Empirical Mode Decomposition for Intrinsic-Relationship Extraction in Large Sensor Deployments

Romain Fontugne
University of Tokyo /
JFLI, CNRS

Jorge Ortiz, David Culler
Computer Science Division
University of California, Berkeley

Hiroshi Esaki
University of Tokyo

***Abstract*—In this paper we investigate the utility of empirical mode decomposition (EMD) to identify intrinsically correlated usage patterns among sensors in a large deployment. We use data collected from almost 700 sensors in a 12-story building measuring power, pressure, temperature, and other physical phenomena. We discover that doing a correlation analysis on the raw traces does not discriminate well enough to identify meaningful relationships between sensors. We correlate the trace from a pump with the rest of the sensor traces and find that simple correlation filters only 50% of the sensors as being correlated with the behavior of the pump. In contrast, by running the correlation analysis on the constituent frequencies extracted by the EMD process, we filter out over 99% of the sensors as being correlated – with the highest correlation coming from sensors that serve the same room as the pump. We believe our approach can be used to construct inter-device correlation models that can help understand and identify misbehaving or inefficient usage patterns.**

I. Introduction

Buildings consume an enormous amount of energy in countries around the world. In Japan, 28% of the energy produced is consumed in buildings [12] while in the United States it is as high as 40% [16]. Moreover, studies show that between 30-80% of it is wasted [5], [13]. Large commercial buildings are typically instrumented with a large number of sensors measuring various aspects of building operation. Although this data is typically used to assure operational stability, they may also be used to measure, observe, and identify instances of wasted use.

Identifying instances of wasted energy use is non-trivial. System efficiency is defined as the ratio of the useful work done to the energy it consumes. In the case of buildings, we broadly define useful work as the energy used to support occupant activities. From the perspective of the building that means maintaining a comfortable temperature setting, providing power for plug-load devices, and providing adequate lighting conditions; particularly in spaces that are occupied. However, identifying efficient use of resources, *especially* when a space is occupied, is difficult. Typically it involves deep knowledge of the usage scenario and a meaningful understanding of what it takes to support the activity. Furthermore, situations and activities differ greatly. The outside weather changes, varying schedules affect occupancy, rooms have lectures, class, or other office activities. Simply put, the process is time consuming, requires specialized knowledge, and does not scale.

Devices are typically used together in some fashion. For example, in an office setting a person enters their office, turns on their PC and lights, etc. When the person leaves the office, they revert back to the state their devices were in before arrival. If one of the items is not reverted to its pre-arrival state, waste occurs. The same is true about equipment usage. When the outside temperature is low the heater turns on. *Waste occurs when abnormal in-concert usage patterns arise*. Fundamentally, understanding “normal” spatio-temporal usage patterns between devices could help identify problems when devices are not being used correctly. We conjecture that inefficient energy use can be identified through anomalies in the correlation patterns between devices. We examine device correlation patterns in this paper and look specifically at processing raw sensor traces, such that the correlations we find are meaningful.

In this paper, we present early results for correlating usage patterns across a large number of sensors in a single deployment. We analyze data from a 12-story office building at the University of Tokyo. The deployment consists of almost 700 sensors monitoring a broad range of devices inside and outside the building. Our initial observations and results include the following:

1) Raw-trace correlation analysis is too strongly influenced by the common low-frequency trends in the data to identify meaningful relationships.
2) Using a technique called empirical model decomposition (EMD) [7] removes this trend and helps identify truly correlated sensor traces.
3) We can construct clusters of correlated sensors that are spatio-temporally correlated, *without a priori knowledge of their placement*.

In the rest of the paper we explain EMD and how we use it, we show various examples of our technique on real-world traces, and we discuss the implications and future work.

II. Related Work

Recently, there has been increased interest in minimizing building energy consumption. Our approach differs quite substantially from related work. Agarwal et al. [2] present a parameter-fitting approach for a Gaussian model to fit the parameters of an occupancy model to match the occupancy data with a small data set. The model is then used to drive HVAC settings to reduce energy consumption. We ignore occupancy entirely in our approach. It appears as a hidden factor in the correlation patterns we observe.

Bellala et al. [3] look at various buildings to develop a model of efficient power usage using an unsupervised learning technique coupled with a Hidden Markov Model (HMM). They also develop occupancy models based on computer network port-level logs to help determine more efficient management policies for lighting and HVAC. They claim a savings of 9.5% in lighting on a single floor. Kim et al. [8] use branch-level energy monitoring and IP traffic from user's PCs to determine the causal relationships between occupancy and energy use. Their approach is most similar to ours. Understanding how IP traffic, as a proxy for occupancy, correlates with energy use can help determine where inefficiencies may lie.

In each of these studies and others [1], [4], [10], occupancy is used as a trigger that drives efficient resource-usage policies. Efficiency when unoccupied means shutting everything off and efficiency when a space is occupied means anything can be turned on. There is no question that this is an excellent way to identify savings opportunities, however, we take a fundamentally different approach. We are agnostic to the underlying cause or driver for efficient policies to be implemented. More generally, we look to understand *how the equipment is used in concert*. This may help uncover unexpected underlying relationships and can be used in an anomaly detection application to establish "(in)efficient", "(ab)normal" usage patterns. The latter should identify savings opportunities in cases where the space is unoccupied as well as occupied, because it has to do with the underlying behavior of the machines and how they generally work together. Our approach could help achieve both generality and scale for such an application. This article focuses on the first step of this application, the identification of correlated devices.

III. Dataset

The data we used was obtained from a deployment of sensors in a 12-story office building on the campus of the University of Tokyo [6], [14]. The deployment consists of almost 700 sensors monitoring device power consumption, ranging from plug-load devices to components of the heating, ventilation, and air conditioning system (HVAC) and lighting. Sensors also reported temperature, pressure, device-state, and other information. Each sensor reports data on the order of minutes. Over 500 GBs of data was collected over a 2-year span.

For this investigation, we focus on a three-week span in the summer of 2011 (from July 4th to July 24th). The dataset captures regular work days, weekends, and one holiday (July 18th). This timeframe captures the typical usage of the equipment, triggered by occupant activity. For the initial analysis, we focus on three sensors; two water pumps and a light feed. The first pump is an "electric heat pump" and is labled as EHP, the second is a "gas heat pump" and labeled as GHP. The room lighting system serves the same room as the EHP. The GHP serves a different room on the same floor. The expanded portion of our analysis pivots around the EHP and does a pairwise comparison between it and all other sensors in the building. Computationally, this approach does not scale to a large number of sensors. For future work, we will examine various heuristics to narrow the search space before running pairwise comparisons.

IV. Problem statement and Initial approach

In buildings, metadata is poorly and unsystematically managed within a single system domain. Moreover, with the ever growing number of additional sub-meters, it is important to quickly integrate sensor data from multiple systems to understand the full state of the building. It is also important to understand how sensors are used in concert. Anomalies in usage may indicate underlying problems with the equipment or inefficient/incorrect usage.

Figure 2 shows the raw traces for the three devices discussed in the previous section (EHP, GHP, light). All three exhibit a diurnal usage pattern. On weekends, each draw less power. For our initial analysis, we calculated the pairwise correlation coefficient for all sensors in the set. The correlation coefficient for the EHP and light is 0.7715 and the correlation coefficient for the EHP and GHP is 0.6370. Running correlation across them yields high correlation coefficients, mostly due to their underlying daily usage pattern.

Our initial results were not surprising. The diurnal pattern dominates the comparison between the sensors. Weather is the main driver for this behavior and it affects the readings in almost all of the sensors in our dataset. Cross-correlation on raw sensor data is insufficient for filtering intrinsically related behavior. Upon closer examination of the data we assess the following:

- The main underlying diurnal trend occurs in almost all the traces.
- Occupancy and room activities occur at random times during the day and change at a higher frequency than weather patterns.
- Sensors that serve the same location observe the same activities. Therefore, their underlying measurements should be correlated.

In order to uncover these relationships we must remove low-frequency trends in the traces and compare the readings at high frequencies.

V. Methodology

Empirical Mode Decomposition (EMD) [7] is a new technique used for detrending data. Specifically, EMD detrends non-stationary, non-linear timeseries data. A non-stationary signal is a signal whose mean and variance change over time. EMD is a process, not a theoretical tool, and its main use is for removing trends to enable more useful spectral analysis.

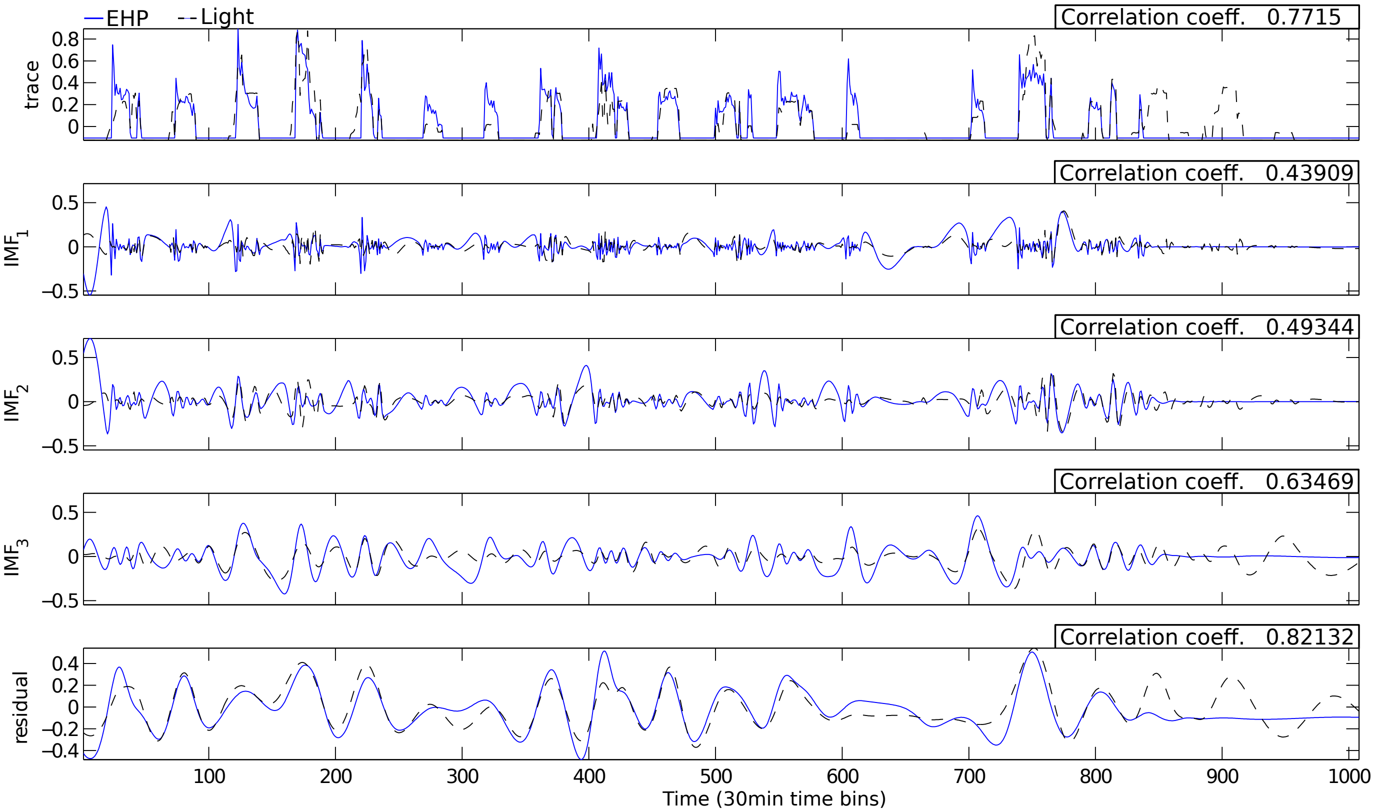


Fig. 1. Decomposition of the EHP and light trace using bivariate EMD. IMFs correlation coefficients highlight the intrinsic relationship of the two traces.

We describe the EMD process as follows: for a signal *X(t)*, let m_1 be the mean of its upper and lower envelopes as determined from a cubic-spline interpolation of local maxima and minima. The locality is determined by an arbitrary parameter.

1) The first component h_1 is computed: $h_1 = X(t) - m_1$
2) In the second sifting process, h_1 is treated as the data, and m_{11} is the mean of h_1's upper and lower envelopes: $h_{11} = h_1 - m_{11}$
3) The procedure is repeated k times, until h_{1k} is a function: $h_{1(k-1)} - m_{1k} = h_{1k}$
4) Then it is designated as $c_1 = h_{1k}$, the first functional component from the data, which contains the shortest period component of the signal. We separate it from the rest of the data: $X(t) - c_1 = r_1$, and the procedure is repeated on $r_j : r_1 - c_2 = r_2, \ldots, r_{n-1} - c_n = r_n$

The result is a set of functions called intrinsic mode functions (IMF); the number of functions in the set depends on the original signal [9]. An IMF is any function with the same number of extrema and zero crossings, with its envelopes being symmetric with respect to zero. We run our correlation analysis on the shared IMF outputs between a pairs of traces. In order to ensure that the IMFs corresponding to two distinct traces are on the same time scale, we use bivariate EMD [15] to decompose two traces at once.

We use EMD to detrend each of the traces and pay particularly close attention to the high-frequency IMFs. Our hypothesis is that correlating at the higher frequencies will yield more meaningful comparisons.

VI. Results

We test our hypothesis in this section by using EMD to remove low-frequency trends in the data and run correlation calculation at overlapping IMF timescales. We discover that EMD allows us to find and compare high-frequency instrinsic behavior that is spatially correlated across sensors. We begin with a small set of three sensors (EHP, GHP, light) and expand our scope to include all the sensors in the dataset.

A. Initial analysis

Lets consider the simple example of Section IV where we would like to know if the EHP trace is correlated with the two other traces. Recall that the correlation coefficients of the raw feeds was 0.7715 and 0.6370, corresponding to the light and GHP, respectively. As stated in previous section this result is correct but not so meaningful, since most of the traces display the same diurnal pattern. Figure 1 and Figure 3 show the EMD decomposition of the three traces. For each trace, EMD has retrieved three IMFs that highlight the higher frequencies of the traces.

Figure 1 shows the normalized raw trace (top) and EMD output IMFs and residual as well as the correlation coefficients

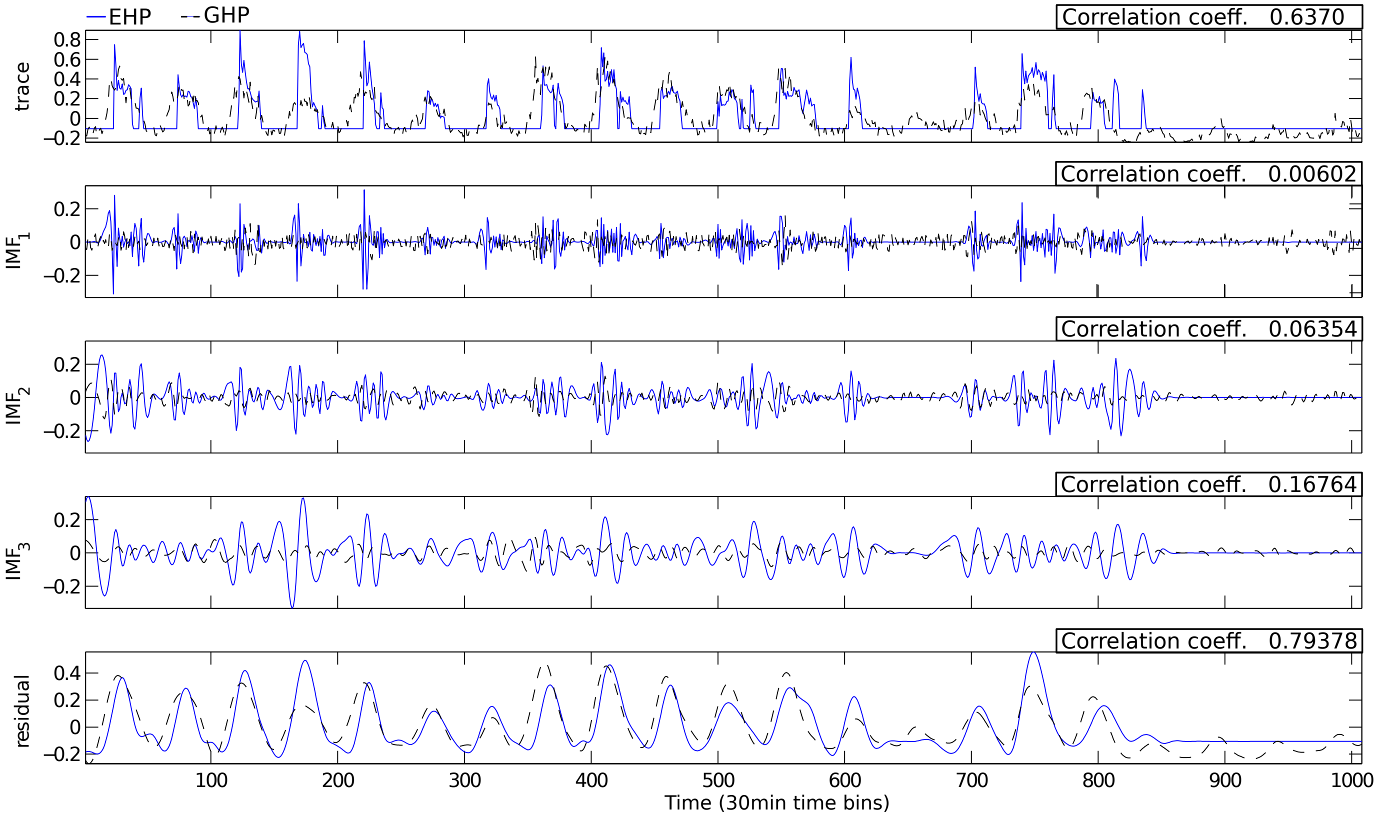


Fig. 3. Decomposition of the EHP and GHP trace using bivariate EMD. IMFs correlation coefficients highlight the intrinsic independence of the two traces.

calculated on the IMFs for the EHP and light traces. The correlation coefficients are 0.43909, 0.49344 and 0.63469 corresponding to the IMF1, IMF2, and IMF3, respectively. Notice the high correlation between the high-frequency IMFs. We know that the light and EHP serve the same room, and their high-frequency IMF correlation corroborates our prior knowledge. Figure 3 shows a complementary result, for the EHP and GHP comparison. The correlation coefficients for the EHP and GHP IMFs suggest that the two may be independent. In fact, they *are* indepdent; they serve completely different rooms in the building!

EMD allows us to remove low-frequency trends that add noise to the original analysis. By comparing IMFs, we see both intrisically correlated and *uncorrelated* behavior. In the next section we expand our analysis and show the effectiveness of our methodology.

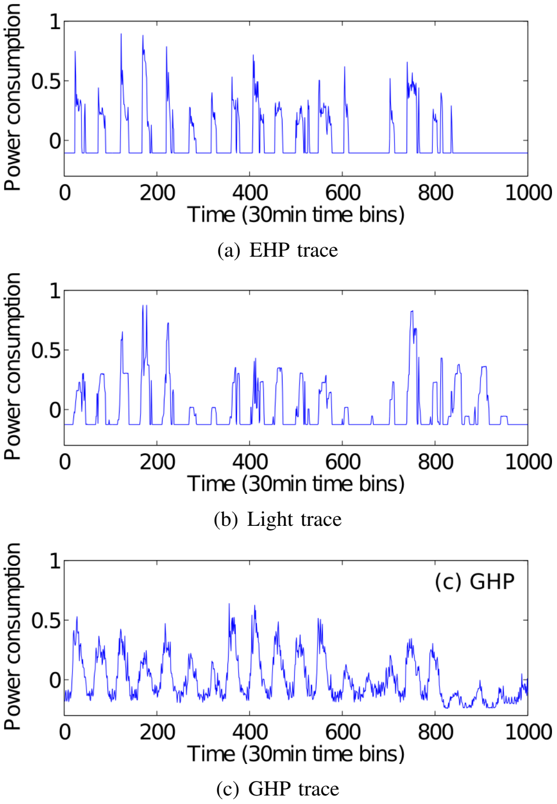
B. Validation

To validate the effectiveness of our approach, we analyze the same three-week time span for *all* 674 sensors deployed in the building. For each trace S we compute two scores: (1) the correlation coefficient between S and the EHP trace and (2) the average value of the IMF correlation coefficients.

Figure 4(a) shows the distribution correlation coefficients. Notice that a large fraction of the dataset is correlated with the EHP trace. *Half* the traces have a correlation coefficient higher than 0.36. As expected, the underlying trend is shared by a large number of device. Although the highest score (i.e. 0.7715) corresponds to the light in the same room that the EHP serves, there are 118 pumps, serving all areas of the building, with a correlation higher than 0.6. Using only these results, it is not clear where the threshold should be set. The distribution is close to uniform, making it difficult to know of how well your threshold discriminates against unrelated traces.

Figure 4(b) shows the distribution of the average correlation value for the IMFs of each trace and the EHP. The number of traces correlated in the high frequency IMFs is significantly smaller than the previous results. It's clear from the distribution that only a small set of devices are *intrinsically correlated* with the EHP. In fact, *only 10 traces out of 674* yielded a score higher than 0.25. This allows us to easily rank traces by correlation.

Upon closer inspection of the 10 most correlated IMF traces, we find that there is a spatial relationship between the EHP and the ten devices. In fact, there is a direct relationship between score and distance from the areas served by the EHP. Figure 5 shows a map of the floor that contains the rooms served by this EHP. The EHP directly serves room $C2$. We introduce a correlation threshold to cluster correlated traces by score. We highlight rooms by the threshold setting on the IMF correlation score. When we set the threshold at 0.5 we see



(a) EHP trace

(b) Light trace

(c) GHP trace

Fig. 2. Traces from three different sensors captured in 2011 from July 4th to July 24th. Data is normalized and aggregated into 30 minutes time bins.

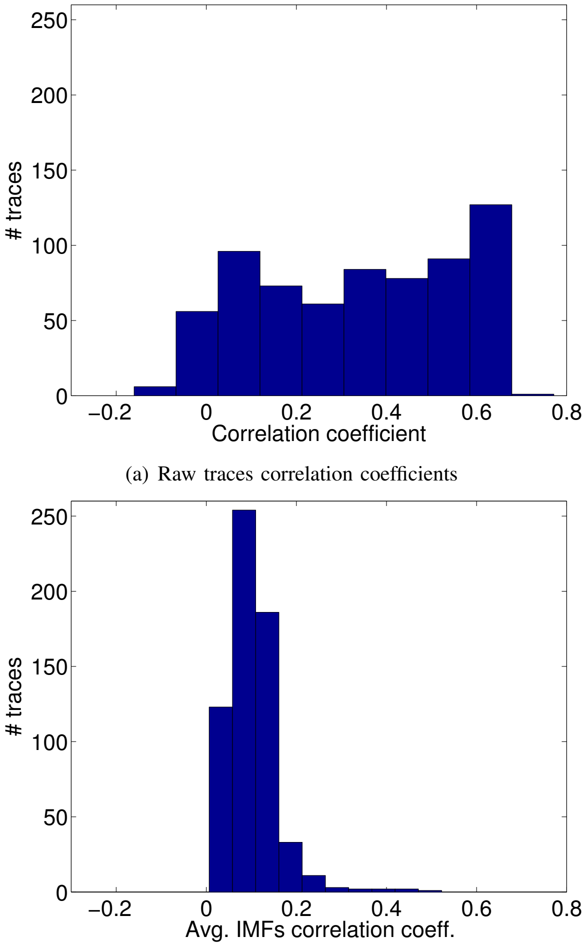
that the sensors that have a correlation higher fall within room $C2$ – the room served directly by the EHP. As we relax the threshold, lowering it to 0.25 and 0.1 we see radial expansion from $C2$. The trace with the highest score, 0.522, is the trace corresponding to the lighting system *in the same room*. The two highest scores for this floor (i.e. 0.316 and 0.279) are the light and EHP traces from next door, room $C1$. Lower values correspond to sensors measuring activities in other rooms that have no specific relationship to the analyzed trace. The results show a direct relationship between IMF correlation and spatial proximity and *supports our initial hypothesis*.

C. Limitations

EMD is useful for finding underlying behavioral relationships between traces of sensor data. However, when we set the timescales smaller than a day, the results were not as strong. The trace has to be long enough to capture the trend. For this data set, the underlying trend is daily, therefore it requires there to be a significant number of samples over many days. Although this was a limitation for this dataset, it really depends on the underlying phenomenon that the sensors are measuring. Its underlying trend is ultimately what EMD will be able to separate from the intrinsic modes of the signal.

D. Discussion

EMD allows us to effectively identify fundamental relationships between sensor traces. We believe that identifying meaningful usage-correlation patterns can help reduce oversights by the occupants and faults that lead to energy waste. A direct application of this is the identification of simultaneous heating and cooling [11]. Simultaneous heating and cooling is when the heating and cooling system either compete with one another or compete with the incoming air from outside. If their combined usage is incorrect, there is major energy waste. This problem is notoriously difficult to identify, since the occupants do not notice changes in temperature and building management systems do not perform cross-signal comparisons. For future work, we intend to run our analysis on the set of sensors that will allow us to identify this problem: the outside temperature sensors, the cooling coil temperature, and the air vent position sensor. If their behavior is not correlated as expected, an alarm will be raised.



(a) Raw traces correlation coefficients

(b) Average IMFs correlation coefficients

Fig. 4. Distribution of the correlation coefficients of the raw traces and correlation coefficients average of the corresponding IMFs using 3 weeks of data from 674 sensors.

We can also apply it to other usage scenarios. In our traces, we found an instance where the pump was on but the lights were off; where, typically, they are active simulatenously. The air conditioning was pumping cool air into a room without occupants. With our approach this could have been identified

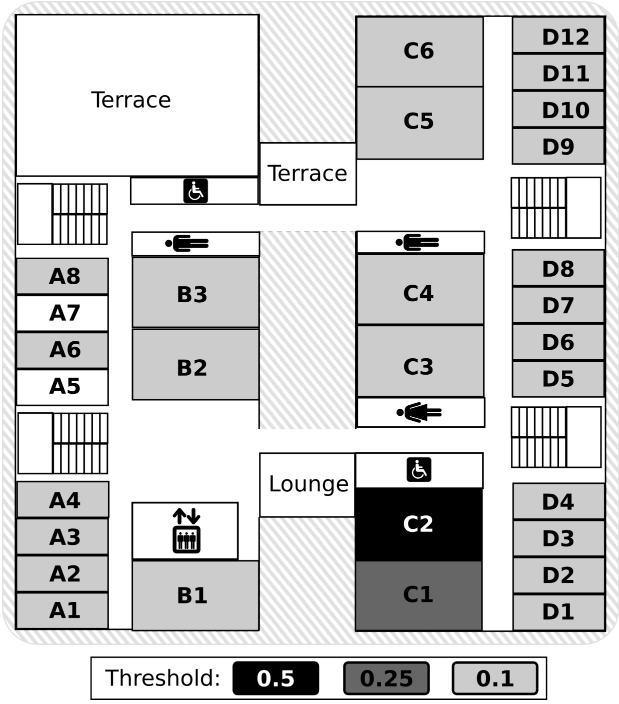


Fig. 5. Map of the floor where the analyzed EHP serves (room $C2$). The location of the sensors identified as related by the proposed approach are highlighted, showing a direct relationship between IMF correlation and spatial proximity.

and corrected. In future work, we intend to package our solution to serve these kinds of applications.

VII. Conclusion

This paper set out to examine the underlying relationship between sensor traces to find interesting correlations in use. We used data from a large deployment of sensors in a building and found that direct correlation analysis on the raw traces was not discriminatory enough to find interesting relationships. Upon closer inspection, we noticed that the underlying trend was dominating the correlation calculation. In order to extract meaningful behavior this trend has to be removed. We show that empirical mode decomposition is a helpful analytical tool for detrending non-linear, non-stationary data; inherent attributes contained in our traces.

We ran our correlation analysis across IMFs, extracted from each trace by the EMD process, and found that the pump and light that serve the same room were highly correlated, while the the other pump was not correlated to either. In order to corroborate the applicability of our approach, we compared the pump trace with *all* 674 sensor traces and found a strong correlation between the relative spatial position of the sensors and their IMF correlations. The most highly-correlated IMFs were serving the same area in the building. As we relax the admittance criteria we find that the spatial correlation expands radially from the main location served by the reference trace.

We plan to examine the use of this method in applications that help discover changes in underlying relationships over time in order to identify opportunities for energy savings in buildings. We will use it to build inter-device correlation models and use these models to establish "(ab)normal" usage patterns. We hope to take it a step further and include a supervised learning approach to distinguish between "(in)efficient" usage patterns as well.

Acknowledgments

The authors thank Gabriel Rilling, Patrick Flandrin, Patrice Abry, Pierre Borgnat for their source code and help on EMD, Hideya Ochiai for sharing the data, Kensuke Fukuda and Young Jon Won for constructive discussions. This research was partially supported by the JSPS fellowship program. This work is also supported in part by the National Science Foundation under grants CPS-0932209 and CPS-0931843.

References

[1] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng. Duty-cycling buildings aggressively: The next frontier in HVAC control. IPSN'11, pages 246–257, Chicago, IL, April 12-14, 2011.

[2] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng. Enabling building energy auditing using adapted occupancy models. Buildsys'11, page 6, Seattle, WA, Nov. 1, 2011.

[3] G. Bellala, M. Marwah, M. Arlitt, G. Lyon, and C. E. Bash. Towards an understanding of campus-scale power consumption. Buildsys'11, page 6, Seattle, WA, Nov. 1, 2011.

[4] G. Gao and K. Whitehouse. The self-programming thermostat: optimizing setback schedules based on home occupancy patterns. BuildSys'09, pages 67–72, Berkeley, California, Nov. 3, 2009.

[5] N. Gershenfeld, S. Samouhos, and B. Nordman. Intelligent infrastructure for energy efficiency. *Science*, 327(5969):3, 2010.

[6] GUTP. Green University of Tokyo Project. *http://www.gutp.jp/*.

[7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[8] Y. Kim, R. Balani, H. Zhao, and M. B. Srivastava. Granger causality analysis on ip traffic and circuit-level energy monitoring. BuildSys'10, pages 43–48, Zurich, Switzerland, Nov. 2, 2010.

[9] Max Lambert and Andrew Engroff and Matt Dyer and Ben Byer. Empirical Mode Decomposition: http://www.clear.rice.edu/elec301/Projects02/empiricalMode/process.html.

[10] S. Meyn, A. Surana, Y. Lin, and S. Narayanan. Anomaly detection using projective markov models in a distributed sensor network. CDC'09, Shanghai, China, December 16-18, 2009.

[11] M. Modera, N. M. andCharlie Huizenga, F. Bauman, E. Arens, and T. Borgers. Efficient thermal energy distribution in commercial buildings final report to california institute for energy efficiency. *Environmental Energy Technologies Division, LBNL Technical Report*, May 1994.

[12] S. Murakami, M. D. Levine, H. Yoshino, T. Inoue, T. Ikaga, Y. Shimoda, S. Miura, T. Sera, M. Nishio, Y. Sakamoto, and W. Fujisaki. Energy consumption and mitigation technologies of the building sector in japan. 6th International Conference on Indoor Air Quality, Ventilation & Energy Conservation in Buildings IAQVEC 2007, Sendai, Japan, October 2007.

[13] Next10. Untapped Potential of Commericial Buildings: Energy Use and Emissions, 2010.

[14] Y. Ogawa, S. Shinoda, and Y. Furui. Faculty-wide information system for energy saving. In Y. Luo, editor, *Cooperative Design, Visualization, and Engineering*, volume 6874 of *Lecture Notes in Computer Science*, pages 162–165. Springer Berlin / Heidelberg, 2011.

[15] G. Rilling, P. Flandrin, P. Gonalves, and J. M. Lilly. Bivariate empirical mode decomposition. *IEEE Signal Processing Letters*, 14(12):936–939, 2007.

[16] U.S. Environmental Protection Agency. Buildings Energy Data Book, 2010.