Energy and Buildings 140 (2017) 81–97

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Data driven prediction models of energy use of appliances in a low-energy house

Luis M. Candanedo*, Véronique Feldheim, Dominique Deramaix

Thermal Engineering and Combustion Laboratory, University of Mons, Rue de l' Epargne 56, 7000 Mons, Belgium

ARTICLE INFO

Article history:
Received 20 September 2016
Received in revised form
26 December 2016
Accepted 27 January 2017
Available online 31 January 2017

Keywords:
Appliances
Energy
Prediction
Wireless sensor network
Statistical learning models
Data mining
Data sets available

ABSTRACT

This paper presents and discusses data-driven predictive models for the energy use of appliances. Data used include measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures. The paper discusses data filtering to remove non-predictive parameters and feature ranking. Four statistical models were trained with repeated cross validation and evaluated in a testing set: (a) multiple linear regression, (b) support vector machine with radial kernel, (c) random forest and (d) gradient boosting machines (GBM). The best model (GBM) was able to explain 97% of the variance (R^2) in the training set and with 57% in the testing set when using all the predictors. From the wireless network, the data from the kitchen, laundry and living room were ranked the highest in importance for the energy prediction. The prediction models with only the weather data, selected the atmospheric pressure (which is correlated to wind speed) as the most relevant weather data variable in the prediction. Therefore, atmospheric pressure may be important to include in energy prediction models and for building performance modeling.

1. Introduction

The understanding of the appliances energy use in buildings has been the subject of numerous research studies [1–8], since appliances represent a significant portion (between 20 and 30% [7,9]) of the electrical energy demand (See Fig. 1). For instance, in a study in the UK for domestic buildings [4], appliances, such as televisions and consumer electronics operating in *standby* were attributed to a 10.2% increase in the electricity consumption. Regression models for energy use can help to understand the relationships between different variables and to quantify their impact. Thus, prediction models of electrical energy consumption in buildings can be useful for a number of applications: to determine adequate sizing of photovoltaics and energy storage to diminish power flow into the grid [10], to detect abnormal energy use patterns [11], to be part of an energy management system for load control [1,12,13], to model predictive control applications where the loads are needed [14], for demand side management (DSM) and demand side response (DSR) [6,15,16] and as an input for building performance simulation analysis [2,17,18].

As indicated in [4], the electricity consumption in domestic buildings is explained by two main factors: the type and number of electrical appliances and the use of the appliances by the occupants. Naturally, both factors are interrelated. The domestic appliances use by the occupants would leave traceable signals in the indoor environment near the vicinity of the appliance, for example: the temperature, humidity, vibrations, light and noise. The occupancy level of the building in different locations could also help to determine the use of the appliances. In this work, the prediction was carried out using different data sources and environmental parameters (indoor and outdoor conditions). Specifically, data from a nearby airport weather station, temperature and humidity in different rooms in the house from a wireless sensor network and one sub-metered electrical energy consumption (lights) have been used to predict the energy use by appliances. Four regression models have been tested, namely (a) multiple linear regression model (lm), (b) support vector machine with radial basis function kernel (SVM-radial), (c) random forest (RF) and (d) gradient boosting machines (GBM) with different combinations of predictors.

The present work mostly deals with the problem of aggregate appliances energy use prediction rather than the topic of modeling of appliances energy loads. Because of that, the literature review focuses on this topic. Nevertheless, research related to studies of appliances loads in buildings and modeling are included as well in this study.

* Corresponding author.
E-mail address: Luismiguel.candanedoibarra@umons.ac.be (L.M. Candanedo).

http://dx.doi.org/10.1016/j.enbuild.2017.01.083
0378-7788/

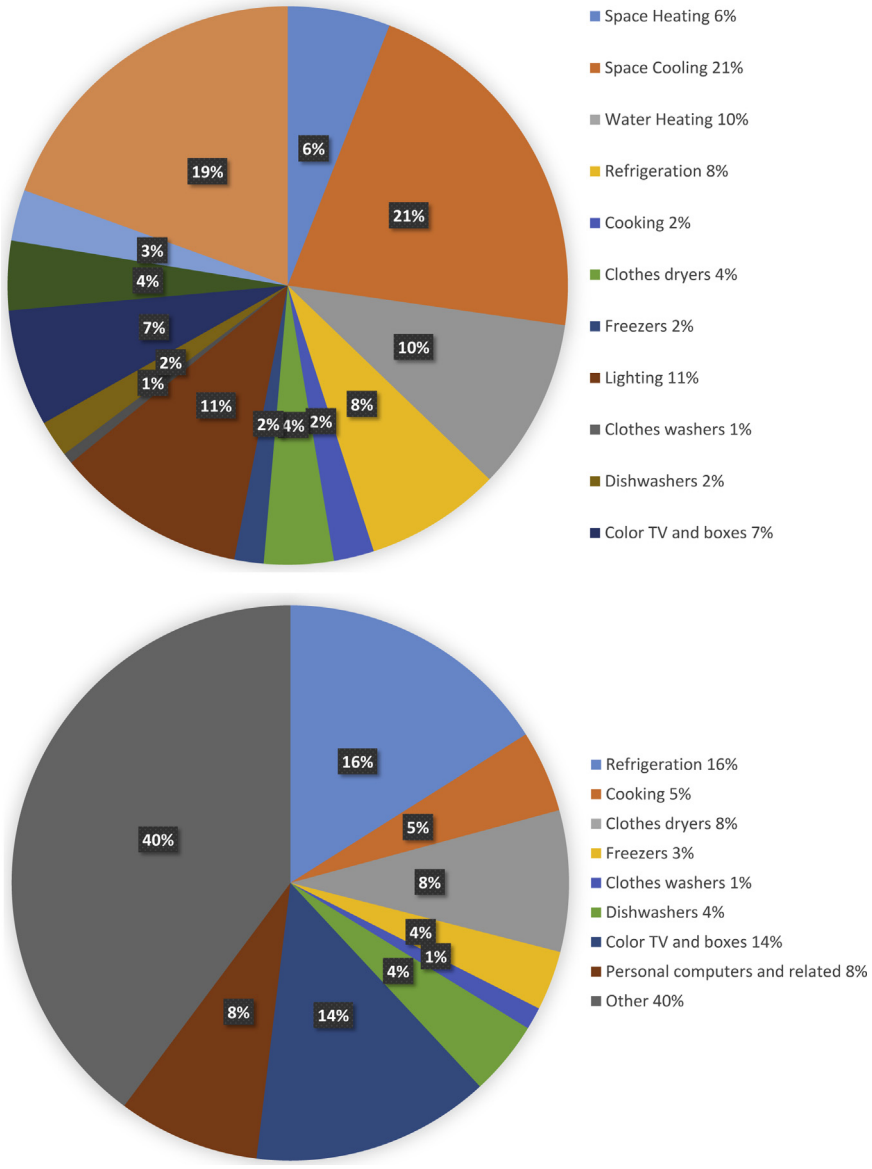


Fig. 1. Residential electrical energy consumption adapted from Table A4 page 139 of annual energy Outlook projections for 2015 [19]. Top chart, percentage of all delivered energy, Bottom chart, after removing Space heating, Space cooling, Lighting and Water Heating contributions.

1.1. Literature review

1.1.1. Appliances' loads in buildings and numerical modeling of their consumption

This section reviews some articles regarding modeling of appliances and other socio economic factors that help to understand the different data and methodologies that have been used in the past to understand appliances' energy use.

Hourly end-use data was collected from 454 houses and 140 commercial buildings in the Pacific Northwest [20]. The data acquisition description for the data analyzed can be found in [21]. Typically, the data acquisition systems monitored between 12 and 16 channels for energy consumption. The study in [20] pointed out that the end-use loads showed large temporal variations, and that although refrigeration and freezer loads tended to be very flat, cooking (food preparation) dishwasher, lights and small appliances showed distinct evening peaks.

A method to generate occupancy data for UK households was presented in [22]. The model uses the Markov Chain Monte Carlo technique to generate synthetic data and is recommended by the authors to estimate the energy demand of appliances, lighting and heating.

A methodology to estimate building energy consumption from EnergyPlus benchmark models was presented in [23]. The EnergyPlus benchmark models were developed by a joint effort between DOE'S Building Technologies Program, Pacific Northwest National Laboratory, Lawrence Berkeley National Laboratory and National Renewable Energy Laboratory [24]. The authors in [23] propose the use of a series of predetermined coefficients to estimate hourly energy consumption from utility bills, in order to relieve the user from performing dynamic simulation of the building.

A Norton equivalent technique was used to model residential loads in [25]. The modeled loads included appliances such as a refrigerator, PC, laptop, TV and washing machines.

A recent study focused on major household appliances (refrigerator, clothes washer, clothes dryer and dishwasher) to find daily energy use profiles for each of them [9]. It was shown that refrigerators have a uniform load profile, while clothes washers, clothes

dryers and dishwashers are very user-dependent and thus vary from household to household and time of day. Research on the potential for demand response for refrigerators, clothes washers, clothes dryers and dishwashers was presented in [6]. The article ranked highest the clothes dryers for the potential of demand response mostly because of their large power demand.

A model was developed by [26] to detect and estimate individual home appliance loads from aggregated power signals. The model employs the so-called explicit duration Hidden Markov model to detect the loads of home appliances.

A literature review was presented in [27] examining socio-economic factors, dwelling characteristics and appliances affecting electricity consumption in domestic buildings. The paper listed several appliances and parameters as having a significant positive effect on domestic electricity consumption: number of appliances, desktop and laptop computers, TV, video player/recorder, video console, electric oven, range hood, refrigerators, freezers, dishwasher, washing machine and tumble dryer among others.

A recent study focused on the thermal modeling of electrical appliances for highly insulated buildings [2]. The study emphasized the necessity of including electrical appliances modeling to obtain more accurate and efficient building energy simulations.

The models presented before are mostly useful for energy building simulation studies to evaluate different buildings designs questions and to try to predict their impact in the buildings energy balances or to estimate future energy bills. The next section will deal with the problem of energy use prediction during the operation phase.

1.1.2. Electricity load prediction

This section presents and discusses research addressing electricity load prediction to identify the parameters, models and other methods that have been useful for energy prediction.

Typically studies have used models such as multiple regression, neural networks, forecasting methods [3,28,29], engineering methods, support vector machines [30], time series techniques [31] and forecasting methods [29] to predict the electricity demand. The models usually have considered parameters such as the time of day, outdoor temperature, month, weekend, holidays, yesterday's consumption, rainfall index, global solar radiation, wind speed and occupancy [32,33].

On a larger scale, the impact of weather variables on the monthly (aggregated) electricity demand in England and Wales between the years 1983–2003 was studied in [34]. The employed method was the parameter multiple regression model for regression. The study used heating degree days, cooling degree days, gross domestic product, and humidity (when available) and was able to explain the variability of the monthly demand between 91 and 95%.

The study of electrical energy use patterns in buildings has been an active area of research in recent years. One minute interval power measurement in 12 houses, including individual devices (furnace, air conditioner, range/cooker, clothes dryer, dishwasher and domestic hot water heater) was presented in [35]. One of the findings was that the daily variation in the temporal distribution of each house was significant. Another study examined electricity data of 1628 households [36]. From the data available to the researchers, an extensive list of variables was studied: weather, location (ZIP code), age of building, ownership, presence of double pane windows, energy efficient light fixtures, floor area, pet ownership, number of refrigerators and entertainment devices, number of occupants and income level were studied. The researchers concluded that the most important variables were weather, location, and floor area. Also the number of refrigerators and entertainment appliances (e.g. videocassette recorders, VCRs) are among the most important determinants of daily minimum consumption. Another study found that being at home during the day correlated with lower appliance efficiency [7]. The provided explanation by the researchers is the lower efficiency was likely due to the increased use of appliances when the house is occupied more often.

A prediction system for the problem of individual appliance prediction was presented in [8]. The system used information such as past consumption, hour, day, season and month. The system is capable of learning from past data. One of the main conclusions was that the last 24 h are the most relevant for prediction.

More recently, 23 houses in Ottawa, Canada were monitored for energy use at 1-min resolution intervals for the air conditioners, furnace fans, some major appliances and other Non-HVAC appliances. In this study the number of occupants was found to be the strongest predictor for Non-HVAC energy use [17]. Also in the UK, appliance ownership and use was studied for 183 dwellings using an odds ratio analysis [5]. The authors reported that households owning more than thirty appliances have an increased probability of having a high electrical energy demand. Some of these households own more than four or more IT devices, more than five entertainment items, an electric oven, hob or range, two or more preservation and cooling appliances or three or more laundry machines. The main limitation of this study was that the data was obtained from a survey and not from measurements.

The review of the published literature highlights the following points:

- The energy consumption of appliances represents a significant portion of the aggregated electricity demand of the residential sector (up to 30% [9] see Fig. 1), and is also important for power management of the grid [3].
- The increased number of appliances owned makes it important to identify which ones are the main contributors to the energy consumption.
- The energy consumption of appliances may be broken down into different contributions and sometimes may include HVAC devices such as air conditioners.
- The patterns for energy use of appliances can vary significantly (e.g., nearly flat for refrigeration equipment, while it is highly variable for devices such as clothes washers, clothes dryers and dishwashers).
- Weather parameters have been proven relevant to predict the electricity energy consumption in buildings. This consumption typically includes the HVAC contributions, which may have different trends with respect to temperature for example. Thus, it would be desirable to know if exterior weather parameters can also help in the prediction of appliances energy use.
- For highly insulated buildings, the thermal influence of appliances on internal gains become more important and relevant in building energy performance [2].

This paper explores several questions. Is the weather data obtained from a nearby weather station representative enough to improve the appliances energy consumption prediction? Can the temperature and humidity measurements from a wireless network help in the energy prediction? From all the data used in prediction models, which parameters are the most important in energy prediction? Since occupancy has a direct effect on the appliances use, what impact does the inclusion of a sub metered energy measurement related to occupancy (light) have on the prediction?

1.2. Research objectives and methodology outline

The purpose of this work is to understand the relationships between appliances energy consumption and different predictors. Also to discuss the performance of different models (linear regression, support vector machines, RF and GBM) to predict

Fig. 2. House pictures.

energy consumption. Furthermore, to rank the influence of predictors/parameters in the prediction.

After the introduction, the paper continues with a description of the house and follows with a description of the energy meters and wireless sensor network. The next section shows profiles for the appliances consumption and exploratory data analysis, followed by a description of the data splits and features. Next, a data filtering is presented to remove any non-relevant variables. After different trained models are presented, the next section corresponds to the model selection and performance. Then, the training of models with different data subsets is shown and the evaluation of trained models in the test sets is presented. The paper ends with a discussion of results and conclusions.

2. House description

The house is located in Stambruges, which is about 24 km from the City of Mons in Belgium (See Fig. 2). The construction was finished on December 2015. All the mechanical systems are new. The low energy house was designed according to the passive house certification [37]. For this certification, the building was designed to have an annual heating load and cooling load of no more than 15 kWh/m^2 per year according to the Passive House Planning Package (PHPP) software design tool. It is important to note that a wood chimney provides most of the heating load in the house. The total kg amount and type of wood has been manually logged monthly. The building air leakage was measured in September 2016 and is 0.6 air changes per hour at 50 Pa. The house was designed with $U < 0.1$ W/m^2 K for the exterior walls, roof and ground. Triple glazed windows are used with $U_g = 0.5$ W/m^2 K and $U_f < 0.9$ W/m^2 K. The ventilation is provided by a heat recovery ventilation unit with between 90 and 95% efficiency. The total floor area is 280 m^2, from which the total heated area is 220 m^2. The façade of the house is oriented +10° (Southwest) from due South. There are usually four occupants of the house, two adults and two teenagers. One of the adults works regularly in the home office.

2.1. Energy meters and ZigBee nodes

The electric energy metering at the passive house was done with M-BUS energy counters. The information from these energy counters is collected every 10 min. The individual electrical load metering includes: the heat recovery ventilation unit, domestic hot water heat pump (COP around 2.7), the energy consumption of appliances, lighting, and electric baseboard heaters. The appliances energy metering includes the energy used by the devices listed in Table 1. The energy information is collected with an internet-connected energy monitoring system where it is stored and then it is reported by e-mail every 12 h. Fig. 3 shows the aggregated energy consumption per month together with the percentage of each contribution. Light energy consumption ranges from 1 to 4% of the total since most of lighting fixtures are LEDs. It can easily be seen that the energy consumption of appliances represents between 70 and 79% of the monthly electrical energy consumption. For comparison, lighting and appliances used about 1.23 more energy than domestic hot water in Spain, 1.28 more than in the EU overall and 1.74 more than in the USA according to data for the residential sector in 2003 [38]. In this study, the appliances and lighting consume between 5.7 and 6.7 more than the DHW energy use.

Table 1
List of appliances in each room or house zone.

Room	Equipment
Laundry	Small Fridge, Upright freezer, Wine Cellar for 160 bottles, Washing machine, Dryer, Internet router, internet hub, Network Attached Storage
Garage*	Rain water pump, electric garage door
Kitchen	Fridge, Induction cooktop, Kitchen hood, Microwave, Oven, Dishwasher, Coffee machine
Dining	WIFI booster, ZigBee coordinator, electrical blinds
Living	TV 138 cm, Hard drive enclosure, DVD player, cable box, laptop, Ink-jet printer, electric blinds
Office	2 desktop computers, 3 computer screens, 1 router, 1 laptop, 1 copier-printer, electric blinds
Ironing	Alarm clock, radio, Iron, electric blind
Room 1	Alarm clock, radio, electric blind, 2 lamps
Room 2	Desktop computer, monitor, alarm clock, electric blind
Room 3	Laptop, alarm clock
Game	93 cm TV, Internet router, DVD player, PlayStation
Bathroom 1	2 electric toothbrushes, hair dryer
Bathroom 2	2 electric toothbrushes
Attic*	Computer, Musical Instruments, Amplifier

Note: The listed equipment in the rooms marked with an * are outside the measurement range of the wireless sensor network.

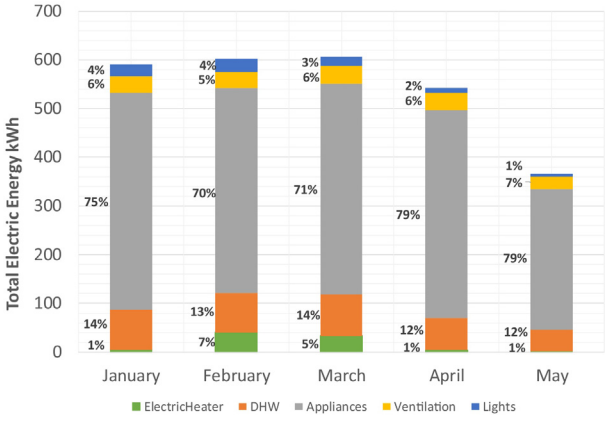


Fig. 3. Aggregated electric energy consumption per month. This graph includes all the energy consumed from January 1st until May 22nd 2016. Note that the energy for the month of May is not complete.

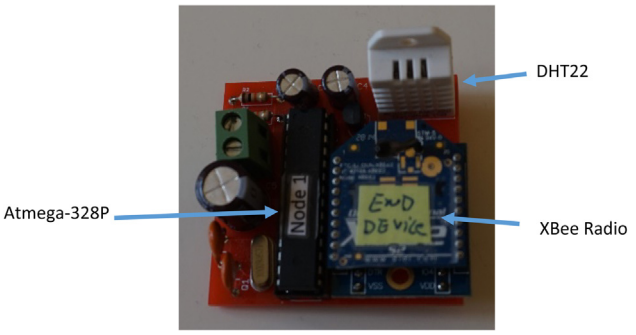


Fig. 4. Photograph of one the wireless sensors. The main components are: the Atmega-328P microcontroller, the DHT 22 sensor for temperature and humidity measurement, and the XBee radio.

Fig. 5. First floor. Temperature and Humidity sensors position. The blue circles indicate the sensor number. The coordinator (C) is positioned around the middle of the house, near the dining room table. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network [39] built with XBee radios [40], Atmega328P microcontrollers [41] and DHT-22 sensors (see Fig. 4). The digital DHT-22 sensors have an accuracy of ±0.5 °C for Temperature and ±3% for relative humidity. The microcontrollers were programmed using the Arduino IDE to read the data from the sensors and then transmit the measured data with the XBee radio. The transmitted information is sent to another XBee radio that acts as the coordinator of the network. Because the house is quite large and it is composed of thick walls and floors, it was necessary to include another two XBee radios that act as routers so there was effective communication from the end nodes to the coordinator. The sensor nodes are powered with batteries. Each sensor node transmits the information around 3.3 min. Figs. 5 and 6 display the location of the sensor nodes. For more information about working with XBee radios and Arduino see [42].

3. Recorded data and description

The energy (Wh) data logged every 10 min for the appliances is the focus of this analysis. The 10 min reporting interval was chosen to be able to capture quick changes in energy consumption.

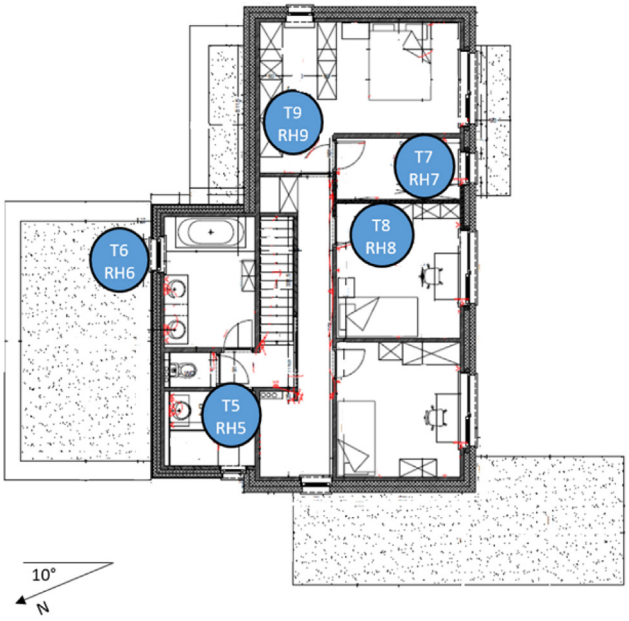


Fig. 6. Second floor. Location of the Temperature and Humidity sensors. The blue circles indicate the sensor number. Sensor node 6 measures the exterior conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Another sub-metered load (lights) is included in the analysis since it has been shown to be a good predictor of room occupancy when combined with relative humidity measurements [43]. The wireless sensor network's temperature and humidity recordings were averaged for the corresponding 10 min periods and merged with the energy data set by date and time. All the data analysis is done in R [44]. The time span of the data set is 137 days (4.5 months). Fig. 7 shows the energy consumption profile for the period. The energy consumption profile shows a high variability. Fig. 8 shows a histogram and boxplot of the data. As can be seen the data distribution has a long tail. In the boxplot, the median is represented with a thick black line inside the blue rectangle, and has a value of 60 Wh. The lower whisker has a value of 10 Wh and the upper whisker has a value of 170 Wh. It also shows that the data above the median is more dispersed and that there are several outliers (marked with the round circles above the upper whisker).

Although there is no weather station outside the house, weather data for the nearest airport weather station (Chièvres Airport, Belgium) is merged by date and time in this study to evaluate its impact on the prediction of the energy consumption of appliances [45]. The Chièvres Airport, Belgium is located about 12 km from the Stambruges house. Since the downloaded weather data is at hourly intervals, linear interpolation is used to have a complete data set (at 10 min intervals). A complete data set is often necessary for statistical regression models to work. Table 2 presents a list of all the variables or features. From the date/time variable other extra features are generated: the number of seconds from midnight for each day (NSM), the week status (weekend or workday) and the day of the week.

3.1. Data sets and exploratory analysis

The combined data set is split in training and test validation using CARET'S create data partition function. 75% of the data is used for the training of the models and the rest is used for testing (See Table 3).

Fig. 9 and Appendices A–C, present pair plots showing the relationships between all the variables with the energy consumption

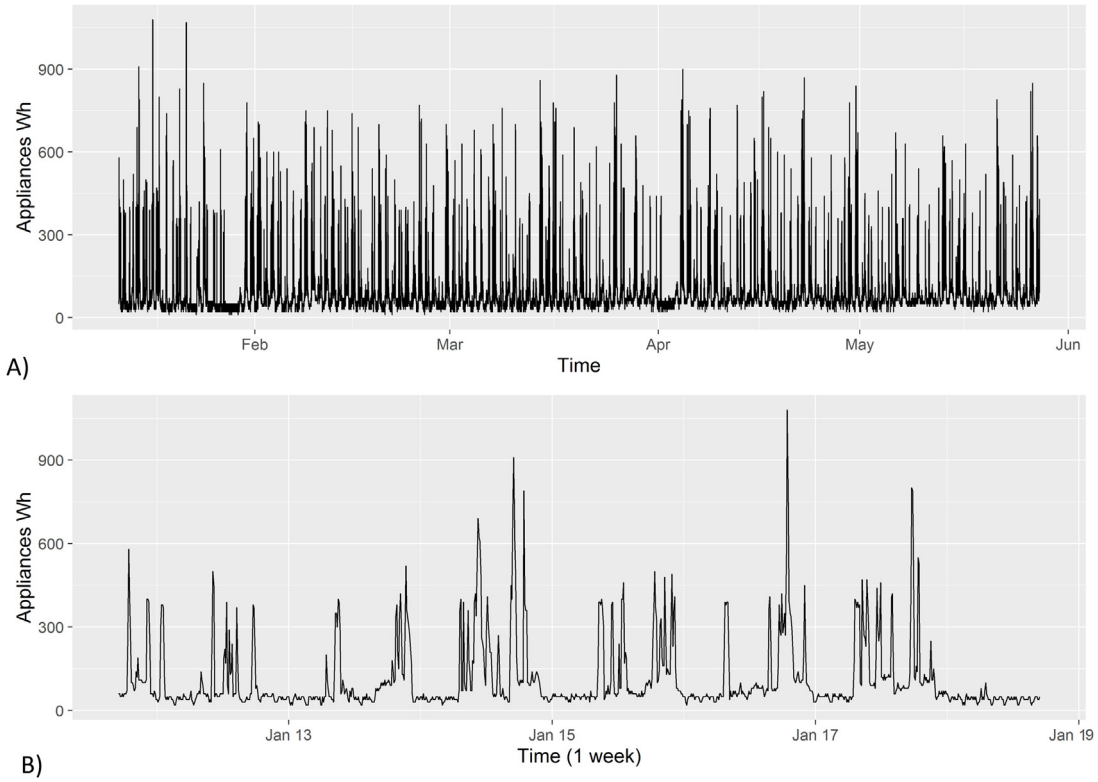


Fig. 7. (A) Appliances energy consumption measurement for the whole period, (B) A closer look at the first week of data.

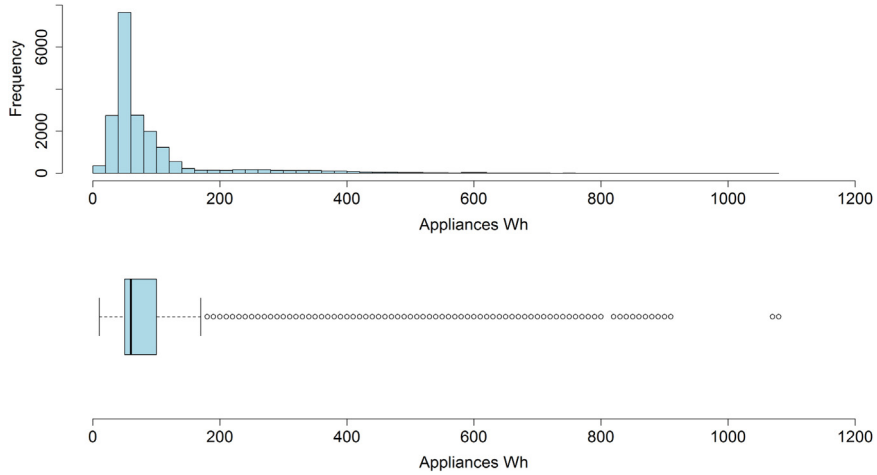


Fig. 8. Appliances energy consumption distribution. Top: histogram, bottom: boxplot. The histogram shows the frequency of energy consumption in the interval (bar width), and the boxplot shows the location of the median with the black line.

of appliances in the training set. These figures were created with the psych package [46]. These Figures show the bivariate scatter plots below the diagonal, histogram plots along the diagonal and the Pearson correlation above it, which is a measure of the linear dependence between two variables. A correlation of 1 is total positive correlation, −1 is total negative correlation and 0 represents no correlation. In red the linear regression fits are shown for each pair.

Fig. 9 shows that there is a positive correlation between the energy consumption of appliances and lights (0.19). The second largest correlation is between appliances and T2. For the indoor temperatures, the correlations are high as expected, since the ventilation is driven by the HRV unit and minimizes air temperature differences between rooms. For example, a positive correlation is found with T1 and T3. For Appendix A, the plot shows that the highest correlation with the appliances is between the outdoor temperature (0.12). There is also a negative correlation between the appliances and outdoor humidity/RH6 (−.09). Appendix B also shows positive correlations between the consumption of appliances and T7, T8 and T9 being 0.03, 0.05 and 0.02 respectively. Appendix C shows the highest correlation between the energy consumption of appliances and NSM with a value of 0.22. A positive correlation of 0.10 is seen between appliances' consumption and outdoor temperature (Tout) that is, the higher temperatures, the

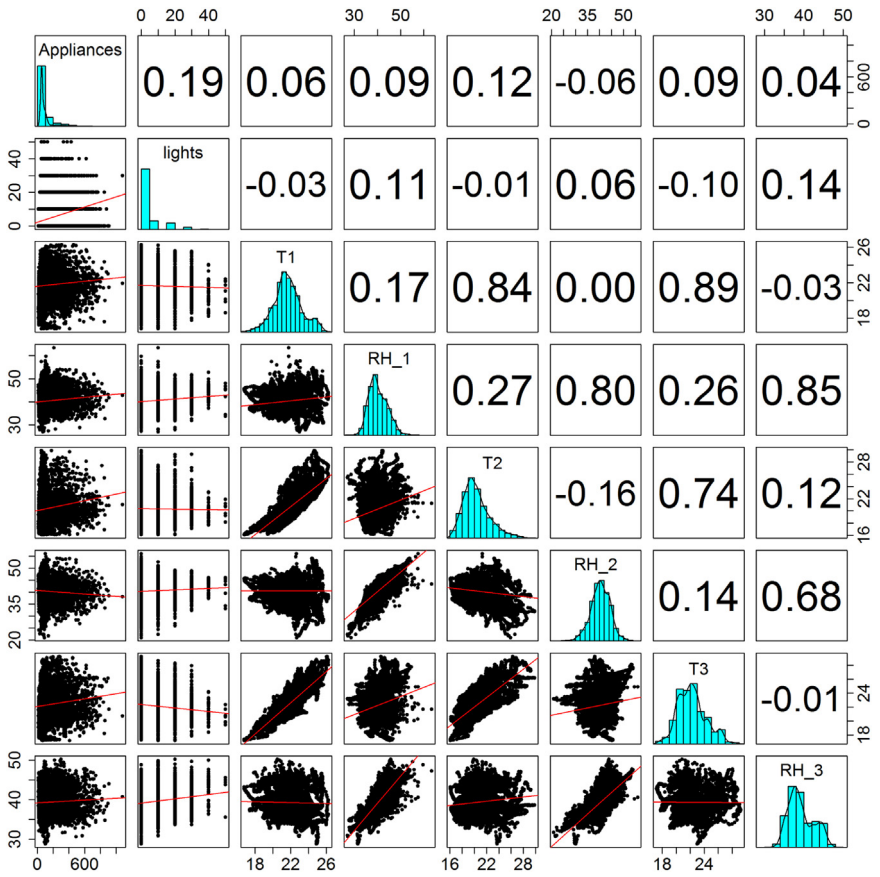


Fig. 9. Pairs plot. Relationship between the energy consumption of appliances with: lights, T1, RH1, T2, RH2, T3, RH3. T1 and RH1 correspond to the kitchen conditions; T2 and RH2 correspond to the living room conditions. (For interpretation of the references to color in text near the reference citation, the reader is referred to the web version of this article.)

higher the energy use by the appliances. Also there is a positive correlation with appliances' consumption and wind speed (0.09), higher wind speeds correlate with higher energy consumption by the appliances. A negative correlation of −0.15 was found with the RHout, and of −0.03 with pressure. Another important and interesting correlation is between the pressure and the wind speed. This relationship is negative (−0.23). The linear trend is with lower pressure the wind speed will be higher.

An hourly heat map was created for four consecutive weeks of data to identify any time trends (see Fig. 10). Fig. 10 was built with the following r packages: ggplot, gridExtra and ggthemes [47–49]. As can be clearly seen, there is a strong time component in the energy consumption pattern. The energy consumption starts to rise around 6 in the morning. Then around noon, there are energy load surges. The energy demand also increases around 6 pm. There is no clear pattern regarding the day of the week.

3.2. Data features filtering and importance

Since the dataset contains several features or parameters and considering that the airport weather station is not at the same location as the house, it is desirable to find out which parameters are the most important and which ones do not improve the prediction of the appliances' energy consumption. For this task the Boruta package [50] is used to select all the relevant variables. Several researchers have used this package for variable filtering [51–53]. To test the Boruta algorithm, two random variables were introduced in the data sets (See [54] for the use of this approach). Moreover, this feature or variable selection helps in model interpretability and reduces complexity of the model [55].

The Boruta package compares importance of attributes with importance of shadow attributes that are created by shuffling original ones [50]. As can be seen in Fig. 11, the Boruta algorithm is capable of detecting the two random variables (boxplots in red) that have no predicting power for the appliances' energy consumption. The two random variables are between the Boruta-created shadow attributes shown in blue: shadowMin, shadowMean and ShadowMax. The algorithm also ranks the variables in order of importance starting with the NSM variable, to the least important, the WeekStatus variable.

Although Fig. 11 provides a lot of insight, it does not tell us about the performance of the selected variables with respect to the RMSE. To test how many variables are optimal to minimize the RMSE the recursive feature elimination (RFE) is used to select the optimal inputs [56]. The Classification and Regression Training package (CARET) [57] has a RFE algorithm and is used in this study. CARET's RFE algorithm needs the factor variables to be cast as dummy variables, for which the r package dummies is used [58]. Then the variables Week status and day of the week are transformed into dummy variables. After doing this, the total number of predictors is 35. The RFE algorithm used the random forest regression method and was trained with 10 cross validation sets. Fig. 12 shows the result of the RFE algorithm. The optimal number of predictors is 34, but the difference between 34 and 35 can be considered negligible as is demonstrated in Fig. 12. Table 4 lists the feature ranking from the RFE.

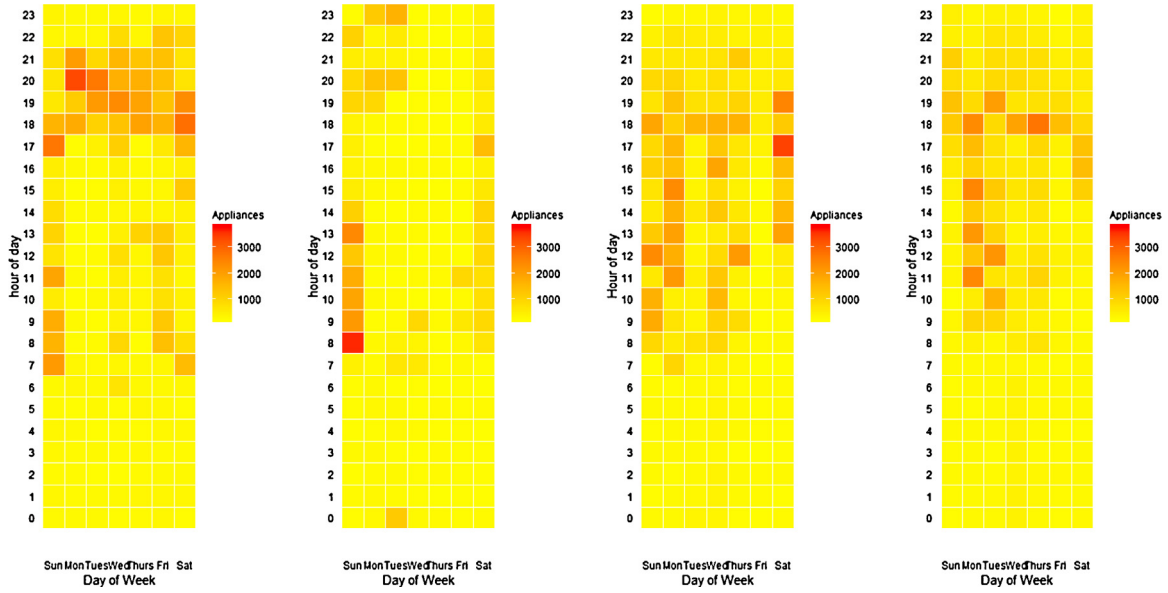


Fig. 10. Hourly energy consumption of appliances heat map for four consecutive weeks.

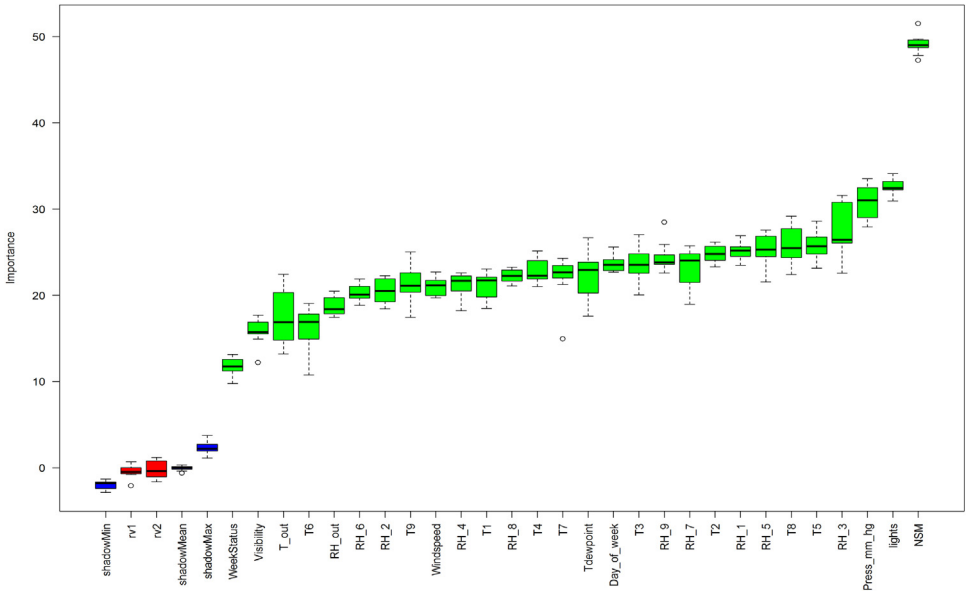


Fig. 11. Variable importance and selection from Boruta's algorithm. (For interpretation of the references to color in text near the reference citation, the reader is referred to the web version of this article.)

Since all the predictors can be considered relevant to minimize the RMSE, they all will be used to test four regression models (lm, SVM-radial, random forest and GBM).

3.3. The performance of regression models

All the regression models were trained with 10 fold cross validation to select the best. To speed up the computations the doParallel package was used [59] for parallel computation. The first model trained was the multiple linear regression. The multiple linear regression uses all the available predictors and finds the appropriate slope quantifying the effect of each predictor and the response [55].

Fig. 13 shows a residual plot for the linear regression model. The residuals were computed as the difference between the real values and the predicted values. From Fig. 13, it is obvious that the relationship between the variables and the energy consumption of appliances is not well represented by the linear model since the residuals are not normally distributed around the horizontal axis.

In order to compare the performance of each of the regression models, different performance evaluation indices are used here: the root mean squared error (RMSE), the coefficient of determination or R-squared/R^2, the mean absolute error (MAE) and the mean absolute percentage error (MAPE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_n^{i=1}(Y_i - \hat{Y}_i)^2}{n}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_n^{i=1}(Y_i - \hat{Y}_i)^2}{\sum_n^{i=1}(Y_i - \bar{Y})^2} \tag{2}$$

Table 2
Data variables and description.

Data variables	Units	Number of features
Appliances energy consumption	Wh	1
Light energy consumption	Wh	2
T1, Temperature in kitchen area	°C	3
RH1, Humidity in kitchen area	%	4
T2, Temperature in living room area	°C	5
RH2, Humidity in living room area	%	6
T3, Temperature in laundry room area	°C	7
RH3, Humidity in laundry room area	%	8
T4, Temperature in office room	°C	9
RH4, Humidity in office room	%	10
T5, Temperature in bathroom	°C	11
RH5, Humidity in bathroom	%	12
T6, Temperature outside the building (north side)	°C	13
RH6, Humidity outside the building (north side)	%	14
T7, Temperature in ironing room	°C	15
RH7, Humidity in ironing room	%	16
T8, Temperature in teenager room 2	°C	17
RH8, Humidity in teenager room 2	%	18
T9, Temperature in parents room	°C	19
RH9, Humidity in parents room	%	20
To, Temperature outside (from Chièvres weather station)	°C	21
Pressure (from Chièvres weather station)	mm Hg	22
RHo, Humidity outside (from Chièvres weather station)	%	23
Windspeed (from Chièvres weather station)	m/s	24
Visibility (from Chièvres weather station)	km	25
Tdewpoint (from Chièvres weather station)	°C	26
Random Variable 1 (RV_1)	Non dimensional	27
Random Variable 2 (RV_2)	Non dimensional	28
Number of seconds from midnight (NSM)	s	29
Week status (weekend (0) or a weekday (1))	Factor/categorical	30
Day of week (Monday, Tuesday... Sunday)	Factor/categorical	31
Date time stamp	year-month-day hour:min:s	–

Table 3
Training and testing data set.

Data set	Number of observations
Training	14,803 and 32 variables
Testing	4932 and 32 variables

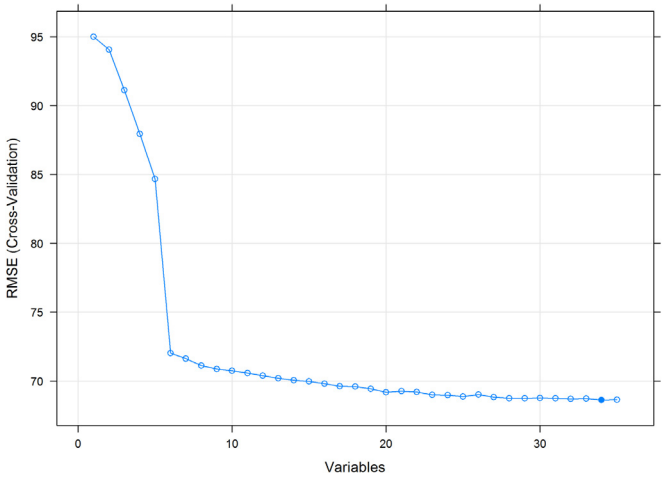


Fig. 12. RMSE using the RFE algorithm. The optimal number of predictors (34) is shown with the filled dot.

$$\mathrm{MAE} = \frac{\sum_{n}^{i=1} \left| Y_i - \hat{Y}_i \right|}{n} \tag{3}$$

$$\mathrm{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left| Y_i - \hat{Y}_i \right|}{Y_i} \tag{4}$$

where Y_i is the actual measurement (energy consumption), $\hat{Y}_i$ is the predicted value and n is the number of measurements.

Support vector machines have been used before to predict building energy consumption in [60]. This work presents a detailed description of the SVM model. As explained in [60], the support vector machine can use different kernels, and the radial basis function kernel has some numerical advantages and was used in that study. A SVM with radial kernel, SVM-radial, is employed in the present

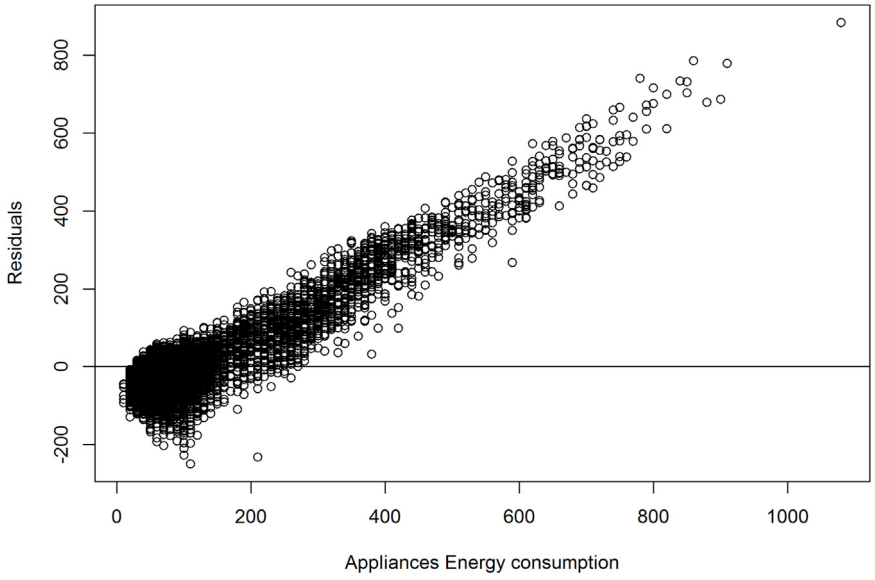


Fig. 13. Residuals and appliances' energy consumption plot of the lm model.

Table 4
Feature ranking from RFE algorithm.

1. NSM	6. RH3	11. RH9	16. RH2	21. WndSpd	26. T6	31. Dy_w_Sun
2. Lights	7. Tdwpt	12. RH7	17. T1	22. RH_8	27. Dy_w_Friday	32. Wstatus_Weekday
3.Press	8. T5	13. T7	18. T2	23. RH_6	28. Dy_w_Mond	33. Wstatus_Weekend
4.RH5	9. T8	14. Visibility	19. T9	24. RHout	29. Dy_w_Satrd	34.Dy_week_Tuesday
5.T3	10. RH1	15. RH4	20. T4	25. Tout	30. Dy_w_Wedn	35. Dy_week_Thursd.

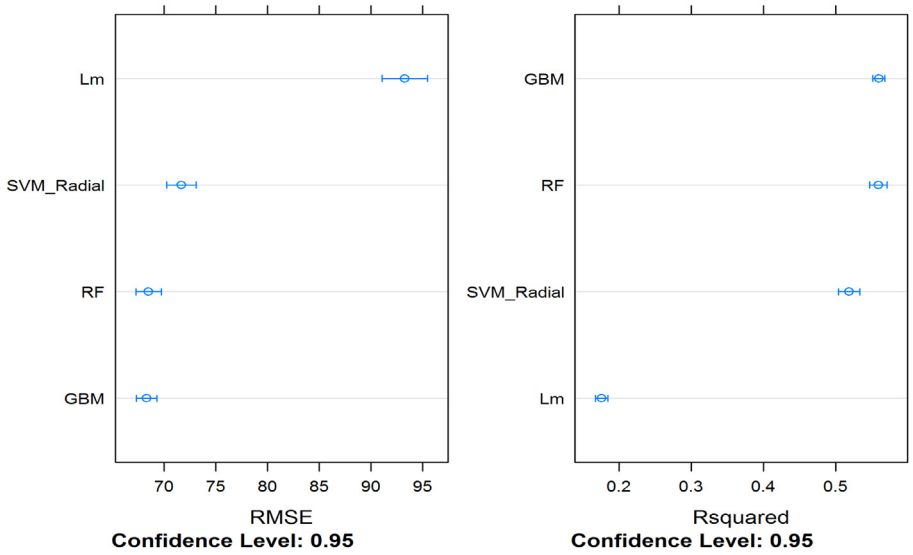


Fig. 14. Trained models comparison of RMSE and *R*-squared values.

research. The SVM-radial model requires two tuning parameters, sigma and cost variables, besides the predictors. The optimal values for sigma (0.4) and the cost (12) variables were obtained with a grid search (see Appendix D).

The random forest model is a Tree-based model [55]. The random forest model is based on the output of several regression trees. However, each tree is built with a random sample of selected predictors. The idea behind this is to decorrelate the trees and improve the prediction. The random forest model requires finding the optimum number of trees and the number of randomly selected predictors (see Appendix E). The RMSE does not appear to change after about 300 trees and the optimal number of random selected predictors is 18 as seen in Appendix E. The GBM models (also known as boosting) try to improve the prediction by using information from the first trees and also require the selection of optimal parameters for the number of trees (10,900) and maximum tree depth (5) (see Appendix F).

3.4. Model selection

After training the models, each model has 30 results from 10-fold cross validation (CV) sets and 3 repeats. This information was used in CARET to plot the RMSE for each model together with the confidence intervals as shown in Fig. 14. This figure also shows the *R*-squared (R^2) from the results and its confidence interval. The best models are the ones that provide the lower RMSE and highest R^2 values.

As can be seen the RF and GBM models have very similar performance based on their RMSE and *R*-squared values and confidence intervals. The SVM radial model also shows a significant reduction of the RMSE compared to the linear regression model.

Table 5 presents the performance of the trained models in the training and testing sets. As can be seen the RMSE of the models in the testing set are within the range suggested in Fig. 14 found during repeated CV.

Fig. 15 shows the relative variable importance for the RF, GBM and SVM-radial models. For the RF and GBM models, the variable importance is measured by the residual sum of squares. For the SVM-Radial model, the relationship between each predictor and the outcome is evaluated and then a linear model is fit and the absolute value of the t value for the slope of the predictor is used [57,61].

3.4.1. Evaluating the prediction with different data subsets

After finding out that the GBM model provided the best RMSE and R^2 in the previous analysis, this model was used to study the prediction performance with different predictors subsets: removing the light consumption, removing the light and no weather data, removing the temperature and humidity from the wireless sensor network and only using the weather and time information. See Table 6 and Fig. 16 for the considered predictors and the models' performance in the training and testing sets.

From Table 6, the performance of the GBM model without the lights predictor is quite accurate in comparison with the GBM model in Table 5. For this model the R^2 is 0.58 in the testing set. The performance for the GBM model without the lights and weather data is slightly less accurate since the R^2 in the testing set was reduced to 0.54. The third model in Table 6 that includes the weather data and the light predictor has an equivalent performance to the fourth model that only includes the weather data since their R^2 are the same (0.49).

3.5. Discussion of results

As seen in Fig. 3, the consumption of appliances represents the highest percentage of electrical consumption, between 70% and 79% of the monthly consumption. The appliances' consumption profile is highly variable as seen in Fig. 7, with periods of almost constant demand followed by high spikes. These results are similar to those

Table 5
Models performance.

Model	Parameters/features	Training				Testing			
		RMSE	R^2	MAE	MAPE %	RMSE	R^2	MAE	MAPE %
LM	Light, T1,RH1,T2,RH2,T3, RH3,T4, RH4,T5,RH5,T6, RH6, T7,RH7,T8,TH8,T9,RH9, To,Pressure,Rho,WindSpd, Tdewpoint, NSM, WeekStatus, Day of Week	93.21	0.18	53.13	61.32	93.18	0.16	51.97	59.93
SVM Radial	Light,T1,RH1,T2,RH2,T3,RH3, T4,RH4,T5,RH5,T6,RH6,T7,RH7,T8,TH8,T9,RH9,To, Pressure,Rho,WindSpeed, Tdewpoint,NSM, WeekStatus, Day of Week	39.35	0.85	15.08	15.60	70.74	0.52	31.36	29.76
GBM	Light,T1,RH1,T2,RH2,T3,RH3, T4,RH4,T5,RH5,T6,RH6, T7,RH7,T8,TH8,T9,RH9,To, Pressure,Rho,WindSpeed, Tdewpoint,NSM, WeekStatus, Day of Week	17.56	0.97	11.97	16.27	66.65	0.57	35.22	38.29
RF	Light,T1,RH1,T2,RH2,T3,RH3, T4,RH4,T5,RH5,T6,RH6,T7, RH7,T8,TH8,T9,RH9,To, Pressure,Rho,WindSpeed, Tdewpoint,NSM, WeekStatus, Day of Week	29.61	0.92	13.75	13.43	68.48	0.54	31.85	31.39

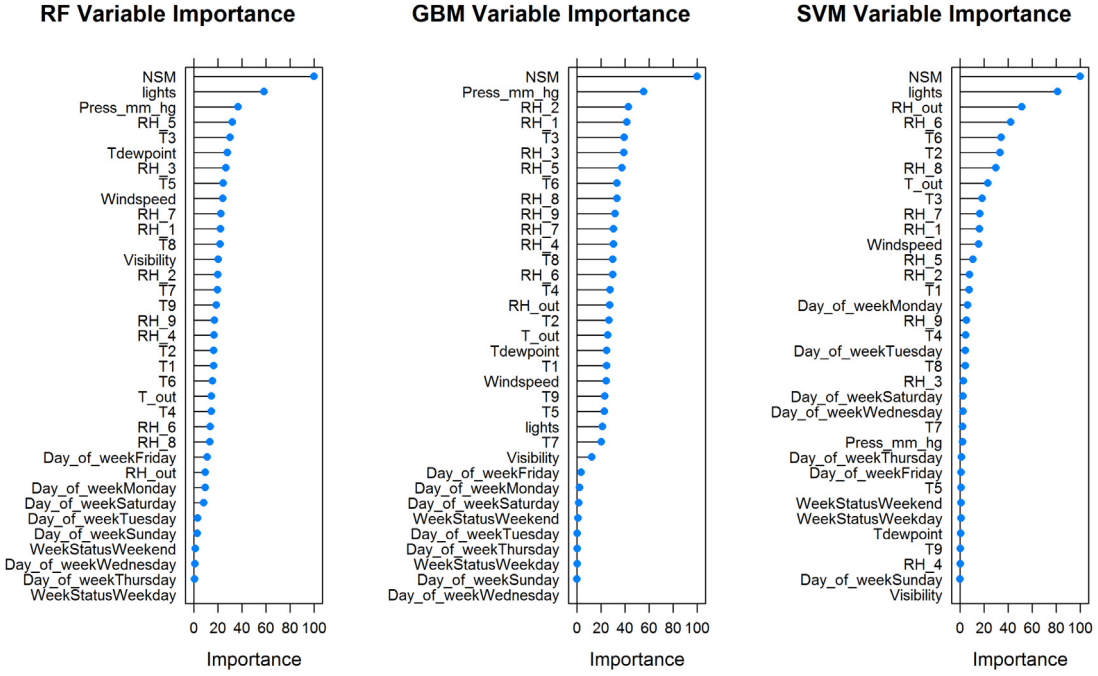


Fig. 15. Variable importance for RF, GBM and SVM radial models.

Table 6
Models performance with different subsets.

Model	Parameters/features	RMSE training				RMSE testing			
		RMSE	R^2	MAE	MAPE %	RMSE	R^2	MAE	MAPE %
GBM – no lights	T1,RH1,T2,RH2,T3,RH3, T4,RH4,T5,RH5,T6,RH6, T7,RH7,T8,TH8,T9,RH9,To, Pressure,Rho,WindSpeed, Tdewpoint,NSM, WeekStatus, Day of Week	17.90	0.97	12.24	16.66	66.21	0.58	35.24	38.65
GBM – no lights and no weather data	T1,RH1,T2,RH2,T3,RH3, T4,RH4,T5,RH5,T6,RH6, T7,RH7,T8,TH8,T9,RH9, NSM, WeekStatus, Day of Week	18.83	0.97	12.85	17.44	68.59	0.54	36.21	39.23
GBM – no Temp and humidity inside house	Light, To,Pressure,Rho, WindSpeed,Tdewpoint,NSM, WeekStatus, Day of Week	27.47	0.93	18.29	23.71	72.64	0.49	40.32	45.33
GBM – only weather and time information	To,Pressure,Rho,WindSpeed, Tdewpoint,NSM, WeekStatus, Day of Week	28.29	0.92	18.85	24.41	72.45	0.49	40.73	46.53

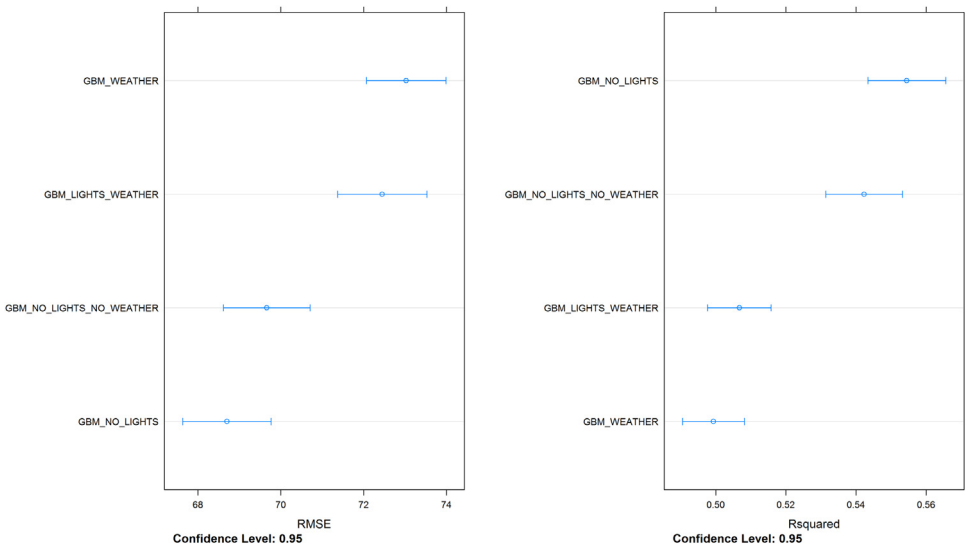


Fig. 16. Trained GBM models comparison of RMSE and *R*-squared values.

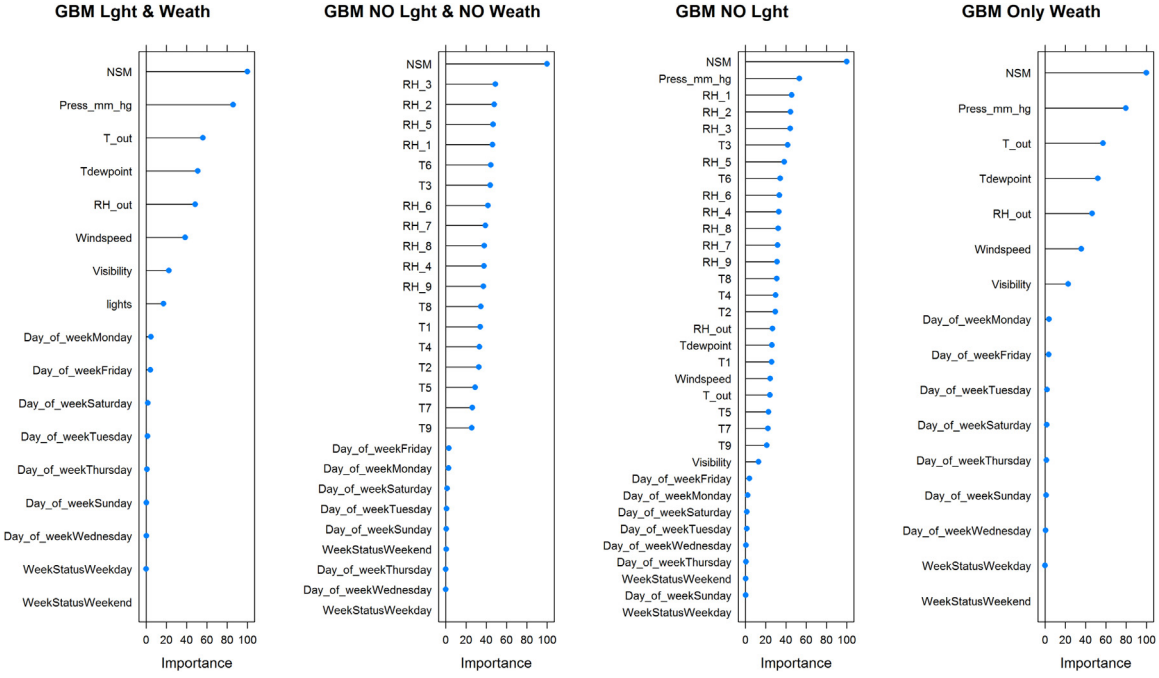


Fig. 17. Variable importance for the different GBM data subsets models.

presented in [9,20]. The boxplot shown in Fig. 8 shows that the data above the median is highly dispersed and skewed.

When doing the exploratory data analysis some paired correlations found were interesting. Fig. 9, Appendices A and B show that high correlations exist between T1 and T2 (0.84), T2 and T3 (0.74), T4 and T5 (0.87), T5 and T6 (0.63), T7 and T8 (0.88), and T8 with T9 (0.86). This is important because it indicates that RC thermal networks are suited for building energy modeling and that measuring the most representative rooms may be representative enough for the energy prediction of appliances. Also it indicates that a one zone thermal model might be adequate for modeling. For RC thermal models, system identification techniques could be used for finding the values of the resistances and capacitances.

The relationships between the humidity in each room and the appliances' consumption are quite telling. From research in [43], the humidity of the room (also the humidity ratio) increased when occupants arrived in the office room and started to drop when the office was empty. Therefore, higher humidity may indicate occupancy.

RH1 (kitchen area) and Appliances are positively correlated (0.06). This could be explained by higher humidity levels in the zone when cooking and human presence. There is a positive correlation between appliances' consumption and RH3 (laundry room humidity) of 0.04. This may mean that the washing machine and/or the dryer are/is in use. There were also small positive correlations of 0.02 and 0.01 between appliances' consumption with RH4 (downstairs office) and RH5 (upstairs bathroom) respectively. A different behavior, negative correlation of −0.06 was found between appliances' consumption and the humidity in the living room (RH2). This may suggest that the energy consumption in this zone is lower with reference to the kitchen and laundry room for example. For

the three bedrooms upstairs the correlations were also negative. Between appliances' consumption and RH7 the correlation was −0.06, between appliances and RH8 it was -0.09, and between appliances and RH9 it was −0.05. An explanation for these results may be that although humidity increases with human presence, when bedrooms are occupied, the devices used in these rooms have lower power demands than the ones in the zones with positive correlations. It can also indicate that the occupants are sleeping and therefore not consuming as much energy in the house.

A small negative correlation (−0.03) was found between the appliances' consumption with atmospheric pressure (see Appendix C). Also a positive correlation of 0.09 between appliances' consumption and wind speed was found. Between the pressure and the wind speed the correlation is negative (−0.23), meaning that lower pressure corresponds with higher wind speeds. Maybe the reason for the relationships between appliances' consumption, pressure and wind speed are related to occupancy changes due to weather, since building occupancy has a strong relationship with energy consumption [62–64]. Fair weather usually corresponds with high atmospheric pressure and pressure drop trends or low pressure may correspond to foul or rainy weather [65]. Wind speed also has been found to correlate with window opening behavior of occupants in buildings [66]. It is possible that the average linear trend of higher consumption of appliances with lower pressure and with higher wind speeds are an indication that the occupants tend to remain indoors in periods of bad weather. Also in results presented by [67] there was a correlation between Cooling degree minutes (CDM) and atmospheric pressure and also a correlation was found for Heating degree minutes (HDM) and atmospheric pressure.

The data filtering shown in Section 3.2 is relevant since it helps to diminish the number of features and predictors that have no effect on the accuracy of the prediction. The Boruta algorithm was able to pick up the inserted two random variables in the data set. It also showed that the weather parameters from the nearby weather station are relevant in the prediction problem.

According to the RFE algorithm, it can be seen in Fig. 12 that six parameters can significantly reduce the RMSE. These parameters are: NSM, lights, pressure, RH5, T3 and RH3 (see Table 4). The least important predictors are the ones related with week status and day of the week.

The plots with the RMSE and R^2 with the associated confidence levels have been proven useful in predicting the performance of each model in the testing set. The best models are the RF and GBM according to the RMSE and R^2 (See Fig. 14). In the training set the GBM model had an R^2 of 0.97 and the RF and an R^2 of 0.92. Regarding the variable importance in Fig. 15, it is interesting to see that for the three models the NSM was the most important predictor. The RF and the SVM models picked the lights as the second most important predictor while for the GBM model it was the atmospheric pressure. The data from the wireless sensor network was ranked highly in the GBM model, especially with information from the living room (RH2), the kitchen (RH1), the laundry room (T3) and bathroom (RH5) in the top positions. Since the wireless data is highly correlated with the other predictors, it is better to study this data subset separately to have a better appreciation for the different ranks. In Table 6 it can be seen that the GBM model without light information is as accurate as the one including light in Table 5 with an R^2 of 0.58 in the testing set. The model only leveraged in the wireless sensor data (without lights and weather information), has a smaller R^2 in the testing set (0.54). The last two models that consider the light and only the exterior weather data in Table 6, have almost the same performance 0.93 R^2 in one and 0.92 in the training sets, with the same R^2 in the testing set (0.49).

When looking at the ranking for the GBM model with no light information (see Fig. 17), it can be seen that the top predictors are the NSM, Pressure, RH1, RH2, RH3, RH5, T6, RH6, RH4, RH9, T8, T4, and T2. This means that information from the kitchen, living room, laundry room, bathroom, outdoors, office, and bedrooms are the most important.

3.6. Research limitations

One of the main limitations of this study is that the analysis was done for only one house. Important information could be found when analyzing several houses, and other relationships can be studied with appliances' energy consumption in combination with: occupant's age, number of occupants, ownership of pets, building's geometry etc. Another research limitation is the length of continuous analyzed data. Different energy use patterns can potentially be found depending on the season of the year. Regarding the weather station, the predictions of appliances energy use could probably be better if the weather station was closer to the house. This research has not looked into the problem of optimal location of the wireless indoor sensors for improvement of the energy prediction. It is also possible that more sensors and better sensor accuracy could help to improve the energy prediction.

4. Conclusion

The statistical data analysis has shown thought-provoking results in both the exploratory analysis and in prediction models. The pairwise plots are useful because they shed light on the different relationships between parameters that could be hidden in final predictive models. The GBM and RF models improve the RSME and R^2 of predictions compared to the SVM-radial and multiple linear regression lm. For all the models, the time information (NSM) was ranked as the most important to predict the appliances' consumption.

The weather data from the nearby weather station was shown to increase the prediction accuracy in the GBM models. The GBM models with only weather data ranked the pressure as the most important weather variable, followed by the outdoor temperature, dew point temperature, outdoor relative humidity, wind speed and visibility. The possible explanation for why the pressure has a strong prediction power may be related to its influence on the wind speed and higher rainfall probability which could potentially increase the occupancy of the house. Research by [67] found that atmospheric pressure is highly correlated with the cooling degree minutes (CDM) and heating degree minutes (HDM). Also, pressure has direct effects on air humidity ratio, density and enthalpy.

Data from a wireless sensor network that measures humidity and temperature has been proven to increase the prediction accuracy. The data analysis showed that data from the kitchen, laundry room, living room and bathrooms had the most important contributions. Data from the other rooms also helps in the prediction. When looking at the appliances in each room in Table 1, it can be seen that the laundry, kitchen and living rooms would be expected to have the highest contributions because of the equipment present (see also Fig. 1). The prediction of appliances' consumption with data from the wireless network indicates that it can help to locate where in a building the main appliances' energy consumption contributions are found.

When using all the predictors the light consumption was ranked highly. However, when studying different predictor subsets, removing the light consumption appeared not to have a significant impact. This may be an indication that other features are correlated well with the light energy consumption.

This study has found curious relationships between variables. Future work could include considering weather data such as solar radiation and precipitation. This has also been recommended in

[62]. Also occupancy and occupant's activity information could be useful to improve the prediction and find its relationship with other parameters (exterior weather for example). The wireless sensors could also measure CO_2 and noise to help in the prediction and to track the occupant's movement from room to room and time spent in each room.

In order to allow for reproducibility of the presented results, and for fellow scientist and researchers to test their regression models, the data and the processing scripts will be made available in the following public repository: https://github.com/LuisM78/Appliances-energy-prediction-data.

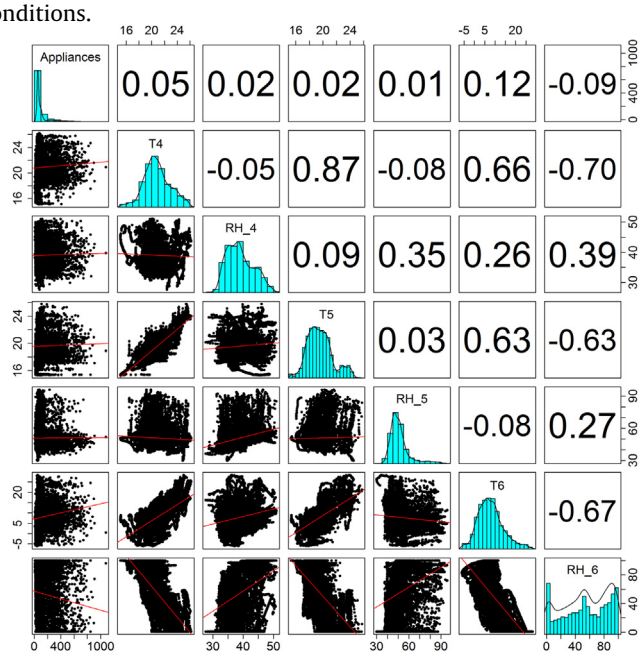
Acknowledgments

This work has received funding from the European Union's Seventh Program for research, technological development and demonstration under grant agreement no. 285173 — NEED4B "New Energy Efficient Demonstration for Buildings".

We would like to express our appreciation and thanks to Marcel Rustin and Christophe Coetsier for their help in the assembly and customization of the temperature and humidity wireless sensors.

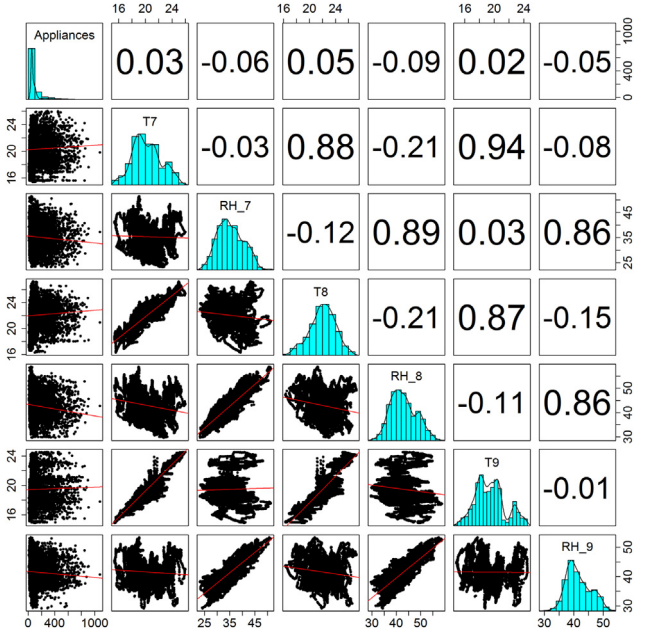
Appendix A.

Pairs plot showing relationships between the energy consumption of appliances with: T4, RH4, T5, RH5, T6, RH6. T4 and RH4 correspond to the office conditions; T5 and RH5 correspond to the bathroom conditions, and T6 and RH6 correspond to the outdoor conditions.



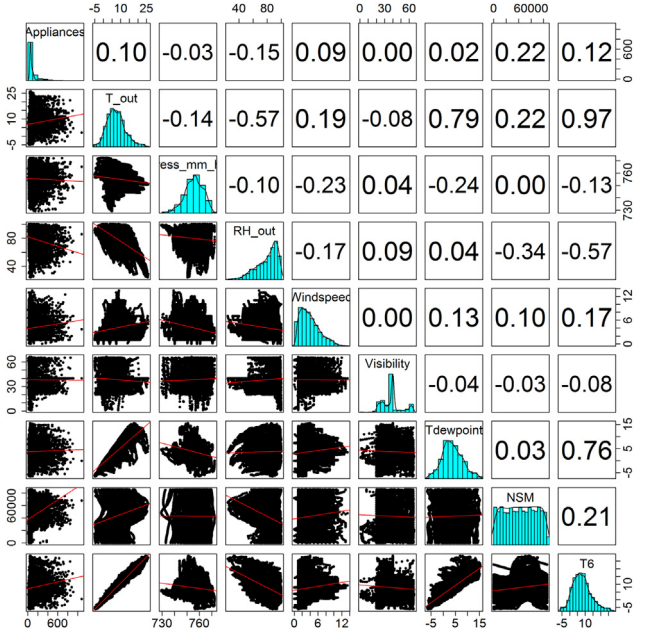
Appendix B.

Pairs plot showing relationships between the energy consumption of appliances with: T7, RH7,T8, RH8, T9,RH9. T7 and RH7 correspond to the ironing room conditions, T8 and RH8 correspond to room 2, T9 and RH9 correspond to the main room.



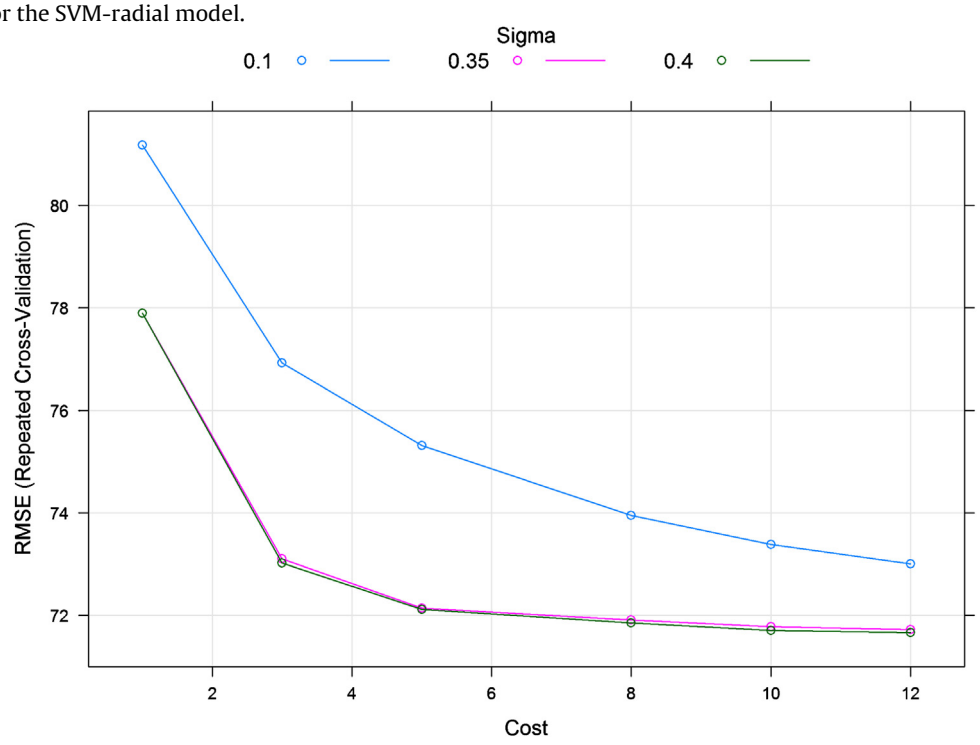
Appendix C.

Pairs plot showing relationships between the energy consumption of appliances with: T_out, Pressure, RH_out, Windspeed, Visibility, TDewpoint, NSM and T6.



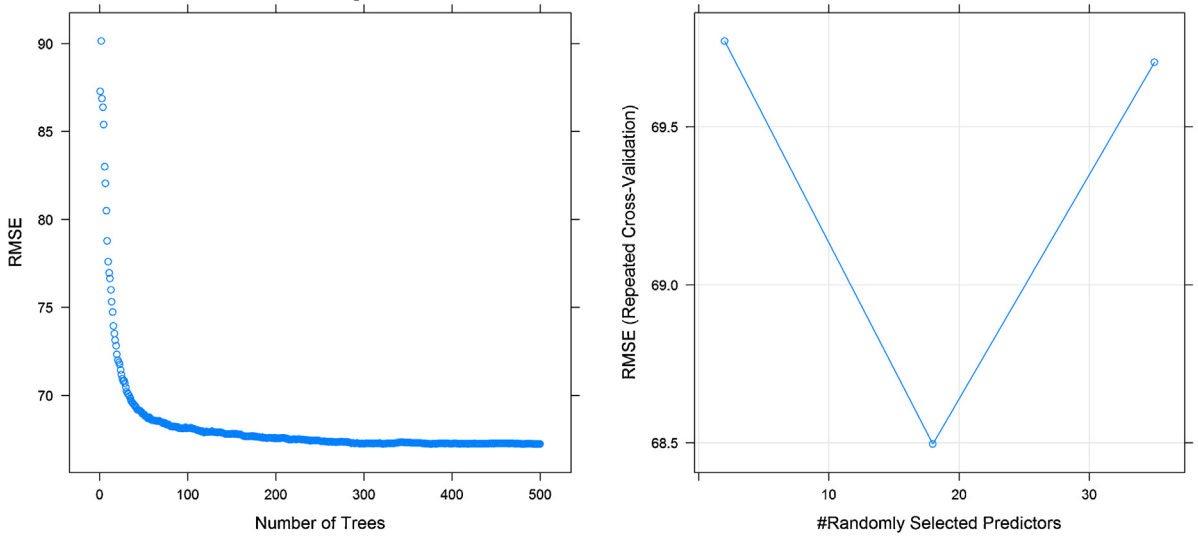
Appendix D.

Grid search results for optimal values of sigma and cost values for the SVM-radial model.



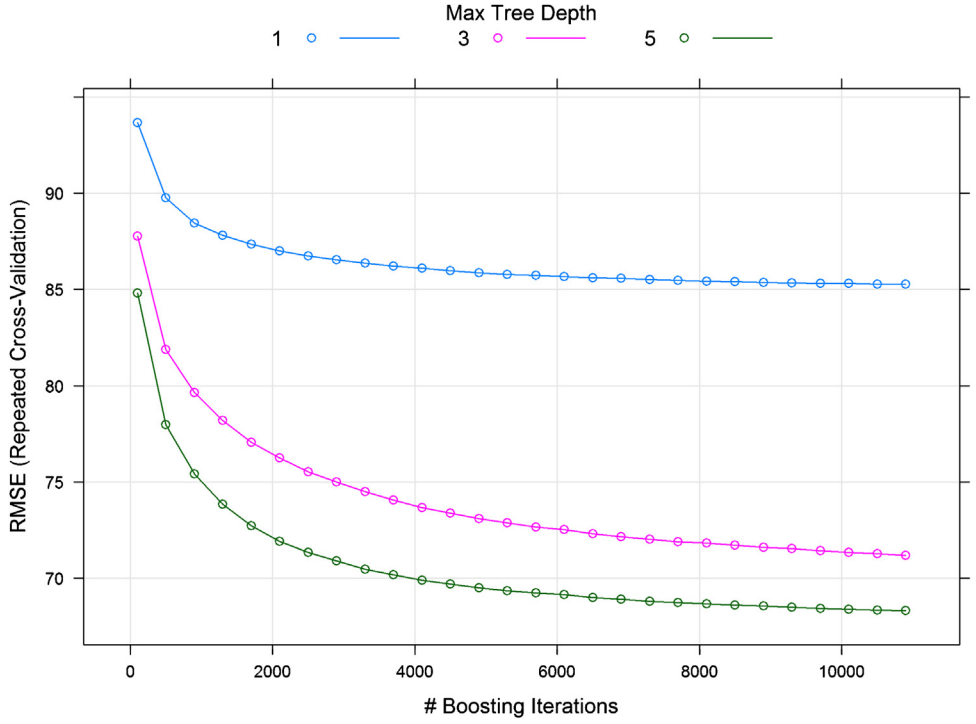
Appendix E.

Random forest model with all the parameters.



Appendix F.

Grid search results for finding the optimal number of trees (boosting iterations) and maximum tree depth for the GBM-model.



References

[1] A. Barbato, A. Capone, M. Rodolfi, D. Tagliaferri, Forecasting the usage of household appliances through power meter sensors for demand management in the smart grid, in: IEEE International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2011, pp. 404–409.

[2] M. Ruellan, H. Park, R. Bennacer, Residential building energy demand and thermal comfort: Thermal dynamics of electrical appliances and their impact, Energy Build (2016) 46–54.

[3] N. Arghira, L. Hawarah, S. Ploix, M. Jacomino, Prediction of appliances energy use in smart homes, Energy 48 (1) (2012) 128–134.

[4] S. Firth, K. Lomas, A. Wright, R. Wall, Identifying trends in the use of domestic appliances from household electricity consumption measurements, Energy Build. 40 (5) (2008) 926–936.

[5] R.V. Jones, K.J. Lomas, Determinants of high electrical energy demand in UK homes: appliance ownership and use, Energy Build. 117 (2016) 71–82.

[6] K.S. Cetin, Characterizing large residential appliance peak load reduction potential utilizing a probabilistic approach, Sci. Technol. Built Environ. 22 (6) (2016) 720–732.

[7] A. Kavousian, R. Rajagopal, M. Fischer, Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings, Energy Build. 99 (2015) 220–230.

[8] K. Basu, L. Hawarah, N. Arghira, H. Joumaa, S. Ploix, A prediction system for home appliance usage, Energy Build. 67 (2013) 668–679.

[9] K.S. Cetin, P.C. Tabares-Velasco, A. Novoselac, Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use, Energy Build. 84 (2014) 716–726.

[10] F. Spertino, P. Di Leo, V. Cocina, Which are the constraints to the photovoltaic grid-parity in the main European markets? Solar Energy 105 (2014) 390–400.

[11] J.E. Seem, Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy Build. 39 (1) (2007) 52–58.

[12] P. Zhao, S. Suryanarayanan, M.G. Simoes, An energy management system for building structures using a multi-agent decision-making control methodology, IEEE Trans. Ind. Appl. 49 (1) (2013) 322–330.

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

[14] J.A. Candanedo, V.R. Dehkordi, M. Stylianou, Model-based predictive control of an ice storage device in a building cooling system, Appl. Energy 111 (2013) 1032–1045.

[15] S. Mitchell, R. Sarhadian, S. Guow, B. Coburn, J. Lutton, I. Christi, D. Rauss, C. Haiad, Residential appliance demand response testing, in: ACEEE Summer Study on Energy Efficient Buildings, Pacific Grove, CA, 2014.

[16] R. D'hulst, W. Labeeuw, B. Beusen, S. Claessens, G. Deconinck, K. Vanthournout, Demand response flexibility and flexibility potential of residential smart appliances: experiences from large pilot test in Belgium, Appl. Energy 155 (2015) 79–90.

[17] G. Johnson, I. Beausoleil-Morrison, Electrical-end-use data from 23 houses sampled each minute for simulating micro-generation systems, Appl. Therm. Eng. (2016).

[18] M. Muratori, M.C. Roberts, R. Sioshansi, V. Marano, G. Rizzoni, A highly resolved modeling technique to simulate residential power demand, Appl. Energy 107 (2013) 465–473.

[19] U.E.I. Association, Annual energy outlook, US Department of Energy, Washington, DC, 2012, Retrieved from http://www. eia. gov, 2012.

[20] R.G. Pratt, C.C. Conner, B.A. Cooke, E.E. Richman, Metered end-use consumption and load shapes from the ELCAP residential sample of existing homes in the Pacific Northwest, Energy Build. 19 (3) (1993) 179–193.

[21] W.F. Sandusky, E.W. Pearson, N.E. Miller, R.S. Crowder, G.B. Parker, R.P. Mazzucchi, G.M. Stokes, J.J. Thomas, R.G. Pratt, G.J. Schuster, M.A. Halverson, J.L. Stoops, F.J. Peterson, R.A. Gillman, R.A. Stokes, S.G. Hauser, ELCAP operational experience, Energy Build. 19 (3) (1993) 167–178.

[22] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (8) (2008) 1560–1566.

[23] N. Fumo, P. Mago, R. Luck, Methodology to estimate building energy consumption using EnergyPlus Benchmark Models, Energy Build. 42 (12) (2010) 2331–2337.

[24] P. Torcellini, M. Deru, B. Griffith, K. Benne, M. Halverson, D. Winiarski, D. Crawley, Proceeding of DOE commercial building benchmark models, 2008, pp. 17–22.

[25] M. Ghorbani, M.S. Rad, H. Mokhtari, M. Honarmand, M. Youhannaie, Residential loads modeling by norton equivalent model of household loads, in: Power and Energy Engineering Conference (APPEEC), Asia-Pacific, IEEE, 2011, pp. 1–4.

[26] Z. Guo, Z.J. Wang, A. Kashani, Home appliance load modeling from aggregated smart meter data, IEEE Trans. Power Syst. 30 (1) (2015) 254–262.

[27] R.V. Jones, A. Fuertes, K.J. Lomas, The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings, Renew. Sustain. Energy Rev. 43 (2015) 901–917.

[28] S.-H. Ling, F.H. Leung, H. Lam, P.K. Tam, Short-term electric load forecasting based on a neural fuzzy network, IEEE Trans. Ind. Electron. 50 (6) (2003) 1305–1316.

[29] A. Veit, C. Goebel, R. Tidke, C. Doblander, H.-A. Jacobsen, Household electricity demand forecasting: benchmarking state-of-the-art methods, in: Proceedings of the 5th international conference on Future energy systems, ACM, 2014, pp. 233–234.

[30] H.-x. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (6) (2012) 3586–3592.
[31] F. McLoughlin, A. Duffy, M. Conlon, Evaluation of time series techniques to characterise domestic electricity demand, Energy 50 (2013) 120–130.
[32] C. Sandels, J. Widén, L. Nordström, E. Andersson, Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data, Energy Build. 108 (2015) 279–290.
[33] P. Bacher, H. Madsen, H.A. Nielsen, B. Perers, Short-term heat load forecasting for single family houses, Energy Build. 65 (2013) 101–112.
[34] C.-L. Hor, S.J. Watson, S. Majithia, Analyzing the impact of weather variables on monthly electricity demand, IEEE Trans. Power Syst. 20 (4) (2005) 2078–2085.
[35] N. Saldanha, I. Beausoleil-Morrison, Measured end-use electric load profiles for 12 Canadian houses at high temporal resolution, Energy Build. 49 (2012) 519–530.
[36] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194.
[37] W. Feist, R. Pfluger, B. Kaufmann, J. Schnieders, O. Kah, Passive House Planning Package, Passive House Institute, Darmstadt, 2007.
[38] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (3) (2008) 394–398.
[39] Z. Alliance, Zigbee Specification, 2006.
[40] X. Digi, XBee-PRO® RF Modules, Digi International Inc., Minnesota, 2009.
[41] Atmel Corporation, ATmega328, 2016.
[42] R. Faludi, Building wireless sensor networks: with ZigBee, XBee, arduino, and processing, O'Reilly Media, Inc., 2010.
[43] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39.
[44] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2014.
[45] rp5.ru, Reliable Prognosis, 2016.
[46] W. Revelle, psych: Procedures for Psychological, Psychometric, and Personality Research, Northwestern University, 2016.
[47] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.
[48] A.B. Jeffrey, ggthemes: Extra Themes, Scales and Geoms for 'ggplot2', 2016.
[49] B. Auguie, gridExtra: Miscellaneous Functions for "Grid" Graphics, 2015.
[50] M.B. Kursa, W.R. Rudnicki, Feature selection with the boruta package, J. Stat. Softw. 36 (11) (2010) 1–13.
[51] T.-T. Nguyen, J.Z. Huang, T.T. Nguyen, Two-level quantile regression forests for bias correction in range prediction, Mach. Learn. 101 (1–3) (2015) 325–343.
[52] S. Strempel, M. Nendza, M. Scheringer, K. Hungerbühler, Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals, Environ. Toxicol. Chem. 32 (5) (2013) 1187–1195.
[53] M. Belgiu, L. Drăguţ, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogram. Rem. Sens. 114 (2016) 24–31.
[54] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567.
[55] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.
[56] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy 127 (2014) 1–10.
[57] M. Kuhn, caret: Classification and Regression Training, 2015.
[58] C. Brown, dummies: Create dummy/indicator variables flexibly and efficiently, 2012.
[59] Revolution Analytics, S. Weston, doParallel: Foreach Parallel Adaptor for the 'parallel' Package, 2015.
[60] B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. 37 (5) (2005) 545–553.
[61] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, 2013.
[62] T. Hong, S.C. Taylor-Lange, S. D'Oca, D. Yan, S.P. Corgnati, Advances in research and applications of energy-related occupant behavior in buildings, Energy Build. 116 (2016) 694–702.
[63] X. Liang, T. Hong, G.Q. Shen, Improving the accuracy of energy baseline models for commercial buildings with occupancy data, Appl. Energy 179 (2016) 247–260.
[64] D. Yan, W. O'Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: Current state and future challenges, Energy Build. 107 (2015) 264–278.
[65] F.K. Lutgens, E.J. Tarbuck, D. Tusa, The Atmosphere, Prentice-Hall, 1995.
[66] Y. Zhang, P. Barrett, Factors influencing the occupants' window opening behaviour in a naturally ventilated office building, Build. Environ. 50 (2012) 125–134.
[67] M.G. Fikru, L. Gautier, The impact of weather variation on energy consumption in residential houses, Appl. Energy 144 (2015) 19–30.