

Inference from Environmental Sensor Time-Series for the Generation of Smart-Space Control Schema

Anit Kumar Sahu^{1,2}, Jonathan Francis^{1,2}, Sirajum Munir², Charles Shelton², Anthony Rowe¹, and Christopher Martin²

¹Carnegie Mellon University, Pittsburgh, PA

²Bosch Research and Technology Center, Pittsburgh, PA

anits@andrew.cmu.edu, {jon.francis, sirajum.munir, charles.shelton}@us.bosch.com, agr@ece.cmu.edu, christopher.martin@us.bosch.com

ABSTRACT

The *disconnect* between inefficient control schedules in conventional Building Automation Systems (BAS) and the subjective needs of occupants calls for “smart” control schema that incorporate occupancy patterns and the relational evolution of different environmental sensor modalities. The *generalizability* of smart-space control schema depends on the usage of data-driven methods to capture the inter-dependencies between different environmental sensor modalities, as opposed to relying on overly-constrained physical building thermal models. In this paper, we develop a data-driven inference scheme based on temperature time-series and occupancy patterns, collected from a 343 square-foot conference room. Moreover, we also comment on the usage of the inferred data-driven model for generating closed-loop control schema, which use the inferred model as a feed-forward input mechanism. We then present insights from our data analysis, demonstrating the efficacy of the proposed data-driven inference model; we show a 27% improvement in temperature time series trend-fitting, when fine-grained occupancy measures are taken into account. Finally, we discuss model trade-offs, along with future steps.

Categories and Subject Descriptors

C.2.4 [Computer-Communication Networks]: Distributed Systems; H.4.1 [Information Systems Applications]: Office Automation

Keywords

Inference, Time-Series, Wireless Sensor Networks

1. INTRODUCTION

Building Automation Systems (BAS) combine hardware

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

BuildSys '16 November 16–17, 2016, Stanford, CA, USA

© 2016 ACM. ISBN 123-4567-24-567/08/06...\$15.00

DOI: 10.475/123.4

units with software agents to perform centralized¹ control of building-wide infrastructure systems (HVAC, lighting, surveillance, etc.), according to specific objectives or pre-defined rules. Because these rules are usually written offline and follow only rigid set-point schedules, the resultant control schema do not adapt to unexpected events, such as natural phenomena or the subjective thermal comfort requirements of building occupants. Moreover, when occupants are incorporated into models at all, we find that the consideration is usually on the basis of only a binary occupancy indicator for a space and not on an exact person count measure. As a result, the majority of HVAC control systems are merely calibrated for maximum occupancy conditions (as defined for fire code compliance and other building standards), leading to significant energy waste when the space is underutilized. Keeping the above challenges in mind, the main contributions of this paper are as follows:

- We develop a data-driven inference model, that predicts the evolution of environmental modalities and aids the generation of control signals, without relying on heavily-constrained room and building-specific physical models.
- We incorporate the occupant count as a pseudo-actuator and show that lower error is incurred in the inference performance, as compared to the case where the occupancy is not taken into consideration.

Several works in literature address the problem of the inference and control for BAS – albeit separately; some works address inference schemes through constrained building-specific models, while others focus on the open-loop control schemes based on those models (see, for example [3, 4, 7, 8]). However, few studies consider the inference step and the control schema generation tied together in a *closed-loop* fashion, i.e., wherein the control scheme is based on the inferred data-driven model and the inferred model changes with time based on the control signals. In this paper, we concern ourselves with just the HVAC system, characterize the evolution of occupancy and temperature as two sensor modalities, outline the interesting inherent trade-offs from our empirical study and their implications for energy-saving versus occupant comfort, and show how our solution not only gener-

¹By centralized, we mean the presence of a fusion center, i.e., a server which has access to all the time-series data collected by different sensors at all times.

alizes to other scenarios, but also addresses the duality of inference and control by minimizing model mismatch.

2. SENSOR DEPLOYMENT AND DATA COLLECTION

We have deployed two distributed wireless sensor networks – one in the Scaife Hall building of Carnegie Mellon University, main campus, in Pittsburgh, PA, USA; and another throughout Bosch Research & Technology Center Pittsburgh. Each deployment includes a set of actuators and environmental sensors, connected via the open *SensorAndrew* internet-of-things platform, developed in a previous work. [9] We specifically consider two classes of sensors in this study: the *Tiny Wireless Sensor for the Internet of things* (TWIST) is a self-contained environmental sensor unit, capable of reporting various modalities (temperature, pressure, humidity, audio, luminosity, accelerometry, magnetometry, gyroscopy) at 10-60 second intervals; the *Fine-Grained Occupancy Estimator Using Kinect* (FORK) is a sensor assembly, covered in concurrent work, consisting of a depth sensor and an embedded computing device – able to publish the exact number of persons that have crossed a geo-fenced threshold in the depth-sensor’s field-of-view. [5] Sensors of both types were strategically deployed in various locations throughout the lab, for detecting a wide range of environmental and occupancy phenomena. We consider the Warhol conference room: a 343 square-foot space in Bosch R&D Pittsburgh, with one large window (cannot be opened), one entrance, and a single air vent from a centrally-controlled HVAC system. For the rest of the paper, we focus on characterizing the temperature time-series in Warhol, with respect to various occupancy and environmental conditions.

In the period from August 2015 to December 2015, we have data from 41 weekdays and 15 weekend days. In the absence of any building thermal models and the external environmental temperature, the weekend data is specifically useful for studying the evolution of temperature without any external disturbances, e.g., occupants.

3. DATA-DRIVEN MODELING

In this section, we describe the data-driven model, that we use to draw inference from the collected time-series data. We denote the environmental state vector at time slot k as $\mathbf{x}[k]$, where $\mathbf{x}[k] = [T[k], H[k], L[k], P[k]]^\top$ and $T[k]$, $H[k]$, $L[k]$, $P[k]$ denote the temperature, heat, light and pressure, respectively. Intuitively, environmental modalities can be seen as states whose values at any given time instant depends on the past values and on the past actuator states. Hence, the state update can be formally written in terms of a linear dynamical system as follows:

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{O}[k] + \mathbf{C}\mathbf{U}[k], \quad (1)$$

where $\mathbf{O}[k]$, $\mathbf{U}[k]$, \mathbf{A} , \mathbf{B} and \mathbf{C} respectively represent the occupancy count, actuator state (control input), the matrix denoting the coupling and dependencies across states, the matrix denoting the dependence on occupancy count, and the matrix denoting the dependence on the actuator state. The state update from (1) captures the effects of occupancy count as a disturbance to the control inputs. We discuss the role of the occupancy count in the dynamical system before proceeding. As compared to other works in literature, which

either do not consider occupancy as a pseudo-actuator or model the occupancy as a binary state, i.e., whether a space is occupied or unoccupied ([1, 3, 7]) or consider occupancy based actuation in a binary sense ([2]), we look at the actual occupancy count. Due to the discrete modelling of the occupancy pattern, the matrix \mathbf{B} (which models the dependence of the environmental modalities on the occupancy pattern) needs to be regime-based. When the occupancy count is sufficiently lower than the maximum allowable occupancy density in an area, the effect of occupancy is expected to be minimal. At the same time, when the occupancy count is close to the the maximum allowable occupancy density in an area, the environmental modalities are expected to be affected more by the occupancy count. Thus, the matrix \mathbf{B} would have values close to zero with low occupancy density. We show an example of the temperature evolution in an instance of high occupancy density, below, in Figure 3.

To get a better understanding and intuition of the environmental modalities evolving as a dynamical system in presence of occupants as a disturbance or pseudo-actuator, we focus on the temperature time-series data. The model in (1) then can be rewritten as:

$$x_t[k+1] = f(x_t[k], x_t[k-1]) + b_t o[k] + c_t u[k] + e_t[k], \quad (2)$$

where $x_t[k]$ represents the temperature at time instant k , $f(x_t[k], x_t[k-1])$ models an auto-regressive (AR) process (possibly AR-1 and AR-2), b_t represents the dependence of the temperature as a modality on occupancy and c_t represents the dependence of temperature on the actuator state. It is to be noted that with the absence of information about other state variables, the possible coupling of the temperature as a modality with other modalities is not captured by the model in (2) and the associated error is abstracted out as $e_t[k]$. Finally, the collected time-series data does not contain information about the actuator states. Hence, we abstract out the actuator state into the error term and hence finally the update can be written as:

$$x_t[k+1] = f(x_t[k], x_t[k-1]) + b_t o[k] + e[k], \quad (3)$$

where $e[k] = e_t[k] + c_t u[k]$. Ideally, we would also be interested in estimating the actuator state from the time-series data. It is to be noted that the model in (3) is essentially a system identification problem. However, the unavailability of the actuator states makes the task at hand even more challenging as the problem involves two levels of abstraction: firstly, finding the system parameters $f(\cdot)$ and b_t in a given regime, secondly, estimating the actuator states based on the inferred model.

We specifically divide the temperature time-series data into weekends and weekdays. The weekend data plays a crucial role for abstracting out $f(\cdot)$ as with absence of occupants, the temperature evolution can be characterized to a better extent with lower error. Also, there needs to be a multi-time scale model for the different regimes of the day based on the occupancy. Technically speaking, during the night (as there would be most likely no occupants in an office space), the temperature prediction can be a long-term prediction unless the occupancy data indicates otherwise, i.e., a coarse-grained simplistic fit (3) would suffice. We use AR-1 fits for the times of the day when most likely there are no occupants. Formally, the AR-1 and AR-2 updates can be

written as:

$$x_t[k+1] = a_1 x_t[k] + e[k] + a_0 \quad (4)$$

$$x_t[k+2] = a_{22} x_t[k+1] + a_{21} x_t[k] + e[k] + a_{20} \quad (5)$$

Due to the data-driven nature of our model, which does not incorporate overly-constrained building-specific information, (see, for example, (3)-(4)) it can be generalized to different scenarios in different buildings.

4. MODEL VALIDATION

In this section, we validate the proposed data-driven model as described in Section 3. We specifically divide the 24 hours into 7 time intervals with the intervals being 12am to 10am, 10 am to 12 pm, 12 pm to 2pm, 2pm to 4pm, 4 pm to 6pm, 6pm to 8 pm and 8pm to 12am. The reasoning behind such a division is best explained by the occupancy pattern in the room under consideration.² The time intervals which are likely to have occupants and hence may have occupant initiated actuator state changes are the ones assigned two hours duration. The granularity of the time scale is set to a minute. We train the initial fits for the different intervals from the 15 days of weekend data, where we expect to abstract out the thermal characteristics of the room. We also incorporate fits for temperature evolution, where the number of occupants in the conference room is more than 10 for an hour or so and we will demonstrate shortly the advantage of including the occupancy count as a pseudo-actuator. We specifically use AR-1 models for interval 1 as fits for longer intervals require adaptivity and infrequent updates. On the other hand, we use AR-2 models for all the other intervals which is in turn dictated by shorter intervals and presence of occupants.

We use boosting algorithms (see [10] for details) for generating fitted models. Due to the dynamic nature of temperature evolution, which changes on a day to day basis, it is unlikely that one fitted model would work for everyday. Hence, we generate different fitted models from the 15 weekend days data and then assign weights to the fitted models so as to come up with a combination of the fitted models. Formally, let the trained fits for an interval i for $i = 1, \dots, 7$ be indexed by $f_i^j(\cdot)$, where $j = 1, \dots, 15$. For every interval i , while generating a fit, the first 1/5-th of the interval is where the predicted data stream from each fitted model is compared with the actual data stream. The normalized absolute error for each predicted data stream is evaluated which is given by

$$NAE_i(j) = \frac{5}{|i|} \sum_{|i|/5} \left| f_i^j(k) - x_{t,i}[k] \right|, \quad (6)$$

where $NAE_i(j)$, $|i|$, $f_i^j(k)$ and $x_{t,i}[k]$ denote the normalized absolute error for the interval i for the fitted model j , the cardinality of the sample points in the interval i , the predicted sample value by the j -th fitted model for the i -th interval at time index k and the actual temperature value at time index k for the i -th interval respectively. We use normalized absolute error (NAE, see for example (6)) for

²In situations where the occupancy pattern is unknown, *a priori*, the initial division can be done based on the occupancy prediction and can then be fine-tuned as and when more data is available.

determining the suitability of a particular fitted model. Normalized mean square error (NMSE) is another natural option for the fitted model error characterization. However, in presence of outliers and other errors such as packet losses for the sensor data, the fitted models penalized by NMSE will tend to be biased by the outliers and hence resulting in fits with larger prediction error. In other words, fitted models penalized by NMSE tend to weigh the sample points incurring higher prediction error which is usually the case with outliers and hence the fitted model fits the outliers better than the other sample points (the trend). The weight given to the j -th fitted model for the final fitted model for the interval i is given by

$$w_i(j) = \frac{\exp(1/NAE_i(j))}{\sum_{l=1}^{15} \exp(1/NAE_i(l))}. \quad (7)$$

The final fitted model, which in turn is denoted by $\hat{f}_i(\cdot)$, is then given by $\hat{f}_i(\cdot) = \sum_{l=1}^{15} w_i(l) f_i^l(\cdot)$. We use $\exp(1/NAE_i(j))$ to generate weights for different fitted models, which was chosen over $\exp(-NAE_i(j))$ as the former weighting model, penalizes models with higher prediction errors more.

It is also to be noted that the incorporation of the occupancy as a pseudo-actuator is through models (see, for example (3)), which are weighed into the final fitted model through (6)-(7).

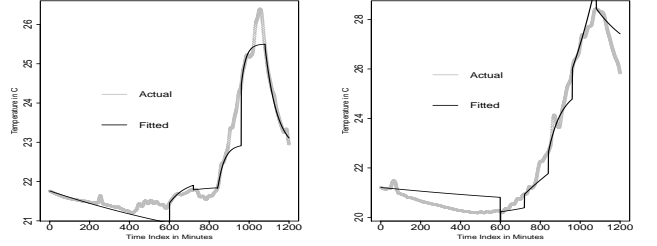


Figure 1: Examples of boosted fits

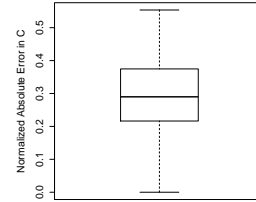


Figure 2: Box plot of Normalized Absolute Error across 41 days

In figure 1, we plot examples of two boosted fits. As the proposed model uses regime based fitting, the resultant fit is piecewise (hence discontinuities at the edges of the intervals). The average NAE across 41 weekdays of temperature data is found to be 0.3°C . In other words, for any sample point, the predicted sample point from our fitted models are on an average off by 0.3°C from the actual sample point. The variation of NAE's across different days is demonstrated through the box plot in figure 2. Figure 3, which we generate from 5 weekdays data demonstrates the importance of including occupancy count as a pseudo-actuator in the models. There is a 27% reduction in the NAE on an average across 41 days, when occupancy count is incorporated into the model

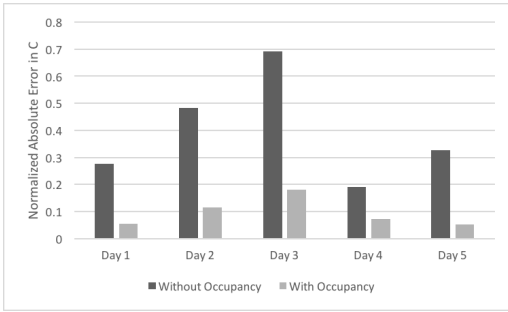


Figure 3: Comparison of Prediction Errors: Occupancy vs No Occupancy

as a pseudo-actuator as compared to the case when occupancy is not considered. The reduction in NAE is calculated as $100 * (NAE(noccu) - NAE(occu)) / NAE(noccu)$. NAE is also an indicator of the model mismatch.³ Model mismatch can lead to significant effects on the performance of the model predictive controller typically used as control schemes for HVAC systems (see, [6] for example). The effects can be seen on the instantaneous control signals generated from the controller which would lead to higher energy control signals thereby, resulting in high energy consumption. Due to the model mismatch, as the environmental modalities would be driven to states which don't exactly correspond to the thermal comfort levels of the occupants, the thermal comfort levels of the occupants is also affected. Hence, reduction of the effects of model mismatch is a crucial task which we address through our robust inference scheme in this paper.

4.1 Trade-offs

In this subsection, we discuss about the various intricate trade-offs involved in the proposed model.

- **Sampling interval:** While selecting smaller sampling intervals yields better inference, this calls for the use of more fine-grained control algorithms (which might not be feasible).
- **Length of intervals for generating fits:** While longer intervals better suit periods of little to no occupancy, the fits for longer intervals are likely to miss abrupt changes in the temperature data streams. Also, the control schedule for a long interval needs little to no updates; smaller intervals which are more suitable for periods with higher occupancy and are more likely to track abrupt changes in the data streams. However, the control schedules would need more frequent updates. For example, in figure 1, the abrupt changes in the temperature data stream are not captured perfectly by the fitted model for the interval 1 – 600.
- **AR model orders:** Higher-order AR models ensure better prediction performance, but training them requires more data and also more inputs⁴. For example in figure 1, the AR-1 fitted model (interval 1 – 600) has higher prediction error as compared to the AR-2 fitted models for other intervals.

³State-space model mismatch to be specific, i.e., mismatch of the AR coefficients and mismatch of the occupancy dependence coefficient. See (4) for details.

⁴An AR- n model needs n inputs.

5. CONCLUSION

In this paper, we proposed a data-driven model to characterize and model the evolution of temperature as a modality, with the discrete occupancy count as a pseudo-actuator. The proposed data-driven model is basically an auto-regressive linear dynamical system, which does not require building-specific models and hence can be generalized to various scenarios. A natural direction for future research includes coming up with control schema based on the proposed inference schemes extended to other environmental modalities, acting in closed-loop. In this paper, we focused on temperature as a modality to come up with data-driven models, which, in turn, would facilitate smart control schema. The incorporation of all the different modalities as part of the inference model would be an interesting future direction to consider. The estimation of actuator states from the different time-series data related to environmental modalities is also of interest as far as future work is concerned.

6. REFERENCES

- [1] R. V. Andersen, B. W. Olesen, and J. Toftum. Simulation of the effects of occupant behaviour on indoor climate and energy consumption. In *Proceedings of Clima*, volume 2007, page 9th, 2007.
- [2] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal. Sentinel: occupancy based hvac actuation using existing wifi infrastructure within commercial buildings. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 17. ACM, 2013.
- [3] J. R. Dobbs and B. M. Hency. Model predictive hvac control with online occupancy model. *Energy and Buildings*, 82:675–684, 2014.
- [4] B. Dong and B. Andrews. Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings. In *Proceedings of building simulation*, pages 1444–1451, 2009.
- [5] S. Munir, R. S. Arora, C. Hesling, J. Li, J. Francis, C. Shelton, C. Martin, A. Rowe, and M. Berges. Real-time fine grained occupancy estimation using depth sensors. In *BuildSys (Submitted)*, 2016.
- [6] A. Nafsun and N. Yusoff. Effect of model-plant mismatch on mpc controller performance. *J. Appl. Sci*, 21(11):3579–3585, 2011.
- [7] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*, 45:15–27, 2012.
- [8] F. Oldewurtel, D. Sturzenegger, and M. Morari. Importance of occupancy information for building climate control. *Applied energy*, 101:521–532, 2013.
- [9] A. Rowe, M. Berges, G. Bhatia, E. Goldman, R. Rajkumar, J. Garrett, J. Moura, and L. Soibelman. Sensor andrew: Large-scale campus-wide sensing and actuation. *IBM Journal of Research and Development*, 55(1.2):6:1–6:14, Jan 2011.
- [10] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651–1686, 1998.