Towards Automatic Spatial Verification of Sensor Placement in Buildings

Dezhi Hong[1], Jorge Ortiz[2], Kamin Whitehouse[1], David Culler[2]
[1]Department of Computer Science, University of Virginia, USA
[2]Computer Science Division, University of California Berkeley, USA
{hong, whitehouse}@cs.virginia.edu, {jortiz, culler}@cs.berkeley.edu

ABSTRACT

Most large, commercial buildings contain thousands of sensors that are manually deployed and managed. These sensors are used by software and firmware processes to analyze and control building operations. Many such processes rely on sensor placement information in order to perform correctly. However, as buildings evolve and building subsystems grow and change, managing placement information becomes burdensome and error-prone. An automatic verification process is needed. We investigate empirical methods to automate spatial verification. We find that a spatial clustering algorithm is able to classify relative sensor locations – for 15 sensors, spread across five rooms in a building – with 93.3% accuracy, 13% better than a k-means clustering-based baseline method. Analysis on the raw time series data has a classification accuracy of only 53%. By decomposing the signal into intrinsic modes and performing correlation analysis, an observable, statistical boundary emerges that corresponds to a physical one. These results may suggest that automatic verification of placement information is possible.

Categories and Subject Descriptors

C.3 [**Special-Purpose And Application-Based Systems**]: Real-time and embedded systems

General Terms

Performance, Experimentation, Verification

Keywords

Sensor Placement, Empirical Mode Decomposition, Correlation Coefficient, Clustering

1. INTRODUCTION

Buildings have become a prime target for cyber-physical systems research, as they consume 40% of the energy in the U.S. [4], are poorly understood, and offer a rich sensing infrastructure. Thousands of sensors are embedded throughout the building and produce periodic physical measurements. In order to interpret the information, metadata describing the placement of sensors is recorded. However, deployments and their metadata are configured manually. As such, they are prone to human error. Moreover, over time sensors are replaced and the physical configuration of the building changes – walls removed, new offices set up – but the metadata describing the new locations are not. This leads to analytical errors in processes that rely on the metadata when interpretting sensor feeds. For example, model-predictive control processes rely on the sensors in a specific room or floor [16]. Because of the size and distributed nature of the deployment, it is cumbersome, error-prone, and impractical to maintain accurate metadata about sensor placement over time. An automatic processes is needed.

Typically, placement information is embedded in the name or associated metadata for each sensor in the building. These are used to group sensors by location. For example, in our building data, all sensors that contain the string '410' in their name are in room 410. Processes typically group streams in this fashion: using regular-expression matching or field-matching queries on the characters in the sensor name or metadata. If these are not updated to reflect changes then such group-by query results will not accurately represent true spatial relationships. Fontugne et al. [6] observe that spatial associations can be derived empirically. We start with this approach in our work and explore, more deeply, the extent to which it can be used as a verification tool for corroborating the groups constructed from character-matching queries. We refer to this process as *spatial verification*.

Prior work [6] makes use of a technique called Empirical Mode Decomposition (EMD) [10] to statistically cluster correlated usage patterns. Sensors close to each other show strong statistical correlations while sensors further apart show weaker correlations. The main parameter in their approach, the correlation threshold, is explored to demonstrate how it relates to characteristic spatial patterns in the sensor feeds. However, they do not characterize the threshold as it relates to physical configuration. Fontugne et al. [7] expand the work by applying EMD to uncover functional device patterns. They develop an unsupervised learning method to model normal usage patterns and apply an anomaly detection algorithm to alert when patterns have deviated from the norm. The methodology used in their work divides raw signals into four separate frequency bands and shows the medium band to carry the most spatial information.

BuildSys'13, November 14–15, 2013, Rome, Italy

In this paper, we explore the threshold parameter in [6] more deeply, in order to move towards automatic spatial clustering, to be used as a form of verification. We use EMD and the intrinsic mode function (IMF) re-aggregation methodology described in [7], with some modifications, to statistically analyze the threshold parameter and its relationship to spatial separation in a building. We explore the hypothesis that *a statistical boundary, analogous to a physical one, exists and is empirically discoverable*. We conduct an empirical analysis on the data collected from 15 sensors in 5 rooms over a one-month period. Our study makes the following contributions:

- We corroborate the results in [6], verifying the spatial correlation pattern in a very different building.
- We characterize the correlation coefficient (corrcoeff) distribution of sensors in the same room and different rooms and validate our existence hypothesis for this preliminary sample.
- We demonstrate that the statistical boundary between sensors in various rooms converges to a similar value and this value generalizes across rooms in this study.
- We show the tradeoff between the true and false positive rate inherent to threshold selection. We also show that our method improves the classification accuracy from 80% to 93.3%.

Our results are promising yet preliminary. We are able to find a statistical separation across a small number of rooms, quite well. Our study, however, does not explore the extent to which the physical separation affects the results. Certainly for rooms that are far apart we observe a statistical distinction using our methodology. However, we also find that in some cases, our approach does not work as well. We discuss the approach and results in the rest of the paper, followed by a short discussion and future work.

2. RELATED WORK

There has been much research work on sensor stream clustering and trace analysis. Chen and Tu [2] investigate how to cluster data streams in real-time using a density-based approach with a two-tiered framework. The first tier captures the dynamics of a data stream with a density decaying technique and then maps it to a grid. The second tier computes a grid density based on how it clusters the grid. Their approach differs from ours in that they focus on decreasing algorithm complexity for real-time sensor stream clustering. We run our analysis on historical traces and use correlation analysis in our clustering algorithm.

Kapitanova et al. [13] describe a technique to monitor sensor operations in the home and identify sensor failures. The classifier is trained on historical sensor data to obtain the relationship between sensors, assuming the number and location of sensors is known. When a failure or removal of a sensor occurs, the classifier's behavior deviates and the event is captured. Our method does not require any prior knowledge and instead tries to cluster feeds to discover their relative placement.

Lu and Whitehouse [15] formulate a new algorithm, particularly leveraging the semantic constraints interpreted from sensor data to determine sensor locations. The algorithm identifies how many rooms are present using motion sensors and determines room position based on physical constraints. Finally, it maps each sensor into the associated room. Our efforts focus on using intrinsic patterns typically pre-existing in building system sensor feeds to uncover physical relationships.

Fontugne et al. [6] propose a new method to decompose sensor signals with EMD. They extract the intrinsic usage pattern from the raw traces and show that sensors close to each other have higher intrinsic correlation. However, they do not explore the observation more deeply by answering whether there is a statistically discoverable boundary between sensor clusters in different rooms, or if there is a uniform threshold in the correlation coefficients able to be generalized to different rooms.

Fontugne et al. [7] carry on the work and propose an unsupervised method to monitor sensor behavior in buildings. They constructed a reference model out of the underlying pattens, obtained with EMD, and use it to compare future activity against it. They report an anomaly whenever a device deviates from the reference. This work exploits EMD as a method to detrend the signals and capture the inter-device relationships.

Much work utilizes EMD on medical data [1], speech analysis [8], image processing [17] and climate analysis [14]. Our method adopts EMD to determine whether a discoverable statistical boundary exists in sensors traces from sensors in different rooms and whether such a boundary can be generalized across rooms with various kinds of sensors.

3. METHODOLOGY

We start our analysis by extending the methodology used in SBS [7], based on empirical mode decomposition (EMD). In our analysis, we collect traces from several sensors and run EMD on them. This produces a set of constituent subsignals called "intrinsic mode functions" (IMF), which we separate by frequency range and re-aggregate into distinct bands. Then, we inspect the relationship between the sensors by computing the corrcoeff within a particular band, which gives us the spatial information we are interested in. Finally, we separate the result set into sub-sets, and closely examine their statistical characteristics. Before describing our methodology in detail, we introduce some definitions and notation.

3.1 Correlation

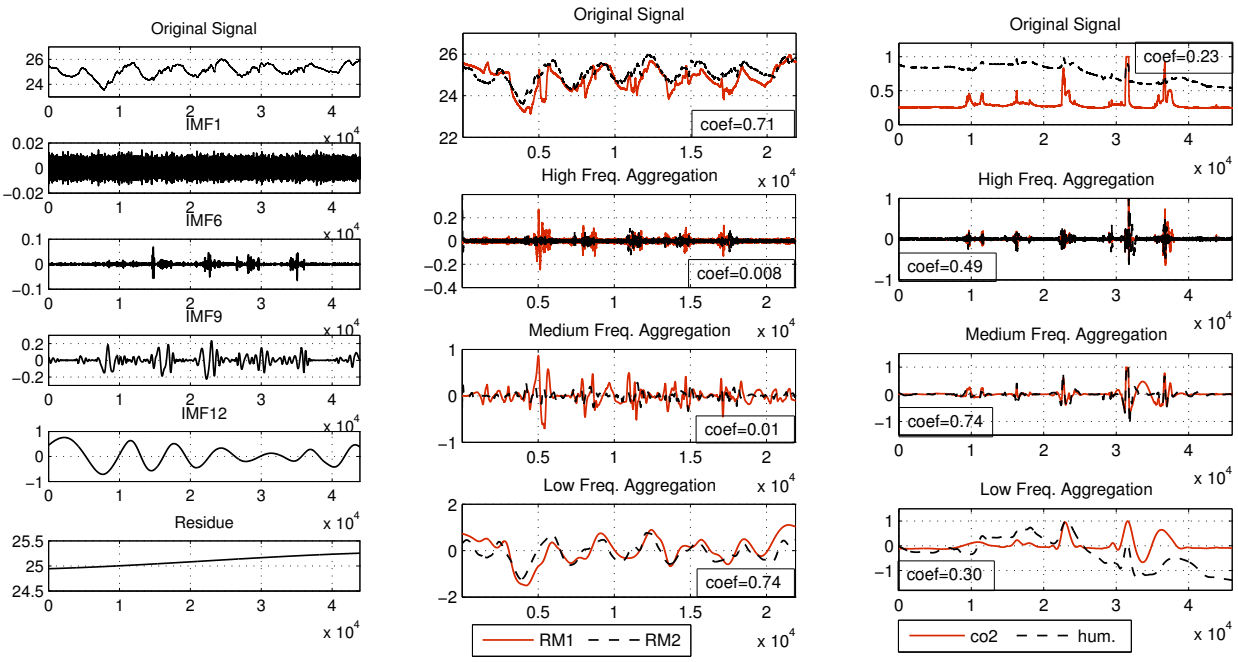
We make extensive use of the correlation coefficient function defined as:

$$r(X,Y) = r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

where X, Y are separate sets of values, n is the total number of sample points in each set, and $\overline{X}$ is the mean value of X (same for $\overline{Y}$ and Y). For each pair of sensors, we compute the corrcoeff to ascertain the relationship between them.

3.2 EMD Basics

Non-stationary signals refer to those whose frequencies change over time. The data generated in buildings is naturally non-stationary, since physical readings are highly influ-



(a) An example of EMD

(b) IMFs aggregation

(c) IMFs aggregation

Figure 1: (a) EMD decomposes a signal and exposes intrinsic oscillatory components; (b) Aggregation of IMFs within a predefined frequency range makes seemingly similar signals from different locations more distinguishable; (c) IMF aggregation makes seemingly distinct signals of different sensors in the same room show high correlation.

enced by the dynamics of physical properties in the immediate surroundings of the sensor. Empirical mode decomposition [10] is a method designed for non-linear, non-stationary signal analysis. We use it to detrend our sensor data and re-aggregate output components within a specific frequency band, based on the SBS methodology [7]. We give a quick overview of EMD and present observations from our data analysis to show a threshold for discovering the boundary between sensor feeds.

Algorithm 1: Empirical Mode Decomposition

Give signal $X(t)$:
while *the # of maxima in* $X(t) > 3$ **do**
  (1) identify all the local extrema in $X(t)$;
  (2) perform a cubic spline interpolation of maxima to get the upper envelope;
  (3) repeat (2) on minima to get the lower envelope;
  (4) $h(t) = X(t) - mean((2),(3))$;
  (5) repeat (2)-(4) until $h(t)$ is an IMF;
  (6) $X(t) = X(t) - h(t)$, and return the IMF;
end

EMD is similar to Fourier transform (FT). However, for FT to be useful the system (or signal) must be linear and the data must be strictly periodic or stationary. In contrast, EMD directly extracts components associated with different energy from the signal and generates a collection of intrinsic mode functions at different time scales. IMFs are extracted locally and normalized to fluctuate around zero. An IMF is a function with equal number of extrema and zero crossings (or differ by one at most), with its envelopes being symmetric with respect to zero. A summary of the process of EMD is depicted in Algorithm 1 and the reader is referred to [10] for further reading on EMD. The number of IMFs depends on the original signal and is automatically determined by a pre-determined stoppage criteria.

3.3 Re-aggregation

In Figure 1a, we present example of IMFs extracted using EMD. The original data is generated by a thermometer during a week deployed in a classroom in our testbed building. The graphs following show the result of the decomposition and re-aggregation methodology in SBS [7] on this signal. EMD is able to extract the predominant diurnal pattern (IMF12), induced by occupant activity, from the signal and separate distinct flows (IMF9) from other components. EMD yields distinct components in different time scales and we compute the instantaneous frequencies [11] of IMFs using Generalized Zero-Crossing [9]. We break the time scales into four frequency bands:

- High Frequency: a time scale smaller than 30 minutes, mainly reflecting the operation characteristics of devices and noise in system.
- Medium Frequency: a time scale between 30 minutes and 6 hours, which is within the time span of daily activities inside a building.
- Low Frequency: a time scale between 6 hours and 7 days.

- Residue: everything has a time scale longer than 7 days and shows long-term patterns, such as seasonal changes.

Figure 1b shows a comparison of two temperature sensor feeds from different rooms and their respective decomposition. Despite strong correlation in the raw time series, the medium frequency IMF shows little correlation. Only the low frequency diurnal pattern is correlated. Alternatively, Figure 1c shows a CO_2 trace and a humidity trace.

While the raw signals appear to be very different, and indeed have modest correlation, the medium frequency components are strongly correlated. We conjecture that the medium frequency band "records" local activity. Occupants and movement in the space affect the levels of various physical phenomenon, namely temperature, humidity, CO_2 levels, etc. Over shorter time spans, noise in the system hides the effects of local activity. Longer time-spans capture long-term trends related to weather or building operation schedules. The medium frequency band captures activities such as meetings and office occupation times. These examples illustrate the basis for an automated process. By isolating a particular component of the signal we seek to strip away common diurnal factors and also eliminate differences in the response of various sensors to environmental factors. We combine this observation with a simple classifier to derive colocation.

3.4 Distribution

Let $ts^i_{j,t}$ be a time-series for sensor j in room i observed over some time interval t. For simplicity, we ignore t in defining subsequent functions and re-introduce it where necessary. For each trace we run EMD and obtain a set of n IMFs, denoted as follows:

$$\Phi^i_j = EMD(ts^i_j) = \{IMF_{1\sim n}\}$$

IMFs are traces themselves, so we divide and re-aggregate them into the four bands, B, further described in Section 3.3.

$$B = \{H(igh), M(edium), L(ow), R(esidue)\}$$

Let the re-aggregation of the bands be denoted as:

$$Aggr(\Phi^i_j) = \left\{IMF^i_{f,j}\right\}$$

where $f \in B$. We pick the *medium* frequency band (M) to compute the pairwise corrcoeff of the sensor traces. In order to understand and characterize the boundary between sensors we consider two sets of corrcoeffs for each room; the "intra"-room set and "inter"-room set, as defined:

$$R^i_{intra,t} = \left\{r(IMF^i_{M,j,t}, IMF^i_{M,k,t})\right\},\ s.t. \forall j,k \in S_i$$

The intra set only contains pairs of sensors in the same room, so both $ts^i_{j,t}$ and $ts^i_{k,t}$ are traces from sensors in room i.

$$R^i_{inter,t} = \left\{r(IMF^i_{M,j,t}, IMF^{i'}_{M,k,t})\right\},$$

$$s.t. \forall j \in S_i,\ \forall k \in S'_i,\ i \neq i'$$

By contrast, the *inter* set contains pairs across rooms, meaning $ts_{j,t}$ is a trace from a sensor in room i and $ts_{k,t}$ is a sensor trace from some other room i'. Note the use of t in the definitions. We re-introduce t here to denote that the construction of each set is performed with respect to a specific time interval.

Finally, we examine populations, R^i_{intra} and R^i_{inter}, across multiple time intervals (in days):

$$R^i_{intra} = \bigcup_{\forall t} R^i_{intra,t},\ s.t.\ t \in \{1,3,5,7,14,21,28\}$$

$$R^i_{inter} = \bigcup_{\forall t} R^i_{inter,t},\ s.t.\ t \in \{1,3,5,7,14,21,28\}$$

We generate a CDF for each of the two populations with respect to each room. This allows us to closely examine the statistical characteristics of the relationship between sensors in the same space and those in different spaces. Each room offers a potentially different perspective on this relationship.

3.5 Threshold Analysis

In order to understand the statistical properties, we generate two corrcoeff distributions by computing the corrcoeff between pairs of traces within and across each room, as detailed in the previous section. Figure 4 shows how we divide the corrcoeff values into two sets. The figure shows two intra and two inter sets. Specifically, we examine how a choice in cut-off threshold affects the ability to separate the sets, when their separation is not known a priori, relative to each room. Our hypothesis is that there exists a computable, statistical boundary between sensors in different rooms.

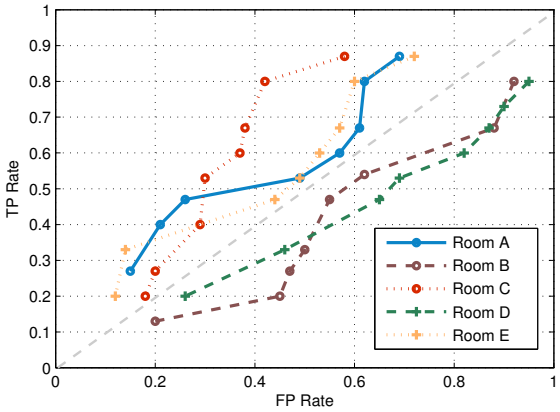
To test our hypothesis, we choose a threshold value relative to the distribution of corrcoeffs. All pairs with a corrcoeff larger than the threshold will be classified as being in the same room. To closely analyze the threshold parameter, we generate a receiver operating characteristic (ROC) curve by varying the threshold value. Then, we look for a good tradeoff point between the true-positive and false-positive rate; one that maximizes the difference between TPR and FPR. We compare the ROCs generated for our "medium" frequency band IMFs against raw-signal, cross-correlation values, in order to ascertain the extent to which the SBS [7] methodology is advantageous for discovering a statistical separation, analogous to a physical one. We also examine whether there is a uniform boundary between clusters across all the rooms.

4. EXPERIMENTAL RESULTS

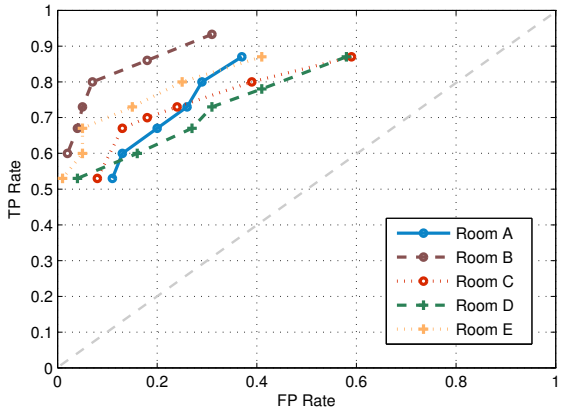
We conduct two sets of experiments. First, we quantify the sensitivity of our method for different threshold values and examine the effect of different time spans on the threshold. We then cluster the traces based on our threshold analysis and compare it with a baseline approach using multidimensional scaling and k-means.

4.1 Experimental Setup

We perform an empirical study on sensor data collected from 15 sensors across 5 rooms on 4 different floors of a large building, as detailed in Table 1. Each room has three sensors: a temperature sensor, a CO_2 sensor, and a humidity sensor. The data from these is reported to an sMAP [3] archiver. The data set used comes from a deployment [18] lasting over 6 months on several floors in Sutardja Dai Hall (SDH) at UC Berkeley, where one sensor box – which contains a thermometer, a humidity sensor and a CO_2 sensor



(a) Correlating the raw signals.



(b) Correlating the re-aggregated IMFs in the "medium" frequency band.

Figure 2: The ROC curves depict the sensitivity of the raw signal and mid-frequency IMFs to the threshold value. We choose the 0.2 FPR point as the boundary threshold for each room.

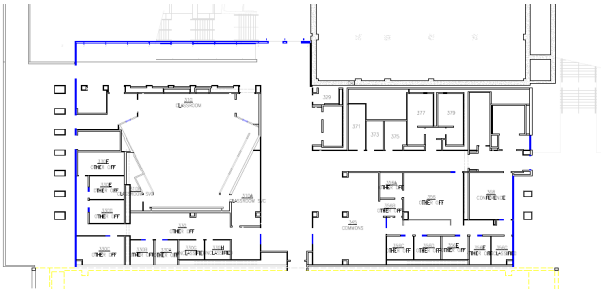


Figure 3: We collect data from 15 sensors in 5 rooms sitting on 4 different floors. This is a map of a section of the 3rd floor in Sutardja Dai Hall.

– is placed in each room. The box reports data over 6LowPAN [12] to a sMAP archiver every 15 seconds. Due to intermittent data loss, we pick a time span without interruption, starting in January until mid-Feburary, 2013, for evaluation.

Table 1: Room Specs

Room#	Orientation	Floor	Type
A	West	2	Computer Lab
B	South	4	Conference Room
C	No Window	2	Classroom
D	North	7	Conference Room
E	South	5	Conference Room

4.2 Baseline and Metrics

As a baseline, after we generate the two distributions described previously, we apply multidimensional scaling (MDS) to the corrcoeff matrix, in order to transform the original high-dimensional relative space to a 3-D space with an absolute origin, and run the k-means clustering algorithm. We choose the true-positive rate (TPR, also known as recall rate) and false-positive rate (FPR) as metrics to evaluate the performance of our method versus the naive approach, which correlates the raw traces. A true-positive (TP) is when a sensor pair in a room is classified as being co-located while a false-positive (FP) is when a sensor that is not in room is classified as being so.

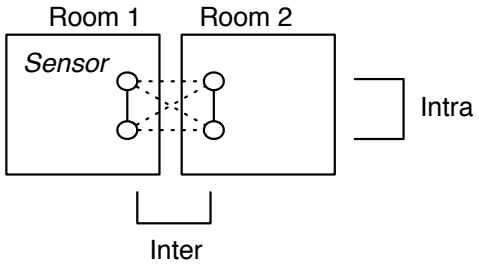


Figure 4: Two populations are examined for our threshold analysis. A solid line connects sensors in the same room while a dotted line connects to a pairs in different rooms.

4.3 Characterizing the Boundary

To corroborate our boundary-existence hypothesis, we first need to characterize the boundary between sensors in different rooms. We compute the pairwise correlation coefficients (corrcoeffs) between sensor traces in both of populations depicted in Figure 4, over different time spans – ranging from one day to one month. After generating points over different time spans for each room, we accumulate the corrcoeffs to obtain distributions as shown in Figure 5, for each of the five rooms.

The dashed vertical lines in Figure 5 represent an arbitrary threshold that partitions the distribution into two sets. Pairs of sensors to the right of the line are classified as being in the same room. Pairs of sensors to the left are classified as being in different rooms. The CDFs on the left column show the distribution of corrcoeffs for pairs known to be in the same room and the CDFs on the right show the distribution of corrcoeffs in different rooms. Note in the figure, we set the threshold to the same value to both the left and

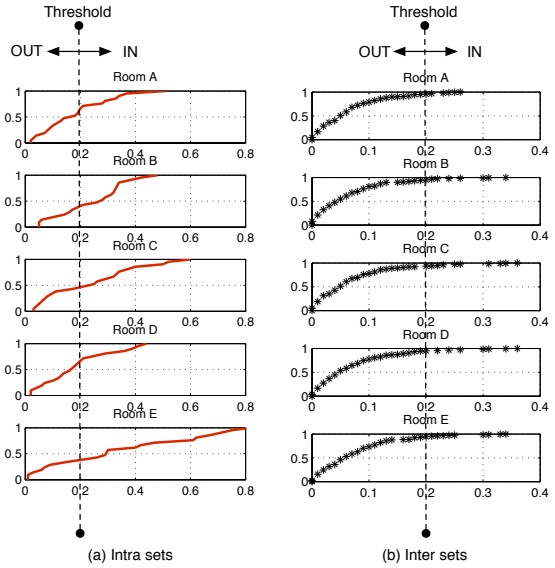


Figure 5: CDF of correlation coefficients between IMFs of sensor feeds: the dotted lines point to some threshold which divides the distribution and produces a TPR and FPR.

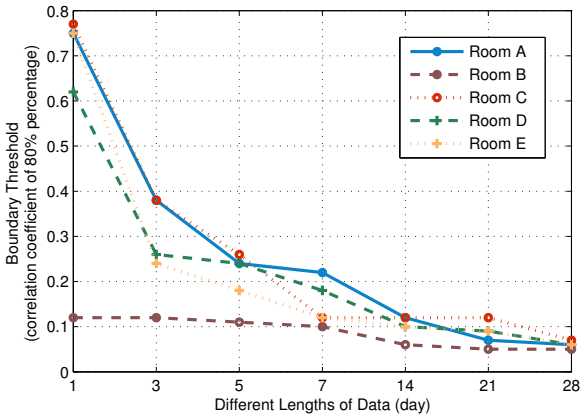


Figure 6: The threshold values all converge to a similar value and we can derive the optimal value with as minimal as 14 days data.

right side, in order to observe the effect of the true/false positive rates. By adjusting the threshold, we get different TPRs/FPRs parameterized by the threshold. Figure 2 captures the range tradeoff in a corresponding ROC curve.

Figure 2 illustrates the TPR/FPR sensitivity to different threshold values for our method and the naive approach. A good cluster achieves a high TPR and a low FPR. As we vary the threshold, we see that our approach achieves a TPR between 52%–93% and a FPR between 5%–59%. We can see that the average TPR for the ROC graph on the right is higher than the ROC graph on the left. Moreover, the corresponding average FPR is lower on the right than on the left. In general, as the TPR rises, the FPR also goes up – *a tradeoff exists between maximizing TPR and maintaining a lower FPR.*

The "boundary" is represented as the corrcoeff that produces a "good" TPR with an "acceptable" FPR. In Figure 2b, we choose 0.2 FPR as the boundary threshold. This point represents the largest difference between TPR and FPR – an acceptable tradeoff point. Looking at Figure 5, the 0.2 FPR corresponds roughly to the 80th-percentile correlation coefficient, on the "inter" set (the set of CDFs on the right). The recall rate for each room – using a 80th-percentile corrcoeff threshold value – ranges between 62%-86% and the threshold value falls into a narrow interval between 0.1 to 0.12. This shows that *we are able to choose a uniform value for all the rooms regardless of the sensor type.*

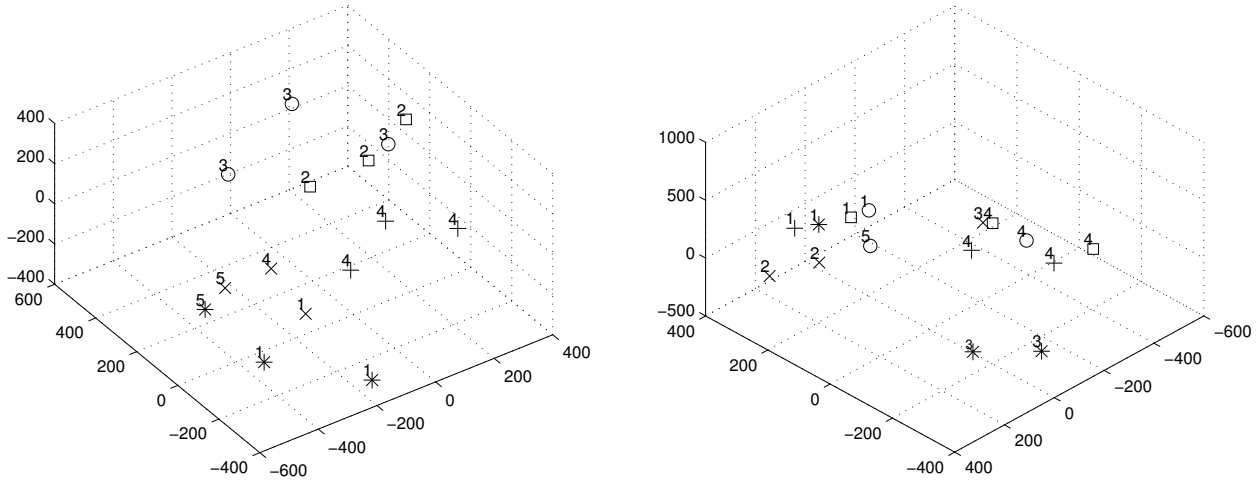
4.4 Convergence over Time

Using the threshold the roughly 80th-percentile corrcoeff corresponds to in the distribution, we examine how it affects the classification rate across traces that span different lengths of time. Convergence and consistency across different time spans is critical to automate the parameter selection process. Observe how the threshold values differ quite significantly in Figure 6. However, the threshold values gradually converge, as the length of training data increases from one day to one month. The values derived after 14 days of data are approximately the same as the final convergence value (around 0.07). In other words, we can determine a threshold from two weeks of data.

4.5 Clustering Results

We cluster the sensor traces over the entire one-month period, and use the roughly 80th percentile corrcoeff (0.07) as the boundary threshold. A sensor is classified into the cluster with the largest corrcoeff. The clustering result is shown in Table 2. A "1" means the sensor is classified as inside the corresponding room. In general, after obtaining the sensor clusters, we don't know which room each cluster corresponds to without further information such as the metadata of sensors. The labels "A-E" in Table 2 are used to indicate the ground truth of where each sensor is physically placed since we have such information. Overall, the classification accuracy is 93.3%. We do not cluster on the corrcoeffs obtained among raw signals because the 80%-percentile corrcoeff values do not converge across rooms. The reason that we are able to get such a high accuracy, which is seemingly different from the statistics in Figure 5 and Figure 2, is because the statistics in the two figures are generated out of the corrcoeffs accumulated over different time spans (the same intervals in Figure 6) while the clustering here is performed on the corrcoeffs from the entire one-month period.

To compare with our threshold-based method, we also cluster using a baseline approach. The pairwise corrcoeff for sensors in different rooms can be interpreted as a "distance" between them. A larger coefficient indicates a closer "distance", and vice versa. However, since the distances between pairs is relative, we use multidimensional scaling [5] to find a common basis in three dimensions, re-map the relative distance metric (feature vector) into this three-dimensional grid and use k-means to classify the traces. We set k to equal the number of rooms, since the goal of the approach is to verify spatial placement at room-level granularity. Generally, we believe that k should equal the number of rooms you wish to classify the sensors into. The clustering results are shown in Figure 7. Ground truth is shown through different markers (x, o, +, star, box). Each marker stands for



(a) Clustering on corrcoeffs from our method.

(b) Clustering on corrcoeffs from the naive approach.

Figure 7: Clustering with k-means on the corrcoeff matrix after applying multidimensional scaling (MDS): The EMD-based set achieves an accuracy of 80% while the results with raw-trace is only 53.3% classification accuracy.

	A	B	C	D	E	
$Sensor A_1$	1	0	0	0	0	✓
A_2	1	0	0	0	0	✓
A_3	1	0	0	0	0	✓
B_1	0	1	0	0	0	✓
B_2	0	1	0	0	0	✓
B_3	0	1	0	0	0	✓
C_1	0	0	1	0	0	✓
C_2	0	0	1	0	0	✓
C_3	0	0	1	0	0	✓
D_1	0	0	0	1	0	✓
D_2	0	0	0	1	0	✓
D_3	0	0	1	0	0	×
E_1	0	0	0	0	1	✓
E_2	0	0	0	0	1	✓
E_3	0	0	0	0	1	✓

Table 2: Clustering result using the thresholding method: a "1" means the sensor is classified as inside the room. We get the "✓" and "×" by comparing the clustering results with ground truth.

one room. The cluster each sensor assigned to is denoted with a number. The classification accuracy of the baseline approach on corrcoeffs matrix of re-aggregated IMFs is 80%. For raw traces, the baseline approach achieves an accuracy of only 53.3%.

4.6 Discussion

Bi-modal Distribution.

From the results illustrated in Figure 5, we observe a bi-modality in the corrcoeff distribution for the two population sets. Sensors in the same room correlate to each other more (typically a corcoeff of 0.4 or higher) than sensors in different rooms. This bi-modal distribution may provide insight for us to understand the boundary and search for an effective discriminator more broadly.

Across Different Sources.

To further validate the effectiveness of the proposed method, we should consider using data from different sources. For example, in room B in Sutardja Dai Hall, there are two different sets of temperature sensors reporting data at different rates and granularities. We demonstrate our ability to classify sensor streams on the same platform (recall the sensor box we used to collect data). It would be more convincing to verify the effectiveness of our method with sensor streams generated from devices on different systems – since separate systems are independent. For instance, we can use temperature data from the second deployment and use the CO_2 and humidity sensor data from the first deployment and compare the results to what we have gathered.

Generalizability.

In our results, the boundary threshold parameter converges to a narrow interval, as the data set expands over a longer time range. This may suggest that our method generalizes across rooms in a building, although further validation in a larger, more representative data set is necessary. This study looked at 5 different rooms with a large physical separation from one another. A more representative data set would consider all the rooms and pay special attention to rooms that share a common orientation and are separated by a single wall or floor slab.

Based on this study, and the previous, related one [6], we conjecture that local activity modulates various types of physical signals – captured by the various kinds of physical sensors embedded throughout the building – and that those signals are attenuated over distance and physical boundaries (such as walls). We believe that this is what drives our observations. If the conjecture is true, the effects will be less pronounced in larger rooms, such as an auditorium or a large laboratory space.

As our approach performs slightly better than traditional learning techniques, we must further evaluate its robustness

versus the baseline method; across the entire building and across multiple buildings. In future work, we will examine the two approaches across larger intra-building data sets and compare results across multiple buildings. A key factor is the variance of classification accuracy – smaller variance demonstrates robustness.

5. CONCLUSION

We present a new method for spatial placement clustering. We first characterize the corrcoef distribution of medium frequencies IMFs between sensors in the same/different room(s), and then we learn the tradeoff between achieving a higher TPR and maintaining a lower FPR by manipulating a discriminator parameter within these two distributions. For a preliminary sample of relatively well separated rooms, we find that there is a clear boundary between sensor clusters in terms of their spatial placement and the boundary can be probed statistically. We also find a uniform discriminator can be learned and generalized across these rooms. For this initial study, our method is able to classify the sensors of 93.3% accuracy, which is 13% higher than a tradition k-means approach, with a TPR between 62%-86% and a FPR less than 20%.

We believe that our results come partly from the fact that these rooms contain such distinct characteristics and activities that they can be perfectly separated with our approach. However, given a more complete and representative data set, such as rooms facing the same side of the building on the same floor, it is not clear how effective the method might be. In future work, we will examine how far this method takes us and explore how it may be used in combination with different techniques to improve the results more generally. Automated metadata verification is important to include in the lifecycle of building data management.

6. ACKNOWLEDGEMENT

Thanks to our shepherd, Kin Cheong Sou, and the anonymous reviewers for helpful comments. Thanks to Albert Goto, for making the first author's stay at UC Berkeley comfortable. Sincere gratitude goes to Kamin and David, for letting this collaboration happen. This work was partly funded by the NSF grants EFRI-1038271 and CPS-1239552.

7. REFERENCES

[1] A. Arafat and T. Hasan. Automatic detection of ECG wave boundaries using empirical mode decomposition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[2] Y. Chen and L. Tu. Density-based clustering for real-time stream data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, 2007.

[3] S. Dawson-Haggerty, X. Jiang, G. Tolle, J. Ortiz, and D. Culler. sMAP: a simple measurement and actuation profile for physical information. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, SenSys '10, 2010.

[4] Department of Energy. 2011 Buildings Energy Data Book. `http://buildingsdatabook.eren.doe.gov/`.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[6] R. Fontugne, J. Ortiz, D. Culler, and H. Esaki. Empirical mode decomposition for intrinsic-relationship extraction in large sensor deployments. In *Workshop on Internet of Things Applications*, IoT-App'12, 2012.

[7] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler, and H. Esaki. Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. In *Proceedings of the 12th international conference on Information processing in sensor networks*, IPSN '13, 2013.

[8] H. Huang and J. Pan. Speech pitch determination based on Hilbert-Huang transform. *Signal Processing*, 86(4):792 – 803, 2006.

[9] N. E. Huang. Computing frequency by using generalized zero-crossing applied to intrinsic mode functions, January 2006.

[10] 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*, 1998.

[11] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chen, and K. Blank. On Instantaneous Frequency. *Advances in Adaptive Data Analysis*, 1(2), 2009.

[12] J. W. Hui and D. E. Culler. Extending IP to Low-Power, Wireless Personal Area Networks. *Internet Computing, IEEE*, 12(4), July-Aug. 2008.

[13] K. Kapitanova, E. Hoque, J. A. Stankovic, K. Whitehouse, and S. H. Son. Being SMART about failures: assessing repairs in SMART homes. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, 2012.

[14] T. Lee and T. B. M. J. Ouarda. Prediction of climate nonstationary oscillation processes with empirical mode decomposition. *Journal of Geophysical Research: Atmospheres*, 116(D6), 2011.

[15] J. Lu and K. Whitehouse. Smart blueprints: automatically generated maps of homes and the devices within them. In *Proceedings of the 10th international conference on Pervasive Computing*, Pervasive'12, 2012.

[16] Y. Ma and F. Borrelli. Fast stochastic predictive control for building temperature regulation. In *American Control Conference (ACC), 2012*, 2012.

[17] H. Mohammadzade, F. Agrafioti, J. Gao, and D. Hatzinakos. BEMD for expression transformation in face recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[18] J. Taneja, A. Krioukov, S. Dawson-Haggerty, and D. E. Culler. Enabling Advanced Environmental Conditioning with a Building Application Stack. Technical Report UCB/EECS-2013-14, EECS Department, University of California, Berkeley, Feb 2013.