Building and Environment 138 (2018) 250–261

Contents lists available at ScienceDirect

Building and Environment

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

Reconstruction of the indoor temperature dataset of a house using data driven models for performance evaluation

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

Keywords:
Learning curves
Multiple linear regression
Random forest
Passive house
Temperatures
Sample size

ABSTRACT

Whenever the long term monitoring of a building is attempted it is likely that specific sensors or the whole monitoring system used may experience long-term failure therefore creating important gaps in one or more variables of special interest. These long gaps may not be addressed using simple linear interpolation. The option of only using the available data for descriptive statistics would produce results that are biased towards the season of measurement. In addition discarding the incomplete data represents a significant waste of time and effort in the research study. A work around to reduce the bias problem is to predict the missing data from other measured variables using machine-learning techniques. Some questions that follow are: How much data is necessary to be able to train a regression model? What is the expected error of such prediction? What is the best model for such a task? This paper addresses the problem of completing a data set for the interior temperatures inside a passive house using different monitored predictors such as exterior temperature, humidity, wind speed, visibility, pressure and electrical energy use inside the building. Two regression models, multiple linear regression and random forest are compared using learning curves for the training and testing sets for visualizing the so-called bias-variance trade off. The learning curves help to answer the question of optimal sample size for training, model selection and expected error. Finally, descriptive statistics such as median, maximum, minimum, and room temperature averages are presented before and after completing the data sets.

1. Introduction

Data loggers fail often during the time of monitoring building performance data and this situation usually leaves the engineers and research scientists involved in a difficult position for answering research questions. It would be possible to avoid such a problem by implementing some system redundancy (multiple loggers monitoring the same parameters) in order to prevent data loss during failures, but this would come at a significantly higher cost of equipment and maintenance. However, for this strategy to work, the loggers would have to be independent of one another. This would not be the case if they were to be powered by the same source. Also having redundancy does not necessarily increase reliability for different reasons e.g. complexity, human neglect [1].

An alternative to deal with the occurrence of missing data is to use machine-learning techniques in order to create models to predict the variable of interest from other available variables. In this work, we use learning curves to visualize and understand the effect of sample size on the training and testing set's root mean square errors. These types of learning curves have been used many times to appreciate the effect of sample size on the prediction error of the variable of interest [2–5].

In this paper, we investigate the interior temperatures inside a passive house. The temperatures were monitored using a wireless sensor network built with ZigBee radios and Arduino microcontrollers. The network architecture is such that a concentrator node receives all the incoming information at an Arduino base station that processes and records the information onto an SD card. The base station also acts as a server to access the data through the internet. This base station may experience data logging failure due to: power loss, a writing error to the memory card or malfunction of the ZigBee coordinator radio.

Since this work mostly deals with monitoring of temperatures inside a passive building and the use of data driven regression models for prediction, these are the main interests of the literature review. As a reminder to the reader, regression models estimate relationships among different variables for prediction of a continuous variable. For related research about prediction of building energy consumption regression models, building energy model calibration and artificial neural networks the reader is referred to [6–16]. A more recent publication dealing with the problem with imputation of missing values in building sensor data is presented by Chong, Lam et al. [17].

* Corresponding author.
E-mail addresses: Luismiguel.candanedoibarra@umons.ac.be, luismiguel_78@hotmail.com (L.M. Candanedo).

https://doi.org/10.1016/j.buildenv.2018.04.035
Received 11 January 2018; Received in revised form 16 April 2018; Accepted 26 April 2018
Available online 30 April 2018

1.1. Literature review

1.1.1. Regression models for indoor air temperature

Artificial neural networks ANNs were used for modeling the internal temperature of a building by Gouda, Danaher et al. [18]. The model employed outdoor temperature, solar irradiance, heating valve position, and the interior temperature. The research also evaluated the impact of hidden neurons on the mean squared error in training, testing and validation data sets.

The simulated indoor greenhouse temperature was modeled using neural networks by Uchida Frausto and Pieters [19]. The data set included outside air temperature, outdoor relative humidity, global solar radiation, and cloudiness of the sky. The research also investigated the optimal number of neurons.

Linear and neural network models to predict indoor temperature of a building were presented by Mechaqrane and Zouak [20]. The neural network model was able to outperform the linear model significantly. The model considered hourly values for variables for the indoor temperature, outdoor temperature, the solar radiation, and the auxiliary heating power.

A comparison of a multimode physical model and data-driven models using neural networks was presented by Ruano et al. in Ref. [21]. The neural network model used indoor air temperature, outside solar radiation, air temperature and relative humidity. The neural network model was shown to be better than the physical model for the prediction of the interior temperature.

In other studies, Thomas and Soleilmani [22,23] compared different regression models for predicting the indoor temperature in two buildings. The nonlinear artificial neural network models (ANN) were able to outperform the linear models. In one of the buildings, the data used included the outdoor and indoor temperatures, heating power, wall temperatures, ventilation flow rates, time of day and sun radiation. Two of the main conclusions were that the nonlinear combination of sun radiation and time of day are good predictors for the indoor temperatures and that the indoor temperature is non-linear dependent on the ventilation rate.

The indoor temperature of a tropical greenhouse was modeled by Patil, Tantau et al. [24] using autoregressive and neural network models. The models used outdoor temperature, solar radiation, relative humidity and cloud cover as input variables for the developed prediction models.

Neural network and nonlinear (genetic algorithm) models were compared in Ref. [25] to predict indoor temperature and relative humidity in a test house. The research used time, outdoor temperature, indoor temperature, outdoor relative humidity and indoor relative humidity. The two models developed were considered satisfactory for indoor temperature and for relative humidity based on the high correlation coefficients between real and predicted values which ranged from 0.997 to 0.998.

Data obtained over three months was used to develop linear parametric autoregressive (ARC) models and a neural network based nonlinear autoregressive model (NNARX) was used by Mustafaraj et al. to predict the room temperature and relative humidity of an office room in Ref. [26]. The models used outdoor temperatures, outside relative humidity, room temperature, room relative humidity, supply air relative humidity, supply air temperature, supply air flow rate, chilled water temperature, hot water temperature and room CO_2 concentration. The authors evaluated the models performance with different steps ahead predictions.

Multiple linear regression and artificial neural models for the prediction of daily mean indoor temperature and relative humidity were presented in Ref. [27] by Ozbalta et al., The models employed the day of year, the outdoor temperature, outdoor relative humidity, wind speed and indoor temperature and humidity. The artificial neural models performed better in the testing data set based on their higher R squared values compared to those of the multiple linear regression models.

Auto-regressive neural network models were employed by Marvuglia, Messineo et al. [28] to predict indoor temperature of a building. The models used outdoor temperature, air relative humidity, wind speed and interior air temperature as inputs. The authors used training, testing and validation sets to avoid overfitting the neural network. In this work, the data used samples at 1 h intervals.

Hourly indoor temperature and humidity were predicted using an artificial neural network by Mba et al. in Ref. [29]. The researchers established the optimal structure for the number of hidden neurons and activation functions. The models employed outdoor temperature and humidity, indoor temperature and humidity and a series of derived time lagged values.

1.1.2. Internal temperatures in passive buildings

Mean average indoor hourly temperatures during the winter and summer seasons for more than 100 passive units are presented in Ref. [30] by Schnieders and Hermelink. The study used data for passive houses in different locations across Europe. For the houses in Hannover, the mean indoor temperatures in winter ranged from about 13.5 °C to about 22.5 °C with an average for all the houses of 20 °C. The mean temperature for unoccupied homes was below 17 °C. During the summer, the mean indoor air temperatures ranged from about 20.5 °C to about 25.5 °C. In some instances, the 95th percentile of hourly mean values was higher than 27 °C.

Strategies to prevent overheating in a residential passive house in Slovenia were presented by Mlakar et al. in Ref. [31]. The research work employed monitored data to develop a mathematical model for the internal temperature. The research found that the indoor temperature of the house could be controlled by opening windows during hot summer days, with shading of southern and western windows and with the minimization of internal energy gains.

A comparison of indoor temperatures for two passive and two low energy buildings in Vienna was done by Mahdavi et al. in Ref. [32]. The study found that the use of mechanical ventilation in the passive houses was better for controlling the CO_2 levels during the cold periods. In addition, it was found that overall the indoor conditions satisfied the occupants.

Temperature and humidity profiles in small scale passive building blocks were presented in Ref. [33] by Mlakar and Strancar. The study compared three different design constructions to evaluate the effect of moisture transport within walls and estimate the risk of mold damage due to high relative humidity within it.

Table 1 presents a summary of the reviewed literature highlighting the models used by the authors, model inputs and some relevant findings from their research.

The review of the published literature highlights the following points:

- Artificial Neural Network models have been used extensively for indoor temperature prediction in buildings and greenhouses [18–21,24].
- The different regression models have used data such as: exterior and interior air temperature and humidity, solar radiation, cloud cover, wind speed, energy use, ventilation rates and CO_2 levels among others.
- In general, the literature reviewed has shown that the errors in the interior temperature are small when using autoregressive models. The only caveat is that the models decrease in accuracy when the predictions ahead increase.
- The reviewed literature did not use measured occupancy information as an input for the regression models. However, CO_2 levels have been used as an input which indirectly reflects occupancy.
- The different works have addressed the problem of the optimal number of neurons, derived features (lagged values), and variable selection to minimize the prediction error.

Table 1
Summary of previous research work.

Source	Models	Inputs	comments
Gouda, Danaher et al., 2002 [18]	Artificial neural networks ANNs	Outdoor temperature, solar irradiance, heating valve position, interior temperature	Tested error in training, testing and validation sets. Error typically between −1 and 1 °C
Uchida Frausto and Pieters [19]	Auto regressive model and Neural network	Outside air temperature, outdoor relative humidity, global solar radiation, sky cloudiness	Average absolute error lower than 1 °C
Mechaqrane and Zouak [20]	Neural network auto regressive with exogenous input (NNARX)	Heating power, indoor temperature, outdoor temperature, solar radiation,	Typical error from −0.5 to 0.5 °C
Ruano, Crispim et al. [21]	Neural network and physical models	Indoor air temperature, solar radiation, outside temperature and relative humidity	Neural network model was better than the physical model. The RMSE of the neural network was 0.0493 vs 0.1777 for the physical model
Thomas and Soleimani-Mohseni [22]	Linear and artificial neural network models for 1 h and 2 h ahead predictions	Outdoor and indoor temperature, heating power, wall temperatures, ventilation flow rates, time of day and sun radiation.	The mean absolute error in the training set ranged from 0.1449 to 0.2639 For the testing set, MAE ranged from 0.1458 to 0.2769
Patil, Tantau et al., 2008 [24]	Autoregressive and neural network models	External temperature, solar radiation, relative humidity, cloud cover	The outside temperature was found to have the greatest impact on the inside temperature
Lu and Viljanen [25]	Neural networks, and autoregressive models. Used information from previous step to predict the next	Time, outdoor temperature, indoor temperature, outdoor relative humidity and indoor relative humidity	The best model for temperature had a mean squared error of 0.239 with one-step ahead prediction in the testing set.
Mustafaraj, Lowry et al. [26]	Neural network based nonlinear autoregressive model	outdoor temperatures, outside relative humidity, room temperature, room relative humidity, supply air relative humidity, supply air temperature, supply air flow rate, chilled water temperature, hot water temperature and room CO_2 concentration	The MSE for temperature using a step ahead size of 6 (30 min) was 0.0162.
Özbalta, Sezer et al. [27]	Artificial neural network, multiple regression	Day of the year, outdoor temperature and humidity, wind speed, indoor temperature	RMSE ranged from 0.56 to 1.02 for the Artificial neural network models for indoor temperature prediction
Marvuglia, Messineo et al. [28]	Auto-regressive neural network with external inputs (NNARX)	Outdoor air temperature, relative humidity, wind speed and interior air temperature	Research evaluated the impact of number of neurons and delay time. The best MSE was 0.217.
Mba, Meukam et al. [29]	Artificial neural network	Outdoor temperature, humidity, indoor temperature and humidity and a series of time lagged values	High correlation coefficient between predicted and measure temperature of 0.9850.

Fig. 1. Pictures of the house.

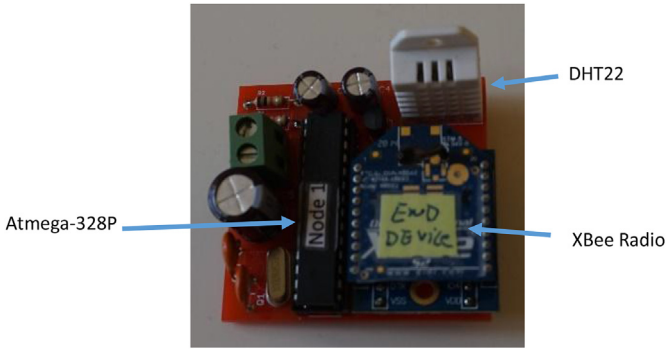

Fig. 2. Photograph of one the wireless sensors. The main components are: the Atmega-328P microcontroller, the DHT 22 sensor for temperature and humidity measurement, and the XBee radio.

- The reviewed data driven regression models have so far not considered the effect of training sample size on the prediction errors.

As can be seen from the literature, the review has not found research work in the area of temperature prediction in buildings addressing the problem of optimal sample size and the corresponding model performance. The main reasons for this could be the relatively small data sets available before, the fact that training of the models can take a significant amount of time and that computing power was not easily available. This paper examines different questions such as: What is the optimal sample size to train models? What is the expected error in the testing set? What is the effect of completing the data sets on the descriptive summary statistics?

1.2. Research objectives and methodology outline

The main purpose of this study was to be able to complete the indoor temperature of a building for a year long data set using the available sampled data. This includes the use of training regression models and the evaluation of their performance in a testing set using

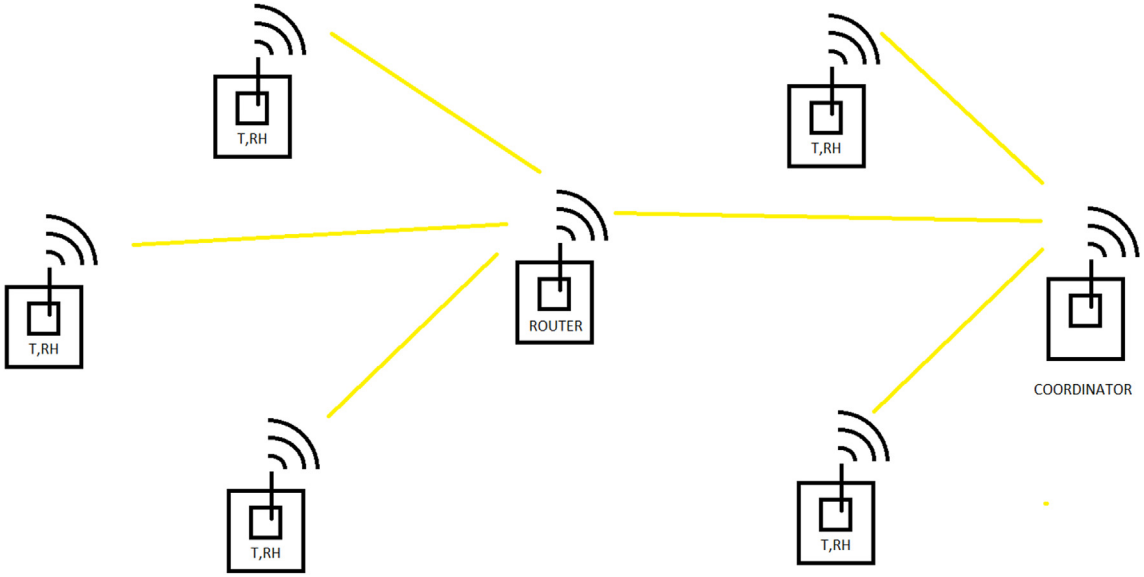

Fig. 3. Conceptual representation of the wireless network architecture with the XBee radios. As shown in the diagram, the router extends the number of end devices and range of the network and forwards the information to the coordinator node.

Fig. 4. First floor. Position of Temperature and Humidity sensors. The blue circles indicate the sensor number. The coordinator (C) is placed near the middle of the house. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

different sample sizes to be able to understand the effect of sample size. Having a complete data set will also enable us to use it for more realistic building simulation calibration studies. Finally, this work will compare descriptive statistics before and after completing the data set.

The rest of the paper is organized as follows: first, a short description of the monitored house and the wireless sensor network. The next section describes the data sets used and the training and testing of two regression models. Then, a description of the data set processing methodology is presented. A discussion of the relative variable importance for the prediction model is included. A comparison of descriptive statistics for the complete and incomplete datasets follows. The paper ends with a discussion of results and conclusions.

2. The passive house

The passive house is located in Stambruges about 24 km from the city of Mons in in Belgium [34,35] (see Fig. 1). The house was designed using the Passive House Planning Package (PHPP) which is an energy simulation tool implemented in Excel [36]. The simulation results from PHPP must show that the building will have an annual heating load and cooling load of no more than 15 kWh/m^2 per year. The house is highly insulated. The U value for the exterior walls, roof and ground are lower than 0.1 W/m^{2}K. The windows are triple glazed with Ug = 0.5 W/m^{2}K and U_f < 0.9 W/m^{2}K and can be operated by the occupants. The total floor area is 280 m^2, from which the total heated area is 220 m^2. The main source of space heating is a wood chimney. In addition, there was no cooling system installed in the house during the time the data was collected. Therefore, the indoor temperature could not have been controlled with a thermostat to provide a specific set point. Information about building occupancy and windows status (open/not) was not recorded.

2.1. Electrical energy metering and ZigBee wireless sensor network

M-BUS energy counters were used to monitor the electric energy use. The information was collected every 10 min. The sub metering includes the energy used by the appliances, lighting, domestic hot water (heat pump), electric heater and ventilation heat recovery.

The house interior temperature and humidity in different rooms were monitored with a custom made ZigBee wireless sensor network [37]. The sensor nodes were made of XBee radios [38], Amega328P microcontrollers [39] and DHT-22 sensors. Fig. 2 shows the main components for the sensor nodes. The accuracy for the temperature is ± 0.5 °C and ± 3% for the relative humidity. The sensor nodes reported the temperature and humidity around every 3.3 min. In the ZigBee architecture, one radio acts as the network coordinator. There can be multiple end nodes (sensors) and routers to extend the amount of end nodes and the range of the wireless network. The coordinator radio receives all the incoming information from the end sensor nodes and those relayed by the routers. Please refer to Fig. 3, Fig. 4 and Fig. 5 for operation and position of the end sensor nodes.

Fig. 5. 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 colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Data Variables and description.

Data Variables	Units	Number of Features
Total Electrical Energy use (Appliances + lights + domestic hot water + ventilation + electric heater)	Wh	1
T1, Temperature in kitchen area	°C	2
T2, Temperature in living room area	°C	3
T3, Temperature in laundry room area	°C	4
T4, Temperature in office room	°C	5
T5, Temperature in bathroom	°C	6
T7, Temperature in ironing room	°C	7
T8, Temperature in teenager room 2	°C	8
T9, Temperature in parents room	°C	9
T_{AVG}, Average indoor temperature	°C	10
To, Temperature outside (from Chièvres weather station)	°C	11
Pressure (from Chièvres weather station)	mm Hg	12
RHo, Humidity outside (from Chièvres weather station)	%	13
Windspeed (from Chièvres weather station)	m/s	14
Visibility (from Chièvres weather station)	km	15
Tdewpoint (from Chièvres weather station)	°C	16
Date time stamp	year-month-day hour:min:s	–

3. Recorded data

All the data analysis and modeling presented here was done in R [40]. R is a programming language and environment that has been developed for statistical computing and is similar to the S language developed at Bell laboratories. R is available as free software. The weather data for the nearest airport weather station (Chièvres Airport, Belgium) was merged by date and time in this study and used to predict the average interior air temperature (T_{AVG}) [41]. The Chièvres Airport is located about 12 km from the passive house. The weather data from the Chièvres Airport was found to have prediction power in Ref. [34]. The downloaded weather data is at hourly intervals, and linear interpolation was done to complete the data set at 10 min intervals. The additional T_{AVG} variable is a volume weighed derived variable based on the measured indoor temperatures [42]. Mathematically, it can be expressed as:

$$T_{AVG} = \frac{\sum_{i=1}^{n} T_i \cdot Vol_i}{\sum_{i=1}^{n} Vol_i}$$

Where T_i is the room temperature in degrees C in the room i and Vol_i is the corresponding volume in the room i.

Table 2 lists all the variables or features used in this work to build the regression models. The total electrical energy used in the house represents the energy used by the appliances, lighting, domestic hot water, ventilation and the electric heater. The measured air temperatures by the ZigBee Wireless Sensor network were also used in the prediction models. The specific weather data from Chièvres is detailed in Table 2.

Fig. 6. Data processing flow chart.

3.1. Data sets processing

The monitored data for 2016 was merged with the weather data from Chièvres using the date and time stamp. The resulting data has some missing values in the variable of interest (T_{AVG}). In order to train the regression models it is preferable to work only with a complete data set (free of missing data). After removing the missing rows, the resulting complete data set was split into training and testing sets. Later on, two regression models were trained: multiple linear regression (LM) and random forest with different sample sizes to examine the effect on the error. Afterwards, using one model the T_{AVG} was predicted. Fig. 6 shows the work flow for the data processing. In Fig. 6 D&T refers to date and time stamp.

Finally, a similar analysis was done to predict the missing data for each of the interior temperatures (*T1, T2, T3* … etc) using T_{AVG} to obtain linear regression models.

Fig. 7 shows the maximum, average and minimum indoor temperatures for the house together with the outdoor temperature. The monitoring discontinuities can be easily appreciated for the monitored indoor data. There is significant data missing during the first days of January, between May and June, between mid-July and mid-August, and in the month of November for several days.

3.2. Model learning curves

From the available data, only the complete data set was used for training and testing the models. Please note that a complete data set refers to a data set free of any missing observations. The data set with missing data has 52704 entries for 2016 (at 10 min intervals), the complete data set (free of missing data) has 37026 entries or information for 257 days. The complete data set was split into training and validation randomly using CARET'S[1] create data partition function. 75% of the data was used for training of the models and the remaining was used for testing [43].

Two regression models, multiple linear regression and random forest, were trained to predict the average indoor temperature using different sample sizes for the training set to observe how the error in the training and testing sets behave as the sample size increases. The predictor variables for the models are: To, RHo, wind speed, visibility, pressure, Tdewpoint and Total electric energy use (Please refer to Table 2). The random forest model is a tree based model which uses the output of several regression trees [55]. In this model, each tree is built with a random sample of selected predicting variables.

The root mean squared error (RMSE) [44] is used to evaluate the performance of the models against the sample size.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2}{n}} \quad (1)$$

Where Y_i is the actual measurement (T_{AVG}), $\widehat{Y_i}$ is the predicted value and n is the number of measurements (sample size).

Fig. 8 shows the performance of the two models with sample size. The behavior of the RMSE curves is similar for both models. At the beginning, the RMSE in the training set (black curve) is smaller but then it stabilizes as the sample size increases. Also, as the training sample size increases, the RMSE in the testing set tends to decrease. However, for the LM model, after a sample size of about 15300 the testing set

[1] The Classification and Regression Training package (CARET) has a set of functions to facilitate the creation and testing of regression models.

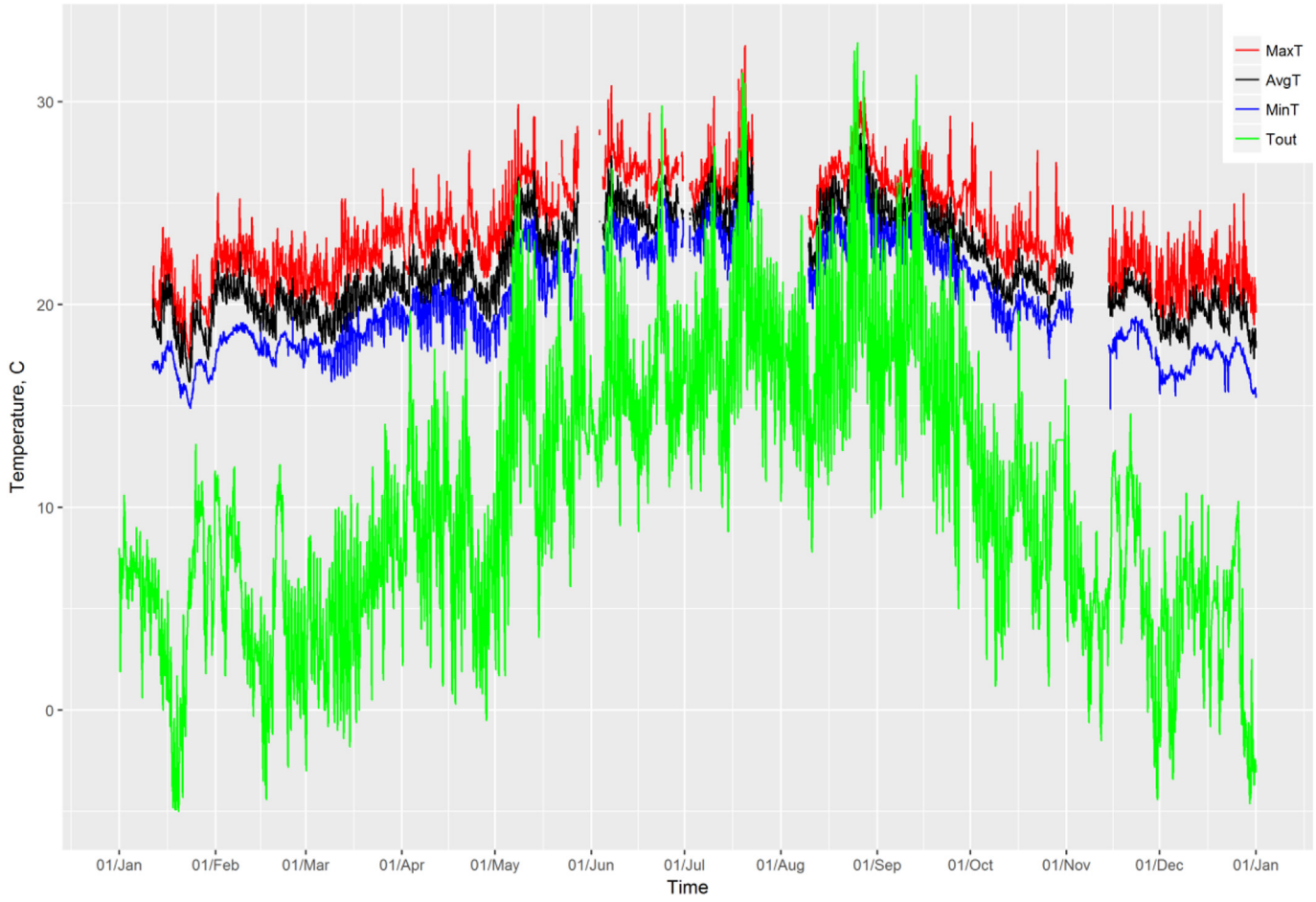

Fig. 7. Maximum, average and minimum indoor temperature and outdoor air temperature for the passive house in 2016.

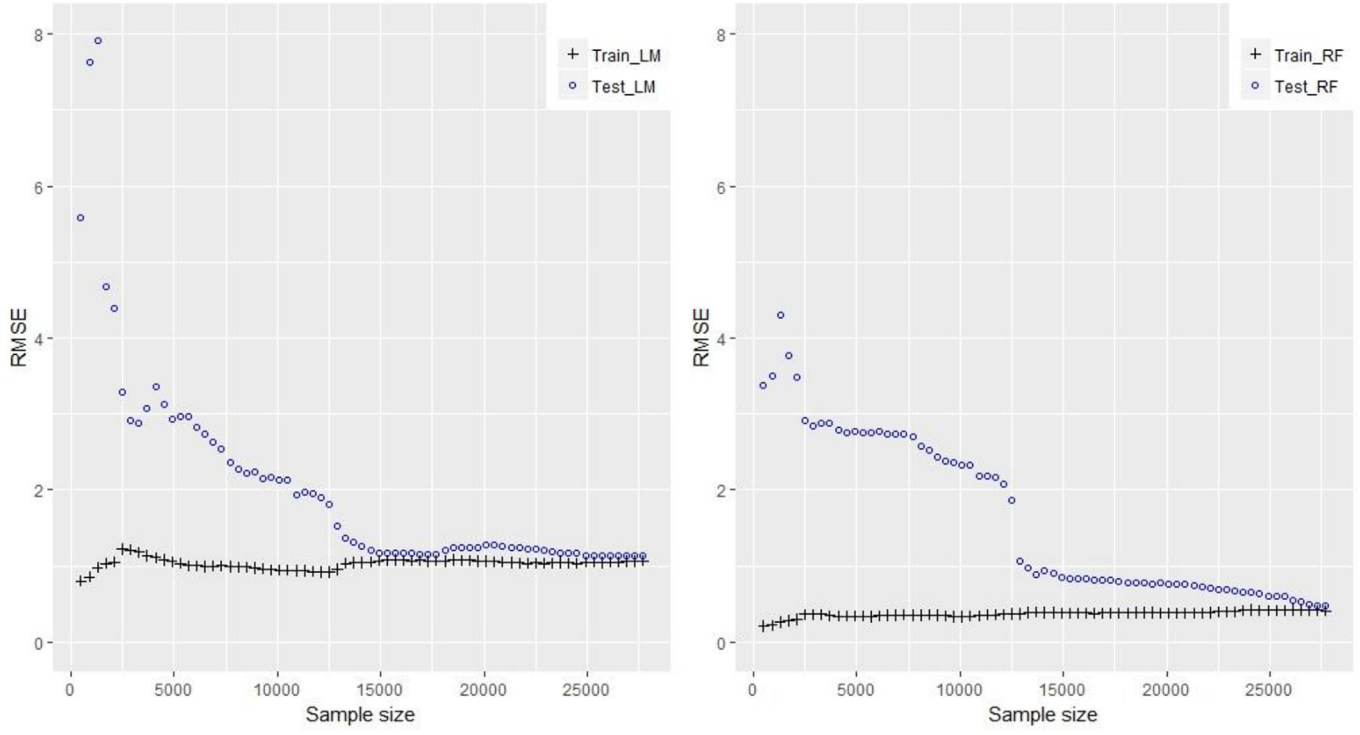

Fig. 8. Learning curves for the LM (left) and RF (right) models.

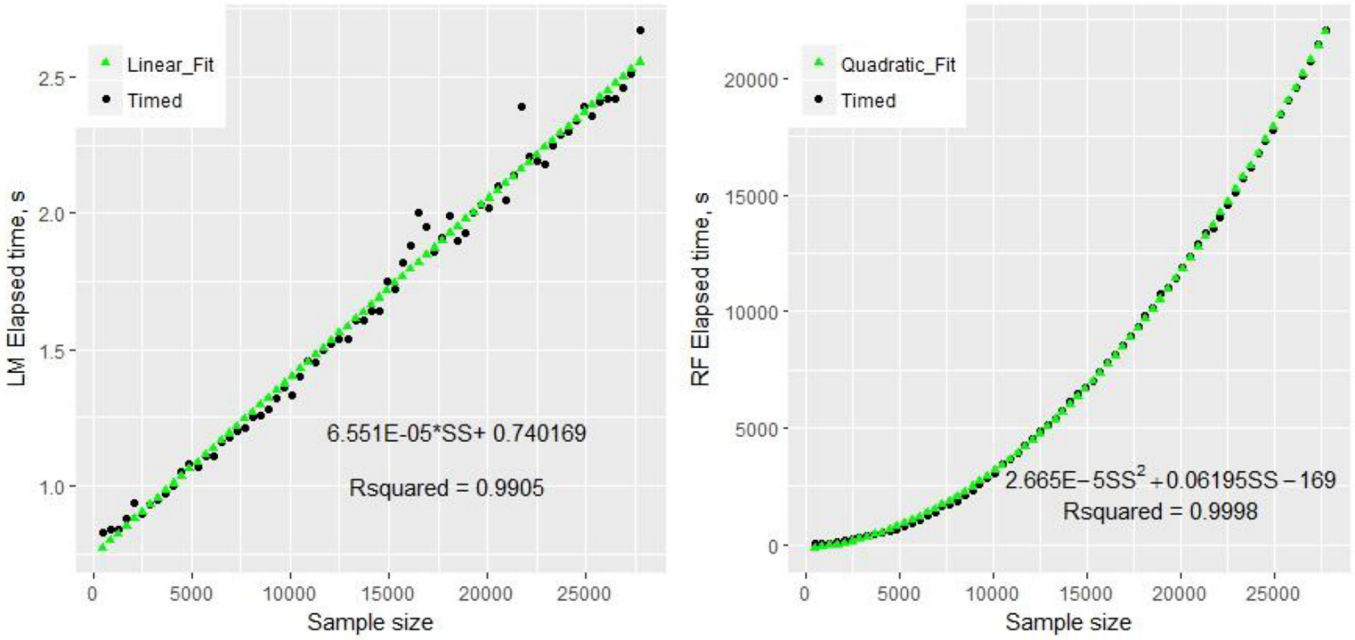

Fig. 9. Elapsed time for the LM and RF models versus sample size.

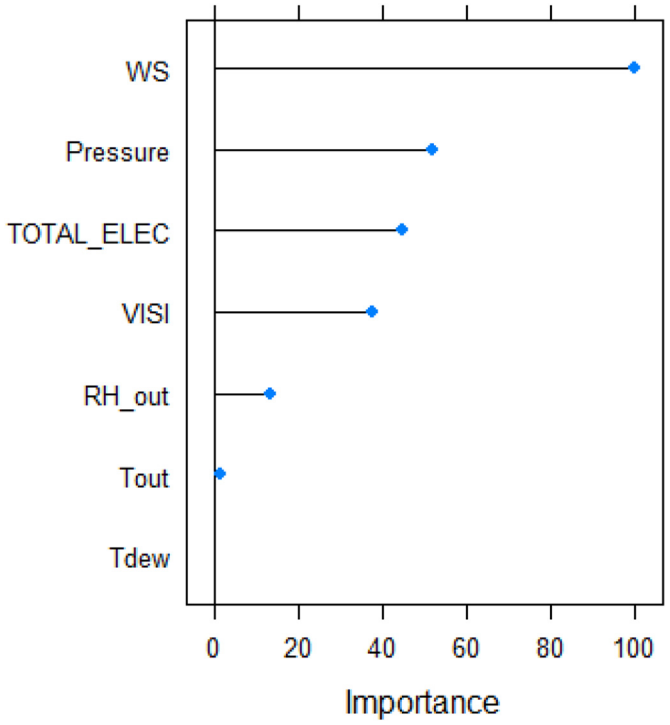

Fig. 10. Variable importance for the RF model.

RMSE (1.17 °C) stabilizes and becomes as close as the RMSE in the training set (1.07 °C). For the RF model, the RMSE (0.48 °C) in the testing set approaches the RMSE of the training (0.42 °C) set but only after a sample size of about 27300. In other words, the regression models do not improve their performance significantly after those samples sizes have been reached.

Another important aspect of performance to be considered besides the error is the elapsed time to train the models. During training and testing of the models, the elapsed time for each training cycle was recorded and the values are presented in Fig. 9. The LM model was significantly faster to train than the RF model. For example at a large sample size of 27300, the LM model only uses 2.51s versus 8549s for the RF model. In other words, the LM model is 8500 times faster to train at that sample size. The training time for the LM is a linear function of the sample size. For the RF, the elapsed time is a quadratic function.

Now that the sample size question was considered, a RF model using all the training data was trained using repeated 10 fold cross validations repeated 3 times to select the best RF model. The RF model was preferred to the LM model because it is more precise as seen in Fig. 8. To speed up the computation time during cross validations, the doParallel package was used [45].

Fig. 10 shows the relative variable importance for the trained RF model. For the RF model, the variable importance was measured by the residual sum of squares. As seen in the figure, the wind speed and pressure are important variables for the prediction. Similar results for variable importance were also found in Ref. [34]. The influence of the wind speed on the interior temperature can be understood due to its direct impact on the exterior convective heat transfer coefficient [46–48]. Also Pressure is correlated to wind speed [34].

Besides the root mean squared error (RMSE), other error metrics used here are: the coefficient of determination or R-squared/R^2, the

Table 3
RF model performance.

Model	Parameters/Features	Training				Testing			
		RMSE	R^2	MAE	MAPE %	RMSE	R^2	MAE	MAPE %
RF	T_{AVG}, To, Pressure, Rho, Windspeed, Visibility, Tdewpoint	0.15	0.996	0.10	0.45	0.35	0.978	0.23	1.06

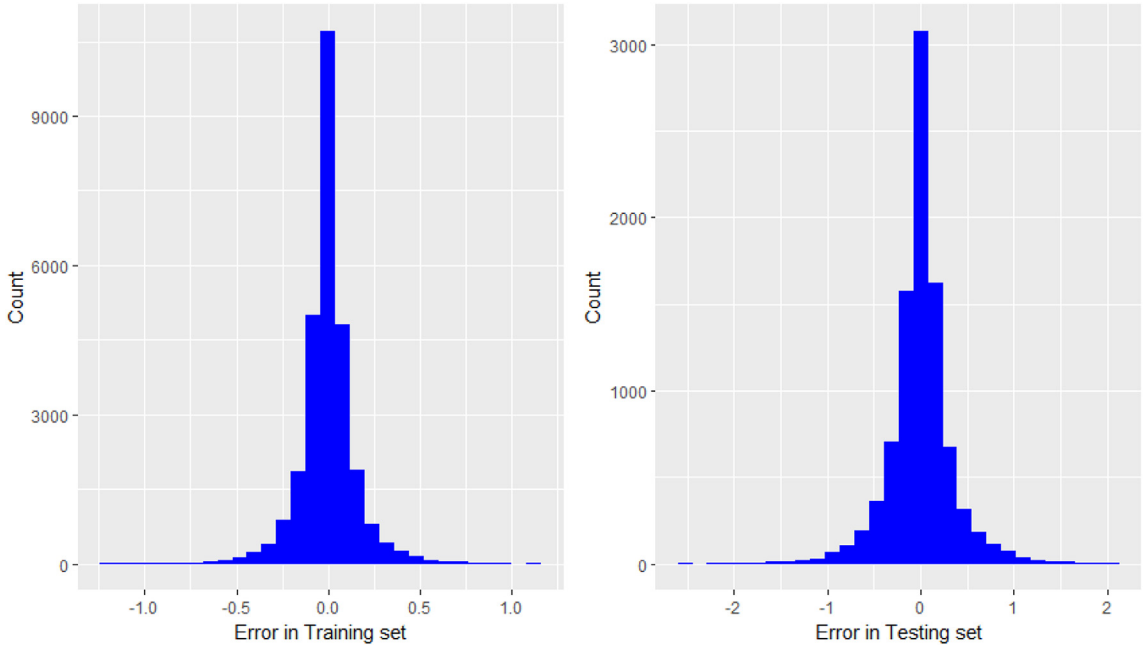

Fig. 11. Histograms for error in the prediction of T_{AVG} in the training set (left) and testing set (right).

Table 4
Linear models for interior temperatures.

Model	Parameters/Features	Training				Testing			
		RMSE	R^2	MAE	MAPE %	RMSE	R^2	MAE	MAPE %
lm	T_{AVG}, *T1*	0.438	0.956	0.333	1.491	0.438	0.956	0.333	1.494
lm	T_{AVG}, T2	0.870	0.887	0.707	3.344	0.874	0.887	0.707	3.340
lm	T_{AVG}, T3	1.03	0.848	0.771	3.414	1.02	0.851	0.764	3.373
lm	T_{AVG}, T4	0.765	0.905	0.573	2.700	0.762	0.906	0.574	3.373
lm	T_{AVG}, T5	0.814	0.907	0.643	3.120	0.817	0.907	0.644	3.127
lm	T_{AVG}, T7	0.969	0.866	0.747	3.603	0.966	0.866	0.744	3.589
lm	T_{AVG}, T8	0.995	0.830	0.789	3.535	0.997	0.827	0.789	3.537
lm	T_{AVG}, T9	0.810	0.914	0.639	3.167	0.813	0.913	0.640	3.175

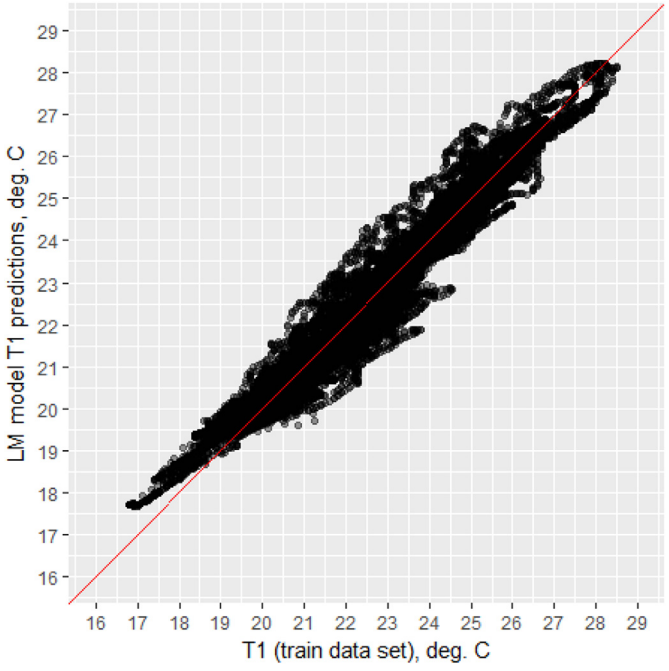

Fig. 12. Linear regression prediction for T1 and training data.

mean absolute error (MAE) and the mean absolute percentage error (MAPE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - \widehat{Y_i}|}{n} \quad (3)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i - \widehat{Y_i}|}{Y_i} \quad (4)$$

The R-squared can be interpreted as the amount of variability in the data explained by the regression model [49]. The mean absolute error measures the difference between the measured T_{AVG} and the predicted temperature value $\widehat{Y_i}$. Similarly, the MAPE measures the average or mean absolute error as a percentage.

Table 3 shows the RF model performance metrics for the training and testing sets. As can be seen, the RMSE is small for both the training (0.15 °C) and testing sets (0.35 °C). The R^2 is high for both training (0.996) and testing (0.978) sets. In the training set, the mean absolute error is 0.10 °C and 0.23 °C in the testing set.

Fig. 11 shows histograms for the prediction error for both sets. For the training set the error is usually within −0.5 to 0.5 °C and for the testing set within −1 to 1 °C.

After the final RF model was obtained, it was used to predict the

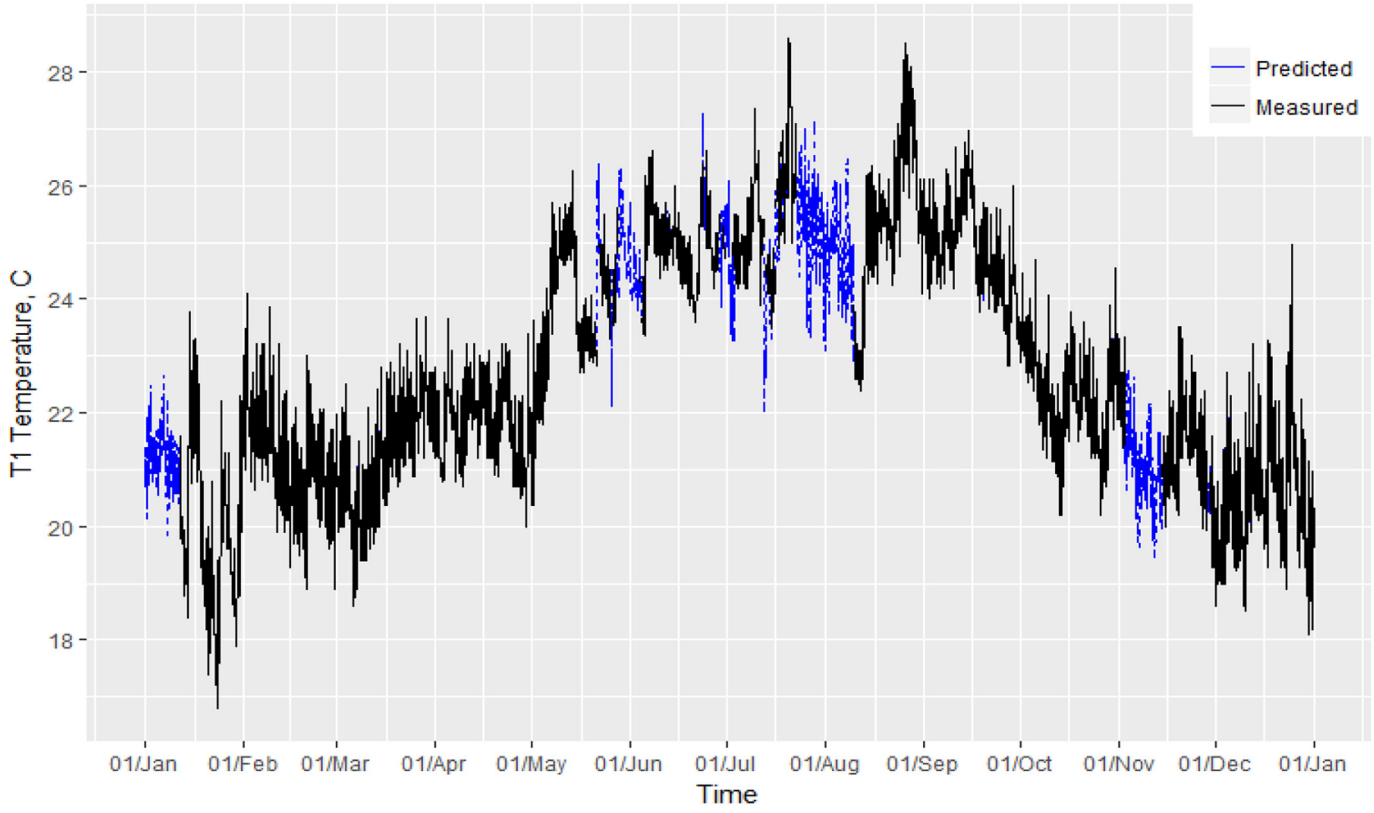

Fig. 13. T1 profile for the whole year completed with predictions for missing data.

Table 5
Summary of descriptive statistics for interior temperatures.

		T1 (Kitchen)	T2 (Living Room)	T3 (Laundry)	T4 Office	T5 Bathroom	T7 Ironing	T8 Boys bed	T9 Parents
Min.	Missing	16.8	16.1	15.5	14.9	15.3	15.4	16.3	14.9
	Completed	16.8	16.1	15.5	14.9	15.3	15.4	16.3	14.9
Median	Missing	22.2	20.7	22.7	21.7	20.2	21.2	22.7	20.3
	Completed	22.4	20.9	22.9	22.0	20.3	21.4	22.8	20.4
Mean	Missing	22.5	21.3	23.0	22.0	20.7	21.5	22.8	20.6
	Completed	22.8	20.9	23.2	22.3	20.9	21.7	23.0	20.9
Max	Missing	28.5	30.8	29.4	28.4	28.0	28.3	30.4	28.1
	Completed	28.6	32.8	29.4	28.4	28.0	28.3	30.4	29.1

average indoor temperature for the missing data. Then, what remained was to find relationships between the average indoor temperature and the temperatures for each zone (T1, T2, T3 …). For this task, simple linear regression models were trained using 10 fold cross validations repeated 3 times.

As can be seen in Table 4, the linear regression models were able to reproduce the temperatures for each of the building zones for both the testing and training sets very well. For instance, the MAE is lower than 1 °C for all the models for both training and testing sets. In addition, the errors in the training and testing sets were very close which is an indication that they were trained with enough data. Fig. 12 shows the predictions for the linear model for T1, and it can be appreciated that the predictions fall around the 45-degree red line curve, which is also a good indication that it is adequate to use a linear model. Fig. 13 shows the yearlong temperature profile for T1 with the predicted and monitored data set. It can be seen that the predictions behave similarly to the measurements.

3.3. Descriptive statistics and boxplots

Table 5 lists some descriptive statistics for the incomplete and the complete data sets to appreciate the effect of using the regression models. As seen in Fig. 7, significant data was missing during the months of July and August. Therefore, it is expected to find that the median temperatures for the complete dataset in Table 5 are slightly higher than for the incomplete data set. The lowest room median temperature in the completed data set was found in the Bathroom (T5), which is not surprising since this room is on the north side of the building. The highest completed median temperatures happens in the laundry room, which in spite of not facing the south also has a high concentration of electrical devices that release heat e.g. small fridge, upright freezer, wine cellar, washing machine, dryer, internet router, internet hub and network attached storage [34]. Overall, the highest maximum temperature was found in the Living room which is close to large south facing windows. The relationships between the room temperatures and the difference between the completed and incomplete data sets can be appreciated in the boxplots in Fig. 14.

3.4. Research limitations

The results of the study completely depend on the quality of the regression models. It is likely, that also having different variable measurements (solar radiation, cloudiness and occupancy information) could provide information to improve the prediction of the regression

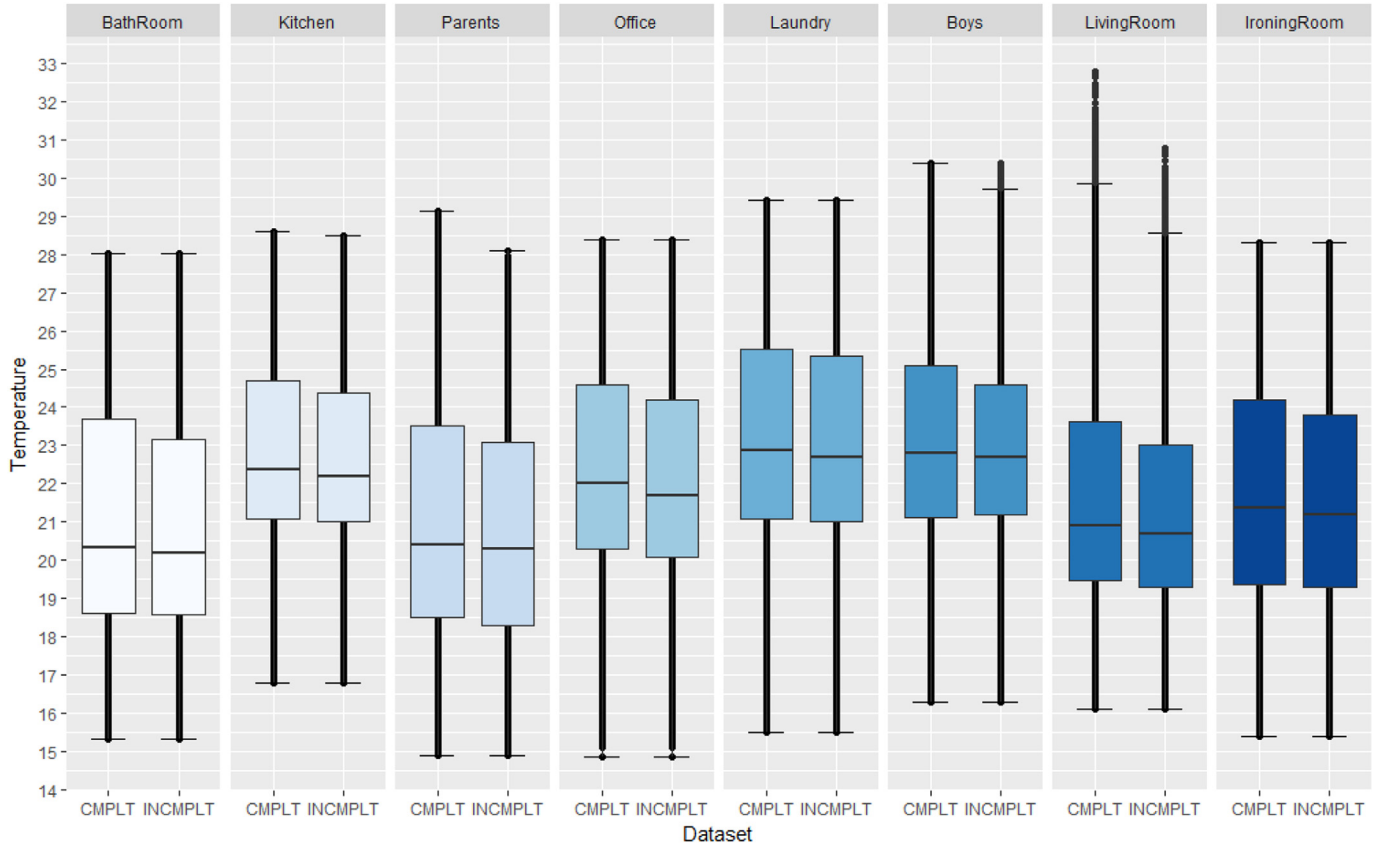

Fig. 14. Room Temperatures Boxplot Comparison for 2016 per room for the complete and incomplete datasets.

models. Moreover, creating the regression curves can be a time consuming task depending on the chosen regression model. However, this problem will be less of a barrier as the computing power continues to increase.

4. Conclusion

As shown in the paper, the use of learning curves as seen in Fig. 8, is a powerful method to understand the effect of sample size on training and testing errors of data driven prediction models. This has practical implications when doing monitoring studies to evaluate how much data it is required to collect to train a prediction model and to compare the performance of difference regression models. It must be noted that for optimal performance the regression models would need to be trained with samples from different seasons to be able to perform well throughout the year. In this study, the random forest model shows better performance compared to the multiple linear regression model as seen in Fig. 8. The optimal sample size for the trained linear model was about 15300, while for the random forest model it was about 27300. The reader must bear in mind that these sample sizes are specific to the studied house. Learning curves for other regression models for other data sets and monitored buildings would display similar behavior and have specific optimal sample sizes.

As seen in Fig. 10, wind speed, pressure, and the total electrical energy used are important variables for an accurate prediction of the interior temperature for the specific trained random forest. It must be noted that other regression models (e.g. Support Vector Machines, Neural Networks etc.) might re order the ranking of importance.

We expect that the completed data sets would be less biased when producing descriptive statistics compared to the missing data. They also have the effect of slightly rising the median temperatures of the rooms because the missing data mainly corresponds to summer months.

The results also indicate that since the studied passive house is so insulated, internal gains have a direct impact on the indoor temperatures, as demonstrated by the fact that the highest median temperature happens in the laundry room where there is a high concentration of electrical equipment. The monitored temperatures in the living room can be quite high as seen in Fig. 14. The highest measured temperature there was 30.8 °C and the largest predicted 32.8 °C. This has the implication that better control of solar gains is necessary in this zone to ensure comfortable temperatures. The minimum temperature in the house was seen in the office (14.9 °C).

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

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".

References

[1] S.D. Sagan, Learning from normal accidents, Organ. Environ. 17 (1) (2004) 15–19.
[2] C. Perlich, F. Provost, J.S. Simonoff, Tree induction vs. logistic regression: a learning-curve analysis, J. Mach. Learn. Res. 4 (Jun) (2003) 211–255.
[3] W. Chen, F.W. Samuelson, B.D. Gallas, L. Kang, B. Sahiner, N. Petrick, On the assessment of the added value of new predictive biomarkers, BMC Med. Res. Meth. 13 (1) (2013) 98.
[4] C.L. Chen, A. Mahjoubfar, L.-C. Tai, I.K. Blaby, A. Huang, K.R. Niazi, B. Jalali, Deep learning in label-free cell classification, Sci. Rep. 6 (2016) 21471.
[5] A. Ng, Machine Learning, Stanford University, Coursera, 2016.
[6] S.A. Kalogirou, Artificial neural networks in energy applications in buildings, Int. J. Low Carbon Technol. 1 (3) (2006) 201–216.

[7] R.H. Dodier, G.P. Henze, Statistical analysis of neural networks as applied to building energy prediction, J. Sol. Energy Eng. Trans. ASME 126 (1) (2004) 592–600.
[8] 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.
[9] M. Braun, H. Altan, S. Beck, Using regression analysis to predict the future energy consumption of a supermarket in the UK, Appl. Energy 130 (2014) 305–313.
[10] J. Kneifel, D. Webb, Predicting energy performance of a net-zero energy building: a statistical approach, Appl. Energy 178 (2016) 468–483.
[11] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Appl. Energy 123 (2014) 168–178.
[12] H.-x. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (6) (2012) 3586–3592.
[13] P. Raftery, M. Keane, J. O'Donnell, Calibrating whole building energy models: an evidence-based methodology, Energy Build. 43 (9) (2011) 2356–2364.
[14] L. Norford, R. Socolow, E.S. Hsieh, G. Spadaro, Two-to-one discrepancy between measured and predicted performance of a 'low-energy'office building: insights from a reconciliation based on the DOE-2 model, Energy Build. 21 (2) (1994) 121–131.
[15] A.H. Neto, F.A.S. Fiorelli, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy Build. 40 (12) (2008) 2169–2176.
[16] G. Mustafaraj, D. Marini, A. Costa, M. Keane, Model calibration for building energy efficiency simulation, Appl. Energy 130 (2014) 72–85.
[17] A. Chong, K.P. Lam, W. Xu, O.T. Karaguzel, Y. Mo, Imputation of missing values in building sensor data, ASHRAE and IBPSA-USA SimBuild 2016 Building Performance Modeling Conference, Salt Lake City, UT, 2016.
[18] M.M. Gouda, S. Danaher, C.P. Underwood, Application of an artificial neural network for modelling the thermal dynamics of a Building's space and its heating system, Math. Comput. Model. Dyn. Syst. 8 (3) (2002) 333–344.
[19] H. Uchida Frausto, J.G. Pieters, Modelling greenhouse temperature using system identification by means of neural networks, Neurocomputing 56 (2004) 423–428.
[20] A. Mechaqrane, M. Zouak, A comparison of linear and neural network ARX models applied to a prediction of the indoor temperature of a building, Neural Comput. Appl. 13 (1) (2004) 32–37.
[21] A.E. Ruano, E.M. Crispim, E.Z. Conceiçao, M.M.J. Lúcio, Prediction of building's temperature using neural networks models, Energy Build. 38 (6) (2006) 682–694.
[22] B. Thomas, M. Soleimani-Mohseni, Artificial neural network models for indoor temperature prediction: investigations in two buildings, Neural Comput. Appl. 16 (1) (2007) 81–89.
[23] M. Soleimani-Mohseni, B. Thomas, P. Fahlén, Estimation of operative temperature in buildings using artificial neural networks, Energy Build. 38 (6) (2006) 635–640.
[24] S.L. Patil, H.J. Tantau, V.M. Salokhe, Modelling of tropical greenhouse temperature by auto regressive and neural network models, Biosyst. Eng. 99 (3) (2008) 423–431.
[25] T. Lu, M. Viljanen, Prediction of indoor temperature and relative humidity using neural network models: model comparison, Neural Comput. Appl. 18 (4) (2009) 345.
[26] G. Mustafaraj, G. Lowry, J. Chen, Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office, Energy Build. 43 (6) (2011) 1452–1460.
[27] T.G. Özbalta, A. Sezer, Y. Yıldız, Models for prediction of daily mean indoor temperature and relative humidity: education building in Izmir, Turkey, Indoor Built Environ. 21 (6) (2012) 772–781.
[28] A. Marvuglia, A. Messineo, G. Nicolosi, Coupling a neural network temperature predictor and a fuzzy logic controller to perform thermal comfort regulation in an office building, Build. Environ. 72 (2014) 287–299.
[29] L. Mba, P. Meukam, A. Kemajou, Application of artificial neural network for predicting hourly indoor air temperature and relative humidity in modern building in humid region, Energy Build. 121 (2016) 32–42.
[30] J. Schnieders, A. Hermelink, CEPHEUS results: measurements and occupants' satisfaction provide evidence for Passive Houses being an option for sustainable building, Energy Pol. 34 (2) (2006) 151–171.
[31] J. Mlakar, J. Štrancar, Overheating in residential passive house: solution strategies revealed and confirmed through data analysis and simulations, Energy Build. 43 (6) (2011) 1443–1451.
[32] A. Mahdavi, E.-M. Doppelbauer, A performance comparison of passive and low-energy buildings, Energy Build. 42 (8) (2010) 1314–1319.
[33] J. Mlakar, J. Štrancar, Temperature and humidity profiles in passive-house building blocks, Build. Environ. 60 (2013) 185–193.
[34] L.M. Candanedo, V. Feldheim, D. Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy Build. 140 (2017) 81–97.
[35] L.M. Candanedo, V. Feldheim, D. Deramaix, A methodology based on Hidden Markov Models for occupancy detection and a case study in a low energy residential building, Energy Build. 148 (2017) 327–341.
[36] W. Feist, R. Pfluger, B. Kaufmann, J. Schnieders, O. Kah, Passive House Planning Package 2007, Passive House Institute, Darmstadt, 2007.
[37] Z. Alliance, Zigbee Specification, (2006).
[38] X. Digi, XBee-PRO® RF Modules, Digi International Inc, Minnesota, 2009.
[39] Atmel Corporation, ATmega328, (2016).
[40] R Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2014.
[41] rp5.ru, Reliable prognosis, http://rp5.ru/Weather_archive_in_Chievres_(airport), (2016).
[42] H. Akay, N. Paydar, A. Bilgic, Fatigue life predictions for thermally loaded solder joints using a volume-weighted averaging technique, J. Electron. Packag. 119 (4) (1997) 228–235.
[43] M. Kuhn, Caret: Classification and Regression Training, (2015).
[44] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer New York, 2013.
[45] Revolution Analytics, S. Weston, doParallel: Foreach Parallel Adaptor for the 'parallel' Package, (2015).
[46] F.L. Test, R.C. Lessmann, A. Johary, Heat transfer during wind flow over rectangular bodies in the natural environmental, Trans. ASME J. Heat Tran. 103 (1981) 262–267.
[47] S. Sharples, P.S. Charlesworth, Full-scale measurements of wind-induced convective heat transfer from a roof-mounted flat plate solar collector, Sol. Energy 62 (2) (1998) 69–77.
[48] J.A. Palyvos, A survey of wind convection coefficient correlations for building envelope energy systems' modeling, Appl. Therm. Eng. 28 (8–9) (2008) 801–808.
[49] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, John Wiley & Sons, 2010.