Implementation of multivariate statistical techniques for prediction of energy use

Zhengxu Li

June of 2022

Abstract– This project concerns the development of multivariate statistical techniques for the prediction of energy use, based on the main factors that influence energy consumption. Several strategies were studied to elaborate models that are capable to forecast dependent variable (energy use) using various independent variables. The proposed prediction models are relatively simple and uncomplicated to implement, therefore, can play an important role in the maintenance of facilities' energy consumption and be applied as an universal method for industrial world. A primary concern of the work is that the raw data was time series data and contained huge amount of independent variables, as a result, Principal Component Analysis (PCA) was used to reduce the number of parameters to make the model more simple. Then, implementation of two types of linear regression models was performed: Multiple Linear Regression model (MLR) and Generalized Linear Model (GLM). In order to determine the parameters located in these two models, the Ordinary Least Squares (OLS) and Maximum likelihood estimation (MLE) were utilized and their differences were studied. In addition, a machine-learning algorithm named Random Forest (RF) was also investigated to make an comparison between classical methods and machine-learning model. The proposed models present the following characteristics: simplicity, large applicability, good match with obtained data and easy deployment.

Keywords– Regression analysis, forecast method, ordinary least squares, maximum likelihood, GLM, MLR, RF, energy consumption, buildings, appliance.

1 Introduction

The EU has been progressing well in the field of climate and energy, the EU greenhouse gas emissions were reduced by 24% between 1990 and 2019, while the economy grew by around 60% over the same period. A considerable positive change can be seen, however, emissions not covered by the EU Emissions Trading System (ETS), such as emissions from non-ETS industry and buildings remained unchanged between 2018 and 2019, emissions from these sectors have been stable for several years [7].

As part of the European Green Deal, the Commission proposed in September 2020 to raise the 2030 greenhouse gas emission reduction target to at least 55% compared to 1990 [6]. To achieve such target, the European Commission established potential policies, and one of these measures is to reduce emission from buildings in order to improve the energy performance. It also mention that energy demand management in facilities is an important component in the climate system and researches of energy use patterns in buildings could, consequently, play a positive role in emission reductions.

As a result, energy demand management of buildings are expected to be of great importance in infrastructures and has considerable potential to reduce power demand. Recently, researchers have shown interests in forecasting energy use in buildings [3] although it is a challenging work due to the large amount of related factors and difficulties in acquiring appropriate data. By applying prediction methods in facilities, the whole energy utilization could be easily forecast and controlled, and it provides insights about the energy performance of the building. The regression models have the ability that helps to analyse the relationships between variables and to understand their influence, probably even identify the root causes of some abnormal behaviors. These mentioned functions give reasonable stimulation for putting modeling techniques into practice for energy efficiency improvement.

Furthermore, predictive models of building energy consumption can be applied for numerous applications: changing size of Photovoltaic (PV) to control electricity bills

• E-mail adress: zhengxu.li@autonoma.cat
• Work supervised by: Walter Andrés Ortiz Vargas (Department of Mathematics, UAB)
• Course 2021/22

[17], improving energy efficiency and reducing energy costs using semi-centralized decision-making methodology [22], proposing energy system to manage the storage and to allow the user to know their electrical power and energy balances [4].

Regarding the development of the work, an Principal Components Analysis (PCA) was used firstly to reducing the dimensionality of the datasets while preserving as much of the variation as possible. After that, relevant variables were selected to build the Multiple Linear Regression (MLR) and Generalized Linear Model (GLM) prediction models. When we determine the unknown parameters located in the model, a comparison between Maximum Ordinary Least Squares (OLS) and Likelihood Estimation (MLE) model was performed. It is worthy to mention that the validation of measured data is also an important part of this article, where the prediction made by the model were tested with the measured data. Moreover, a Random Forest (RF) model was developed and discussed. Finally, all the models were evaluated with different numeric metrics including MSE (mean squared error) and R-squared (coefficient of determination). A detailed analysis showed that the Random Forest model presents a satisfactory accuracy (correlation coefficient of 0.56) and has better performance among all models.

The objective of this work is to get an adequate understanding of the relation between appliances energy consumption and different variables (for example, the correlation between the energy use and temperature and relative humidity), then mploy and train different methods for energy use prediction as a result of several independent variables and compare the performance of these algorithms in testing process, finally obtain a ranking of the influence of predictors in these models and get superior knowledge related to this topic.

2 PREPARATION

2.1 Literature Review

In the recent years, a number of researchers have sought to model energy loads prediction, different genres include engineering/physical methods, numerical algorithms and artificial neural network approach. For instance, [13] employed a physical model for energy use prediction introducing dynamic approach which is simple but accurate (the validation stage shows satisfactory result of 0.9 R^2). [20] developed a engineering hour-based method for estimation heating demand, showed that temperature and relative humidity have strong influence in heating demand forecasting and the model gave satisfactory results as well.

A numerical model based on multiple linear regression [10] had the objective of making statistical tool to calculate energy usage and then be adopted to benchmark swimming pools, the authors concluded that, for benchmarking purposes, the energy use of facilities, should be normalised with the number of visitors. Besides, the study of [16] presents a statistical model for predicting the time-aggregated power consumption of an indoor building which could be employed for supervision of the facility's energy performance that can quickly detect possible irregularities and thus minimize overall energy use.

Numerous attempts have been made to build AI-based models about this topic. [21] presented ANN (Artificial neural network) algorithm to elicit relationship among variables and project energy usage and thermal level of an indoor swimming pool, the method formed the basis of optimization-based control system for swimming facilities. A study in [1] benchmark state-of-the-art methods for forecasting electricity demand, applied neural networks, exponential smoothing and ARIMA algorithms. Energy consumption was considered and modeled as a stochastic process in [2], the historical data was clustered using k-means algorithm to improve the precision of the prediction, and this study also suggests that a significant part of electricity demands of the buildings comes from consumption of appliances, in fact, the increasing number of family devices makes it more and more important to detect the main contributor to the energy bill, which is also one of the components of this work.

The aforementioned models has the capacity of modeling energy usage for buildings and endeavor to study importance/influence of concerned dependent variables, with expectation of estimating future energy demands. After review those published papers and other posts from different sources, the following interesting points could be highlighted.

- There exist some obstacle to build a universal and efficient model since there exist significant differences between different devices, the performance and pattern of distinct appliance can vary largely from on to another.

- Factors like season and time have been proven to make a relevant influence to the pattern of energy use. Furthermore, the weather parameters also have certain impact on the consumption of energy [5].

- While selecting the algorithm to employ, it is a wise idea to trade off between simplicity and precision according to the applied scenario. When the priority is accuracy, it is beneficial to apply complex model although the implementation could be a taxing work, while the simplicity and transferability are more important aspects to pursue, it is advisable to use simple model like linear regression.

- When we develop the model, it is essential to take a series of assumption in to consideration such as the acquired information are representative enough to the building, in order to determine whether the variables are able to contribute to the improvement of the quality of prediction.

2.2 Methodology

This study investigated the effect of a series of independent variables on energy use. In order to develop a reliable energy prediction model, several multivariate methods have been applied for the analysis. Figure 1 illustrates the research workflow that was followed to reach the main goal of this study, the main steps are identified in the graph.

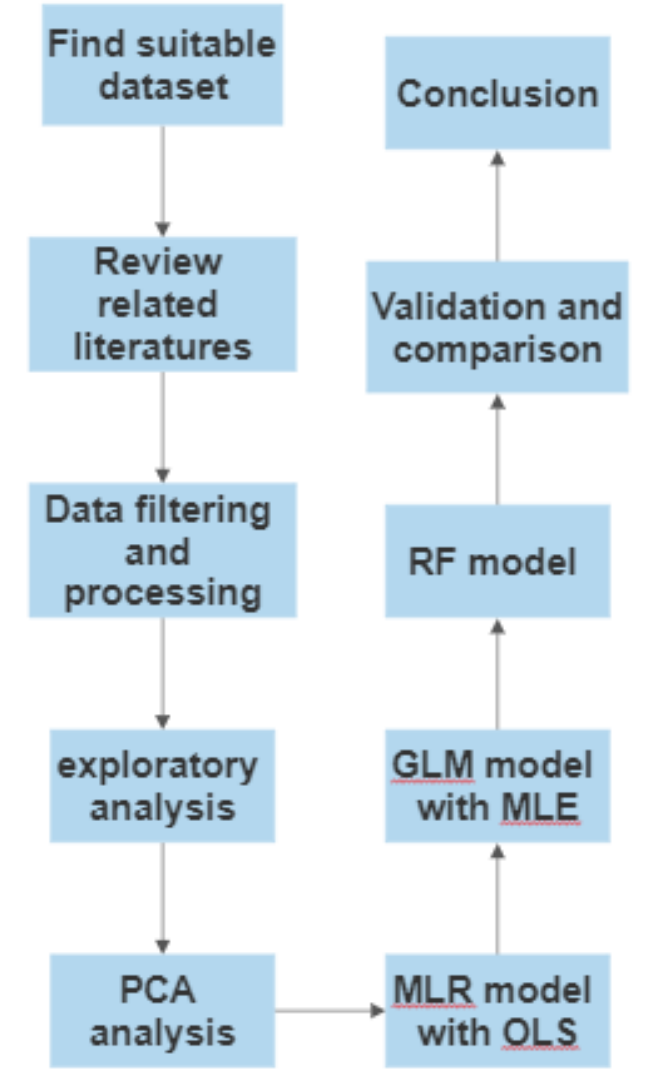

Fig. 1: Diagram of the workflow of this study

3 Data description and variables

This part of the document describes some characteristics of the investigated building and a gives details of the wireless sensor network that was build to capture the data. The following subsections illustrate profiles for all the incorporated variables and an in-depth exploratory data analysis.

3.1 Data acquisition

The investigated house where the data was measured is located in Stambruges, approximately 24 km from Mons (capital of Hainaut province in Belgium). It is located at 50.3 N, 3.43 E, 135 feet above the average sea level, with barometric pressure of 101KPa. The building was constructed and commissioned at 2016, it has a total usable area of 278 m^2, with 220 m^2 of heated area. The Fig.2 shows a photograph of the south-oriented ($+10°$ Southwest from due South) façade for the building. The building has a large list of appliances in each zone/room (living room, laundry, kitchen, bathroom, etc.) of the house which represent a significant part of the overall energy usage, for example, fridge, oven, computers, TV, dishwasher, etc. The building was addressed and designed in accordance with the standard of Passive House Planning Package [8], which supports the development in the field of highly efficient energy use in order to increase energy efficiency in residential buildings.

Fig. 2: Building Image

Household appliance energy consumption was registered with energy counters that is called M-BUS, which is a new European standard for reading of instruments and different types of sensors.The use of M-bus allows fully electronic reading of data. There is no need of maintenance and cleaning and it guarantee 10 years of reliable work, in our case it serves as energy count meter, with original resolution of 10 minutes time steps.

The measurement of the temperature and humidity information was realized through a wireless sensor network (WSN), a internet-connected energy monitoring system of small embedded devices (sensors) that communicate wirelessly following an ad hoc configuration. Specifically, the WSN used for environmental monitoring is called ZigBee, which consists of a coordinator that collect temperature and humidity data from several nodes that are responsible to provide those data. Each sensor node is developed from an Arduino based microcontroller Atemega-328P [11], radio Xbee [14] and basic digital temperature and humidity sensor DHT22 [15]. The Atemega-328p is a, high performance yet low power consumption (powered by batteries), 8-bit microcontroller used in Arduino board. Xbee refer to modules that provide wireless end-point connectivity to devices, with 2.4 GHz frequency. The DHT22 sensor has relatively good specification compared with other similar sensor, its temperature measuring ranges from -40°C to +125°C with ±0.5 degrees accuracy, humidity measuring ranges, from 0 to 100% with 2-5% accuracy. The programmed Atemega-328p microcontroller read the DHT22 sensor data and deliever it to XBee radio and then transmitted to another XBee that serves as coordinator since it is necessary to guarantee high-quality communication inside the large building.

3.2 Exploratory analysis

The raw data of energy use for the appliances (measured in Wh) were registered with a 10-minutes resolution, this variable is the focus of this analysis, also the response variable that will be predicted by the model. The reporting interval of 10 minutes was an adequate resolution to be able to detect any abnormal performance during the whole period. All

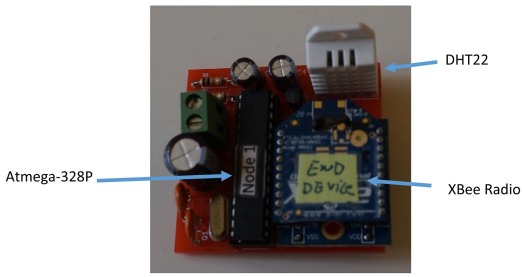

Fig. 3: Image of wireless sensor with three components: microcontroller Atmega-328P, temperature and humidity sensor DHT22, radio XBee

related operations (data cleaning, visualization, implementation and validation of the models, etc.)was done in RStudio [12]. The total time period of the data set has a duration of 137 days (from January to March). The Fig. 4 depicts the appliance energy use of a fraction the studied period (in fact, the data of the second week), it can be seen in the line chart that our dependent variable (energy usage) is wildly fluctuating (ranges from 10 to 1080 watt-hour), although there exist some flat time following the spikes, the curve represents a very variable tendency and lots of peaks are shown, these could be explained by that few equipment like refrigerators that continuously consume energy have stable energy pattern while other appliances like dishwashers and TV have significant temporal variation.

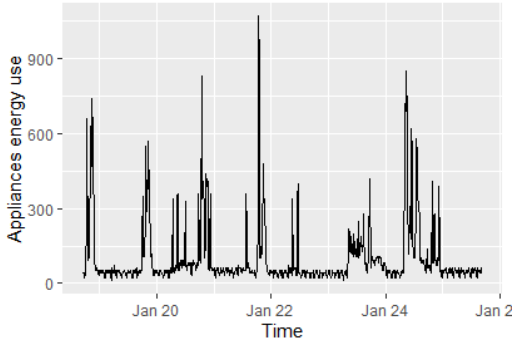

Fig. 4: Energy consumption

A box-plot (Fig.5) is made in order to understand better the distribution of the energy consumption, where the median was located in the bottom side on the box-plot(the median was indicated using the black line inside the orange rectangle), the higher whisker has a value of 170 watt-hour, a number of 10 watt-hour can be seen as the lower hinge of the box-plot. Further, it can be deduced from the box-plot that the distribution of data is quiet disperse and does not follow a Gaussian distribution as a large amount of data were situated above the median, ranged from 200 watt-hour to 900 watt-hour, which were marked with round circles above the upper edge, being interpreted as outliers.

To make the model representative, it is crucial to work with high-quality data, hence, the dataset was cleaned manually by detecting abnormal registered data (probably caused by operational disruption, mechanical flaws, software errors, etc [16]).

As mentioned beforehand in the literature review section, time could be an useful pattern to forecast the energy use for

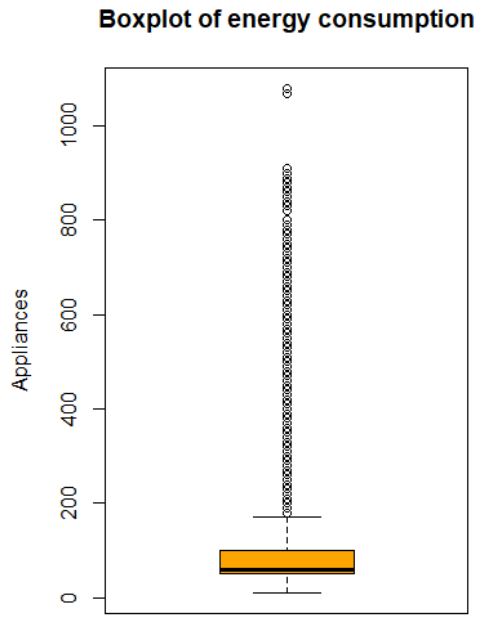

Fig. 5: Box-plot

the reason that it can reflect the human activity habits (the energy use is likely to be high when there are occupancy acting in the building), therefore, it can be useful to generate extra factors based on the time such as TOD (time of the daym, expressed by number of seconds from the 0 a.m) and the DOW (day of the week, Monday, Tuesday, Wednesday...).

To summarize, the considered variables of these study can be categorized in the following groups:

- Response/dependent variable, which is the variable that will be predicted using different models, which names: Appliance.
- Variable that reflects occupancy information of the building: Lights.
- Temperature data measured from different room of the building: Temperature in kitchen (T1), living (T2), laundry (T3), office (T4), bathroom (T5), outside (T6), ironing (T7), teenager (T8), parents room (T9).
- Relative humidity measured from different room of the building: Humidity in kitchen (RH1), living (RH2), laundry (RH3), office (RH4), bathroom (RH5), outside (RH6), ironing (RH7), teenager (RH8), parents room (RH9).
- Meteorological data obtained by a weather station in the vicinity of the building, including: outside temperature (To), atmospheric pressure (Pressure), outside humidity (H0), wind speed, visibility, dew point temperature (Tdewpoint).
- Data derived from the data time: time of the day (TOD, expressed in seconds), day of the week (DOW, refers to Monday, Tuesday, Wednesday...)

The raw data was based on time series at the beginning, and owing to the fact that the order of the autoregressive process were not known, it was critical to investigate autoregressive properties using PACF (partial autocorrelation function) to guarantee that the collected data are independent and to not violate the usual assumption of independent errors made in regression models, so that each observation in the dataset was then treated as independent [18].

An interesting step to take during the process the analysis is calculating the cor-relationship between parameters/variables, correlation coefficients are indicators of the linear relationship between two different variables. A linear correlation coefficient greater than zero suggests a positive relationship, a correlation of 1 is total positive correlation. A value less than zero signifies a negative relationship, -1 is total negative correlation. Finally, zero correlation means that the two variables are independent. Two correlation matrices (Fig.6 and Fig.7) were made to depict the relationships between all pairs of variables with the energy consumption of appliances. These figures show ones along the diagonal since the correlation coefficient respect to the variable itself always equals to one, the grids has scaled coloration where blue color indicates positive correlations and red color represent negative relationship. Besides, the dark the color imply a strong positive/negative correlation while a light color indicates a weak correlation.

Taking a careful look at the first correlation matrix, some interesting finding can be explored, first of all, it illustrates that energy consumption of appliances and lights (0.2) are positively correlated and it is the largest positive values between energy use and other variables in the first plot. The second highest value is with T2 and T6 which are both equal to 0.12. As we expected, the temperature variables have high correlations among them, for instance, T1 have a positive high correlation coefficient 0.89 with both T3 and T5, this finding may indicate that various rooms in the building are heated at the same time or share a same heating/cooling system.

Continue looking at the first correlation matrix and focusing on the relative humidity information, RH2 (living room) and RH6 (outside) are negatively correlated with the appliance energy use with values of coefficient -0.06 and -0.08 respectively, which implies that energy consumption in these areas is lower compared to another zones, so when people stay in these places, the humidity increase (due to human presence) but the appliance energy usage tends to decrease. While for another rooms, distinct behaviours were found, for example, there exist a positive correlation between appliance energy consumption and RH1 (relative humidity in the kitchen), maybe the reason for such relationship is related to the equipment in the kitchen area such as oven, dishwasher and microwave have high energy consumption, therefore, once occupants operates these appliances, the energy use begin to increase. There exist, as well, some slight positive correlation of 0.01 and 0.02 between appliances energy use with respect to RH5 (bathroom) and RH4 (office).

When it comes to the second correlation matrix, a attention-catching value can be seen when look at the first row (the first row or the first column is of greatest interest since one of the purpose of this work is to quantify the impact that different variables have upon the target variable: energy consumption), the highest correlation is the one respect to time of the day (TOD), it make sense as the behavior of occupants shows temporal pattern, more energy would be used when residents were home during the evening and large number of appliances were turned on. Therefore, it can be useful to take the time variable TOD into consideration while implementing the model, this finding can also be shown in the Fig.8 where a heat map was made to show the time component.

Outdoor temperature (Tout) is positively correlated with appliances' consumption with a value of 0.10, which means the higher the outdoor temperature is, the more energy is consumed, it is possible that cooling system works more when outdoor temperature is high. Some negligible correlation with energy use were found for T7 (ironing room, 0.03), T8 (teenager room, 0.04) and T9 (parents room, 0.01), which may inference that energy consumption in this area is irrelevant in relation to the other rooms.

Regarding the humidity data in the second matrix, it shows negative correlation between energy consumption and RH7, RH8 and RH9 being -0.06, -0.09 and -0.05 respectively. It is possible that when the relative humidity increase (normally because of human presence), due to the low energy (compared with other devices) consumption of appliances located in these bedrooms, the power demands tend to decrease.

It is worthy to discuss some interesting finding in the exploratory data analysis considering the two correlations matrix as a whole. One thing that catch attention is that temperatures are highly correlated (0.84 for T1 and T2, 0.89 for T1 and T3, 0.88 for T1 and T4, 0.89 for T1 and T5, 0.94 for T7 and T9, etc.), this may indicate that the rooms share a same heating/cooling system in the design level, and one universal thermal method may be suitable for thermal modeling, where further investigation could be conducted.

4 MODELS IMPLEMENTATION

In this section of elaborating the prediction model, there first step was Principal Component Analysis (PCA), it is important since it helps to diminish the dimentionality, which refers to reducing the number of input variables so that we can consider only a small portion of variables and make the model as concise as possible. After selecting relevant variables, different models (MLR,GLM and RF) are trained, two types of linear regression models (MLR and GLM) are determined using two different parameter optimization techniques (OLS VS MLE). Then, the trained models (including RF) are shown and a comparison of performance of different models will be realized.

4.1 Reduction of dimension using PCA

Principal component analysis (PCA), is a statistical procedure that helps summarizing the information content of a large volume of variables by means of a smaller set of principal components, it helps us to identify patterns in data based on the correlation between features.The goal of PCA is to identify the directions where the data varies the most, the PCA method is especially useful when the variables are highly correlated, which suites well our situation since the temperature/humidity data of different rooms are correlated

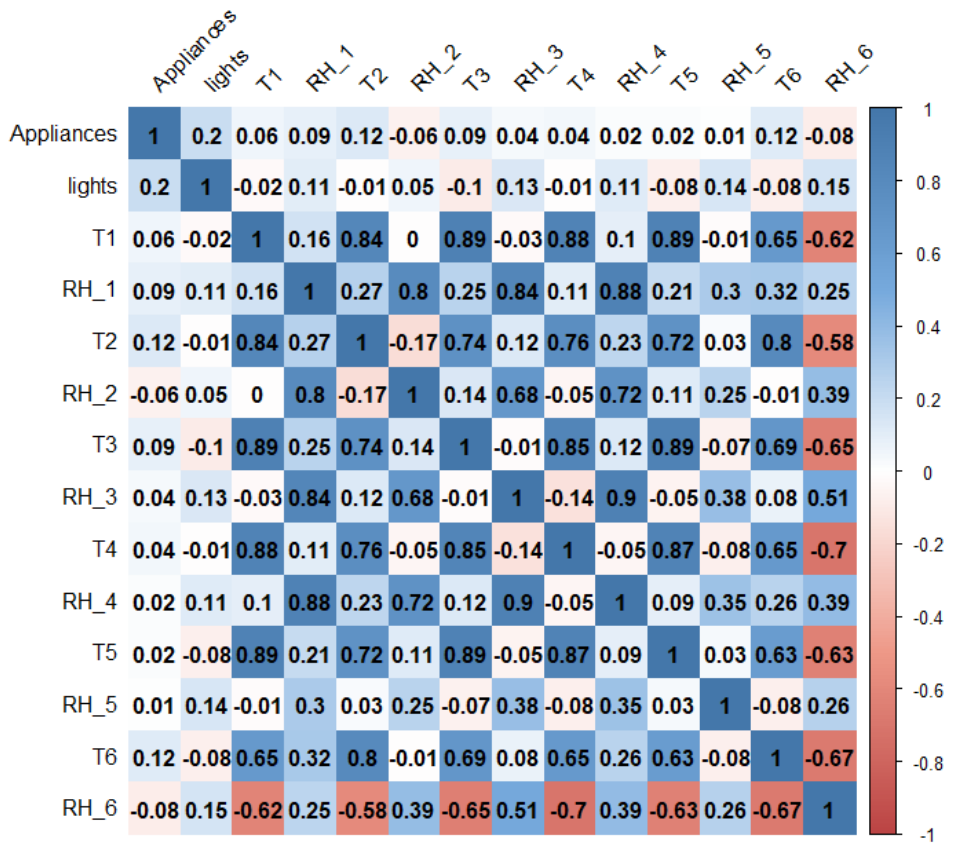

Fig. 6: Correlation matrix showing relationships between the energy consumption of appliances with some variables

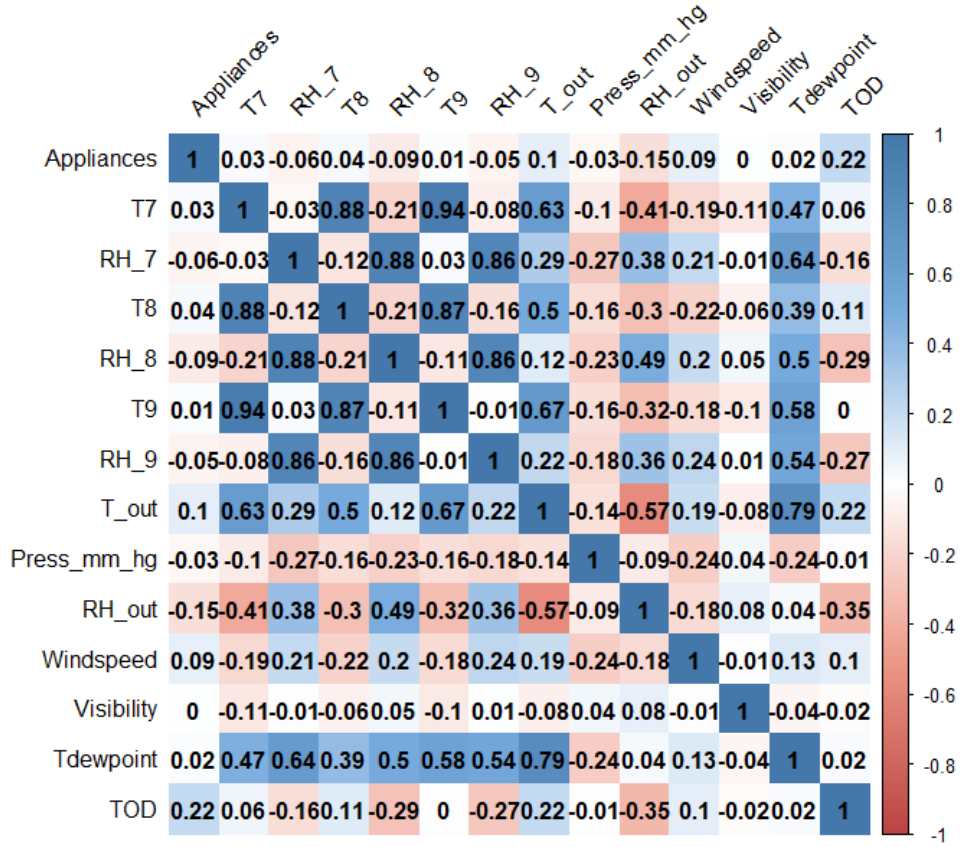

Fig. 7: Correlation matrix showing relationships between the energy consumption of appliances with other variables

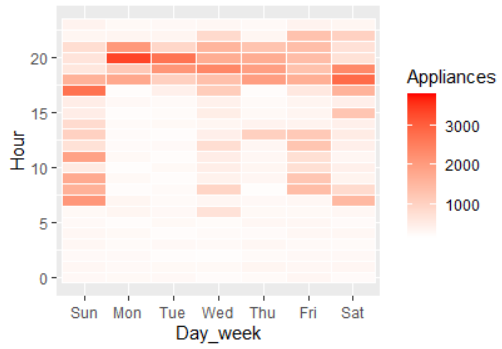

Fig. 8: Hourly energy consumption of appliances of the first week, a time pattern can be observed.

as shown in the part of exploratory analysis. Specifically, in this context, it contributes to the feature selection and gets an initial approximation of the most statistically significant variables [19].

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	9.344663e+00	3.337380e+01	33.37380
Dim.2	7.086321e+00	2.530829e+01	58.68209
Dim.3	2.038417e+00	7.280061e+00	65.96215
Dim.4	1.956591e+00	6.987827e+00	72.94998
Dim.5	1.353608e+00	4.834315e+00	77.78429
Dim.6	1.046823e+00	3.738652e+00	81.52294
Dim.7	9.778976e-01	3.492491e+00	85.01543
Dim.8	8.421726e-01	3.007759e+00	88.02319
Dim.9	6.964005e-01	2.487144e+00	90.51034
Dim.10	5.574105e-01	1.990752e+00	92.50109
Dim.11	4.896125e-01	1.748616e+00	94.24971
Dim.12	3.727034e-01	1.331084e+00	95.58079
Dim.13	1.984266e-01	7.086665e-01	96.28946

Fig. 9: Proportion of variance explained by each eigenvalue

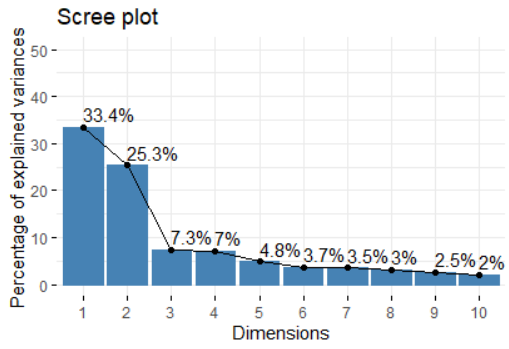

Fig. 10: Scree plot

The proportion of variance explained by each eigenvalue is shown in the second column of Fig.9, in which we can see that 33.37% of variance can be explained by principal component 1, 25.3% of variance can be explained by principal component 2, 7.2% of the variance can be explained by principal component 3. Accumulating the first 3 components, 65.96% of the variance can be explained, which is not a relatively bad result considering all the 28 variables that we take in to account, the aforementioned information is also reflected in the scree plot, which reveals how much variation each PC captures from the data. The y axis refers to the eigenvalue, which essentially stand for the amount of variation. The purpose of using a scree plot is to select which principal components to keep, an ideal curve should be steep, then bends at an “elbow” (cutting-off point), and after that begins to maintain stable. In our case (Fig.10), PC 3 could be an adequate ”elbow” that we look for, as a consequence, the first, second and third principal components are sufficient to describe the data.

In order to get better understanding of the constitution of these most important principal components, further steps of PCA analysis were performed. The Fig.12 (in Appendix) shows the quality of representation for variables on the factor map. It is calculated as the squared coordinates (coordinates of variables along each dimension). A high value indicates a good representation of the variable on the principal component meanwhile a low value give an indication that the variable is not perfectly represented by the PCs. We can deduce that variables of temperature can be well explained by the first principal component while the humidity data can be explained mainly by the second PC.

Finally, the factor map (see Fig.13 in Appendix), which control colors according to variable contributions, shows the relationships between all variables, where positively correlated variables are grouped together. Since all variables related to temperature are closed to each other, being orthogonal to the axis of dim1, variables related to humidity are grouped in another cluster and tend to be perpendicular to dim2, it can be concluded that temperature and humidity are relevant in this analysis.

All those graphs regarding PCA analysis described above are information-telling as they suggest that variables regarding temperature and humidity are of good quality and contribute great proportion to the most relevant principal component such as the first and the second PC. This helpful clue can guide us to select adequate variables to implement our prediction models in the next section.

To build our models, the filtered data was split into two subsets: training and test, where three quarter of the data (14803 rows) was used for the training of the models and the rest 25% accounted for testing set.

4.2 Implementation of MLR models

The first implemented model was the Multiple linear Regression. Multiple linear regression (MLR) is a statistical technique that uses a series of explanatory/independent variables to predict the response/dependent variable. MLR has the potential to be an adequate model in the matter of implementation complexity, it is easy to follow and it is adaptable to the different cases within a acceptable margin of error [9]. The equation that relates y-variable to x-variables is commonly expressed by:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \epsilon_i.$$

where y_i represents the response variable, β_i refers to the slope coefficient in the predictors, x_i is the independent variables and ϵ_i stands for error term.

It is essential to diagnose several fundamental assumptions for linear regression model, for example, examine spread-location plot to check the homogeneity of variance, visualize the QQ plot of residuals to estimate the normality assumption.

In this part, MLR model based on Ordinary Least Squares(OLS) was trained, OLS is a technique used for modelling of linear relationships between a response variable and one or more predictor variables, the MLR model is trained in a manner that the sum of squares of differences of prediction and observed values will be minimized.

MLR models based on different subset of variables were trained and validated using the previously separated data sets, the comparison of results for training and testing sets among these models (built with different sets of variables) is displayed in the table 1.

TABLE 1: MLR MODELS WITH OLS

Variables	MSE_Tr	R2_Tr	MSE_Te	R2_Te
Temperature	99.52	0.08	95.71	0.09
Humidity	98.33	0.094	94.53	0.11
All	94.35	0.16	90.39	0.18

The tabla contains 4 columns, where the first two columns refer to the training set and the last two columns show the testing set. The first row imply training the model with all the variables related to temperature (T1, T2, ... , T9), while the second row means training the model with all the variables concerning humidity (RH1, RH2, ... , RH9). The last row refers to the model that trained with all the independent variables. It can be seen that the last model, as the more complete one, has the best performance than the first two MLR models based on only a subset of the variables. However, that does not necessarily conclude that models based on only temperature/humidity variables have inferior performance, since they employed less number number independent variables and were trained more rapidly, these models are worthwhile in cases where precision is not the first priority.

4.3 Fit Generalized Linear Regression models using MLE

Apart from determine the parameters only using the default method (OLS) integrated in R, we also fit our linear model using MLE technique to make a comparison, in this case, a Generalized Liner Model (GLM) is discussed. As mentioned before, a common method to estimate coefficient located in MLR model is Ordinary least squares, however, OLS makes assumptions about the elements of the error term follows Gaussian distribution, which is not always appropriate for every problem, it is possible that they are independent to each other. Generalized linear model is a flexible generalization of ordinary linear regression and can be used as a method to ”relax” the assumption of errors term.

The technique that the generalized linear model uses to estimate parameters (i.e.intercept and slopes) is called maximum likelihood estimation (MLE). This technique is different than the technique used for OLS regression model which sought to minimize the sum of squared residuals, although it turns out that OLS regression technique is the maximum likelihood estimate for a simple linear model. Actually, if we suppose the error distribution as a Gaussian error distribution with an identity link function, we are estimating a traditional linear model (MLR).

It seems that they are identical methods, but in fact, Generalized linear models (GLM) extends the basic idea of MLR model to incorporate more diverse outcomes and to specify more directly the data generating process behind our data. The reason why set up this more complex framework is that by changing the error distribution and link function, it is possible to accommodate a broad set of models that are cannot be estimated well by OLS regression techniques. It can be seen that in our model, there are categorical type variables (for example, DOW which represents the day of the week), which is why, to make such prediction, a generalized linear model can be used, taking into account that the distribution family in this case is Poisson.

To make a proper comparison, we establish the three GLM models using the same subset of variables (variables related to temperature, variables related to humidity and all registered variables) as the three MLR models developed before.

TABLE 2: GLM MODELS WITH MLE

Variables	MSE_Tr	R2_Tr	MSE_Te	R2_Te
Temperature	98.3	0.095	95.72	0.08
Humidity	98.5	0.0927	94.52	0.11
All	94.4	0.165	96.10	0.18

GLM models based on different subset of variables were trained and the difference between these models is displayed in the table 2. Similarly to MLR, the model with highest number of variables tends to be more precise. Comparing with the table 2 of the MLR model, a small improvement can be seen as the MSE tend to be lower and R-squared tend to be slightly higher, which correspond to our theoretical expectation that GLM models are more suitable for this data set.

4.4 Machine Learning Algorithm – Random Forest

This part of the work discuss a machine learning algorithm, the random forest (RF) models were implemented for prediction of energy consumption using different set of predictors and were compared to investigate the importance of these variables and identify the most influential features.

Random Forest Regression is a popular machine learning algorithm that belongs to the supervised learning techniques, it uses ensemble learning method for regression, can perform both classification and regression tasks for Machine Learning by constructing a multitude of decision trees.

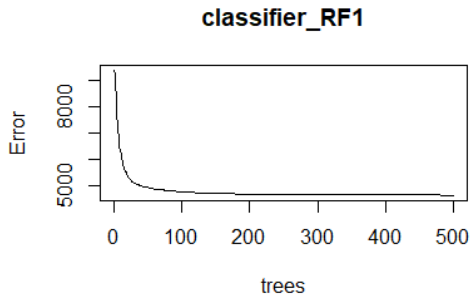

Fig. 11: Training process for Random Forest

The Fig.11 shows the relationship between the error and

number of decision trees that form the forest, the error decrease rapidly until when number of trees approximates 80, suggesting that optimal number of tress is around 80 which corresponds to the 'elbow point'.

Analogy to the prediction model established before, three RF models using three sets of variables was developed and the results are shown in the table 3, we arrive the same situation as before, in which the model deployed with all variables had lowest MSE and highest R-squared compared with other two models that used less predictors.

TABLE 3: RF MODEL

Variables	MSE_Tr	R2_Tr	MSE_Te	R2_Te
Temperature	33.14	0.92	66.02	0.57
Humidity	33.34	0.93	68.73	0.54
All	30.28	0.94	65.95	0.57

The relative importance of variables for the random forest model was also studied, being quantified by the residual sum of squares (RSS that mesures the level of variance in the error term). According to the RF algorithm, the most important parameters are: TOD, RH5, lights and atmospheric pressure. The random forest model represent a MSE of 60.46 and has ability to explain 57.55% the variance (according to the R-squared value), which shows a significant improvement than all the models of linear regression class (MLR and GLM).

4.5 Performance and comparison of models

The experimental data shows that the RF models present notable advantages in energy consumption prediction and is the optimal prediction model compared to other designed models (MLR and GLM). The best model of type RF was trained with all dependent variables, it is capable of explaining 94% of the variance in the training split and is able to explain 57% of the variance in the testing split (while MLR and GML had merely less than 20% R-squared value), despite the difficulties related to prediction modeling of energy consumption.

5 Conclusion

It can be concluded that for this specific dataset the ideal model was based on the random forest method, while the GLM model also provide considerably good result. MLR methods is able to explain the building energy consumption using an linear combination of the influencing variables in a simple way although the precision in considerably low compared with other models. Random forest algorithm was proven to be reliable and effective, outperform linear regression in this case, the explanation could be the RF has more larger number of parameters and can therefore fit more easily than regression models that has few number of parameters.

It is worth to point out the TOD was the most influential predictor that dominate the forecasting task, which can possibly explained by that the energy use is mostly determined by human activities since the occupants have their own daily routine and repeat on daily basis.

The study of data has shown intriguing findings in both the exploratory data analysis and model implementation. The correlation matrices demonstrate the different relationships between parameters. PCA analysis can be especially useful when we process a huge amount of collected variables and find the relevant parameters.The MLR and GLM models have similar performance when evaluating the model with MSE and R-squre, RF models improve the precision considerably compared with these two linear based model. The implemented models could be practical for energy prediction and energy building simulation studies thanks to their precision and simplicity, additionally, it could have potential capability to serve as a great support tool for industry as it can be customized within a short time to find energetic solutions.

The results from the analysis of the dataset stressed the significant importance of studying the data before training the model. It is indispensable to filter and clean the dataset after an in-depth investigation, in fact, a big part of data engineer's job consist of data cleaning. It is necessary to guarantee the quality of data before employ the prediction model.

Employing analysis based on data regarding only one house is one of the limitation of this work, it would be beneficial if more scenarios were investigated. Moreover, the study could have been robuster if more data were given, since short period of data may not be able to detect seasonal pattern or tendency of consecutive years. The prediction of energy use could possibly more precise if more sensors were installed to collect more variety of data.

Future work could be done in order to improve the model, for example, considering more meteorological information such as raining, more human activity information such as occupant's behaviours, even design factors like morphology of the building and thermal inertia could be helpful to identify relationship with other parameters, provide better understanding and hence improve the accuracy. More advanced study could be addressed to improve robustness of the model and transfer to other similar buildings.

Acknowledgments

I would like to thank the tutor Walter Andrés Ortiz Vargas for proposing the idea of establishing prediction model and initial framework of the paper, as well as supervising me passionately and patiently during the whole process. I would like to show my gratitude to the his pearls of wisdom concerning the development of this article, and sharing his valuable insight.

We are also immensely grateful to Dr. Luis Candanedo for upload publicly the dataset to the UCI machine learning repository so that we have the precious opportunity to perform our analysis and study.

References

[1] V. Andreas, G. Christoph, T. Rohit, D. Christoph, and J. Hans-Arno. Household electricity demand forecasting – benchmarking state-of-the-art methods. 04 2014.

[2] J.A. Candanedo, V.R. Dehkordi, and M. Stylianou. Model-based predictive control of an ice storage device in a building cooling system. *Applied Energy*, 111:1032–1045, 2013.

[3] L.M. Candanedo, V. Feldheim, and D. Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

[4] M. Castillo-Cagigal, E. Caamaño-Martín, E. Matallanas, D. Masa-Bote, A. Gutiérrez, F. Monasterio-Huelin, and J. Jiménez-Leube. Pv self-consumption optimization with storage and active dsm for the residential sector. *Solar Energy*, 85(9):2338–2348, 2011.

[5] G. Ciulla and A. D'Amico. Building energy performance forecasting: A multiple linear regression approach. *Applied Energy*, 253:113500, 2019.

[6] European Commission. Communication from the commission to the european parliament, the european council, the council, the european economic and social committee and the committee of the regions. `https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF`, 2022.

[7] European Commission. Eu progress made in cutting emissions. `https://ec.europa.eu/clima/eu-action/climate-strategies-targets/progress-made-cutting-emissions_en`, 2022.

[8] W. Feist, R. Pfluger, B. Kaufmann, J. Schnieders, and O. Kah. Passive house planning package 2007. *Specifications for Quality Approved Passive Houses, Technical Information PHI-2007/1 (E), Darmstadt, Passivhaus Institut (December 2007)*, 2007.

[9] A.A. Gassar and S.H. Cha. Energy prediction techniques for large-scale buildings towards a sustainable built environment: A review. *Energy and Buildings*, 224:110238, 2020.

[10] W. Kampel, S. Carlucci, B. Aas, and A. Bruland. A proposal of energy performance indicators for a reliable benchmark of swimming facilities. *Energy and Buildings*, 129:186–198, 2016.

[11] D. Kumar, A. Beckers, J. Balasch, B. Gierlichs, and I. Verbauwhede. An in-depth and black-box characterization of the effects of laser pulses on atmega328p. pages 156–170, 2018.

[12] L.Jan. An r and s-plus companion to multivariate analysis. *Journal of Statistical Software*, 14, 10 2005.

[13] L.Tao, L.Xiaoshu, and V.Martti. A new method for modeling energy performance in buildings. *Energy Procedia*, 75:1825–1831, 2015. Clean, Efficient and Affordable Energy for a Sustainable Future: The 7th International Conference on Applied Energy (ICAE2015).

[14] Product Manual. Xbee znet 2.5/xbee pro znet 2.5 oem rf modules. *Digi International Inc*, 2008.

[15] M.Bogdan. How to use the dht22 sensor for measuring temperature and humidity with the arduino board. *Acta Uiversitatis Cibiniensis–Technical Series*, 68:22–25, 2016.

[16] O.Ø. Smedegård, T. Jonsson, B. Aas, J. Stene, L. Georges, and S. Carlucci. The implementation of multiple linear regression for swimming pool facilities: Case study at jøa, norway. *Energies*, 14(16), 2021.

[17] F. Spertino, P. Di Leo, and V. Cocina. Which are the constraints to the photovoltaic grid-parity in the main european markets? *Solar Energy*, 105:390–400, 2014.

[18] G. Tunnicliffe Wilson. Time series analysis: Forecasting and control,5th edition, by george e. p. box, gwilym m. jenkins, gregory c. reinsel and greta m. ljung, 2015. published by john wiley and sons inc., hoboken, new jersey, pp. 712. isbn: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37:n/a–n/a, 03 2016.

[19] N. Wermuth. *Multivariate Statistical Analysis*, volume 24, pages 915–. 01 2011.

[20] L. Westerlund, J. Dahl, and L. Johansson. A theoretical investigation of the heat demand for public baths. *Energy*, 21(7):731–737, 1996.

[21] B. Yuce, Haijiang L., Y. Rezgui, I. Petri, B. Jayan, and Chunfeng Y. Utilizing artificial neural network to predict energy consumption and thermal comfort level: An indoor swimming pool case study. *Energy and Buildings*, 80:45–56, 2014.

[22] P. Zhao, S. Suryanarayanan, and M.G. Simoes. An energy management system for building structures using a multi-agent decision-making control methodology. *IEEE Transactions on Industry Applications*, 49(1):322–330, 2013.

APPENDIX

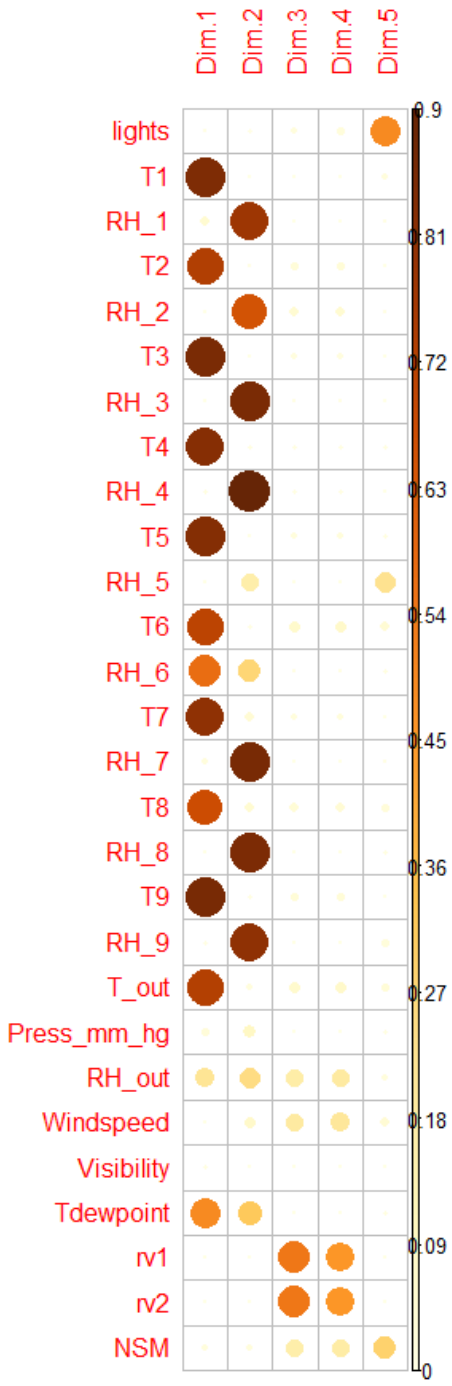

Fig. 12: Quality of variables

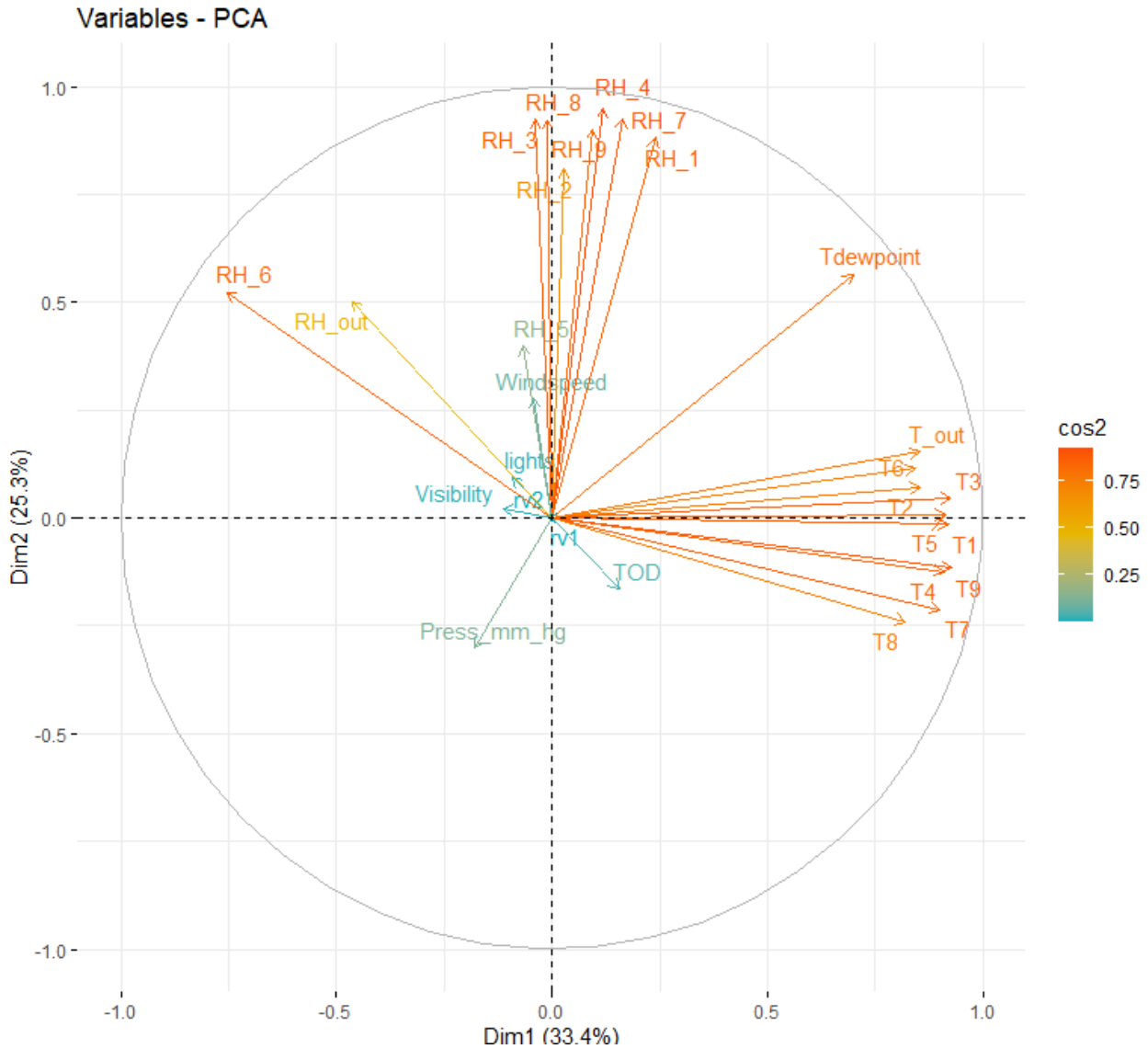

Fig. 13: Variables Factor Map