Short-Horizon Probabilistic Models For Whole Home Motion Prediction Using Connected Thermostat Data

Brent Huchuk\textsuperscript{1}, Scott Sanner\textsuperscript{1}, William O’Brien\textsuperscript{2}
\textsuperscript{1}University of Toronto, Toronto, Canada
\textsuperscript{2}Carleton University, Ottawa, Canada

Abstract

As residential thermostats have become increasingly capable of sensing occupancy, reliably predicting future occupancy over short-term control horizons becomes a key mechanism for improving the efficiency of heating, ventilation, and air conditioning (HVAC) systems. Past research has established successful conditional dependencies in modeling and simulating future occupancy states for residential applications based on manually collected or sensed data. These studies often are small-scale tests with limited use of commercially available data. Using a large dataset representing the relatively new class of devices called connected thermostats, two probabilistic methods were identified and implemented in hopes of identifying candidates for future control and simulation applications. Logistic regression and a Markov model were used to conduct inference on a defined model of motion detection. Testing was done on actual device data from 1000 real homes. Both methods were found to have average accuracies above 80\% in next timestep prediction. The Markov model performs better over a longer horizon (of up to 3 hours). In classification error types, the Markov model outperforms the logistic model both in terms of false positive and false negative rates. The results of the logistic regression are shown to improve with some simple encoding strategies.

Introduction

Thermostats are often considered a mechanism for both reducing energy and maintaining comfort. One basic method for achieving heating and cooling savings within the home, while not sacrificing comfort, is to engage in setback (or setup) periods when the home is unoccupied (Peffer et al., 2011). Traditionally this has been accomplished by programming a schedule in a thermostat that reflects the occupancy pattern of the residents. A correctly programmed thermostat has been shown to save 25\% of natural gas consumption for heating while maintaining comfort for the occupants (Nelson and MacArthur, 1978). Achieving this level of savings relies on the diligence of the homeowner or property manager to set up and maintain these schedules. Unfortunately, it has been found this ideal situation is often not the case (Nevius and Pigg, 2000; Pritoni et al., 2015; Peffer et al., 2011).

Previous research has attempted to adapt HVAC controls to the patterns of occupants. The ‘PreHeat’ algorithm (Scott et al., 2011) sought to predict occupancy and reduce the heating energy requirements for a home by finding the most similar days based on occupancy patterns in the home’s history. These occupancy patterns were based on observations from a combination of radio frequency and motion detection sensors. The algorithm was capable of reducing gas consumption in addition to reducing time spent in temperatures not in the desired range when occupied. Panagopoulos et al. (2015) developed the ‘AdaHeat’ system for heating control using the probabilistic methods from Scott et al. (2011) but using their own sensing data for occupancy state. Krumm and Brush (2011) utilized a GPS network to be able to generate probabilities that an occupant was home at certain times every day. These investigations rely on methods difficult to scale with hundreds of thousands of devices particular with the resources of edge computing (i.e., on the device). Further they often rely on sensor networks or devices which go beyond those currently seen in common residential thermostat and sensing network configurations.

The latest generation of thermostats, connected thermostats, maintain a link to external and complementary systems through an active wireless internet connection. With this communication ability, connected thermostats attempt to maximize the savings by simplifying the scheduling task and allowing control of setpoints, system modes, and current occupancy state of the home from access points other than the thermostat (Ford et al., 2017). The active adaptation to the current state of the home by the thermostat is done through manual adjustment or the use of various sensing technologies (Ford et al., 2017; Kleiminger et al., 2014). For example, if the home is scheduled to be occupied but no occupancy (or inferred occupancy) is being detected, the system can decide to go into an away mode. Ideally, these reactive controls should incorporate understanding about the individual users occupancy patterns and characteristics of the home to be able to properly weigh the benefits of aggressive strategies while maintaining an
appropriate threshold for user-discomfort. With the longitudinal data available for each connected thermostat device, machine learning can be utilized to train the required lightweight models to be run on the device.

This paper documents an investigation into the design and effectiveness of two basic machine learning models in preforming future prediction of detected motion over short time horizons. The investigation utilized commercially representative interval data from actual connected thermostat devices enrolled in the Donate Your Data (DYD) program (ecobee Inc., 2018). The ecobee devices often have multiple sensor points within the home as these particular connected thermostats support remote sensors which are paired with the thermostat. Use of the real device data exposes the investigation to one of the major technical challenges faced by industry in that the motion capture data from passive infrared sensors (PIR) only infers the occupancy state of the home. With that limitation, testing of the methods will address two fundamental questions:

**Q1** How accurate are predictions made using the Markov and logistic regression models over a short-horizon (a few hours) and how does that compare to baselines of using the previous state, the most frequent state in general, or most frequent state at certain times of day?  

**Q2** How do the Markov and logistic regression models compare in specific classification error rates (i.e., false positives and false negatives)?

The remainder of the paper is structured as follows: the Literature Review section provides an overview of related works, the Methodology section reviews the data source, and a description of the training and testing used in our investigation. The Results and Analysis section presents the findings and discussion regarding **Q1** and **Q2**, and finally the Conclusion section contains the summary of findings from the investigation and identifies future works.

**Literature Review**

It has been previously established by the work of other researchers that the presence of people in buildings can be predicted by a number of different parameters. Often presence in a space (e.g., building, thermal zone, room) is considered a Markovian process (i.e., dependent on the previous state of the system). Markov models have been applied to occupancy patterns in buildings; including in residential applications (Page et al., 2008; Lu et al., 2010; Dong et al., 2018; Dobbs and Hencsey, 2014; Li and Dong, 2018). The order of the Markov property (i.e., how many previous time steps have an effect) can vary. Page et al. (2008) claimed that the future presence should be a first order Markovian system and not possess any history beyond the previous time step. Meanwhile Li and Dong (2018) had a dynamic value based on the prediction length being sought. The appropriate order of the system would be affected by the length of time periods and sensor types. In addition to just previous motion, many developed methods have included both a Markovian component and a temporal component (Dobbs and Hencsey, 2014; Li and Dong, 2018; Page et al., 2008; Shi et al., 2017).

In the implementation of an occupancy prediction model for use with a preheating system of a residential building, Scott et al. (2011) considered only similarity in the temporal patterns for predicting occupancy later in the day. The type of day has also been identified as a driver of people’s patterns of motion and occupancy. Anecdotally, Scott et al. (2011) described it as “weekend chaos” in comparison to weekday occupancy patterns. Similarly, in applications to office spaces the difference between weekdays and weekends are explicitly or inherently differentiated (Zhao et al., 2015; Gunay et al., 2015). Depending on the sensors available, other features such as carbon dioxide, light, or humidity features have also been considered when seeking to predict occupancy states of a home (Candanedo et al., 2017). Shi et al. (2017) implemented a logistic regression with the covariates to the problem being hour of day and previous occupancy.

**Methodology**

The following section outlines the design of the problem for motion-state (or occupancy) prediction based on the conditional relationships found in the literature. Two methods for conducting inference based on the relationships and predicting motion are presented. In addition, the baseline comparison methods and data source are described.

**Data**

The data used in the investigation was taken from the DYD program (ecobee Inc., 2018) administered by the North American connected thermostat manufacturer ecobee Inc. The dataset contains over 25,000 thermostats and up to 33 months of data per device. For each thermostat some metadata is provided and consists of general information such as location and equipment type. Many of these values are user-reported. For each device there is also interval data (at a five-minute frequency) of the measurements recorded by the thermostat and any remote sensors. The thermostat records information including temperatures, motion detections, and the equipment runtimes. The remote sensors are small, battery-operated devices sold with the thermostat. They connect to the thermostats using a radio signal. These sensors contain an additional temperature and PIR sensor. The placement of the remote sensors in the home is unknown, with users free to place them as they wish. The users are provided recommendations...
on best practices with regards to the the placement. A thermostat can support up to 32 remote sensors. Huchuk et al. (2018) provides a more detailed description of the data set.

From the general population of thermostats, a sample of 1000 devices were randomly selected from those who had at least a year of data starting from September 2016. The sample was also limited to generations of ecobee devices which have native motion detection on the device in addition to potentially having remote sensors attached. Figure 1 shows, for the 1000 selected devices, the distribution in the number of remote sensors paired with the thermostat. The figure shows that the majority of people rely on more than the on-thermostat motion sensor and have additional motion sensing within their home. The average number of remote sensors per thermostat in our sample was just over two.

**Figure 1: Distribution in the number of remote sensors for the sample of 1000 thermostats. The vertical line indicates the average number of remote sensors.**

The raw five-minute interval motion data for each thermostat was resampled to a 30-minute frequency. The resampling performed a couple of functions. First, the thermostats already are set on schedules at 30-minute intervals. Second, it helped to reduce noise in the PIR data patterns. In a control implementation, the motion state would exist at the 30-minute intervals while control decisions such as HVAC runtime would remain operating at the five-minute, or even sub-five-minute, intervals. The conversion of five-minute to 30-minute intervals was considered to be a logical-OR condition using all available motion sensors. This method was similar in construction to the mapping done by Kleiminger et al. (2014). If at least one motion event was detected within the 30-minute period, that 30-minute period was considered motion positive. If all of the five-minute data was missing, then the 30-minute period was also considered missing. Missing data can be a result of a number of issues. Most are related to power to the device or the wireless internet connection of the device. Based on the data provided, it is not possible to reliably classify the reason for missing observations.

Two weeks of data for a single user following the occupancy mapping process can be seen in Figure 2. The data spans Sunday to a Saturday. By inspection it appears there is fairly consistent motion between 8:00 and 22:00, but there are clear deviations on days such as 2017-08-08 or in the early morning hours. Also seen is missing data on 2017-08-07, 2017-08-09, and 2017-08-10.

**Predicting Motion**

The features selected for our model were formulated into the graphical model shown in Figure 3. In the model, \( W \) represents a binary value (0 or 1) for being a weekday (1) or a weekend (0), \( H \) is the half-hour bin of the day (0 to 47), and \( M \) is the motion state (1 for motion, 0 for no motion). The decision was made to use only a first-order Markov process \( (M_{t-1}) \). The subscript \( t \) is the time at the current time step while \( t - 1 \) is the previous time step. The arrows and their direction in the figure represent dependence while the double circles (seen around all nodes other than \( M_t \)) indicate that these are observed during inference and used as evidence. The effects of seasonality, another factor to presence (Page et al., 2008), was not included explicitly as a conditional variable. Instead it was accounted for by training the model on data just before the testing period.

**Inference on the Graphical Model**

With the problem in the form of Figure 3, it is possible to train a model representing the joint probability of the system. During training, all of the parameters in Figure 3 are provided at each 30-minute interval of the training period. The joint representation, expressed in Equation 1, once learned can be queried given any combination of available information. By learning the complete joint probability, the model contains an extensive hypothesis space and is able account for various interactions between the parameters. For our inference, the resulting conditional probability was: \( P(M_t \mid M_{t-1}, H_t, W_t) \). For analysis, this model will be referred to as the Markov model. The model was implemented using pgmpy (Ankan and Panda, 2015), a python package designed for work with probabilistic graphical models.

\[
P(M_t, M_{t-1}, W_t, H_t) = P(M_t \mid M_{t-1}, H_t, W_t)P(M_{t-1})P(W_t)P(H_t)
\]

(1)

For predictions over a longer horizon than a single timestep, the general form of Figure 3 can be extended with variables at \( t + 1 \) and beyond given the requirements. Once the network is formed and trained, these individual time steps can be queried, resulting in the probability of there being motion detected or not at each instance. If the probability of a motion detection is greater than 50% it is treated as being detected else it is expected to see no motion. In situations where the sequence of values is of greater concern than the probability of a specific time step, an additional query option exists. In those applications it is possible to perform inference with a maximum a
possible interactions to only a linear weighting to model counterpart limiting its ability to capture all more restricted hypothesis space than the graphical dependencies between current motion prediction ($M_t$) and previous motion ($M_{t-1}$), time of day ($H_t$), and whether it is a weekday or not ($W_t$).

Figure 3: Graphical model indicating conditional dependencies between current motion prediction ($M_t$) and previous motion ($M_{t-1}$), time of day ($H_t$), and whether it is a weekday or not ($W_t$).

posteriori (MAP) query. In the MAP query, the maximum value is taken for the unobserved variables, resulting in a predicted sequence of consecutive values. Both query methods were used during evaluation and their performance compared. Both individual time steps and continuous sequences could be of value depending on the intended implementation.

Inference using Logistic Regression

An alternative interpretation of the problem in general is as a classification problem: solvable using logistic regression. Logistic regression has been regularly used in various occupancy and behavioural applications including occupancy, window opening, blind operation, and lighting (Dong et al., 2018). In the most straight-forward implementation of logistic regression, the same values are implemented as shown in Figure 3. The classifier assigns a category value based on the conditional probability $P(M_t \mid M_{t-1}, H_t, W_t)$. However, this model has a more restricted hypothesis space than the graphical model counterpart limiting its ability to capture all possible interactions to only a linear weighting to those feature values. To extend this model, an additional method was implemented in which the feature combinations of $M_{t-1}$, $H_t$, and $W_t$ were encoded to a single value. This new features was input to the logistic regression using one-hot encoding. This provides logistic regression with a larger hypothesis space to deal with more complex relationships. Both the encoded and non-encoded logistic regression were implemented using Scikit-learn (Pedregosa et al., 2011).

Baseline Methods

To compare with the logistic regression and Markov model inference, three baseline methods were constructed. These were intended to be simple and easily implemented solutions. The most frequent state baseline predicts the most frequent state in the training data for an individual. For example, if a user is away a majority of the time, then predictions were always of the away state. The most frequent state in that time bin baseline took the most frequent state in that particular 30-minute bin for that user in the training set regardless of day type. Finally, the previous state baseline was propagating the last observation of motion to the next queried state.

Training and Testing

The training and testing of the methods was conducted in the same way for each model. Four 10-week periods were selected; one in each season. The first eight weeks of each test period were used for training while the last two weeks were used for testing. For each thermostat and with each model, predictions were made from time $t$ until $t + 5$. The decision of a three hour prediction falls in line with similar investigations on predictive control. Other predictive HVAC applications have used various differing lengths including 60 minutes (Panagopoulos et al., 2015; Dobbs and Hencey, 2014), 90 minutes (Scott et al., 2011), or even 24 hours (Li and Dong, 2018). The three hour prediction horizon would be useful in most thermal planning applications deviating away from a schedule; however depending on the home’s construction this could reasonably be too much (in a small, low mass home) or too little time (in a large, thermally massive home). All the models and baselines could be
extended to longer durations than three hours simply by extending in a similar fashion past $t + 5$. For the logistic regression (both encoded and non-encoded), the training incorporated a three-fold cross-validation to prevent over-fitting. During the cross-validation, the regularization strength was tuned for each model using a grid search with values from $10^{-4}$ to $10^4$ at orders of magnitude in between. The logistic regression was trained with a unique model for each time prediction (i.e., $M_t$ and $M_{t+5}$ were found using their own models).

### Results and Analysis

An example of prediction during one test using the non-encoded logistic regression, on the same individual as in Figure 2, at the $M_t$ timestep can be seen in Figure 4. Visually it can be seen that the model is consistently predicting occupancy from 9:00 to 20:00 which roughly looks like the actual motion patterns but often is falsely predicting motion. The non-encoded logistic regression does not predict any of the early morning (0:00 to 7:00) motion seen in Figure 2. In contrast, the results from the Markov model for the same user at the same prediction of $M_t$ are seen in Figure 5. While the majority of time between 8:30 and 20:00 is being predicted as having motion, it does identify some periods of no expected motion. It also predicts at least a few motion states in the early morning.

To answer Q1, an average daily accuracy score for each user was tabulated for each method using the compiled data from all 1000 thermostats and all tests. Figure 6 shows the distributions of this value for both versions of the logistic regression along with both query types on the Markov model and the baselines. The accuracy is calculated at each 30-minute timestep prediction and then averaged each day per thermostat. The median value for the Markov model (using either query option) appears to the be the most accurate across the test cases, with a median value of 82%. It can be seen that the encoded logistic regression outperforms the standard one, especially once the horizon increases past $M_t$. Finally, the previous state is found to be a relatively good predictor at a single time step ahead, though none of the baselines appear to perform very well compared to the logistic regression or Markov models. The most frequent state in that time bin baseline does have very stable performance across the increasing prediction lengths. In applications of longer predictions horizons than three hours, the conditional dependence to previous motion state is expected to continue to decline and performance of the logistic regression and Markov models would be closer to that of the most frequent state in that time bin baseline.

The median daily average accuracy value of 82% at time $t$ is comparable to the accuracies achieved by others in residential settings (Candanedo et al., 2017; Kleiminger et al., 2014). Given the inherent variations in individuals schedules and activities (especially if multiple people may inhabit the

![Figure 4: Motion prediction at time $M_t$ for a single thermostats over a two week testing period (starting on a Sunday) using the non-encoded logistic regression.](image)

![Figure 5: Motion prediction at time $M_t$ for a single thermostats over a two week testing period (starting on a Sunday) using the Markov model with the marginal query.](image)
Figure 6: Average daily accuracy for each thermostat and each method at different prediction horizons.

Figure 7: Average daily number of false positives for each thermostat and each method at different prediction horizons for Markov and logistic regression models.

Figure 8: Average daily number of false negatives for each thermostat and each method at different prediction horizons for Markov and logistic regression models.
The capabilities of two simple probabilistic methods, logistic regression and a Markov model, were tested for motion prediction in residential buildings. The investigation used raw, and representative data from 1000 connected thermostat devices. The two methods, in multiple configurations, were both able to beat simple baselines in terms of accuracy and achieve similar accuracies to the reported values from other investigations. The Markov model, given its larger hypothesis space, was able to outperform a simple logistic regression. This advantage was diminished by taking the effort to encode the features and create a similar hypothesis space. In a comparison of error rates both in false positives and false negatives on predictions made up to three hours in advance, the Markov model appears to outperform the standard logistic regression. While the Markov model remained consistently the best performer, logistic regression when properly designed can be similarly effective.

In the future, the number of methods is intended to be expanded. While these models were selected because of their relative simplicity and ability to be scaled or deployed on existing hardware devices, they do still have their limits or assumptions which other methods may address. For example, use of deep learning methods such as the recurrent neural net, which captures past information in its structure, may find dependencies that are deeper in the history or based on more complex relations between variables. Alternatively, a hidden Markov model would better reflect the partial observability of occupancy based only on PIR motion detections and allows for the longer-term propagation of the historical information through a belief state.

Acknowledgements

This work is funded by a research grant provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and ecobee Inc.

References


