Environment Systems and Decisions (2019) 39:281–294
https://doi.org/10.1007/s10669-019-09727-1

Data-driven simulation for energy consumption estimation in a smart home

Stephen Adams[1] · Steven Greenspan[2] · Maria Velez-Rojas[3] · Serge Mankovski[3] · Peter A. Beling[1]

Published online: 1 April 2019

Abstract
Simulation and data-driven models are both tools that can play an important role in reducing the energy consumption of buildings and homes. However, sophisticated control schemes and models are only as good as the data collected by sensors and provided to them. Low-quality or faulty sensor that provide inaccurate data can lead to inefficient buildings. In this paper, we investigate the relationship between sensor quality and the prediction of energy consumption. We first construct a simulation of appliance energy consumption in a smart home and then assess the predictive ability of several data-driven models while varying the quality and function of the simulated sensors. The simulation was constructed using a smart home data set collected by other researchers. We find that the predictive ability is only decreased when noise is added to the appliance energy random variable. We conclude that low-quality sensors that do not monitor the environment as accurately as the devices used in the original study could be used for humidity and temperature without significantly reducing the predictive ability of the data-driven models. The method and findings have implications for how to conduct cost-benefit analyses of IoT device requirements.

Keywords Data-driven models · Energy estimation · Simulation

Abbreviations

α	Temperature parameter for Boltzman distribution of light energy consumption
β	Temperature parameter for Boltzman distribution of appliance energy consumption
GBM	Gradient boosted machines
HVAC	Heating, ventilation, and air conditioning
IoT	Internet-of-things
MAE	Mean absolute error
MAPE	Mean absolute percentage error
μ	Mean of Gaussian distribution
RH#	Relative humidity sensor number #
RMSE	Root mean squared error
σ:	Standard deviation of Gaussian distribution
SVM	Support vector machine
T#	Temperature sensor number #
Z	Random variable for appliance energy consumption

✉ Stephen Adams
sca2c@virginia.edu

Steven Greenspan
steven.greenspan@ca.com

Maria Velez-Rojas
maria.velez-rojas@ca.com

Serge Mankovski
serge.mankovskii@ca.com

Peter A. Beling
beling@virginia.edu

[1] University of Virginia, Charlottesville, VA 22903, USA

[2] CA Technologies, 200 Princeton S. Corp Center, Ewing, NJ 08628, USA

[3] CA Technologies, 3965 Freedom Circle, Santa Clara, CA 95054, USA

1 Introduction

Buildings roughly make up 40% of the world's total energy consumption (Costa et al. 2013; Shaikh et al. 2014). In the United States, publicly owned buildings have stringent goals for the energy consumption of existing and new buildings. Simulation can play an important role in reducing the energy consumption of buildings (see Hong et al. (2018) for an in-depth review of the literature and ten challenges facing the field of building simulation). However, there is no single building simulation tool that can be used for all buildings

(Trčka et al. 2010), so building simulation tools are often combined in the analysis or constructed for a specific purpose. Beyond energy consumption, simulations are widely used in the design and control of buildings to investigate evacuation strategies (Chen et al. 2017; Ding et al. 2015; Sano et al. 2018), model water usage (Xue et al. 2017), and model occupancy (Chen et al. 2018; Wang et al. 2011).

Heating, ventilation, and air conditioning (HVAC) systems are generally the largest consumer of energy within a building and thus offer the greatest pathway for reducing energy consumption (Ahmad et al. 2016). However, these largely automated HVAC systems must rely on data from sensors distributed throughout the building to efficiently function. Malfunctioning sensors, or low-quality sensors that do not accurately collect data, could be detrimental to building efficiency and could cause buildings to operate in a fashion that consumes more energy than a building with less sophisticated HVAC controls. Further, studies have shown that humans mismanaging a building can lead to higher energy consumption even when the smart-building control system is functioning properly (Belafi et al. 2017). Effectively modeling human behavior in a building is crucial to estimating the energy consumption (D'Oca et al. 2014).

Historically, advanced HVAC control systems have primarily been utilized in large-scale buildings due to their high cost. However, as the price of these control systems comes down and internet-of-things (IoT) devices become more prevalent in daily life, automated HVAC control for smart-homes will become the norm, and all of the challenges of reducing energy in large-scale buildings will now need to be addressed on the small scale of single family homes. Allard et al. (2018) have demonstrated that taking into account the uncertainty in the energy evaluation methods decreases the performance gap, the difference between estimated energy consumption during design and the actual energy consumption during use. We hypothesize that utilizing machine learning methods and data-driven simulations will help alleviate many of the problems facing energy consumption in residential buildings.

However, the enhanced operational capability that comes with IoT devices and smart controls can come at the cost of system resilience. As Marchese and Linkov (2017) discuss, finding the optimal balance of smartness and resilience for a particular system can be challenging. Smart systems often have a pyramid-like structure with a centralized decision system while a resilient system has a flat architecture that allows for redundancy and parallel processing. Smart HVAC control systems for buildings and residential housing will have to adequately balance this tradeoff to construct systems that reduce the energy consumption.

While HVAC is generally the largest consumer of energy in a building, appliance energy consumption can represent 20–30% of a household's total energy consumption (Cetin et al. 2014; Kavousian et al. 2015). Models that can estimate appliance energy consumption would have varying uses, for example detecting abnormal usage patterns (Seem 2007) and assisting in modeling and simulating energy consumption of buildings, residential households, and apparments (Johnson and Beausoleil-Morrison 2017; Mirakhorli and Dong 2017; Muratori et al. 2013; Ruellan et al. 2016). Further, these models could be used to assist in control decisions (Barbato et al. 2011; Zhao et al. 2013).

These considerations lead us to hypothesize that an important research area is to understand the impact of sensor quality on the prediction of energy consumption in houses. The main contribution of this paper is to address that research question by constructing and validating a simulation-based method for generating artificial data sets and then assessing the impact of varying sensor quality or sensor failure. One type of sensor failure can be modeled as the addition of a noise term to the true measurement. This type of failure is similar to the type of data that would be collected from a low-quality sensor. There are other types of sensor failures that we do not consider in this study but the proposed method could be used to study their affects in a future study.

We acknowledge that the presented study is limited to a particular type of residential construction and that the results and conclusions, in terms of the predictive accuracy of the data-driven models and sensor correlations, may not generalize to other settings. However, the presented method could be easily repeated and applied to new data sets collected from other types of residential and commercial buildings. Further, future work utilizing the proposed method should address this issue and could utilize U.S. based data sets such as the Pecan Street data set.[1]

In this study, we first construct a simulation of appliance energy consumption in a smart home and then assess the predictive ability of several data-driven models while varying the quality and function of the simulated sensors. The simulation was constructed using a smart home data set collected by other researchers (Candanedo et al. 2017) which is publicly available on the UCI machine learning repository (Lichman 2013). Candanedo et al. (2017) equipped a test house with sensors that collected environmental data (temperature and relative humidity) and energy usage data from lights and appliances. Their ultimate goal was to develop data-driven models for predicting the appliance energy usage using the other sensors and external weather data. Candanedo et al. (2017) performed an analysis of variable importance, evaluated several types of data-driven models, and tuned these models to optimize predictive ability.

[1] https://www.pecanstreet.org/category/dataport/

The presented method has two primary contributions to the literature. The first contribution is the use of data-driven models to establish relationships between IoT sensors in a smart home. This will be useful for accurately estimating sensor readings during the design phase and detecting outliers and faulty sensors during operation. The second contribution is the data-driven simulation. Specifically, we build data-driven models that estimate different variables within the home. The appliance energy consumption variable is simulated by first predicting a latent variable which represents the human behavior in the home. The amount of energy consumption is conditioned on this latent "occupancy" variable. The simulation could be used in future work to aid in the design of efficient buildings and homes and to test the effectiveness of varying control policies. Specifically, the simulation could be used to estimate the energy savings when using different sensors to make control decisions and to estimate energy consumption when comparing control policies. HVAC control decisions can significantly affect the energy consumption of a home, and the proposed simulation could be used to evaluate these types of decisions (Yang et al. 2014). The simulation could also be used as the initial building block of a larger simulation investigating the energy consumption of multiple houses and growing the scale of the simulation to neighborhoods or cities.

This paper is organized in the following fashion. In Sect. 2, we describe the smart home and the data collected in the smart home. In Sect. 3, we outline the data-driven models used to predict the different variables in the data set. The simulation is outlined in Sect. 4, and we provide our conclusions in Sect. 5.

2 Background

In this section, we give background on the house used for collecting the data and provide some visualizations of the data. This data set was collected by Candanedo et al. (2017) and is publicly available on the UCI machine-learning repository (Lichman 2013).

2.1 Description of the House

The house, a new construction, is located in Stamburges, Belgium and is designed to have an annual heating load less than 15 kWh/m^2. A single family occupies the house consisting of two adults and two teenagers. One of the adults regularly works from home. The house is equipped with energy meters that monitor the ventilation system, the hot water heat pump, the appliances, the lights, and the baseboard heaters. The house has two stories and contains fourteen rooms with varying types of appliances. For the full list of appliances, see Candanedo et al. (2017).

Table 1 Sensor locations

Sensor identifier	Room
T1 & RH1	Kitchen
T2 & RH2	Living room
T3 & RH3	Laundry room
T4 & RH4	Office
T5 & RH5	Bathroom
T6 & RH6	Outside
T7 & RH7	Ironing room
T8 & RH8	Teenager's bedroom
T9 & RH9	Parent's bedroom

The temperature and relative humidity were monitored in nine rooms using ZigBee[2] wireless network sensors (Alliance 2006). These sensors are located at the same location in each room. The sensors are run on batteries, but the ZigBee coordinator is located in the dining room and uses energy from the house. The sensor locations by room are listed in Table 1. Temperature is measured in degrees Celsius, and relative humidity is expressed as a percent.

2.2 Description of the data

The data set contains over 19,000 observations. The data were collected every 10 minutes from January 11, 2016 to May 27, 2016. A more complete study would collect data for an entire year to include seasonal dynamics. However, the collected data does include the transition from winter to spring so the data does provide some information on seasonal dynamics. Each observation includes a time stamp down to the second. The energy usage is split into energy consumed by the appliances and energy consumed by the light fixtures. Both of these variables are expressed in Watt hours.

The energy consumed by the appliances over the data collection time period is displayed in Fig. 1. The appliance energy usage appears to vary based on the time of day and could be correlated with the occupancy of the house. This figure also indicates that there are periods at the end of January and the beginning of April where the house appears to be unoccupied or experiencing a power outage. Figure 2 displays the energy consumed by the light fixtures over the data collection period. The suspected unoccupied periods at the end of January and the beginning of April are also evident in this data. Further, the energy consumed by the lights decreases at the end of the data collection period when the days are longer.

[2] ©2017 ZigBee Alliance. ZigBee is a registered trademark of the ZigBee Alliance

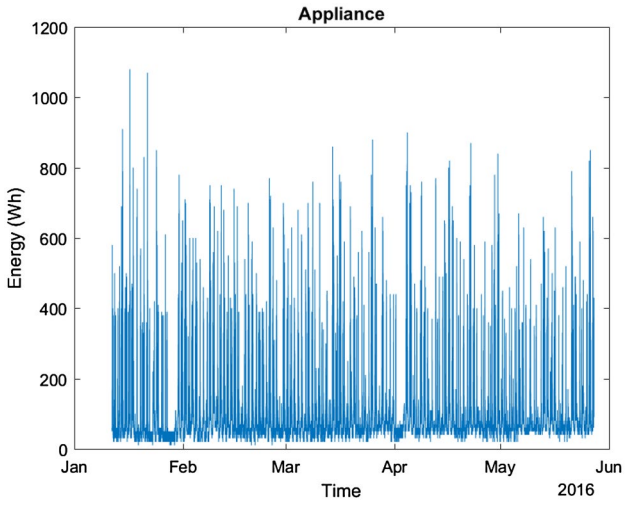

Fig. 1 Energy consumed by appliances

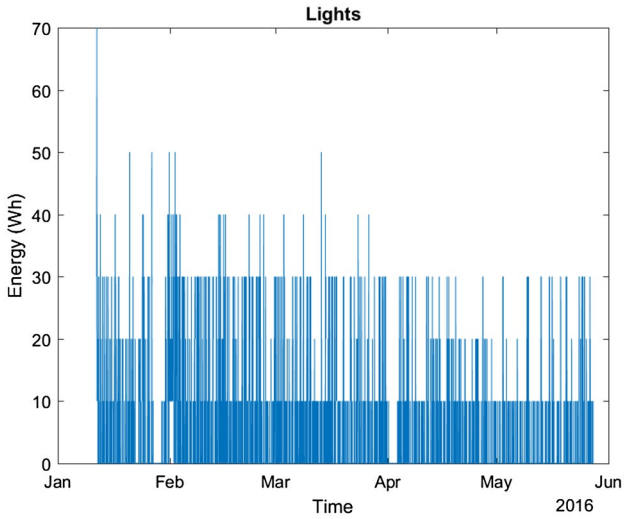

Fig. 2 Energy consumed by light fixtures

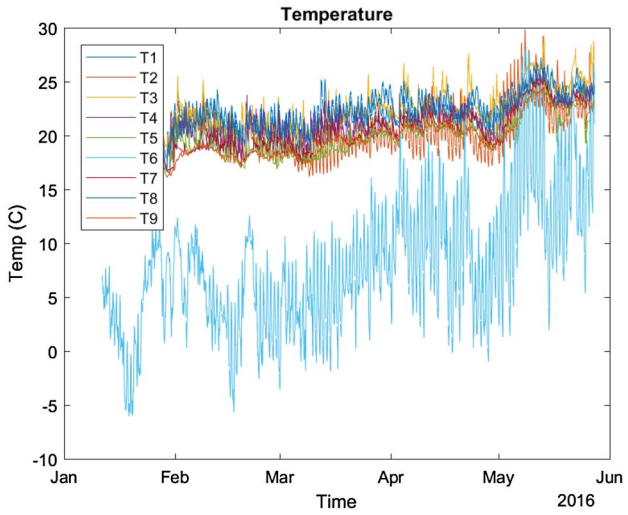

Fig. 3 Temperature data collected from the house. The legend identifies the temperature sensors described in Table 1. Note that T6 is located outside the house

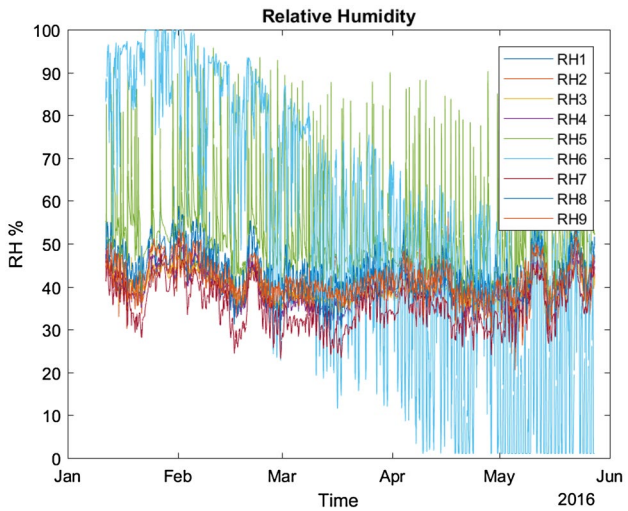

Fig. 4 Relative humidity data collected from the house. The legend identifies the relative humidity sensors described in Table 1. Note that RH6 is located outside the house and that RH5 is located in the bathroom

Figures 3 and 4 display the temperature and relative humidity collected within the house. Note that the relative humidity and temperature sensors at location 6 are actually outside the house and can differ significantly from the data collected inside the house. All of the temperature data appears to be correlated. The relative humidity data from different sensors also appears to be correlated, however, there are large spikes in relative humidity in the bathroom.0

The data set also contains weather data from the Chievres Airport in Belgium, the closest airport to the house. The weather data was collected every hour and then interpolated. The weather data includes the temperature in degrees Celsius, the relative humidity in percent, the pressure in millimeters of mercury, wind speed in meters per second, dew point in degrees Celsius, and visibility in kilometers. Figure 5 displays plots of the weather data.

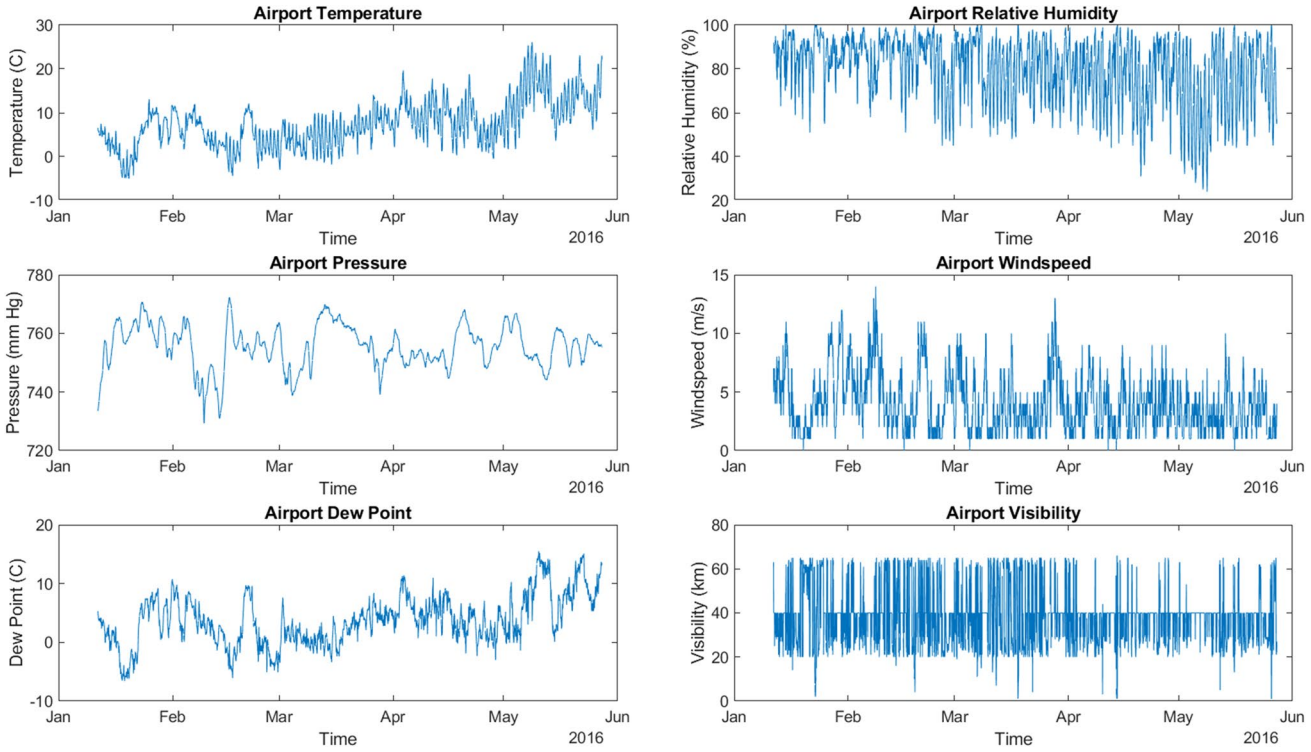

Fig. 5 Weather data from Chievres Airport

3 Data-driven models

In this section, we describe the numerical experiments performed using the data set. MATLAB[3] was used to construct models for predicting appliance energy consumption. Four predictive models were evaluated: linear regression (Murphy 2012), radial support vector machines (SVM) (Dong et al. 2005; Murphy 2012), gradient boosted machines (GBM) (Friedman 2001; Murphy 2012), and the random forest algorithm (Breiman 2001; Murphy 2012). First, the data were randomly divided into testing and training sets. Then, we trained models using the training set and calculate performance on the training and testing sets. The same four performance metrics were used in these experiments as in the original paper: root mean squared error (RMSE), R-squared (R^2), mean absolute error (MAE), and mean absolute percentage error (MAPE). In these experiments, we tested the random forest algorithm using the *TreeBagger* function, the radial SVM using the *fitrsvm* function, and the GBM using the *fitensemble* function.

Before performing the training and testing experiments, we conducted a new feature selection experiment. The *stepwisefit* function was used to find variables that should be included in a linear regression model using stepwise feature selection (Draper and Smith 2014). After performing the fit, 20 features were found to be significant to the prediction of appliance energy usage. These variables were lights, T1, RH1, T2, RH2, T3, RH3, T6, RH6, T7, RH7, T8, RH8, T9, outside temperature, windspeed, visibility, number of seconds from midnight, the day of the week, and the weekend status. This feature selection experiment was only performed on the training set. The training RMSE for the linear model using only the selected variables was 94.32, and the R^2 value was 0.17. These metrics closely match the training RMSE and R^2 in the original paper. However, the selected features do not closely match those of the Boruta algorithm used in the original paper. Notably, the pressure and T5 were excluded from the stepwise analysis, but they were found to be relevant by the Boruta algorithm. From this experiment, we concluded that linear regression does not perform well as a predictive model. However, the linear model appeared to be fairly robust to feature selection because the performance metrics for the feature sets selected by stepwise regression and the Boruta algorithm are similar.

For the random forest algorithm, we grew a forest of 500 trees and sampled 18 predictors for each tree. For the ensemble of regression trees, we used 10,900 trees and had

[3] ©2017 The MathWorks, Inc. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

Table 2 Results for predictive experiments performed in MATLAB

Model	Data Set	RMSE	R^2	MAE	MAPE
Ensemble	Training	68.65	0.56	41.59	0.50
	Testing	80.37	0.35	49.21	0.60
Random forest	Training	46.68	0.80	21.28	0.20
	Testing	67.04	0.55	32.72	0.34
Radial SVM	Training	95.35	0.15	40.63	0.32
	Testing	93.91	0.11	42.16	0.36

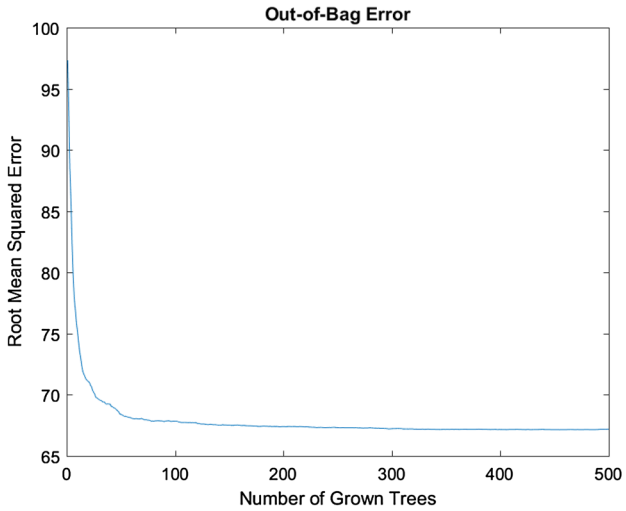

Fig. 6 Out-of-bag error versus number of trees using the random forest algorithm in MATLAB. This out-of-bag error was calculated using the entire data set

a maximum split depth of 5. For the radial SVM, we used the MATLAB option to find the optimal parameters. Three parameters were optimized: box constraint - 973, kernel scale—192, and epsilon—15. Table 2 contains the results for the testing and training sets.

Only the random forest algorithm gives similar results as the original paper. However, it is interesting that the training evaluation under-performs while the testing evaluation over-performs the original results for this model. Both the radial SVM and the ensemble method performed much worse than in the original study. This could be due to several reasons. First, the implementation of the models might be slightly different in the two software packages. Second, the optimal parameters for the ensemble method might not translate. Third, when finding the optimal hyperparameters for the radial SVM, the objective function is slightly different between the two software packages. The original analysis by Candanedo et al. (2017) was performed in the R software package (R Core Team 2013). The CARET package in R minimizes RMSE while

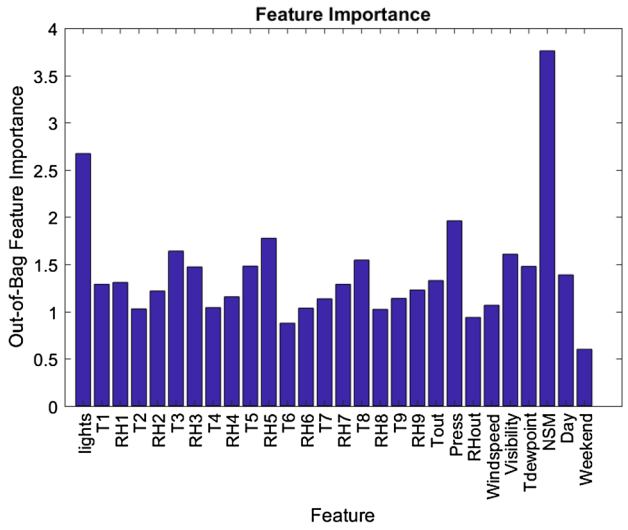

Fig. 7 Variable importance using the random forest algorithm and the entire data set

MATLAB minimizes cross-validation error. The search over the parameter space could also be different.

The random forest algorithm produced the same performance metrics for both MATLAB and R so we performed a little more investigation into the performance of the model. Generally, when using the random forest algorithm, splitting the data into testing and training sets is not required for evaluating generalization error. In this algorithm, the data is bagged and the out-of-bag (OOB) error can be used to evaluate the model. Figure 6 displays the RMSE calculated from the OOB error as the forest increases in size when using the entire data set. Once the forest grew beyond 300 trees, the RMSE converged to the testing RMSE in Table 2. Figure 7 displays the variable importance calculated by the random forest algorithm in MATLAB. The variable importance analysis generally reflects the feature selection analysis in the original paper.

These results may not generalize to other types of residential construction and climate. For example, older houses that do not have modern insulation may be affected more by the outside temperature and relative humidity. However, the method used for estimating the data-driven models could be repeated for data collected in this type of residence.

3.1 Estimating the other variables

In the original paper, the authors focused solely on predicting the appliance energy consumption. We tested if the other collected variables in the data set can be estimated using similar machine-learning techniques. Given the results in the previous section, we decided to use the random forest algorithm and evaluate the generalization error using the OOB

Table 3 RMSE for each response variable

Variable	RMSE
Appliances	67.30
Lights	4.36
T1	0.06
RH1	0.58
T2	0.12
RH2	0.30
T3	0.10
RH3	0.26
T4	0.12
RH4	0.24
T5	0.11
RH5	2.29
T6	0.31
RH6	1.72
T7	0.06
RH7	0.26
T8	0.08
RH8	0.30
T9	0.06
RH9	0.27
Tout	0.19
Press	0.54
RHout	1.02
Windspeed	0.29
Visibility	2.96
Tdewpoint	0.15

error. Specifically, we selected one of the 26 variables in the data set to be the response variable and used the remaining variables in the data set as explanatory variables. A random forest with 300 trees is trained on the entire data set, and RMSE is calculated using the OOB error. Each response variable is treated as a continuous variable.

The RMSE for each response variable is displayed in Table 3. The RMSE for the appliance energy consumption matched that of the previous experiments. The RMSE for the lights is less than 5 Wh. The temperature collected by the installed sensors is on average estimated to within a half a degree Celsius. Generally, the random forest was able to predict the collected relative humidity to within half a percentage point. However, RH5 and RH6 both had a RMSE greater than 1. RH5 was located in the bathroom and randomly experienced spikes, possibly due to showers. RH6 was located outside the house and was, therefore, slightly harder to estimate. The relative humidity from the weather data at the airport also had a RMSE greater than 1. The rest of the weather data were also able to be estimated using the random forest algorithm.

There are only eight unique values for the energy consumption of the light fixtures. Therefore, this problem can be treated as a classification problem with eight classes. A classification random forest was trained, and the OOB classification error was calculated. The classification error for this problem formulation was 0.16.

4 Simulation

In this section, we outline how to construct a simulation for generating a similar data set to the collected data in the original paper. The objective is to use the time information and the weather data from the airport as inputs into the simulation and then to simulate the environmental data and the energy consumption data with varying degrees of randomness. This simulation-based method could be used to create new large-scale simulations of neighborhoods and communities consisting of several houses. Further, the proposed method could be replicated for other types of residential constructions and for different climates. The simulation in this study may not transfer directly to new environments, but new simulations using the proposed method could be constructed if the necessary data is provided.

The simulation utilizes the data-driven models constructed in the previous experiments and contains several steps. First, the time information and the weather information is used to simulate the environmental data within the house. Then, the simulated environmental data is used in conjunction with the time information and the weather information to simulate the light energy consumption. Finally, the previously simulated data (lights and environmental) and the external input information (weather and time) is used to simulate the appliance energy consumption. At each step, randomness or noise can be added to the simulated data. The simulation is constructed in MATLAB.

The simulation can be broken into two stages: training and simulation. The training stage learns the models for simulating the data, and Fig. 8 gives a visual representation of the data used to train each model. The simulation stage generates the data using the trained models, and Fig. 9 shows the flow of data through the simulation.

The environmental simulation uses a regression random forest to generate the data. A separate random forest is trained for each temperature and relative humidity sensor. The weather and time information are used as the explanatory variables and each sensor is the response. This results in 18 separate random forest models for this stage of the simulation. Noise can be added to the simulated data using a Gaussian distribution with zero mean and a standard deviation σ. The value of σ controls the amount of noise injected into the simulated data. In the presented example simulation, 50 trees are grown in each forest and no noise is added to the data. The real and simulated data is displayed in Figs. 10 and 11.

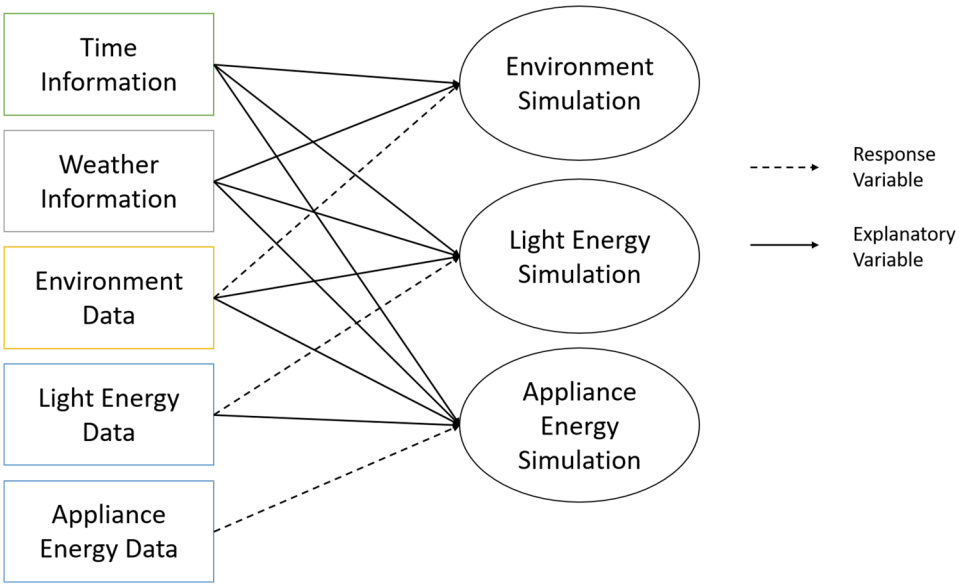

Fig. 8 Visual display of data that is used to train each model. The solid lines represent that the data is used as the explanatory variables. The dashed lines represent the data used as the response variable

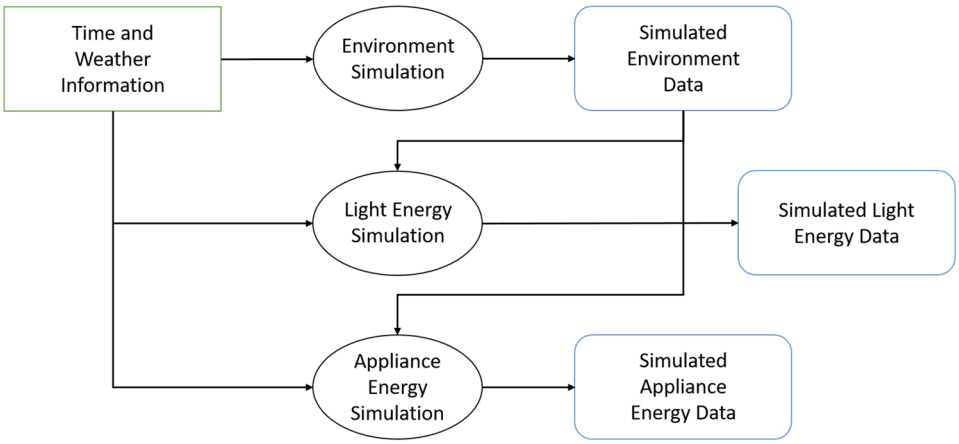

Fig. 9 Visual display of data flowing through the simulation

For the light energy simulation, a classification random forest is used. The model is trained on the time and weather information and the real environmental data. In the presented simulation, 200 trees are grown in the forest. The random forest algorithm is used to generate a posterior distribution over the light energy values. This posterior is then sampled to produce the simulated light energy value. Randomness can be added using a Boltzmann distribution

$$p_i = \frac{e^{x_i/\alpha}}{\sum_{i=1}^{I} e^{x_i/\alpha}}, \tag{1}$$

where p_i is the probability of the i^{th} value of the new distribution to be sampled, x_i is the probability of the i^{th} value of the distribution generated by the random forest classifier, and α is the temperature parameter that controls the amount of randomness. For the presented data, $\alpha = 0.1$. Figure 12 displays the real and simulated data.

The appliance energy data is simulated in two stages. The first stage uses a classification random forest to generate a probability distribution over a binary random variable. As in the light simulation, a Boltzmann distribution with temperature parameter β is used to control the randomness of this distribution. That distribution is randomly sampled for a binary variable Z that defines the appliance energy consumption distribution. In the second stage, an energy value is drawn from one of the Gaussian distributions depending on Z. To ensure that the simulated appliance energy value is positive, the Gaussian distributions are estimated in the log domain. The sampled value is then converted back into the appliance energy domain.

Figure 13 displays a histogram of the appliance energy data in the log domain. A mixture of two Gaussian distributions can be seen in the histogram. Values less than 5 are labeled $Z = 1$, and values greater than 5 are labeled $Z = 2$. This labeling is used to train the classification random forest

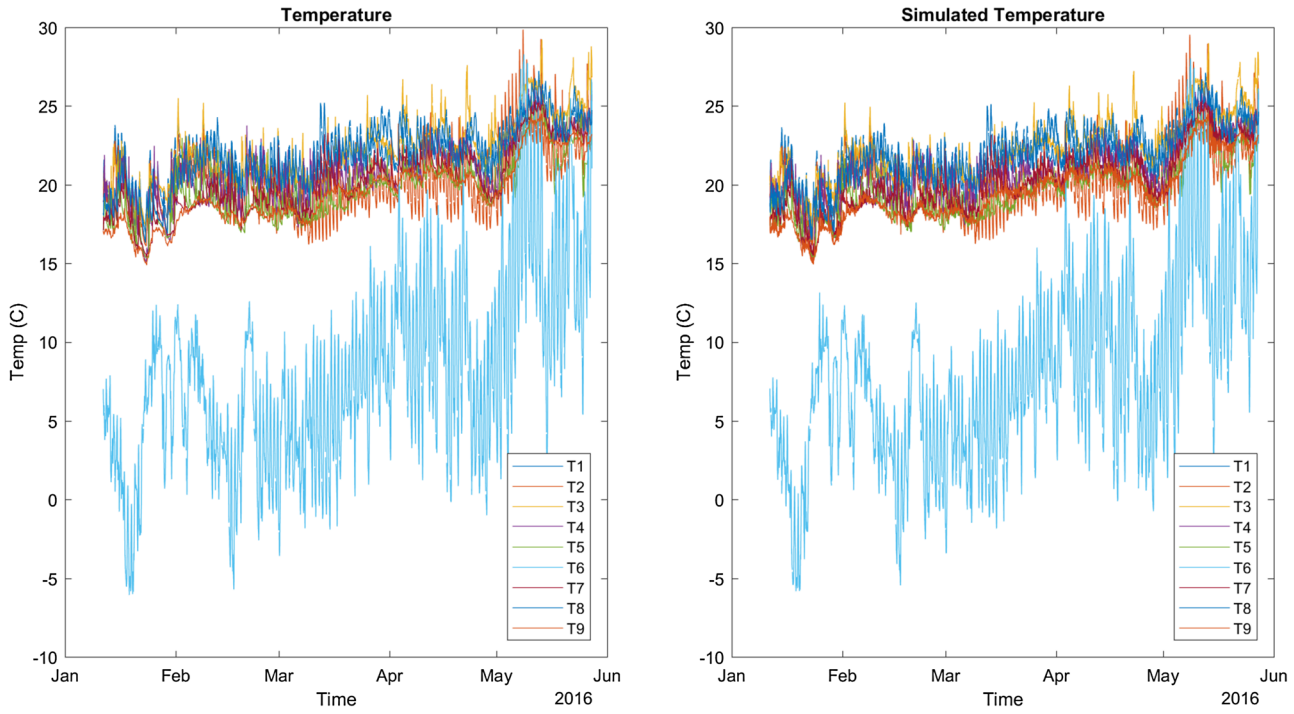

Fig. 10 Real and simulated temperature data

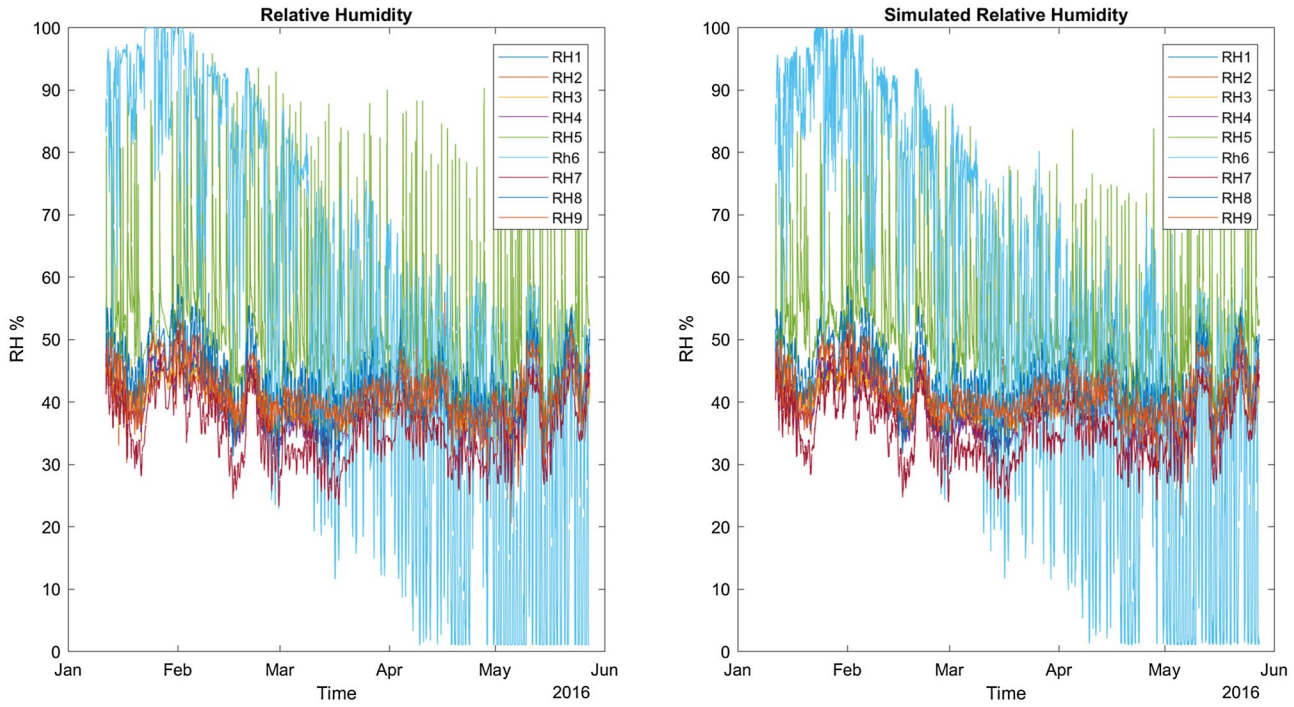

Fig. 11 Real and simulated relative humidity data

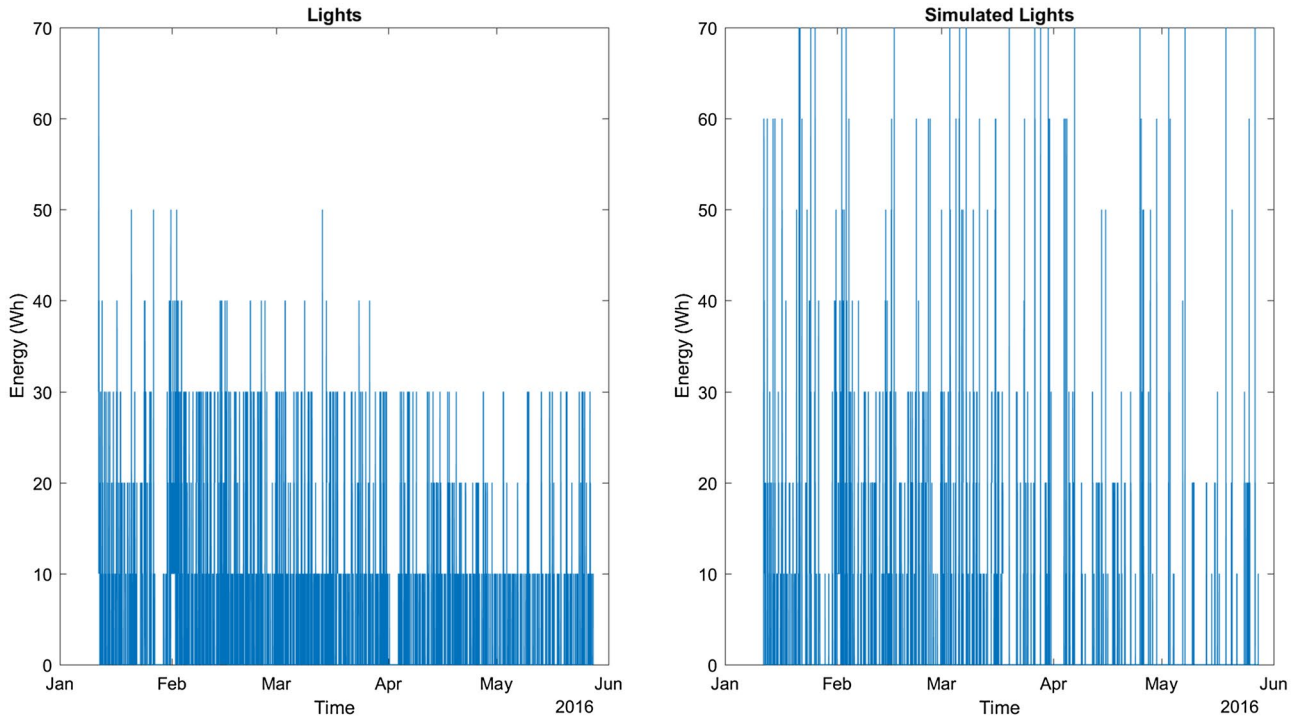

Fig. 12 Real and simulated light data

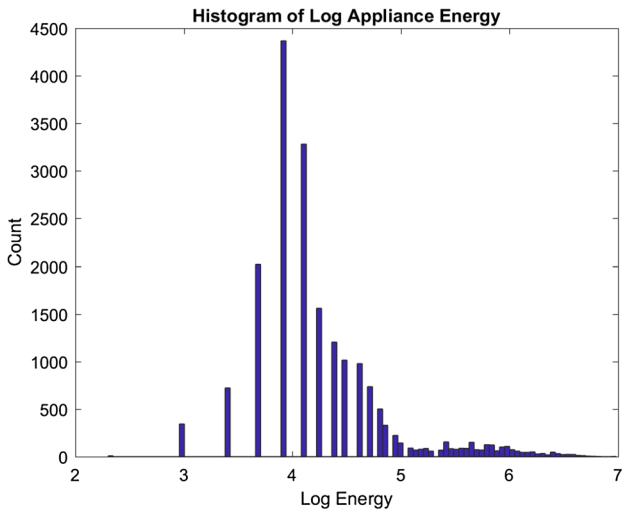

Fig. 13 Histogram of the log of appliance energy consumption data

in the first stage of the appliance simulation. These labels are also used to estimate the parameters of the two Gaussian distributions used to generate the appliance energy value in the log domain.

In the presented simulated data, the temperature parameter in the Boltzmann distribution is 0.1, and the random forest contains 200 trees. The conditional Gaussian distribution for $Z = 1$ has mean $\mu_1 = 4.1$ and $\sigma_1 = 0.39$, and the conditional Gaussian distribution for $Z = 2$ has $\mu_2 = 5.69$ and $\sigma_2 = 0.42$. The real and simulated appliance energy data is displayed in Fig. 14. To help validate the simulation, we trained a random forest on both the real data and the simulated data. The RMSE and variable importance was estimated for each random forest and compared (Fig. 15). The RMSE is lower for the simulated data which is not surprising given the relatively low amount of noise added to the simulated data. The variable importance measures are similar. The time of day and the energy consumed by the lights are the two most important features in both the real setting and the simulated data.

In our final experiment with the simulation, we sequentially increase the amount of noise in the simulation and evaluate the predictive ability of data-driven models. This experiment represents one of the many uses of the simulation. Increased error in collected data can stem from many sources. As sensors degrade, the error in the collected measurements may increase. As another example, low-quality sensors may have less fidelity. This experiment evaluates the ability for the data-driven models to cope with increasingly noisy measurements.

Each parameter that controls randomness or noise in the simulation is increased sequentially from 0.1 to 1 in increments of 0.1. A cross-validation experiment is conducted at each level using a GBM model. The simulated data is generated in MATLAB, but the cross-validation experiments are

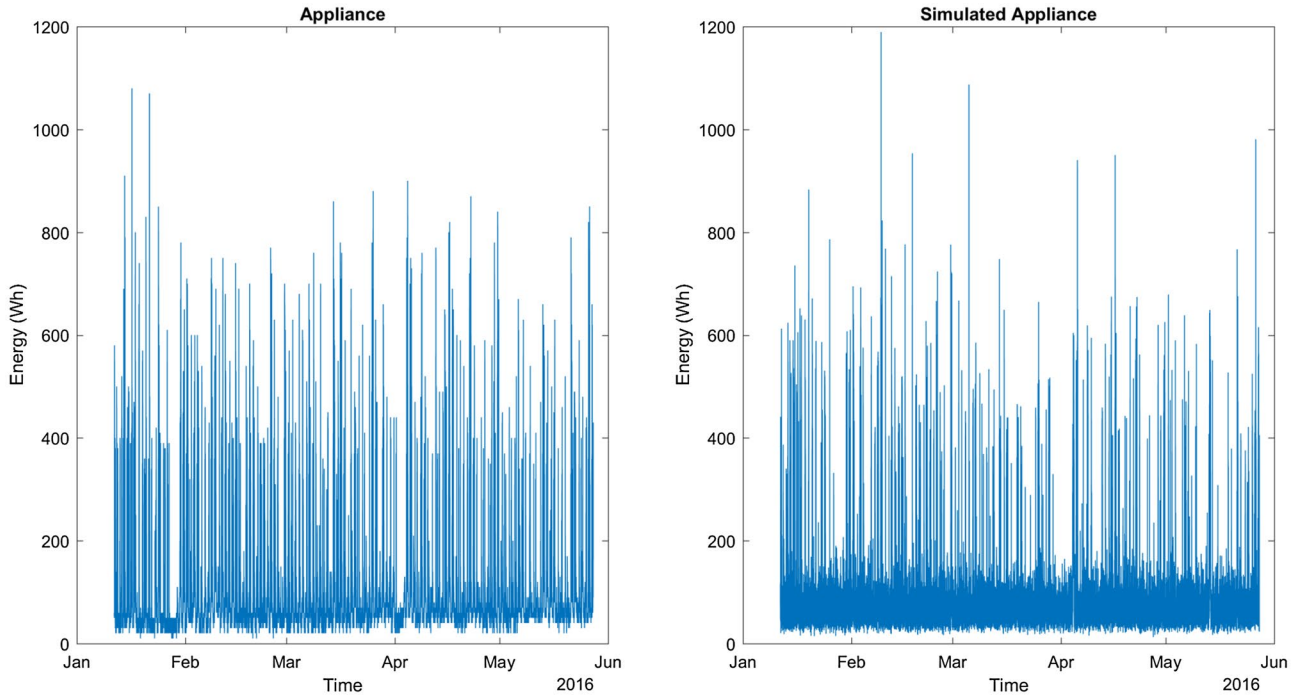

Fig. 14 Real and simulated appliance data

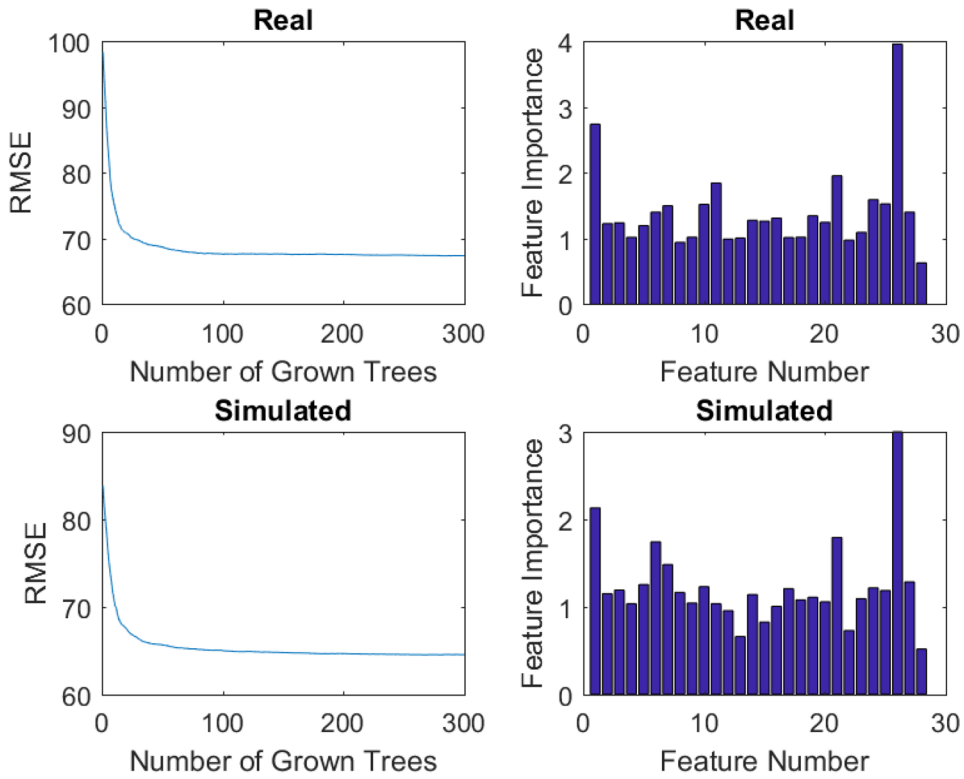

Fig. 15 RMSE and variable importance for real and simulated data

conducted in R. Figures 16, 17, 18, 19 contain the results of these experiments. In these figures, "beta" represents the temperature parameter for the appliances, "alpha" represents the temperature parameter for the lights, "sigmaT" represents the standard deviation for the temperature, and "sigmaRH" represents the standard deviation for the

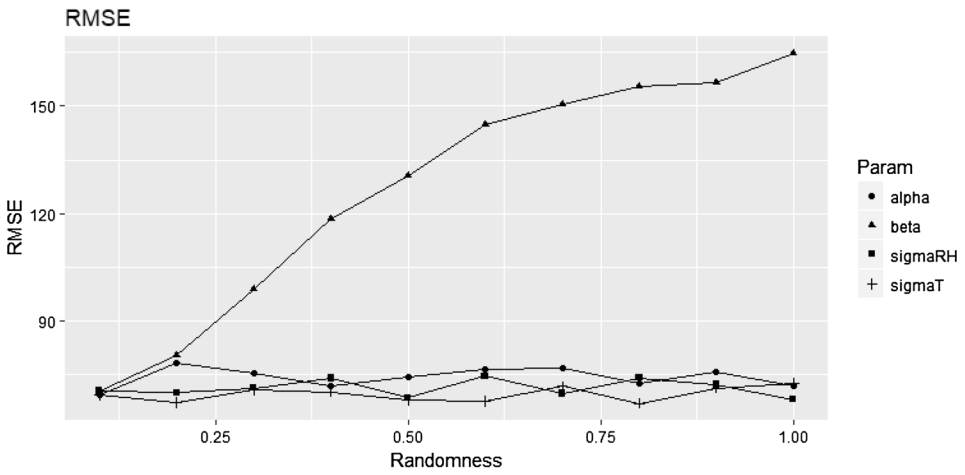

Fig. 16 RMSE metric as randomness is increased for each parameter in the simulation

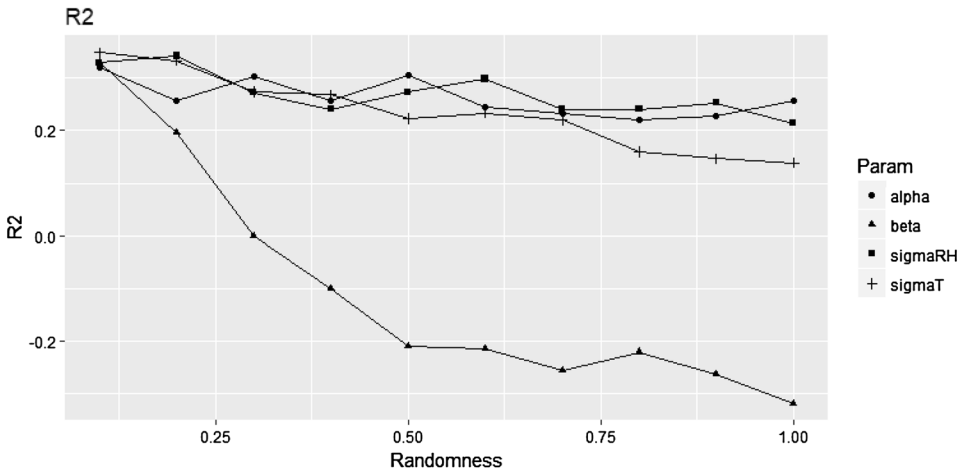

Fig. 17 R2 metric as randomness is increased for each parameter in the simulation

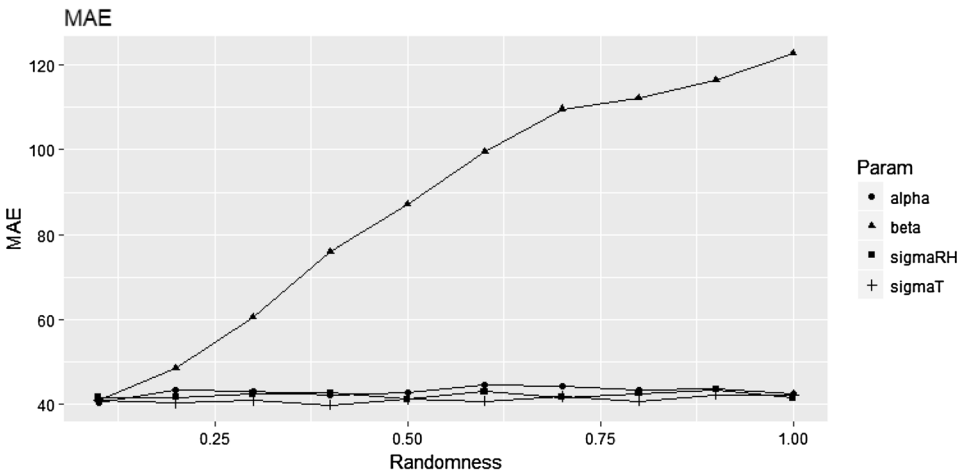

Fig. 18 MAE metric as randomness is increased for each parameter in the simulation

relative humidity. The plots show that performance is only decreased by increasing the amount of noise in the appliance random variable. However, we only test on a limited range, and increasing the noise outside of the example could have a greater impact on the simulation results.

5 Conclusion

In conclusion, we have developed and validated a simulation-based method for evaluating sensor quality in

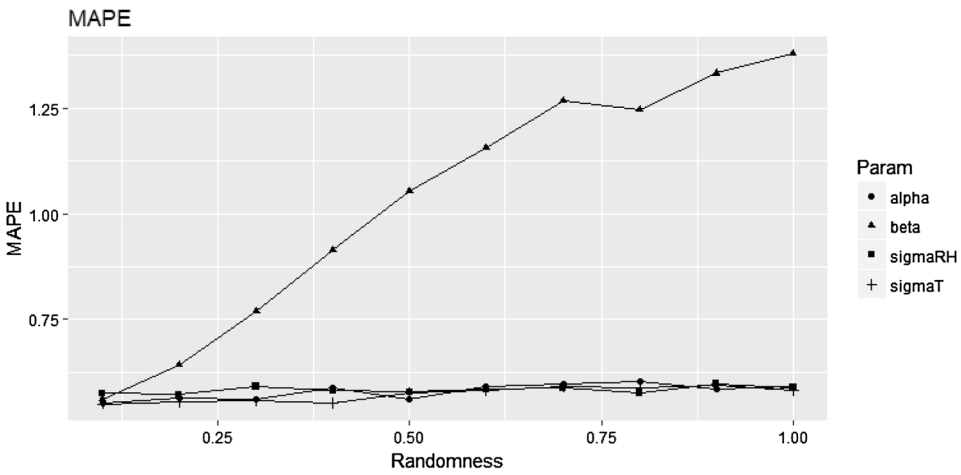

Fig. 19 MAPE metric as randomness is increased for each parameter in the simulation

a smart-home setting. The proposed method uses data-driven models as the underlying models in a simulation that can generate artificial data sets with varying amounts of noise. In addition, we perform new data-driven modeling in MATLAB with slightly different predictive models and perform a new feature selection analysis. Further, we demonstrated that the other collected variables in the data set could be modeled using similar data-driven techniques, specifically the random forest algorithm.

Simulation data with increasing amounts of noise is used to test the predictive ability of the data-driven models as randomness is added to the data. The predictive ability is only decreased when noise is added to the appliance energy random variable. This type of information could be useful when selecting IoT devices for monitoring a smart home. Low-quality sensors that do not monitor the environment as accurately as the devices used in the original study could be used for humidity and temperature without significantly reducing the predictive ability of the data-driven models.

In future work, the presented method should be tested and validated on data sets from other climates and residential constructions. One possible data set is the previously mentioned Pecan Street data set. Other areas of future work could use this simulation to produce test data for several homes. Larger simulations using this simulation process could model neighborhoods, cities, states, or nations. However, the data used as the basis for the presented study was from a single home with a fixed number of occupants. Future work involving scaling the simulation would need to account for other types of dwellings and varying sizes of residences. Further, any future work on scaling the simulation would need to account for a varying types of occupancy, including the number of occupants and diverse schedules.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. CNS: 1650512. This work was conducted in the NSF UICRC Center of Visual and Decision Dynamics, through the sponsorship and guidance of CA Technologies. We thank people on the team who choose not be authors for their thoughts on this paper and the overall project.

References

Ahmad MW, Mourshed M, Yuce B, Rezgui Y (2016) Computational intelligence techniques for HVAC systems: a review. Build Simul 9(4):359–398

Allard I, Olofsson T, Nair G (2018) Energy evaluation of residential buildings: performance gap analysis incorporating uncertainties in the evaluation methods. Build Simul 11(4):725–737

Alliance Z (2006) Zigbee specification

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

Belafi Z, Hong T, Reith A (2017) Smart building management vs intuitive human controllessons learnt from an office building in Hungary. Build Simul 10(6):811–828

Breiman L (2001) Random forests. Machin Learn 45(1):5–32

Candanedo LM, Feldheim V, Deramaix D (2017) Data driven prediction models of energy use of appliances in a low-energy house. Energy Build 140:81–97

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

Chen J, Ma J, Lo S (2017) Event-driven modeling of elevator assisted evacuation in ultra high-rise buildings. Simul Model Pract Theory 74:99–116

Chen Y, Hong T, Luo X (2018) An agent-based stochastic occupancy simulator. Build Simul 11(1):37–49

Costa A, Keane MM, Torrens JI, Corry E (2013) Building operation and energy performance: monitoring, analysis and optimisation toolkit. Appl Energy 101:310–316

Ding Y, Yang L, Weng F, Fu Z, Rao P (2015) Investigation of combined stairs elevators evacuation strategies for high rise buildings based on simulation. Simul Model Pract Theory 53:60–73

D'Oca S, Fabi V, Corgnati SP, Andersen RK (2014) Effect of thermostat and window opening occupant behavior models on energy use in homes. Build Simul 7(6):683–694

Dong B, Cao C, Lee SE (2005) Applying support vector machines to predict building energy consumption in tropical region. Energy Build 37(5):545–553

Draper NR, Smith H (2014) Applied regression analysis. Wiley, Hoboken

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Hong T, Langevin J, Sun K (2018) Building simulation: ten challenges. Build Simul 11:1–18

Johnson G, Beausoleil-Morrison I (2017) Electrical-end-use data from 23 houses sampled each minute for simulating micro-generation systems. Appl Therm Eng 114:1449–1456

Kavousian A, Rajagopal R, Fischer M (2015) 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 99:220–230

Lichman M (2013) UCI machine learning repository. UNiversity of California, Irvine

Marchese D, Linkov I (2017) Can you be smart and resilient at the same time? Environ Sci Technol 51(11):5867–5868

Mirakhorli A, Dong B (2017) Occupant-behavior driven appliance scheduling for residential buildings. BUild Simul 10(6):917–931

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

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press, Cambridge

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Ruellan M, Park H, Bennacer R (2016) Residential building energy demand and thermal comfort: thermal dynamics of electrical appliances and their impact. Energy Build 130:46–54

Sano T, Ronchi E, Minegishi Y, Nilsson D (2018) Modelling pedestrian merging in stair evacuation in multi-purpose buildings. Simul Model Pract Theory 85:80–94

Seem JE (2007) Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy Build 39(1):52–58

Shaikh PH, Nor NBM, Nallagownden P, Elamvazuthi I, Ibrahim T (2014) A review on optimized control systems for building energy and comfort management of smart sustainable buildings. Renew Sustain Energy Rev 34:409–429

Trčka M, Hensen JL, Wetter M (2010) Co-simulation for performance prediction of integrated building and HVAC systems-an analysis of solution characteristics using a two-body system. Simul Model Pract Theory 18(7):957–970

Wang C, Yan D, Jiang Y (2011) A novel approach for building occupancy simulation. Build Simul 4(2):149–167

Xue P, Hong T, Dong B, Mak C (2017) A preliminary investigation of water usage behavior in single-family homes. Build Simul 10(6):949–962

Yang Z, Li N, Becerik-Gerber B, Orosz M (2014) A systematic approach to occupancy modeling in ambient sensor-rich buildings. Simulation 90(8):960–977

Zhao P, Suryanarayanan S, Simões MG (2013) An energy management system for building structures using a multi-agent decision-making control methodology. IEEE Trans Ind Appl 49(1):322–330