July 2018

Data Mining of Smart WiFi Thermostats to Develop Multiple Zonal Dynamic Energy and Comfort Models of a Residential Building

Kefan Huang
Department of Mechanical and Aerospace Engineering / Renewable and Clean Energy, University of Dayton, Dayton, OH, 45469, United States of America, huangk2@udayton.edu

Abdulrahman Alanezi
Department of Mechanical and Aerospace Engineering / Renewable and Clean Energy, University of Dayton, Dayton, OH, 45469, alanezia1@udayton.edu

Kevin Hallinan
Department of Mechanical and Aerospace Engineering / Renewable and Clean Energy, University of Dayton, Dayton, OH, 45469, kevin.hallinan@udayton.edu

Robert Lou
Department of Mechanical and Aerospace Engineering / Renewable and Clean Energy, University of Dayton, Dayton, OH, 45469, louy01@udayton.edu

Data Mining of Smart WiFi Thermostats to Develop Multiple Zonal Dynamic Energy and Comfort Models of a Residential Building

Kefan Huang[1*], Abdulrahman Alanezi[1], Kevin P. Hallinan[1], Robert Lou[1],

[1]Department of Mechanical and Aerospace Engineering / Renewable and Clean Energy, University of Dayton, Dayton, OH, 45469

[*]Corresponding Author: huangk2@udayton.edu

ABSTRACT

Smart WiFi thermostats have gained an increasing foothold in the residential building market. The data emerging from these thermostats is transmitted to the cloud. Companies are attempting to use this data to add value to their customers. This overarching theme establishes the foundation for this research, which seeks to utilize smart WiFi thermostat data from individual residences to develop a dynamic model to predict real time cooling demand and then apply this model to 'what-if' thermostat scheduling scenarios. The ultimate goals of these efforts are to reduce energy use in the residence and/or demonstrate the ability to respond to utility peak demand events. A regression tree approach (Random Forest) was used to develop models to predict the room temperature as measured by each thermostat and the cooling status. The models developed, when applied to validation data (e.g., data not employed in training the model) yielded R-squared values of greater than 0.98. The results from the 'what if' scenarios show a huge opportunity for quantifying cooling energy consumption reduction through the use of more aggressive non-occupied temperature setpoint schedules, as well as the total time that cooling/heating could be interrupted in responding to a high demand event while maintaining thermal comfort within acceptable ranges.

1. INTRODUCTION

About 40% (or about 39 quadrillion British thermal units) of total U.S. energy consumption was consumed by the residential and commercial sectors (eia, 2017). More concerted efforts to increase energy efficiency in commercial and residential buildings have led to substantial reductions in energy use per square foot. Yet while new buildings can immediately take advantage of state-of-the-art energy efficient technologies, existing buildings would require comprehensive, deep retrofits to achieve similar savings. Climate control systems for managing heating and cooling systems for buildings (heating, ventilation and cooling, or HVAC systems) have been in existence for decades. In residential buildings, these controllers are referred to as thermostats. At the most basic level, a thermostat includes a means to allow a user to set a desired temperature, sense actual temperature, and control the heating and/or cooling devices in order to maintain the actual temperature to be nearly equal the desired setpoint temperature. Today, these thermostats use solid-state devices such as thermistors or thermal diodes to measure temperature, and humidity sensors for measuring humidity, and microprocessor-based circuitry to control room temperature and to store and operate based upon user-determined protocols for temperature versus time. Smart WiFi thermostats communicate this information to the cloud, where additional processing is possible. Some thermostats may recognize the pattern of use of residents and manages the comfort based upon the recognized pattern.

Since the energy consumption in buildings depend upon so many different energy characteristics and user behaviors, development of comprehensive physics-based models is difficult at best (Kesik, 2016). On the other hand, Machine Learning Techniques have comparatively offered better model accuracy (Ben-David & Frank, 2009). Recently, one team from Lille University used the Artificial Neural Network machine learning algorithm to forecast indoor temperature (Attoue, Shahrour, & Younes, 2018). The input parameters for their indoor prediction model consisted of outdoor temperature, outdoor humidity, solar radiation, outdoor temperature history (previous outdoor temperature), time and facade temperature history. This team demonstrated that the indoor temperature forecasting could be conducted with good precision considering outdoor temperature and indoor facade temperature history. The results of

their work showed the performance of indoor temperature predictions were relatively good for up to two hours. But, the four-hour prediction was unsatisfactory. Their predictions likely suffered from the absence of indoor humidity, indoor temperature history (previous indoor temperature) and thermostat setpoint in their models.

Another team from University of Valencia and Jaume I University tried to use different kinds of regression methods including Multiple Linear Regression (MLR), Autoregressive (AR), Extreme Learning Machine (ELM), Non-linear Autoregressive Exogenous models (NARX) and Multilayer Perceptron with Non-linear Autoregressive Exogenous (MLP-NARX) to forecast temperature in buildings and compared results to select the best method performance. The parameters they used to develop their prediction model were month, day of the month, official time, room relative humidity (%), outside temperature, room setpoint temperature, total thermal power, and current room temperature. Figure 1 shows the prediction of the temperature with MLP NARX models for one room on January (Mateo, et al., 2013). It is clear that the accuracy of the MLP-NARX prediction was exceptional, but it wasn't clear how the model would perform for variable setpoint schedules.

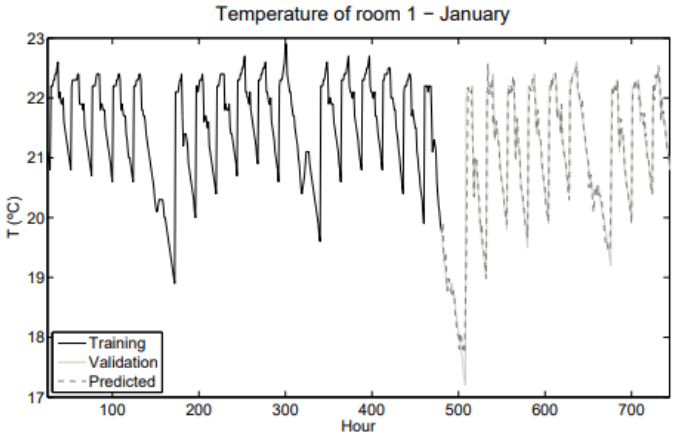

Figure 1: Prediction of the temperature with MLP NARX models for one room on January (Mateo, et al., 2013).

There is real opportunity to improve the accuracy of the models by accounting for additional exterior weather factors such as temperature, relative humidity, wind speed and direction, cloud cover, etc… or other internal factors such as the current and prior relative humidity, cooling and/or heating system status, and internal temperature and humidity for other zones when there are multiple thermostats in a residence. The latter can help to account for cross-zone interactions. There is also opportunity to manage for comfort rather than just temperature. Comfort accounts for not at least a combination of temperature and humidity (Simion, Socaciu, & Unguresan, 2016). For example, if the humidity is lower in the summer, the internal temperature can be increased to maintain constant comfort.

2. METHODOLOGY

2.1 Data used

The indoor characteristic data and outdoor weather data used for development of dynamic and comfort models was collected from smart WiFi thermostats present in 50 homes located in the Midwest of the US, all managed by a local university. The housing set analyzed includes a diversity of houses, with construction years ranging from the early 1900s to current and with square footages ranging from 75.5 to 177.7 m^2. While residences less than five years old have been built to U.S. Energy Star specifications, older homes have received variable attention relative to insulation upgrades. Some have been upgraded to Energy Star specifications. Each home has a two stage central air conditioning and natural gas heating. Thus, the dynamics associated with the homes have significant variation.

Smart WiFi data from these homes has been collected continuously and archived since July 2017. Additionally, local weather station data has been accessed and archived. The data and analysis for only one home is presented here.

2.2 Data processing

The acquired variables require preprocessing before the construction of model construction and validation (Makridakis, Wheelwright, & Hyndman, 1998). First, the WiFi thermostat data had to be synched with the outdoor weather data.

Sample data from one of the WiFi thermostats are shown in Table 2 below. The cooling status variable was indicated as 0 when the compressor was off, 50 at compressor partial load, and 100 at compressor full load.

Table 2: Sample merged WiFi and other acquired data

Indoor Temp (C)	Cooling Setpoint (C)	Cooling Status	Indoor Humidity (%)	WiFi Strength	Human Time	Outdoor Temp (C)	Outdoor Humidity (%)
20.73	20.6	0	60	41	9/16/2017 23:17	16.6	97
21	20.6	100	63	42	9/16/2017 23:26	16.6	97
20.625	20.6	0	60	42	9/16/2017 23:37	16.6	97
…	…	…	…	…	…	…	…

Second, feedback variables (previous indoor temperature and previous outdoor temperature and humidity, previous cooling status, and previous thermostat setpoints) were added to each observation, as previous research had shown the importance of using feedback (Drucker, Shahrary, & Gibbon, 2001). Third, the time since the last reading was added to each observation. This was essential because the WiFi thermostats only record data when there is a change in the cooling or heating status or setpoint temperature.

Figure 2 shows a histogram of the cooling setpoint temperature for one house over the study period. It is clear that there is not a uniform distribution. Thus, any model developed using the obtained data would likely bias the most prevalent setpoint conditions. Thus, the third step in the process was to upsample the data in order to have an equal number of observations for each cooling setpoint value.

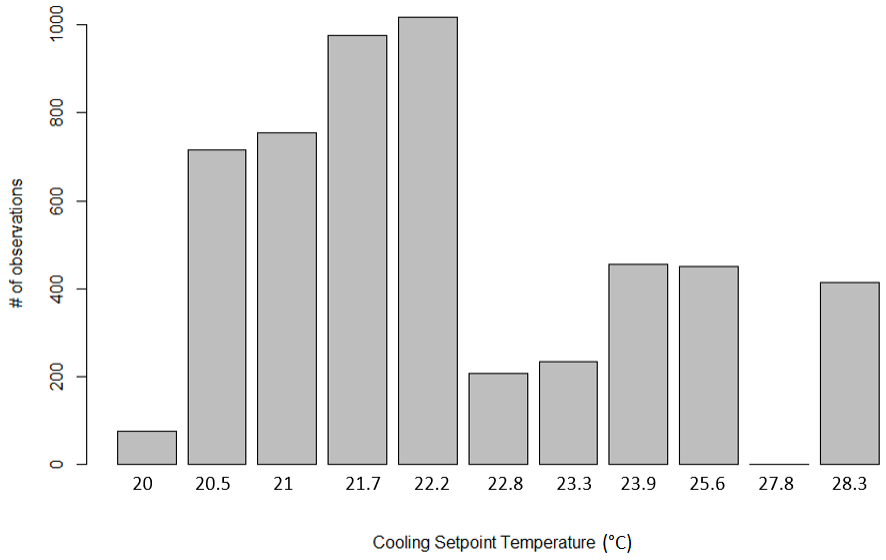

Figure 2: Probability density plot of the cooling setpoint temperature for a representative house

2.3 Model development

The goal was to develop separate models to predict the indoor temperature and the cooling status for each thermostat in each house using a random sampling of the training data (70% of the complete data was used for training). A Random Forest approach was employed for both models. Random Forest is a classification and regression tree approach, where an ensemble of trees are developed, with each tree accounting for a random selection of the training

observations and input factors (Breiman, 2001). The resulting model effectively averages the models developed, weighted by the error in prediction. This approach has been shown to yield excellent performance in settings where the number of variables is much larger than the number of observations, can cope with complex interaction between factors as well as highly correlated variables, minimize oversampling, and return measures of variable importance (Boulesteix, Janitza, Kruppa, & R. König, 2012). Tables 3 and 4 show the model inputs and output variables for respectively the indoor zonal temperature prediction. Moreover, the structure of the Random Forest models is described in Figure 3.

Table 3: Variables used in indoor temperature model

Variables	Input	Output	Ignore
Indoor Temperature (C)		X	
Indoor Relative Humidity (%)	X		
Cooling Status (0,50,100)	X		
Cooling Setpoint (C)	X		
WiFi Strength			X
Human Time			X
Seconds (s)			X
Time Since Last Reading (s)	X		
Previous Indoor Temperature (C)	X		
Previous Cool Status (0,50,100)	X		
Previous Cooling Setpoint (C)	X		
Outdoor Temperature (C)	X		
Outdoor Relative Humidity (%)	X		

Table 4: Variables used in the cooling status model

Variables	Input	Output	Ignore
Indoor Temperature (C)	X		
Indoor Relative Humidity (%)	X		
Cooling Status (0,50,100)		X	
Cooling Setpoint (C)	X		
WiFi Strength			X
Human Time			X
Seconds (s)			X
Time Since Last Reading (s)	X		
Previous Indoor Temperature (C)	X		
Previous Cool Status (0,50,100)	X		
Previous Cooling Setpoint (C)	X		
Outdoor Temperature (C)	X		
Outdoor Relative Humidity (%)	X		

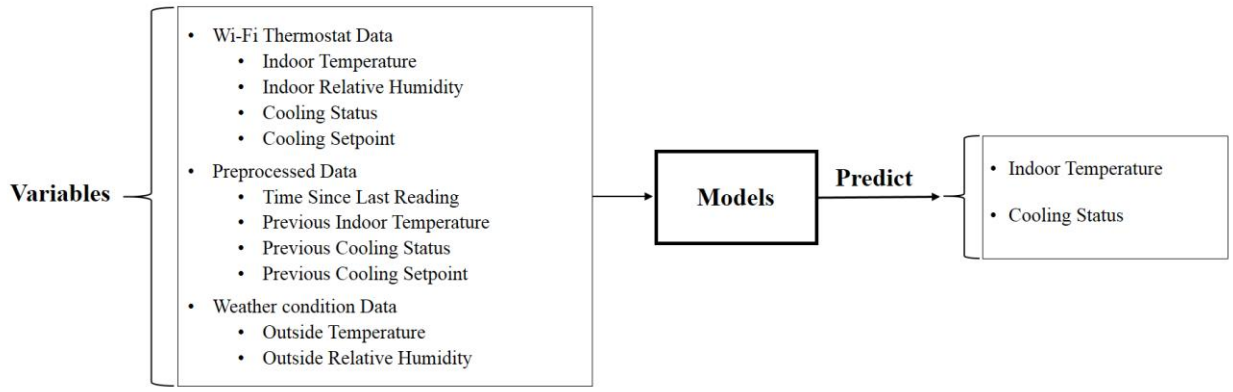

Figure 3: Dynamic and comfort models structure

All models utilized 500 trees and 5 interrogation variables.

2.4 What if analysis

What-if analysis is a data intensive simulation with the goal to inspect the behavior of a complex system under some given hypotheses called scenarios. In particular, what-if analysis measures how changes in a set of independent or control variables impact a set of dependent variables with reference to a given simulation model (Rizzi, 2009). A what-if analysis first requires the establishment of a model. With a model developed, new independent or control variable values can be interrogated.

The ultimate goal of the what-if thermostat scheduling scenarios implemented here is to reduce energy use in the residence or to estimate the effect of responding to high demand events while maintaining desired comfort within acceptable bounds. The effect of different setpoint schedules, ideally linked to zonal occupancy schedules and desired comfort on energy consumption can be evaluated. Table 5 shows three different cooling setpoint schedules for one day considered in the what-if analysis. Case 1 represents the baseline case where there is no setpoint variation. The other cases are associated with setpoint scheduling to achieve energy savings. The intent is to show the value of setpoint scheduling in reducing the energy consumption. The exterior temperature condition for one day of testing is used in these what-if scenarios. The exterior condition considered is associated with a roughly sinusoidally varying temperature with a mean of 21.3 C, and amplitude of 11.1 C. It should be noted that in this house, as is obvious from Figure 2, the residents living there did in fact establish a cooling setpoint of 20.6 C a significant amount of time.

Table 5: Three different cooling setpoint schedules for one day

	Cooling setpoint schedules for one day (C)		
	12:00 am to 7:00 am	7:00 am to 5:00 pm	5:00 pm to 12:00 am
Case 1	20.6	20.6	20.6
Case 2	20.6	22.2	20.6
Case 3	20.6	23.9	20.6

A what-if analysis was also conducted to simulate demand response to grid requests to reduce demand, while ideally staying within a minimum tolerable comfort zone. For this case (Case 4), the outdoor temperature varied from 25.6C to 26.7C over a 14 hour period. The cooling status was set to off during a period of setpoint increase to simulate a demand response event. This case was actually implemented in one of the houses to see the response of the house temperature to a demand response event. The actual data will permit validation of the predicted data using the developed model.

RESULTS

3.1 Model results

In developing the model, 70% of the data obtained from implementation of the test matrix shown in Table 1 was used for training the model. The remaining data was used for validation of the model. The R-squared coefficient of determination (R-squared value) was utilized to evaluate the model accuracy. The R-squared value in machine learning has the capacity to give definite elucidation with respect to rightness of the model (D. Rajeswara Rao, 2016). Figure 4 shows the performance of the indoor temperature model based upon R-squared value. This figure shows exceptional correlation between the model predictions and actual temperatures, yielding an R-squared value of 0.9905.

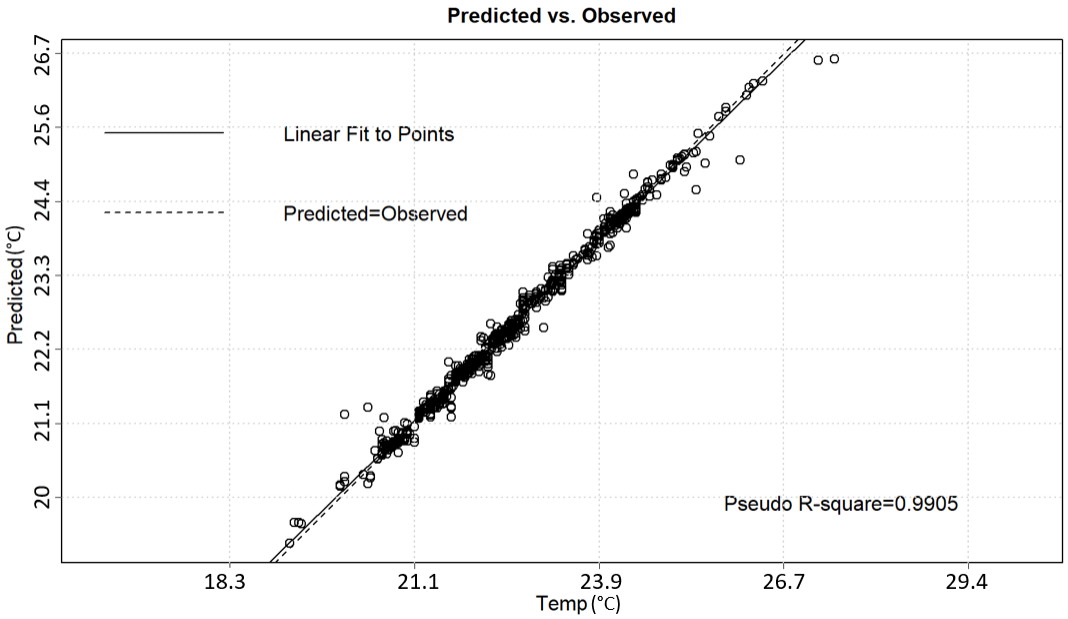

Figure 4: Predicted versus observed values and R-squared value of indoor temperature model

Table 7: Summary of model performance for temperature and cooling status

R-squared Value	
Temperature Model	Cooling Status Model
0.9905	0.9835

3.2 What-if analysis results

With models developed, the what-if scenarios described in Section 2.4 could be conducted. The what-if analysis inputs include only knowledge of the independent or control parameters at all future times and initial conditions for the target variables (temperature and cooling status). The dependent variables (zonal temperature and cooling status) are not known for the remainder of the times. These have to be computed at each time step. Then the computed values at a given time step are used as feedback values for the next time step. As a result, the error in predicting the next time step values is additive. Expectedly, the model predictions should worsen with time.

Figures 5 - 7 show the predicted room temperature for the first three what-if cases considered. Figure 5 shows the predicted temperature, cooling status, and exterior temperature for the constant setpoint schedule associated with Case 1. This figure shows both plausible indoor temperatures and cooling status, and it also shows that the cooling system is undersized. It is not able to maintain the desired setpoint temperature, despite a 100% duty cycle for the air conditioner. In comparison, Figure 6 shows the same parameters for equivalent exterior weather conditions and with moderate temperature scheduling. It's clear that the cooling duty cycle required has been reduced. Figure 7 shows the same parameters for the most aggressive non-occupied set-point temperature. Again, there is a clear reduction in cooling required. Also interesting is that at the higher setpoint temperature, the cooling status remains at half power for much of the time.

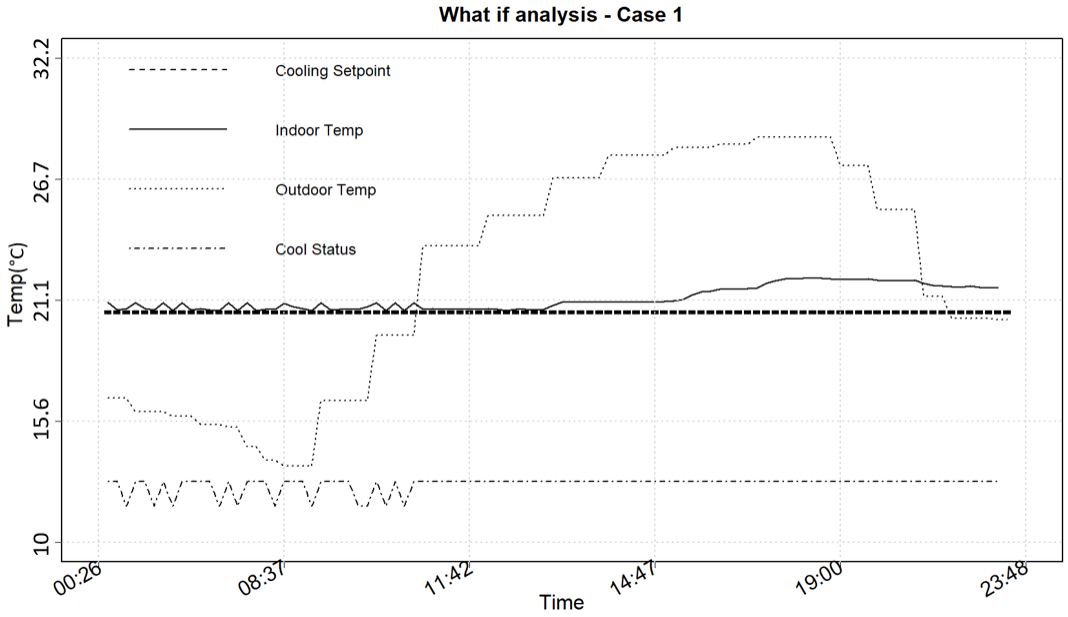

Figure 5: Case 1- What if analysis of original cooling setpoint (constant 20.6 C)

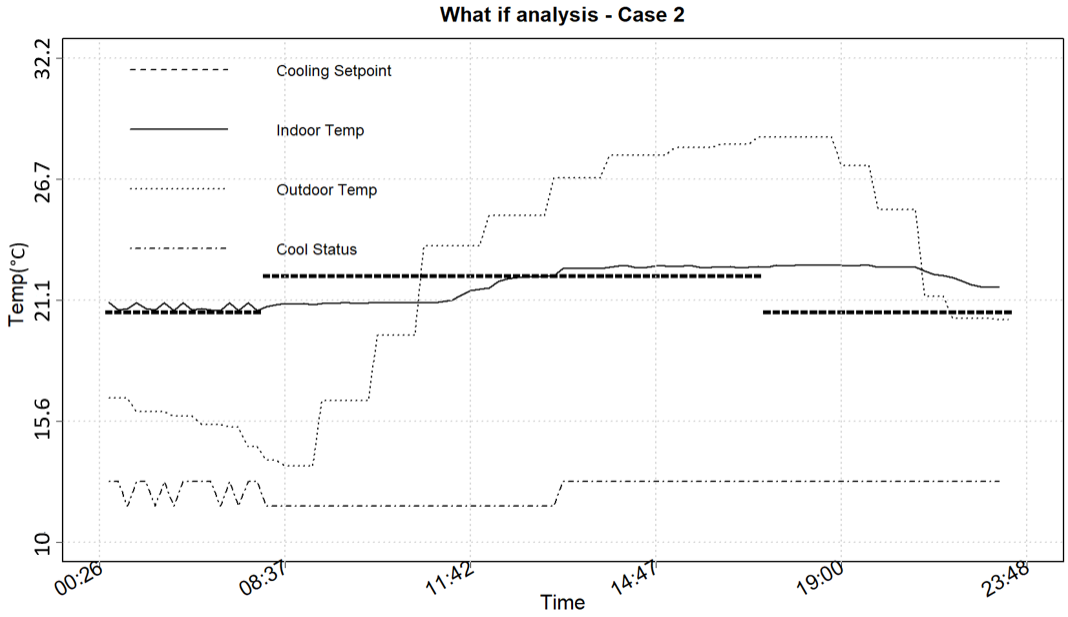

Figure 6: Case 2- What if analysis of small scale of cooling setpoints (20.6 C and 22.2 C)

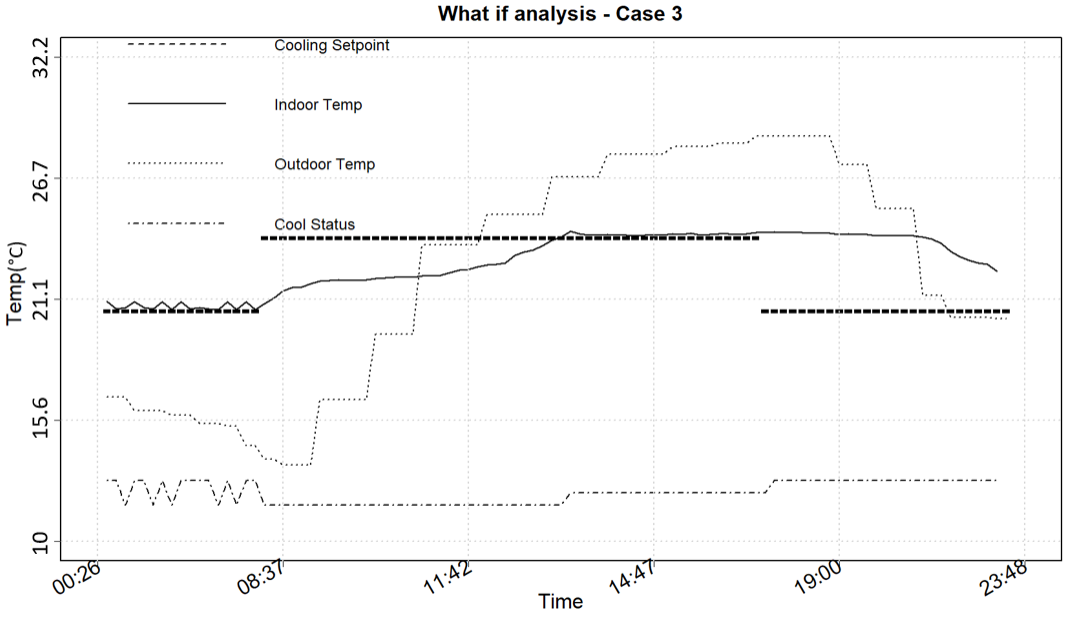

Figure 7: Case 3- What if analysis of large scale of cooling setpoints (20.6 C and 23.9 C)

The effect of the what-if temperature set-point scheduling on total energy consumption can be determined by calculating the effective duty cycle for cooling over a 24 hour period. Equation 1 shows how this effective duty cycle is calculated.

$$\text{effective duty cycle} = \frac{\sum(\text{cool time})\ x\ (\frac{\text{cool status}}{100})}{\sum \text{time}} \quad (1)$$

Table 8 shows the effective duty cycle and the respective cooling energy savings for each of the cases relative to the baseline case for constant setpoint temperature (Case 1). Case 2 renders a significantly lower effective duty cycle than the baseline case resulting in 18.4% reduction in cooling energy. Case 3 yields even more improvement, with a reduction of 28% relative the baseline case. It should be noted that the savings are high, in part because of the low baseline cooling setpoint temperature considered.

Table 8: Summary of what if analysis three cases results

	Total time of a day when cooling status is ON (sec)	Effective Duty Cycle (%)	Cooling energy savings (%)
Case 1	77,340	91.9	-
Case 2	60,780	73.5	18.4
Case 3	52,860	63.9	28

Figure 8 shows the Case 4 what-if scenario associated with cooling disruption for conditions which would otherwise have called for cooling. Presented in this case is both the predicted and actual room temperature, as well as cooling status and outdoor temperature. For this case, the thermostat setpoint temperature is increased at a time of roughly 16:00, at which time the cooling status is set to zero. Were a minimum comfort condition of 26.7°C established for the residence, the cooling could be interrupted for over eight hours given the weather conditions which would be forecast. This demand reduction could be communicated to the utility by the thermostat manager. Note also in the figure the excellent correspondence between the predicted temperature (thin solid line) and actual temperature (dotted line). This correspondence helps to validate the 'rightness' of this forecasting approach.

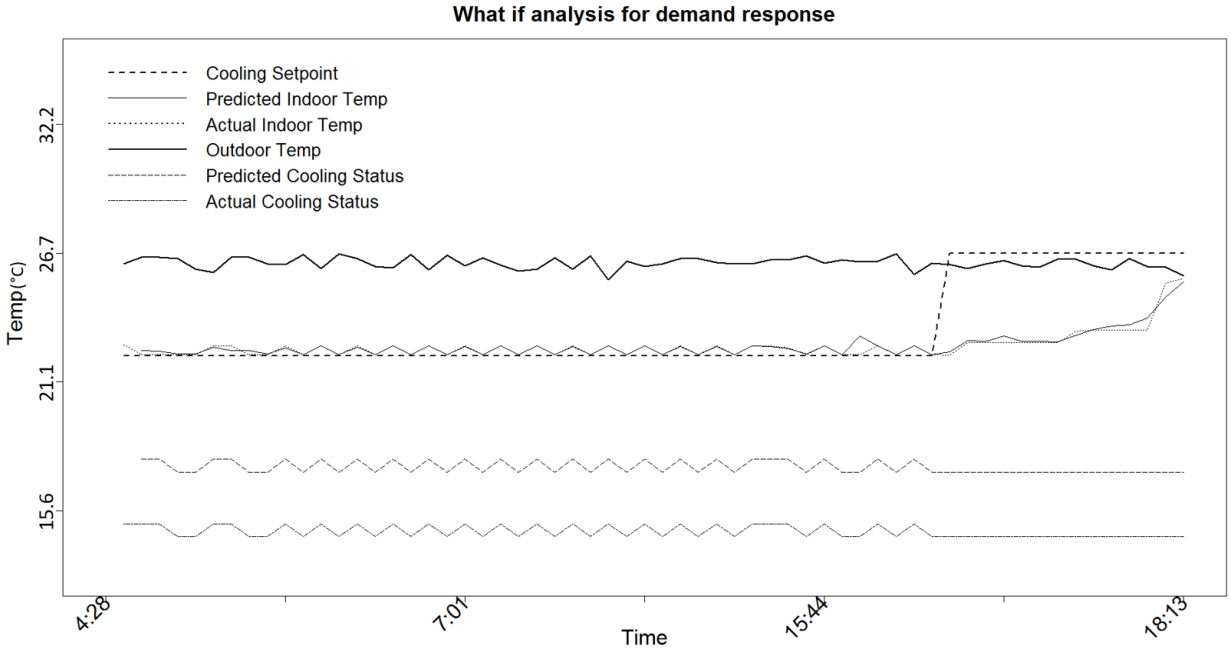

Figure 8: What-if scenario for demand response comfort estimation

3. CONCLUSIONS

This research shows the value of using historical smart WiFi data to model the dynamic response of the residence. Cloud-based calculations combining the thermostat data and historical weather conditions can be used to develop the models. Moreover, the developed models can be applied to interrogate the energy savings benefits of more aggressive set-point scheduling. These potential savings can be communicated to residents. Thus the thermostat manager can add additional value to the resident. As importantly, the developed models can be used to estimate comfort within the residence were the cooling to be curtailed in response to a grid-requested demand reduction event. Using forecasted weather conditions. This service offers a future demand reduction in any residence, while maintaining minimal comfort within a residence, can be estimated. The value to a resident would be that were they to agree to such curtailment within minimum thermal comfort bands, they could benefit from a lower energy cost. Moreover a cloud-based smart WiFi thermostat manager, as a result of this innovation, will be able to provide large-scale grid demand reduction when called for. This ability is especially important as greater renewable energy is brought to the grid.

REFERENCES

Attoue, N., Shahrour, I., & Younes, R. (2018). Smart Building: Use of the Artificial Neural Network. *Energies*, 395.

Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus “hand crafted” expert systems – A credit scoring case study. *Expert Systems with Applications*, 5264-5271.

Boulesteix, A.-L., Janitza, S., Kruppa, J., & R. König, I. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 86–97.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

D. Rajeswara Rao, V. P. (2016). Machine Learning Techniques on Multidimensional Curve. *International Journal of Electrical and Computer Engineering (IJECE)* , 974 ~ 979.

Drucker, H., Shahrary, B., & Gibbon, D. (2001). Relevance Feedback using Support Vector Machines. *01 Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 122-129). San Francisco: Morgan Kaufmann Publishers Inc.

eia. (2017, May 10). *FREQUENTLY ASKED QUESTIONS*. Retrieved from eia: https://www.eia.gov/tools/faqs/faq.php?id=86&t=1

Kesik, T. J. (2016, August 4). *BUILDING SCIENCE CONCEPTS* . Retrieved from Whole Building Design Guide: https://www.wbdg.org/resources/building-science-concepts#refi

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications, 3rd Edition.* .John Wiley & Sons.

Mateo, F., J. Carrasco, J., Mill´an-Giraldo, M., Sellami, A., Escandell-Montero, P., M. Mart´ınez-Mart´ınez, J., & Soria-Olivas, E. (2013). Temperature Forecast in Buildings Using Machine Learning Techniques. *European Symposium on Artificial Neural Networks, Computational Intelligence.* Bruges (Belgium): ESANN.

Nest. (2018, January 16). *Thermostat*. Retrieved from Nest Developers: https://developers.nest.com/documentation/cloud/thermostat-guide

Rizzi, S. (2009). What-if analysis. In L. Liu, & M. T. Özsu, *Encyclopedia of Database Systems* (pp. 3525-3529). Boston: Springer US.

Simion, M., Socaciu, L., & Unguresan, P. (2016). Factors which Influence the Thermal Comfort Inside of Vehicles. *Energy Procedia*, 472-480.

ACKNOWLEDGEMENT

This work has been supported by Emerson Climate Technologies under contract and guidance from Dr. Rajan Rajendran, Mr. Vijay Bahel and Mr. Brian Butler at the Helix Innovation Center, 40 Stewart St. Dayton, OH