Sensing by Proxy in Buildings with Agglomerative Clustering of Indoor Temperature Movements

Daoyuan Li
daoyuan.li@uni.lu

Tegawendé F. Bissyandé
tegawende.bissyande@uni.lu

Jacques Klein
jacques.klein@uni.lu

Yves Le Traon
yves.letraon@uni.lu

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg
Luxembourg

ABSTRACT

As the concept of Internet of Things (IoT) develops, buildings are equipped with increasingly heterogeneous sensors to track building status as well as occupant activities. As users become more and more concerned with their privacy in buildings, explicit sensing techniques can lead to uncomfortableness and resistance from occupants. In this paper, we adapt a sensing by proxy paradigm that monitors building status and coarse occupant activities through agglomerative clustering of indoor temperature movements. Through extensive experimentation on 86 classrooms, offices and labs in a five-story school building in western Europe, we prove that indoor temperature movements can be leveraged to infer latent information about indoor environments, especially about rooms' relative physical locations and rough type of occupant activities. Our results evidence a cost-effective approach to extending commercial building control systems and gaining extra relevant intelligence from such systems.

CCS Concepts

•Information systems → Clustering; •Computer systems organization → Sensor networks;

Keywords

Sensing by proxy; smart buildings; occupancy inference

1. INTRODUCTION

Citizens in a modern society spend a majority of their time everyday inside buildings working or relaxing. In turn, buildings consume a surprisingly large portion of total energy consumption by all sectors. For example, 41% of energy consumption attributes to buildings in the US and buildings consume even more energy than industry in the EU [11]. Many initiatives have thus been proposed to combat the energy consumption issue. As the concept of Internet of Things (IoT) develops, more and more proposals focus on devising more efficient Building Energy and Comfort Management (BECM) systems, which try to fulfill users' comfort requirements while reducing energy footprints for building operations including heating, ventilation, and air conditioning (HVAC), lighting and plug loads. Indeed, research has shown that BECM systems can potentially reduce buildings' energy footprints in both simulated and real-world evaluations.

BECM systems often involves taking advantage of heterogeneous sensors, such as passive infrared (PIR) sensors, cameras, motion and presence detectors, and environmental sensors like temperature, humidity, CO_2, etc., to monitor the status of the building and especially occupant activities, since conservative behaviors can help reducing building operation energy consumption by one-third compared to design point benchmark while careless ones may increase energy footprints by one-third [11]. However, the assumption of tracking individual occupants in real-time is largely impractical and rarely adopted in real-world cases due to technological, construction and maintenance cost and privacy challenges.

To tackle with these challenges, we seek to devise an plug-and-play and cost-effective approach that takes advantage of existing BECM systems and investigates the feasibility of gaining extra intelligence about corresponding buildings and their occupants from such systems. Since user behaviors have a large impact on their indoor environments, we can profile these behaviors by proxy of the resulting impacts they have made. To be exact, sensing by proxy here refers to inferring latent factors with indirect measurements on activity traces rather than directly measuring activities. To make our approach more applicable to different scenarios and cost-effective, we take advantage of mature and widely deployed temperature sensors. We have conducted this study using real-world settings and all our data has been collected from a school building in western Europe, which was planned and constructed around 2000 and is equipped with basic sensors and actuators (e.g., temperature sensors and HVAC system)

SAC 2017,April 03-07, 2017, Marrakech, Morocco

http://dx.doi.org/10.1145/3019612.3019699

to facilitate building operations. The main contributions of this paper include:

1. We adopt a sensing by proxy paradigm to reduce costs and relax users' privacy concerns. We confirm that agglomerative clustering of temperature evolution of indoor environments (with minimal occupant activities) produces accurate adjacency maps with regard to the physical location of each temperature sensor. This adjacency map is helpful and can be complement to indoor floor plan inference and localization.

2. We prove that it is feasible to infer coarse-grained intelligence about occupant activities using agglomerative clustering of temperature evolution of indoor environments with occupant activities even if data have been collected with low frequency in a non-intrusive manner.

3. We provide a data analytics tool that extends off-the-shelf BECM systems for smart homes and smart buildings that helps owners and operators to understand the overall status the buildings with regard to its previous status. Our approach can also be used to track anomalies within buildings.

The remainder of this paper is organized as follows. Section 2 prepares readers with the necessary technical background and Section 3 introduces works that are related to ours. We present our methodology in Section 4 and real-world experiment results in Section 5. Finally we conclude with future research directions in Section 6.

2. BACKGROUND

In this section, we introduce the necessary technical background to facilitate understanding of this paper. Specifically, we present the concept of time series and its typical manipulations, especially time series clustering and the Ward's method, which implements agglomerative clustering based on a sum-of-squares criterion.

We consider sensor readings as time series data. That is, each data point p_i is a (t_i, v_i) tuple where v_i is the numerical value recorded by a sensor at timestamp t_i; a time series T is then a collection of data points that are ordered by their corresponding timestamps, i.e., $T = [(t_0, v_0), (t_1, v_1), ..., (t_{(n-1)}, v_{(n-1)})]$. Thanks to its intrinsic timestamps, time series can be easily resampled, interpolated and extrapolated. Besides, it is also common practice in time series mining community that timestamps can be safely ignored for even-spaced time series sampled at a specific frequency, i.e., those whose timestamps are strictly periodic. In this case, time series are represented as $T = [v_0, v_1, ..., v_{(n-1)}]$. Time series are a common type of data that are frequently found in IoT, medical and health care, and financial applications [8]. Our previous work [6] has also taken advantage of time series to profile household electrical appliancesand this language modeling based approach has been generalized to apply to time series in different domains [9, 7].

When evaluating the similarity or dissimilarity of two time series instances X and Y ($|X| = |Y| = n$), a distance measure $D(X, Y)$ is defined. Popular distance measures for time series include Euclidean distance $D_{Euclidean}(X, Y) = \sqrt{\sum_{i=0}^{n-1}(X_i - Y_i)^2}$ that maps the i-th point in X to the i-th point in Y, and Dynamic Time Warping (DTW) distance, which tries to find the best way to warp the time axis and as a result aligns X and Y differently. As shown in Figure 1, an i-th point in X can be mapped to a j-th point (it is possible that $i \neq j$), and one point in X can be mapped to multiple points in Y. For Euclidean distance, the gray dotted lines indicating the mapping would all be vertical.

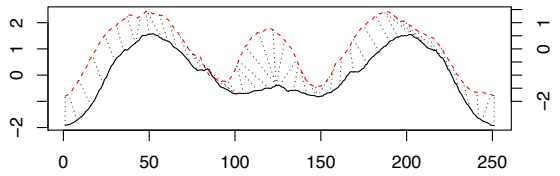

Figure 1: Illustration of how DTW aligns two time series and calculates their distance.

Time series clustering is a common type of unsupervised time series mining task that tries to partition time series into homogeneous groups while maximizing within-group similarity and between-group dissimilarity [10]. Clustering algorithms can be categorized into different families based on their underlying models, for instance hierarchical clustering which is based on connectivity and centroid-based clustering (e.g. k-means) where clusters are represented by a representative point. In this paper, we are especially interested in the former, since hierarchical clusters can be represented as a dendrogram, which depicts the hierarchy arrangement of clusters that can be merged with another at certain distances. Hierarchical clustering do not attempt to generate an arbitrary number of clusters. Instead, it produces a hierarchy that is easier for users to understand and users can set the break points by themselves. Hierarchical agglomerative clustering has recently received great interests in pattern recognition and become especially popular in financial applications. Figure 2 shows an example of hierarchical clustering results, where companies with similar business domain and activities have been grouped together. Note that in dendrograms, the height of a branch indicates how different it is from another while the horizontal orientation is generally irrelevant.

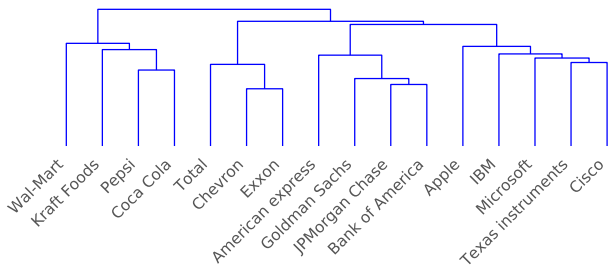

Figure 2: Example dendrogram from hierarchical clustering, using daily stock price variations from January 2012 to January 2016.

Hierarchical clustering can employ either a bottom-up (agglomerative) or top-down (divisive) approach. The former starts with a single instance from the dataset and gradually aggregates instances into clusters until all instances are grouped into a single cluster, while the latter starts with the whole dataset and iteratively divide the dataset into clusters. In general, agglomerative methods are computationally more efficient than divisive ones. Thus we favor the former, and especially the Ward's method [16], which is a popular algorithm used to minimize the total within-cluster variance. Recall that Ward – as an agglomerative approach – works incrementally, the distance (namely the Ward's Linkage) of clusters $I \cup J$ and K are calculated based on a distance update formula $D_{Ward}(I \cup J, K)$ as specified below:

$$\sqrt{\frac{(|I|+|K|)D(I,K)^2 + (|J|+|K|)D(J,K)^2 - |K|D(I,J)^2}{|I|+|J|+|K|}}$$

where I and J are two clusters to be joined into a new cluster and K is any other cluster, and $|*|$ denotes the number of instances in one cluster. The computational complexity of Ward is $O(n^2)$, where n is the size of the dataset. Ward is widely available in many software packages, for example Matlab and Wolfram Mathematica.

3. RELATED WORK

Recent research on BECM systems focuses on collectively taking advantage of both real-time occupancy information and occupant preferences when designing more efficient building control systems. For instance, Chen et al. [3] propose a BECM system that keeps track of occupants' real-time location to enable fine-grained control of ambient environment including lighting, cooling, heating, etc. As sensors and actuators are deployed in buildings and these systems are connected to external networks such as the Internet, occupant security and privacy become a more challenging task since sensor data can be leveraged to make unwanted inferences about occupants and their behaviors [19]. For instance, Yang et al. [17] have conducted empirical experiments using motion sensors in a three-person single-family home and electricity meters in a twelve-person university lab, and shown that data from these sensors can enable inferring real-time occupancy and even occupants' identities. Another approach [18] takes advantage of RFID technologies and implements a localization algorithm that learns about the location of occupants.

Information about indoor environment is important for many applications including indoor localization services, security services like access control and alarms, and privacy protection. However, this information is often either unavailable or obtaining it is time-consuming due to effort-intensive negotiations with building operators. As a result, many approaches have been proposed to explore and infer indoor environments. Earlier approaches takes advantage of laser scanners [12] to infer and reconstruct indoor floor plans, while more recent works leverage mainly commodity sensors available on smartphones [15] and takes a crowd sensing approach. For instance, CrowdInside [1] takes advantage of sensors (including accelerometers, magnetometers, gyroscopes, etc.) on smartphones to construct occupants' motion traces and then infer floor plan as well as room and corridor shapes. [4] introduces another indoor floor plan construction system that takes advantage of Wi-Fi signals to construct room adjacency graphs and leverages user motion data collected from smartphones to to estimate room sizes and orders. Unlike these approaches that involve taking advantage of heterogeneous or *ad hoc* (specific purposed) sensors, our approach does not require any sophisticated sensing hardware and utilizes only indoor temperature sensors, which are commonly found in modern buildings with HVAC control systems.

Indoor occupant activity inference and detection is another research trend since more and more sensors are installed in buildings and occupants are becoming increasingly concerned about their own privacy. For instance, motion sensors and smart meters can be used for detecting whether a room is occupied and even for analyzing occupant identities [17]. A more recent work [14] explores the resonance effect of rooms and devise models to infer the number of occupants by observing changes in the ultrasonic spectrum reflected back from a centrally located ultrasonic chirp transmitter. Our work has been largely influenced by that of Jin et al. [5], where the authors try to infer implicit factors by indirect measurements based on the physical environment. They argue that occupancy can be inferred by indoor CO_2 concentration. Since CO_2 sensors are not as widely available as temperature sensors, we try to investigate the sensing by proxy paradigm using temperature sensor readings. Note that our work mainly concerns inferring indoor environments and occupant activities from sensor data, instead of exploring the vulnerability of networking protocols such as KNX [2].

4. METHODOLOGY

Since different buildings operate with different BECM systems, to extend such systems we have to find a common interface or common type of data when conducting latent sensing. Fortunately, temperature sensors are usually available in the majority of BECM systems because of the requirements by HVAC devices. Besides the availability, we believe indoor temperature movements are largely influenced by both natural factors and occupant behaviors, making temperature sensors a perfect data source for inferring relevant information from buildings and their occupants behaviors. For instance, we have collected data – including temperature measurements and set-points, lighting, alarms, etc. – from the BECM system located in a school building, and indoor temperature records attribute to a significantly large portion in our database. To be exact, around 60 percent of total records are indoor temperature measurements, while in comparison outdoor weather station data contributes around 15 percent. In this section, we introduce the whole pipeline of our approach from collecting data, processing these data so that it fits the agglomerative clustering algorithm, to the validation process.

4.1 Data Collection and Processing

We are interested in collecting indoor temperature data from buildings' BECM systems. Fortunately, majority of BECM

systems are based on open communication standards such as KNX[1] or LON[2]. As a result, it is easy to get a compatible watchdog module and simply attach it to the control bus and start collecting data. We store all collected data in a database for ease of querying and retrieving purposes.

Note that in practice sensor readings generally exhibit different statistical characteristics. For instance, different temperature sensors report temperatures at different frequencies and the amplitude of values may also differ. Besides, abnormal and missing values are very common, making the collected data quite noisy. To proceed processing the data, we have to conduct data cleansing tasks as specified below:

1. Resampling. Specifically, we choose to down-sample data records using a uniform frequency of one hour for temperature readings. This process helps reducing noises as noisy data points can be filtered out. Furthermore, down-sampling can greatly reduces the dataset size and improve computation efficiency.

2. Interpolation. Some sensors may be missing values at certain timestamps even after down-sampling. Missing values are common in our case due to sensor failures and occasional server shutdown. There are many missing value imputation techniques [13], however, we choose to linearly interpolate these missing values since temperatures of indoor environments do not tend to change drastically.

3. Normalization. Temperature sensors in different locations may report values of significantly different amplitude levels. This can result in inaccurate computation especially with distance measures such as the DTW distance. In this paper, we are more concerned with the oscillation development for readings from each individual sensor. As a result, we divide each sensor's readings by their corresponding standard deviation for normalization, i.e., $t'_i = t_i / \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(t_i - \mu)^2}$ for $0 \leq i < n$, where $\mu = \frac{1}{n}\sum_{i=0}^{n-1} t_i$. This step is especially important if we are concerned with the overall temperature movements instead of the absolute terms, which are easy to get using simple algebras and statistical methods.

Afterwards, we calculate the pairwise distance of temperature movements of different rooms (in the form of time series) using both Euclidean and DTW distance and generate a corresponding distance matrix, which is then fed as input to the agglomerative clustering process.

4.2 Baseline Establishment and Validation

Since our intuition is that temperature movements are mainly influenced by natural factors and occupant behaviors, to validate this assumption we have to separate the influence of such two factors. If we can achieve this, then we can continue investigating how each factor contributes to the temperature movements. Separation of natural factors and occupant behaviors can be easy and sometimes maybe trivial since we can just find out when occupants are in the building. For example, offices and schools usually have a consistent schedule that tells us when rooms are occupied (e.g. daytime on weekdays) or not (e.g. nighttime or holidays). For simplicity, we choose to split all temperature data into daytime and nighttime readings.

When investigating how each factor contributes to indoor temperature movements, consider first the natural factors. It is obvious that the physical location can have the biggest impact on room temperature. For instance, rooms facing south (in the Northern Hemisphere) would generally fluctuate more drastically than rooms always in shadows, provided that all rooms in the same building have the same heat insulation features. In this case, we can validate the clustering results against the floor plans in order to evaluate the impact of natural factors. On the other hand, since human activities can greatly impact indoor environments, clustering temperature movements under human influence will tell us more about the actual activities.

In summary, agglomerative clustering of indoor temperature movement data collected when minimal occupant activities are present will likely tell us more information about the physical locations of rooms; while clustering of temperature movement data when occupant activities are present will likely enable us infer how close activities in one room is to those in another room. We continue validating these assumptions with real-world data in the following section.

5. EXPERIMENTAL EVALUATION

In order to validate our assumptions, we have conducted experiments using data from a real-world building that is in daily use. In this section we present the experiments and their results. We present our research questions (**RQs**) and answer then along with experiment results.

5.1 Experiment Subject and Data Collection

All our data has been collected from a single school building in western Europe, which has around 100 classrooms, labs and offices located on five different floors and 86 of these rooms are equipped with a single temperature sensor in each room. This building was planned and constructed around 2000 and most rooms as well outside facades are equipped with sensors and actuators to monitor and control temperature and heating, ventilation, illumination, etc. for the ease of building operations. In total, this building has more than 1000 connected sensors and actuators. All these sensors and actuators (such as light switches and dimming units) have been connected to a KNX bus, which is a broadcast networking protocol where all communication telegrams passes on the bus and pre-configured source/destination pairs may send and receive only relevant telegrams and react. KNX is a very popular building control protocol that has been deployed in several millions of installations worldwide.

We have implemented a system to collect data from the building, as shown in Figure 3. The broadcasting nature of KNX makes it easy for us to simply attach a KNX-to-USB interface to the KNX bus and listens to every telegram on the bus to a gateway server via the USB interface (in our case, it is a Weinzierl KNX USB Interface 311, which

[1]http://www.knx.org/
[2]http://www.lonmark.org/

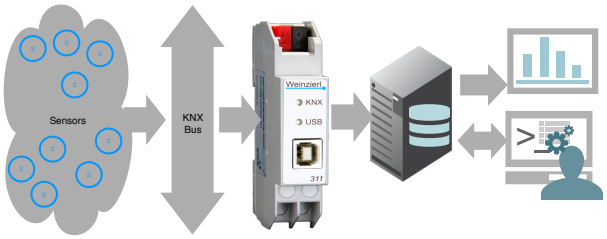

Figure 3: Overview of data collection and management process.

costs around 200 Euros). This gateway then parses and stores all KNX telegrams to a Linux server. We started collecting data from this building from mid February, 2016. In this paper, we have used data collected over a span of five months from February till July, 2016. Each record in our database consists of information such as telegram timestamp, source and destination addresses, KNX telegram type and message (generally a numerical value) parsed from this telegram. The commercial BECM system used by the school provides only a interface to monitor real-time readings from each sensor and no history data were store anywhere. As a result, we have also developed a dashboard (a web application) for the building operators to monitor and view charts about the real-time as well as historical records from sensors and actuators that are interesting to the building operator. In a backend, users and operators may configure a more customized dashboard interface by themselves.

5.2 Inferring Indoor Environment

Following our intuition that rooms located physically together should share similar patterns in room temperature movements due to similar natural influences such as sunshine, rain and wind. Optimally, human occupant impacts need to be ruled out when recording room temperatures that will be utilized for agglomerative clustering. To that end we choose those time periods when there is minimal occupant activities. Since our experiment subject is a school building and no one is in the building during night time, we split temperature readings into two subsets – daytime (07:00 to 19:00) and nighttime (19:00 to 07:00) readings – and explore only the nighttime temperature records. We start our research by investigating **if temperature sensor records can be used to correctly group physically nearby classrooms other than mere temperature fluctuations (RQ1).**

To answer **RQ1**, we extracted all the temperature records and down-sampled them to one-hour frequency to establish a tradeoff between reducing dataset size and lowering the amount of missing value interpolation. After normalization, we then calculate the pairwise Euclidean and DTW distance for all the preprocessed data and feed these distances to agglomerative clustering using Ward's linkage algorithm. We have experimented with data from all five floors. In order to make the readers understand better, we start with the top floor where temperature readings are available in only six classrooms.

Figure 4 shows the simplified floor plan (to protect the privacy of the school) and the temperature readings in each room for around five months. For human eyes, indoor temperature does not fluctuate greatly and these readings from different classrooms look more or less similar along the course. Especially, when the whether gets warmer, the temperature differences among different classrooms become smaller. By generating a distance matrix diagram of temperature movements from different rooms where darker blocks indicate more differences rather than similarities (cf. Figure 5 left), it may take some time (even for experts) to identify that $R6$ and $R4$ are more different than other classrooms. However, such a matrix does not tell us (or building facility management teams) about what can be the potential cause. Finally, note that the distance matrices in our case are symmetric, since $D(X, Y) = D(Y, X)$ for both Euclidean and DTW distance. We present whole matrices in this paper for the sake of straightforwardness.

When applying agglomerative clustering techniques on the temperature movements (cf. Figure 5 right), our approach has produced two bottom level clusters $\{R1, R3\}$ and $\{R2, R5\}$, and moving up from the latter, $R4$ can be attached to $\{R2, R5\}$ to form a larger cluster, which can then be joined by $R6$. These results accurately corresponds with our floor plan in Figure 4, since rooms $R1$ and $R3$ are indeed located next to each other, and the so are the others. Also, $R5$ has been clustered with $R2$ but not with $R1$, probably due to the fact that there is a staircase between $R5$ and $R1$, while $R2$ and $R5$ are physically near each other. As a result, the answer to **RQ1** is indeed positive: room temperatures are good indicators of sensor adjacency (and thus physical adjacency of rooms) inference using agglomerative clustering.

Following **RQ1**, we wonder **how much data is needed for accurate inference of sensor adjacency (RQ2)** and **if Euclidean and DTW distance make a difference to clustering results (RQ3)**. To that end, we repeat our previous clustering process with the time span of room temperature readings using a sliding window with size varying from one to 160 nights and conduct the pairwise distance calculation with both Euclidean and DTW distance. In total, we have generated 26,080 dendrograms, which we programmatically validate if each dendrogram conflicts with our floor plan. A clustering output is considered as an error if any two rooms fall into one cluster in the dendrogram while these two are not strictly next to each other according to the floor plan, for instance, when a dendrogram reads that $R2$ and $R6$ or $R4$ and $R1$ belong to one cluster. Due to wide hallways, we do not consider classrooms located on different side of the hallway as close to each other.

We present the error rate with each time span setting in Figure 6. It is obvious that the more data we use for clustering, the more accurate results are. And in our case with six classrooms, **using room temperature readings (with one-hour sampling frequency) during a span of three months produces clusters strictly correlated with regard to the floor plan (RQ2).** This span may seem long, however, in practice it may still be faster than effort-intensive negotiations with building operators (which took more than six months in our case to get the building floor plans). Besides, in this experiment we use hourly sampled temperature during nighttime, it is thus probable that using temperature readings with higher sampling rate would re-

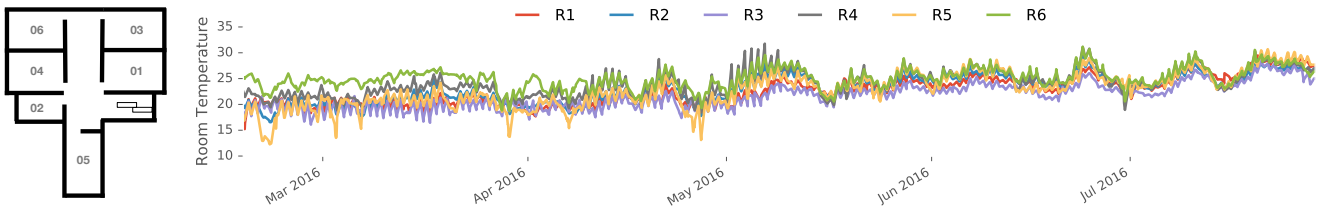

Figure 4: Simplified floor plan (left) and temperature readings during a course of around five months (right).

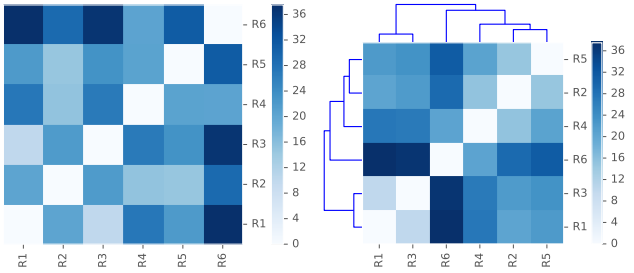

Figure 5: Distance matrix of temperature movements with Euclidean distance (left) and agglomerative clustering clustergram of temperature readings for six rooms with Ward (right).

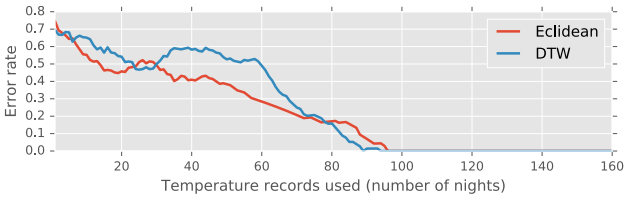

Figure 6: Error rate regarding the amount of data used for agglomerative clustering.

duce the inference time. Regarding **RQ3**, it is obvious from Figure 6 that both Euclidean and DTW distance contributes to better clustering results with more data. Furthermore, DTW seems to be more sensitive to noises since its performance is not as stable as Euclidean with small datasets and smaller datasets are known to have smaller signal to noise ratio (SNR). Due to DTW's computational complexity is a magnitude higher than Euclidean, we find **Euclidean to be a more efficient and accurate distance measure than DTW (RQ3)**.

Next, we seek to find out **if our approach works with data from more classrooms and classrooms from different floors (RQ4)**. We have tested our approach with data from each of the five floors within the school building, results show that agglomerative clustering of room temperature indeed helps inferring the relative physical locations of classrooms even with as many as twenty rooms. Figure 7 presents an example clustering with nighttime temperature movements (during a span of five months) from all 20 rooms on another floor. In order to validate the result, we have assigned a color for each cluster in order to visualize the similarities between different rooms. To investigate if this clustering results corresponding to the physical locations of the rooms, we apply the same color scheme to the floor plan, which is shown in Figure 8. It is obvious that classrooms located next to each other generally fall into the same cluster, with the only exception that $R28$ is colored differently compared to its neighbors, indicating an anomaly. However, when comparing Figure 7, $R28$ is actually attached to the cluster of $\{R25, R27, R29, R31\}$ at a very late phase, suggesting that $R28$ is not so similar with the rest in its cluster. Besides, we have found out that $R28$ is used as a classroom for musical education while others are normal classrooms. This indicates that $R28$ may have special features (e.g., heat insulation or acoustic requirements) in design and construction stage.

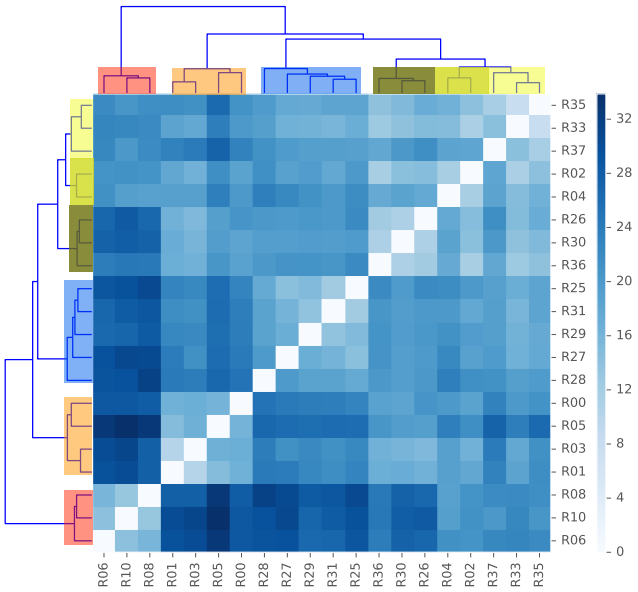

Figure 7: Agglomerative clustering on temperature movements of 20 classrooms.

Furthermore, as shown in Figure 9, our experiments demonstrate that clustering of rooms on higher floors generate more relevant results than rooms in lower ones, indicating that rooms located on higher floors of a building are more impacted by natural factors such as sun, rain and wind which influences buildings' energy performance. Especially, clustering results are terrible with indoor temperature movements in the basement, while better results are achieved above the ground. As a result, our results indeed indicate that **our approach are more sensitive to the floor location of rooms rather than the number of items to cluster (RQ4)**. This result suggest that this approach

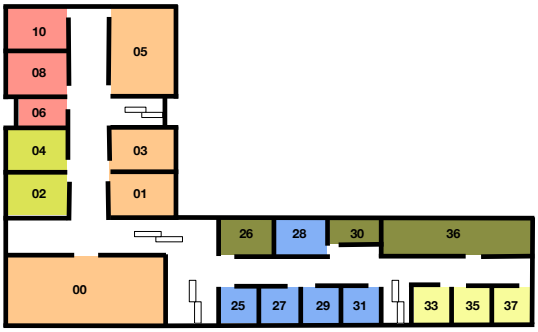

Figure 8: Simplified floor plan with coloring scheme from agglomerative clustering results.

can potentially be performant with rooms in higher tower buildings as their indoor temperature are more impacted by natural factors.

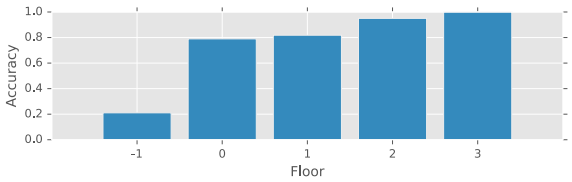

Figure 9: Clustering accuracy (number of correctly clustered rooms divided by total rooms on each floor) among different floors.

5.3 Towards Inferring Occupant Activities

After validating that indoor temperature movements are closely correlated with rooms' physical locations, we set to investigate **the possibility of inferring occupant activities by proxy of temperature movements (RQ5)**. To this end, from the same school building we select temperature movements from 20 rooms that serve different functionality. For instance, some rooms are offices or labs, while others can be normal classrooms or libraries. Note that these rooms are located on different floors of the building and rooms with similar functionality are generally not physically close to each other. In this experiment we have only used temperature readings during daytime for a course of five months, so as to reduce the dilution by readings when no occupant activities are present.

Figure 10 presents the agglomerative clustering results. Much to our surprise, rooms with similar activities or functionality are generally clustered together. For instance, offices seem to have similar temperature movements and science labs do not share much similarity with other type of rooms other than slight similarity with offices. Besides, clustering results suggest that temperature movements in cafeteria, auditorium and reception are quite similar, probably due to the fact that all these rooms see bursts of occupants at specific time slots. Moreover, the rooms hosting kindergartens and preschool classes fall into one cluster, which can be joined by a meeting room, probably indicating that such rooms usually have smaller number of occupants. Last but not least, classrooms for training purposes (art, music and culinary) also have similar temperature movements.

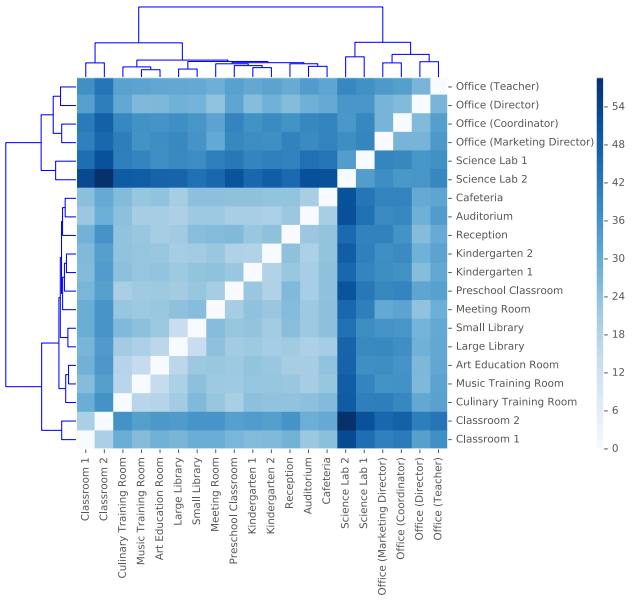

Figure 10: Agglomerative clustering on temperature movements of 20 rooms with different functionality.

While admittedly we are not able to predict what exactly a specific occupant activity is at the moment, by comparing with other activity traces we are still able to tell roughly what such an activity can possibly be. Furthermore, if we setup a database of how different occupant activities can impact indoor temperature movements with higher measuring requirements (e.g., higher sampling frequency, more accurate measurements and larger amount of records), we are confident that finer-grained activity inference can be achieved. As a result, **the answer to RQ5 can be positive when such a activity inference database is established.**

5.4 Discussion

Since most of the temperature sensors in our experiment subject report readings with low frequencies, it has been challenging to infer fine-grained information about indoor environment as well as occupant activities. Despite of dataset limitations, our approach is still able to discover relevant information such as indoor adjacency maps and coarse occupant activities. Furthermore, it is also beneficial to use our system for anomaly detection and building diagnosis. We are able to find groups of rooms whose temperature movements are different from all others. Such anomalies may indicate sensor failures, different HVAC configurations, malfunctioning heat insulation or simply abnormal occupant behaviors. In either case, this kind of information can provide a starting point that helps building owners and operators locating possible issues and fixing them. In addition, since our approach takes advantage of existing building control systems and requires minimal efforts for installation of new hardware, it has the potential to large scale deployment. In turn, when more data are collected, sensing by proxy can become more accurate.

6. CONCLUSION AND FUTURE WORK

It is well established that a thorough understand of indoor environments and occupant activities is a key component in building control systems for better user comfort and more efficient energy usage. Unlike traditional approaches that leverage heterogeneous sensors or crowd-sensing paradigms to monitor indoor environments and activities, we adopt a non-intrusive sensing by proxy paradigm and take advantage of existing infrastructures to be cost-effective. Through extensive experiments with a school building that has 86 rooms equipped with temperature sensors, we are able to apply agglomerative clustering techniques on indoor temperature movements and infer useful information about both the physical features of rooms as well as the functionality of rooms based on traces from occupant activities.

In the future we plan to experiment our approach with more different buildings with respect to geolocations, heights, utility types (office or residential buildings) and number of occupants within rooms. It can also be beneficial to harvest finer-grained indoor temperature movements, i.e., collect temperature data with more accurate sensors and higher frequencies, so that we may infer more detailed information with regard to occupants' exact activities. In addition, other machine learning approaches can be helpful in the context of activity recognition and anomaly detection. Finally, other commonly available sensor and actuator data may also be interesting and adopted to our system.

Acknowledgements

We would like to thank our industrial partner Paul Wurth S.A. for their support in this project.

7. REFERENCES

[1] M. Alzantot and M. Youssef. Crowdinside: automatic construction of indoor floorplans. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 99–108. ACM, 2012.

[2] A. Antonini, F. Maggi, and S. Zanero. A practical attack against a knx-based building automation system. In *Proceedings of the 2nd International Symposium on ICS & SCADA Cyber Security Research 2014*, pages 53–60. BCS, 2014.

[3] H. Chen, P. Chou, S. Duri, H. Lei, and J. Reason. The design and implementation of a smart building control system. In *e-Business Engineering, 2009. ICEBE'09. IEEE International Conference on*, pages 255–262. IEEE, 2009.

[4] Y. Jiang, Y. Xiang, X. Pan, K. Li, Q. Lv, R. P. Dick, L. Shang, and M. Hannigan. Hallway based automatic indoor floorplan construction using room fingerprints. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 315–324. ACM, 2013.

[5] M. Jin, N. Bekiaris-Liberis, K. Weekly, C. Spanos, and A. M. Bayen. Sensing by proxy: Occupancy detection based on indoor co2 concentration. In *The 9th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM'15)*, pages 1–10, 2015.

[6] D. Li, T. Bissyandé, S. Kubler, J. Klein, and Y. Le Traon. Profiling household appliance electricity usage with n-gram language modeling. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 604–609. IEEE, 2016.

[7] D. Li, T. F. Bissyandé, J. Klein, and Y. Le Traon. Dsco-ng: A practical language modeling approach for time series classification. In *International Symposium on Intelligent Data Analysis*, pages 1–13. Springer International Publishing, 2016.

[8] D. Li, T. F. Bissyande, J. Klein, and Y. Le Traon. Time series classification with discrete wavelet transformed data: Insights from an empirical study. In *The 28th International Conference on Software Engineering and Knowledge Engineering (SEKE 2016)*, 2016.

[9] D. Li, L. Li, T. F. Bissyande, J. Klein, and Y. Le Traon. Dsco: A language modeling approach for time series classification. In *12th International Conference on Machine Learning and Data Mining (MLDM 2016)*, 2016.

[10] T. W. Liao. Clustering of time series data – a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[11] T. A. Nguyen and M. Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and buildings*, 56:244–257, 2013.

[12] B. Okorn, X. Xiong, B. Akinci, and D. Huber. Toward automated modeling of floor plans. In *Proceedings of the symposium on 3D data processing, visualization and transmission*, volume 2, 2010.

[13] P. Royston et al. Multiple imputation of missing values. *Stata journal*, 4(3):227–41, 2004.

[14] O. Shih and A. Rowe. Occupancy estimation using ultrasonic chirps. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*, pages 149–158. ACM, 2015.

[15] H. Shin, Y. Chon, and H. Cha. Unsupervised construction of an indoor floor plan using a smartphone. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):889–898, 2012.

[16] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[17] L. Yang, K. Ting, and M. B. Srivastava. Inferring occupancy from opportunistically available sensor data. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 60–68. IEEE, 2014.

[18] Z.-N. Zhen, Q.-S. Jia, C. Song, and X. Guan. An indoor localization algorithm for lighting control using rfid. In *Energy 2030 Conference, 2008. ENERGY 2008. IEEE*, pages 1–6. IEEE, 2008.

[19] T. Zhu, S. Xiao, Q. Zhang, Y. Gu, P. Yi, and Y. Li. Emergent technologies in big data sensing: a survey. *International Journal of Distributed Sensor Networks*, 2015:8, 2015.