for improving the manuscript.

# Bibliography


[1] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[2] Nikolaos G. Paterakis, Elena Mocanu, Madeleine Gibescu, Bart Stappers, and Walter van Alst. Deep learning versus traditional machine learning methods for aggregated energy demand prediction. In *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pages 1–6, 2017.

[3] S.L. Wong, Kevin K.W. Wan, and Tony N.T. Lam. Artificial neural networks for energy analysis of office buildings with daylighting. *Applied Energy*, 87(2):551–557, 2010.

[4] Tehseen Zia, Dietmar Bruckner, and Adeel Zaidi. A hidden markov model based procedure for identifying household electric loads. In *IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society*, pages 3218–3223. IEEE, 2011.

[5] AG Azar, E Olivero, J Hiller, K Lesch, L Jiao, M Kolhe, N Asanalieva, P Ferrez, Q Zhang, R Jacobsen, et al. Algorithms for demand response and load control. *SEMIAH, SEMIAH-WP5-D5*, pages 1–v0, 2015.

[6] Seyedeh Narjes Fallah, Ravinesh C. Deo, Mohammad Shojafar, Mauro Conti, and Shahaboddin Shamshirband. Computational intelligence approaches for energy load forecasting in smart energy management grids: state of the art, future challenges, and research directionsand research directions. *Energies*, 11, 2018.

[7] Dario Pevec, Jurica Babic, Martin A Kayser, Arthur Carvalho, Yashar Ghiassi-Farrokhfal, and Vedran Podobnik. A data-driven statistical approach for extending electric vehicle charging infrastructure. *International journal of energy research*, 42(9):3102–3120, 2018.

[8] Tao Hong, Min Gui, Mesut E Baran, and H Lee Willis. Modeling and forecasting hourly electric load by multiple linear regression with interactions. In *Ieee pes general meeting*, pages 1–8. IEEE, 2010.

[9] Ali Lahouar and Jaleleddine Ben Hadj Slama. Random forests model for one day ahead load forecasting. In *IREC2015 The Sixth International Renewable Energy Congress*, pages 1–6, 2015.

[10] Jinkyu Hong and Won Sup Kim. Weather impacts on electric power load: partial phase synchronization analysis. *Meteorological Applications*, 22(4):811–816, 2015.

[11] Henrique S. Hippert and Carlos Eduardo Pedreira. Estimating temperature profiles for short-term load forecasting: neural networks compared to linear models. In *IEE Proceedings - Generation, Transmission and Distribution*, 2004.

[12] J. Valenzuela, M. Mazumdar, and A. Kapoor. Influence of temperature and load forecast uncertainty on estimates of power generation production costs. *IEEE Transactions on Power Systems*, 15(2):668–674, 2000.

[13] Fahad H. Al-Qahtani and Sven F. Crone. Multivariate k-nearest neighbour regression for time series data — a novel algorithm for forecasting uk electricity demand. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[14] R. Afkhami and F.M. Yazdi. Application of neural networks for short-term load forecasting. In *2006 IEEE Power India Conference*, pages 5 pp.–, 2006.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[16] Vasudev Dehalwar, Akhtar Kalam, Mohan Lal Kolhe, and Aladin Zayegh. Electricity load forecasting for urban area using weather forecast information. *2016 IEEE International Conference on Power and Renewable Energy (ICPRE)*, pages 355–359, 2016.

[17] Tao Hong, Min Gui, Mesut E. Baran, and H. Lee Willis. Modeling and forecasting hourly electric load by multiple linear regression with interactions. *IEEE PES General Meeting*, pages 1–8, 2010.

[18] Muhammad Usman Fahad and Naeem Arbab. "factor affecting short term load forecasting,". *Journal of Clean Energy Technologies vol. 2*, pages 305–309, 2014.

[19] M. Mastouri and N. Bouguila. A methodology for thermal modelling and predictive control for building heating systems. *2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, pages 568–573, 2017.

[20] T Salque, D Marchio, and P Riederer. Neural predictive control for single-speed ground source heat pumps connected to a floor heating system for typical french dwelling. *Building Services Engineering Research and Technology*, 35(2):182–197, 2014.

[21] C. Hansen D. Becker Peter Hirsch D.Paraven, A. Debs and R Golob. "influence of temperature on short-term load forecasting using epri-annstlf". 2002.

[22] P. Sarajcev, D. Jakus, J. Vasilj, and M. Nikolic. Analysis of transformer health index using bayesian statistical models. In *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–7, 2018.

[23] Chai Phing Chen, Sieh Kiong Tiong, Siaw Paw Koh, Fong Yu Chooi Albert, and Fauzan K. Mohd Yapandi. Gas emission prediction for environmental sustainability via heterogeneous data sources correlation with support vector regression. In *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–7, 2018.

[24] Oleg A. Yakimenko and William W. Anderson. Challenges in modeling wind power generation based on available weather data. In *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–9, 2018.

[25] Morten Goodwin and Anis Yazidi. A pattern recognition approach for peak prediction of electrical consumption. In Lazaros Iliadis, Ilias Maglogiannis, and Harris Papadopoulos, editors, *Artificial Intelligence Applications and Innovations*, pages 265–275, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[26] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory — ICDT'99*, pages 217–235, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[27] Dharmendra Kumar Patidar, Bhavin C. Shah, and Manoj R. Mishra. Performance analysis of k nearest neighbors image classifier with different wavelet features. *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, pages 1–6, 2014.

[28] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, 1995.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] Nils Jakob Johannesen, Mohan Kolhe, and Morten Goodwin. Comparison of regression tools for regional electric load forecasting. In *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–6, 2018.