DISTANCE AND SIMILARITY MEASURES FOR SENSORS SELECTION IN HEAVILY INSTRUMENTED BUILDINGS: APPLICATION TO THE INCAS PLATFORM

Cédric Gouy-Pailler, PhD / Research Engineer, cedric.gouy-pailler@cea.fr
Hala Najmeddine, PhD / Research Engineer, hala.najmeddine@cea.fr
Anthony Mouraud, PhD / Research Engineer, anthony.mouraud@cea.fr
Frédéric Suard, PhD / Research Engineer, frederic.suard@cea.fr
CEA, LIST, 91191 Gif-sur-Yvette CEDEX, France
Clara Spitz / PhD student, clara.spitz@cea.fr
Arnaud Jay / Research Engineer, arnaud.jay @cea.fr
Philippe Maréchal / Research Engineer, philippe .marechal@cea.fr
CEA, LITEN, INES, Savoie-Technolac, 73377 Le Bourget-du-Lac CEDEX, FRANCE

ABSTRACT

Energy management in residential buildings is taking an increasing role in the construction workflows. It entails understanding thermal processes at stake in the buildings and quantifying energy consumption, which meets inhabitants comfort requirements. Experimental platforms such as INCAS aim at providing experts with a practical way to study such problems in real conditions. These heavily equipped buildings yield huge amounts of real-time data (sampling rates, number and types of sensors) for which new automatic approaches could be useful to thermal experts. Generic similarity measures from data-mining could therefore provide comprehensive analysis tools to thermal experts.

This paper focuses on the ability of some distance and similarity measures to organize millions of data from homogeneous and heterogeneous sensors into coherent clusters. Simplifying data interpretations to thermal experts in highly equipped buildings, this approach could also stand as a basis for studying smart grids of less equipped domestic houses studies.

Two types of similarity measures are explored. The first one consists of a set of three distances, and accounts for differences in terms of amplitude scaling and shifting between pairs of measurements. It relies on the comparison of homogeneous sensors by quantifying the relative proximity of their amplitude in terms of mean value, variance and time shift. The second type of similarity measure employs a pre-processing step transforming continuous signals into binary events. The resulting spike trains are then compared by quantifying the amount of unitary transformations (events moves or events deletions/additions) needed to align pairs of events sequences.

These proximity measures are computed on real data from experimental buildings of the INCAS platform. It comprises three experimental buildings (with different construction types) dedicated to testing various approaches regarding systems, control and energy-saving policies. These geometrically identical buildings are equipped with hundreds of sensors measuring temperature, humidity, differential pressure, and others data at various positions of the structures with sampling rates of one measurement per minute. Simulation-based temperatures are integrated in the sensors set providing a comparison between real and simulated data.

Results illustrate the contribution of the applied methods when dealing with large amounts of measurements related to instrumented buildings behaviors. Actually results show that coherent clusters regarding distinct signal properties are automatically generated. These clusters can be used for dimensionality reduction (clusters of sensors could be summarized by a single virtual measurement), or relative comparisons between sensors or between real and simulated datasets.

Keywords: data mining, energy saving policies, sensor selection, INCAS, experimental buildings.

1. INTRODUCTION

Energy consumption devoted to residential buildings accounts for up to forty three percents of the global energy produced (ADEME 2009). Current trends therefore aim at drastically reducing consumption by introducing new restrictive standards. For example the resulting “low-energy buildings” do not have any conventional HVAC systems. The “low-energy” approach also implies using innovative construction techniques and materials. For example thermal bridges and infiltration have to be reduced to minimum, and thermal exchanges between internal and external areas should be tightly controlled. It thus entails deep changes in the construction workflow. First the design of residential buildings have to be reconsidered to take into account the various constraints involved by these approaches. Then energy consumption evaluation standards might be evaluated or modified to calculate energy consumption associated to new materials and construction techniques. Finally real-time monitoring in real conditions should be encompassed in the design process. Avoiding energy wastes indeed involves a tight control of the internal environment (temperature, humidity) to fulfill two contradictory criteria: energy savings and inhabitants comfort. Ideally monitoring should be the key point of an active system able to regulate the living environment. Its training would only involve little human intervention and should be done by the system in a transparent way to adapt users’ wishes.

These considerations were the root of the INCAS platform construction, built at the French National Institute of Solar Energy (INES). This platform comprises buildings equipped with plenty of sensors able to measure temperature, humidity, differential pressure, air velocity and other data at various points of the structures. With a sampling rate of one measurement per minute, the sensors provide millions of data per month. Located inside and outside buildings, in different materials, and along various orientations, sensors provide a detailed overview of the thermal processes at stake in both each of these human-free houses. While usual approaches employed by thermal experts are possible with a few sensors, the amount of data provided by the INCAS platform proves difficult to explore. Specifically, sensors can be chosen by experts to study various physical thermal mechanisms, but studying global phenomena and interdependence would surely benefit from large scale data mining techniques.

This article presents similarity-based techniques designed to ease interpretation by highlighting similarities and dissimilarities between sensors. The general approach underlying these techniques consists in computing a measure of dissimilarity between pairs of sensors. The resulting distance measurement yields a real number, which is zero if the two sensors measure the same quantity during the considered period of time. It increases when the difference between the two time series measurements amplifies. Distances are usually defined by the expert after the pursued goal. For example a distance can be designed for rendering differences between the absolute values measured by the sensors, or the dissimilarities related to the amount of variations, namely the variance. A distance matrix is obtained by evaluating dissimilarities between all pairs of sensors. It is then used for clustering purposes, which consists in gathering sensors into groups of coherent clusters. The resulting clusters can be interpreted by considering that two sensors belonging to a same cluster measure similar data (relatively to the chosen distance). This article describes a few distances able to gather sensors into sets of coherent clusters.

The remainder of this article is organized as follows. The real dataset as well as the simulated data will be described in section 2. Distances able to characterize various properties of the measurement time series will then be described. In a following section results will be illustrated using data recorded in February 2011. Methods will be applied to both real data measurements as well as simulated temperatures. Finally results will be discussed in section 5. Specifically, comparisons between real data and simulated temperatures will be discussed.

2. EXPERIMENTAL DATA

The dataset used in this article corresponds to measurements acquired in February 2011. During the considered periods, the human activities inside the houses were negligible and internal loads were well known. A local meteorological station reported the external conditions (temperatures, sun direct horizontal exposure).

Real measurements

While hundreds of sensors are installed in houses, the analysis focuses on the first floor of the house. The external envelope of the building is composed of two separate layers (agglomerate concrete of 15 cm each) separated by a 20-centimeter insulator. In order to ease interpretation of the work, only temperature sensors are considered. The dataset therefore corresponds to twenty-five sensors ($N_s = 25$) recording temperatures between the 16th and the 25th of February. The sample rate is one measurement per minute for each sensor. They are located in three distinct materials (air, windows and walls). All of them are protected from the direct sun illumination. Figure 1 shows positions of the sensors in the numerical 3D model of the house. Table 1 provides a summary of the positions of the sensors. It mentions sensors names, implantation materials and situation in the house. In the following of the article a three dimensional view of the house will be used to illustrate positions of sensors. It should be reminded that the big balcony is south-oriented.

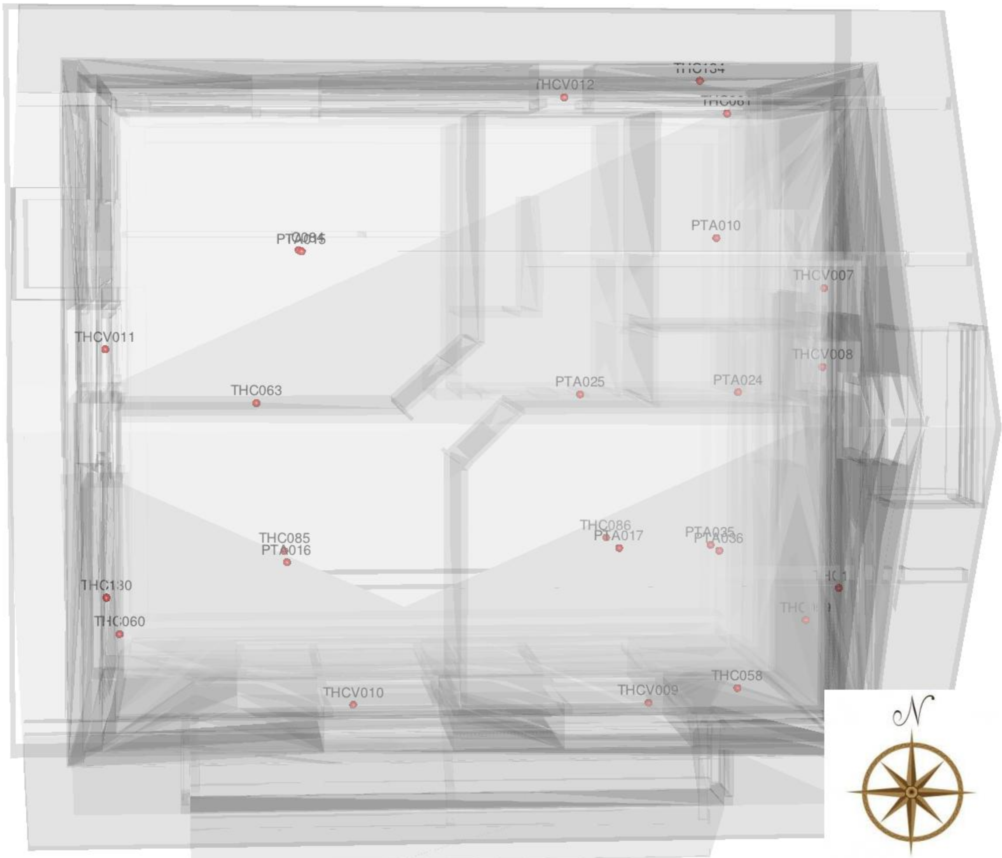

Figure 1: Experimental house from the INCAS platform. Upside view. Bottom of the image is south-oriented. Sensors are all located at the first floor of the building.

Table 1: Description of sensors. Orientations are as follow: Room 1 (West-North), Room 2 (South-West), Room 3 (South-East), Bathroom (North-East), Toilets (East), Landing (North).

Sensors names	Material	Room	Sensors names	Material	Room
THC058	Wall	Room 3	PTA010	Air	Bathroom
THC059	Wall	Room 3	PTA015	Air	Room 1
THC060	Wall	Room 2	PTA016	Air	Room 2
THC061	Wall	Bathroom	PTA017	Air	Room 3
THC063	Wall	Room 1	PTA024	Air	Toilets
THC084	Wall	Room 1	PTA025	Air	Landing
THC085	Wall	Room 2	PTA035	Air	Room 3
THC086	Wall	Room 3	PTA036	Air	Room 3
THC112	Wall	Outside	THCV007	Window	Bathroom
THC130	Wall	Room 2	THCV008	Window	Toilets
THC134	Wall	Bathroom	THCV009	Window	Room 3
			THCV010	Window	Room 2
			THCV011	Window	Room 1
			THCV012	Window	Landing

Simulated temperatures

Simulation is done with EnergyPlus (e.g. Crawley et al, 2001). It is a multi-zonal dynamic simulation tool, used by a very broad panel of engineers, architects and researchers.

The house was simulated with 4 thermal zones: 1) two heated, ground floor and the first floor; 2) two free of heating systems, basement and roof space. In that first step, the model is voluntarily quite simple: inter-zone aeraulic exchange are not modeled, the heat recovery HVAC system is modeled with a limited energy-equivalent ventilation flow rate, heating sources are replaced by two heating ideal systems (one in each thermal zone), thermal bridges were not taken into account, only solar shading due to structures at the south of the house are considered (roof overhang and balcony) and heat convection transfer coefficients are constant and depend of type of wall (vertical, horizontal, internal, external).

In the house the internal loads are emitted by the data acquisition system. The internal gains are spread uniformly through both zones. 160W of heat per thermal zone is assumed, exchanged by convection. One key point of a simulation which will be compared to measures is the creation of the weather file in a EnergyPlus format. This was done thanks to the outside measured conditions. The time step of this file is one sample per minute. Among the various simulated variables provided by EnergyPlus, we only consider the simulated temperature corresponding to the first floor of the house. The comparisons between real and simulated measurements is thus not obvious (one simulated time series against twenty-five measurements). Evaluation of the fitness between simulated and real observations involves choosing among the various potential sensors. Hence the issue of choosing comparison criteria able to characterize the simulated measurements regarding some predefined signal properties (amplitude, variance or time shifts).

3. METHODS

We begin by stating the notations used throughout this paper. Let $\mathbf{x}(\mathrm{t})$ denote the vector of data acquired for all sensors at time t. $\mathbf{x}(\mathrm{t})$ is extended as the concatenation of the real measurements with the simulated temperature corresponding to the first floor thermal zone $\mathbf{x}(\mathrm{t}) = [x_{\mathrm{PTA010}}(t)\ \ x_{\mathrm{PTA015}}(t)\ \ ...\ \ x_{\mathrm{THCV012}}(t)\ \ ...\ \ x_{\mathrm{simu}}(t)]^T$, where T denotes the transpose operator. This work amounts to partitioning the sensors into coherent clusters according to various signal properties. Due to the complex nature of the time series, many techniques have been proposed in the literature for the classification of time series into a reduced dimension space such as Principal Component Analysis PCA (Korn et al. 1997). However, all of them suggest that simple nearest neighbor classification is difficult to beat. This argument opens the question on which distance measure to use. In the following condition, one tested several methods to measure the distance between time series such as the Euclidian Distance ED (Wang and Wang, 2000), the Complexity Invariant Distance CID (Batista et al., 2011), Dynamic Time Warping DTW (Yi et al., 1998) and a spike train based distance (Victor, 2005 and Victor and Purpura, 1997).

Similarity measures for continuous data

Euclidean distance. Euclidean Distance is the most popular distance for data classification (Wang and Wang, 2000). Supposing we have two time series x_i and x_j, the Euclidean distance is computed by

$$ED\left(x_i, x_j\right) = \sqrt{\sum_{t=1}^{N}\left(x_i(t) - x_j(t)\right)^2} \quad (1)$$

where N stands for the number of time points used to compute the distance (in our case it corresponds to approximately one week of data). Each time point from x_i and x_j contributes an equal amount to the total error. Euclidean distance in figure 2 (left) between these two signals is large nevertheless both signals are very similar. The Euclidian distance has several advantages; the complexity of evaluating this measure is linear and it is a very competitive method according to large datasets. Yet the Euclidean Distance can lead to absurd results when the purpose is to classify time series of raw data. In fact, the major drawback of Euclidean Distance is its high sensitivity to error, outliers, missing data and time distortion. To overcome these shortcomings, many researchers have advised that time series should be transformed before computing distance measure (Keogh and Kasetty, 2002). Such transformations include data normalization and smoothing, complexity invariant integration, shifts removing of non-matching parts and distortion correction in time axis (Batista et al., 2011).

Complexity Invariance Distance. The CID introduces information about Complexity Invariance of a couple of given time series (Batista et al, 2011). Applied on temperature data measurements, it can help users to identify sensors with big variation and outlier's data.

$$CID\left(x_i, x_j\right) = \sqrt{\sum_{t=1}^{N}\left(x_i(\mathrm{t}) - x_j(\mathrm{t})\right)^2} \times \frac{max\left(CE(x_i), CE(x_j)\right)}{min\left(CE(x_i), CE(x_j)\right)} \quad (2)$$

Where $CE(x) = \sqrt{\sum_{t=1}^{N-1}(\mathrm{x}(\mathrm{t}+1) - \mathrm{x}(\mathrm{t}))^2}$ represents the complexity estimation of a time sequence stretched represented on figure 2 (middle). Appling CE requires that the two times sequence should have the same length. Unfortunately, the other problems remain unresolved. Using the DTW can help to avoid distortion problem.

Dynamic Time Warping. Dynamic Time Warping DTW (Sankoff and Kruskal, 1983) is an algorithm that aligns and matches two time series or sub-time series with different dimension and solve the distortion time axis, problem usually identified with temperatures time series. This method allows non-linear alignments (Keogh and Pazzani, 2001), (Ratanamahatana and Keogh, 2004) between two time series locally out of phase as shown in figure 2 (right). To align two sequence ($Q_{(1\times n)}$ and $C_{(1\times m)}$) using DTW, we construct an n-by-m matrix corresponding to the squared distance $\mathrm{d}(\mathrm{i},\mathrm{j}) = \left(\mathrm{q_i} - \mathrm{c_j}\right)^2$. To find the best match between these two sequences, we can find a path through the matrix that minimizes the total cumulative distance $\gamma(i,j)$ between them.

$$\gamma(i,j) = d(i,j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (3)$$

We refer the interested reader to Kruskall and Liberman (1983) for more details. One drawback of these methods is their high computational cost.

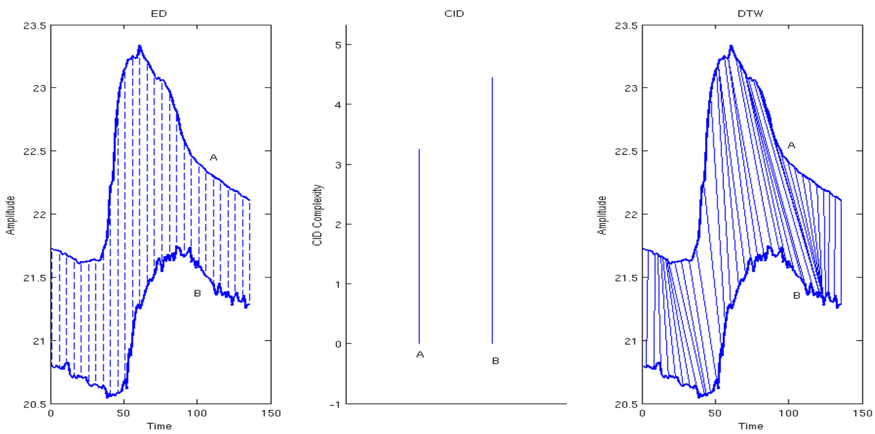

Figure 2: Left: Euclidian Distance. The distance is computed by comparing pairs of amplitude at the same instant between two signals. The comparisons are illustrated in the figure by dashed lines linking compared points. Middle: Complexity Invariance Distance. The Euclidean distance is normalized using the ratio of complexities between the two series. Right: Dynamic Time Warping Distance. On the figure, straight lines linking the two series illustrate the associations realized by the algorithm. Such links amount to time shifts of points such as the best alignment between two series is achieved. The distance between two time series accounts for the total amount of time shift transformations necessary to perform the alignment.

Similarity measures for successive events comparison

From signals to events. Instead of computing similarities between continuous signals acquired from sensors, we also propose a metric based on a transformed version of the signals. In a first step, a basic transformation is applied to the signal in order to extract crucial events available in the signal. While various such transformations can be chosen according to the considered purpose, we here demonstrate the method using a peak-detection algorithm. This pre-processing step aims at extracting local maxima. The rationale behind this approach is that temporal events yield a rich source of information to characterize signals signatures. It should be noticed that the retained information does not anymore accounts for amplitude of initial measurements. It rather focuses on precise events timings. The major advantage of this approach is that resulting similarity measures will only focus on the temporal synchronization between successive events. Our pre-processing will thus illustrate how phase differences can be efficiently rendered. Figure 3 (left) describes the pre-processing step that transforms continuous signals into a train of events.

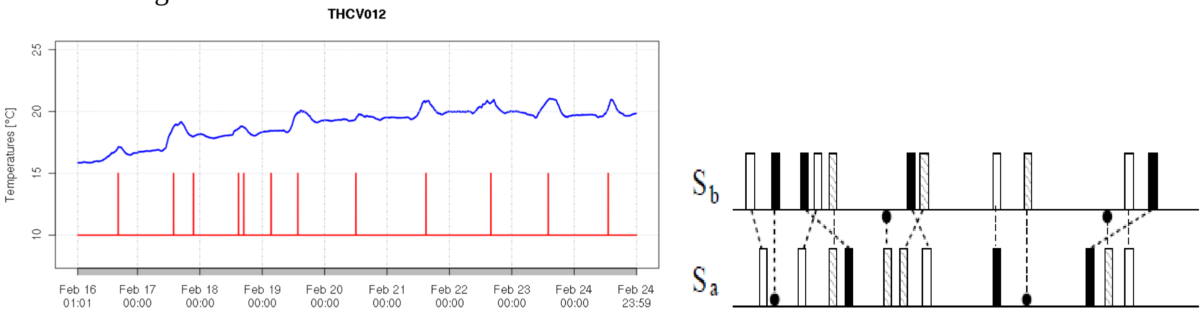

Figure 3: (left) The continuous line represents the temperature measurements from sensor THCV012. The spiky signal accounts for the transformed version of the temperatures as successive events. (Right) Illustration of the spike trains alignment transformations.

Event-based similarity measure. At this point, we thus have successive events corresponding to each sensor. Measuring resemblance between these spikes trains can be done by using unitary operations to transform a pike train into another one. This has been proposed by Victor and Purpura, 1997. In this article, it has been proposed that spikes trains alignment can be realized using

- A spike temporal shift. The cost of such a transformation is set to be proportional to the temporal shift used to align a spike of one train onto the spike of a second train.
- A spike deletion/addition. This transformation will be applied a unitary cost of 1 each time a spike is deleted from a train to match another train.

While an infinite number of transformations can be used to align two spike trains, it can be shown that dynamic programming techniques are able to find the minimum cost for aligning two spike trains (with q fixed). This minimal cost is defined as the event-based distance between the two spike trains.

It should be noted that the maximum distance between two spike trains is achieved by transformations consisting in removing all spikes of the first train and adding spikes to the first train at temporal positions of the second train. In this case the global cost is the sum of the number of spikes in the first and second train. As a consequence there always exists a transformation, which achieves a cost lower or equal to this sum. We use this observation to normalize the distances by dividing the computed cost by the total amount of spikes presents in both trains. The normalized similarity measure is therefore comprised between 0 and 1. The interpretation of the measure is therefore relatively straightforward. Identical spike trains yields a distance of 0, whereas completely dissimilar spike trains would amount to a distance close to 1. Figure 3 (right) illustrates the use of temporal shifts and deletions/additions of spikes to align trains of events.

While some applications require fine-tuning of the q parameter, its interpretation simplifies the choice in our case. It should be observed that a shift transformation is advantageous over a deletion/addition if and only if $q|\Delta t| > 2$. The cost of a deletion of a spike in the first spike train, succeeded by an addition of a spike in this same train at the correct time instant cost 2. The analogous transformation using temporal shift costs $q|\Delta t|$. In our application we will set q such that temporal shift are not allowed between spikes belonging to distinct days. This constraint has a straightforward justification, it indeed means that local maxima can only be associated in an intra-days basis.

Classification based on similarity measures

Distances are computed for all pairs of sensors, thus yielding $\frac{N_s.(N_s-1)}{2}$ distances (because of symmetries). Various methods can then be used to interpret the set of distances in terms of clusters (groups of homogeneous sensors). In this work we will focus on a simple clustering approach, in which distances matrices will be used to gather sensors into coherent clusters. This clustering approach will be based on a hierarchical clustering approach with a fixed number of clusters. As the clustering method is not the goal of this paper, the number of clusters will be arbitrarily set to five. This has been chosen in order to obtain relatively small and interpretable clusters (corresponding to an average of around five sensors per classes). Automated procedures are able to automatically determine an optimal number of clusters according to some pre-defined criterion. We will use 3D model of the house to represent results. The virtual sensor consisting of the simulated temperatures will be placed in the middle of the house.

4. APPLICATION

Our analysis focuses on one week of data recorded last February, from the 16th to the 25th. Four distances have been presented. Among these, three are based on the continuous amplitudes of temperatures (ED, CID and DTW), and one is measuring similarities between spike trains representing local maxima of the temperatures (EV). Matlab has been used to compute distances ED, CID and DTW of the 325 pairs of signals, while R was used to perform the pre-processing steps (peak detection) and compute the event-based distance between pre-processed signals.

The distance matrices are then provided to the clustering algorithm, which yields five clusters of similar (according to the considered distance) sensors. Results of the clustering approach for ED and EV are depicted in figure 4 and 5. On these figures two point of view are presented. On the one hand time series grouped are represented in five distinct plots. Each plot corresponds to one cluster and represents temperature time series. On the other hand, a 3D schematic of the house presents the spatial distribution of the sensors. Figure 4 shows results obtained using the Euclidean Distance.

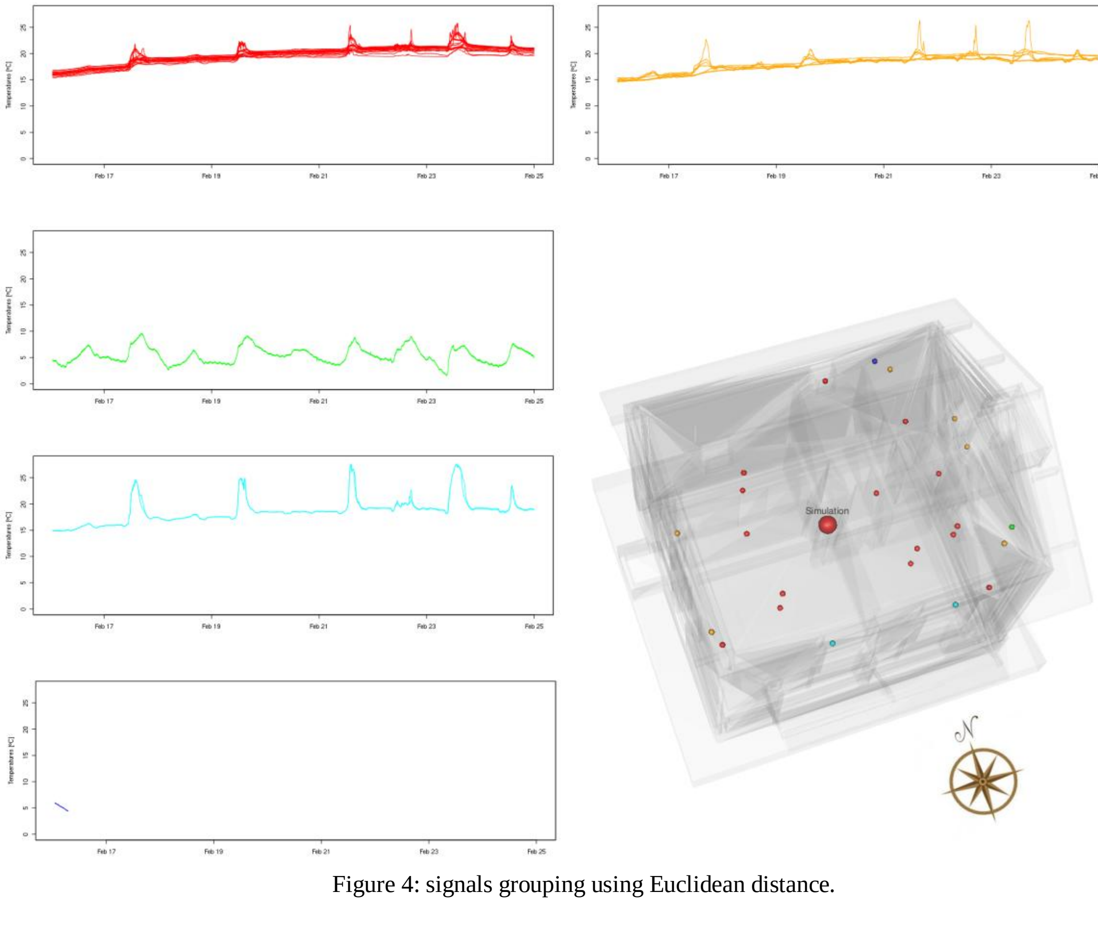

Figure 4: signals grouping using Euclidean distance.

Such results demonstrate that ED focuses on temperature amplitudes in order to gather sensors into coherent clusters. The predominant group (top left of figure 4) contains sensors placed inside the house, in the centers of the rooms. Other groups of sensors are mainly located on the windows or outside the house. THC112 (figure 4, left, second plot) is logically isolated from the other sensors as it registers temperatures drastically lower than inside the house. Interestingly the algorithm successfully gathers window sensors from the east/west orientation and window sensors from the south orientation together. It should also be noticed on figure 4 that window sensors at the north of the house are gathered with inside air sensors. This method thus seems to behave quite well to isolate heat phenomena such as direct sun light exposure.

Results corresponding to the EV-based algorithm are depicted on figure 5. Unlike previous results with ED, results demonstrate that this method focuses on phase synchronization between sensors rather than amplitude of measurements. As an example the predominant group (top left, figure 5) contains sensors located inside the house but also THC112 which is located outside the building. This can be explained because inside temperatures mainly synchronize with outside ones (with some small temporal shift). As we applied the algorithm on a complete week, we are not able to distinguish between all sensors orientations but the same algorithm could be applied at finer time scale to highlight differences of phases synchronizations in different parts of the house.

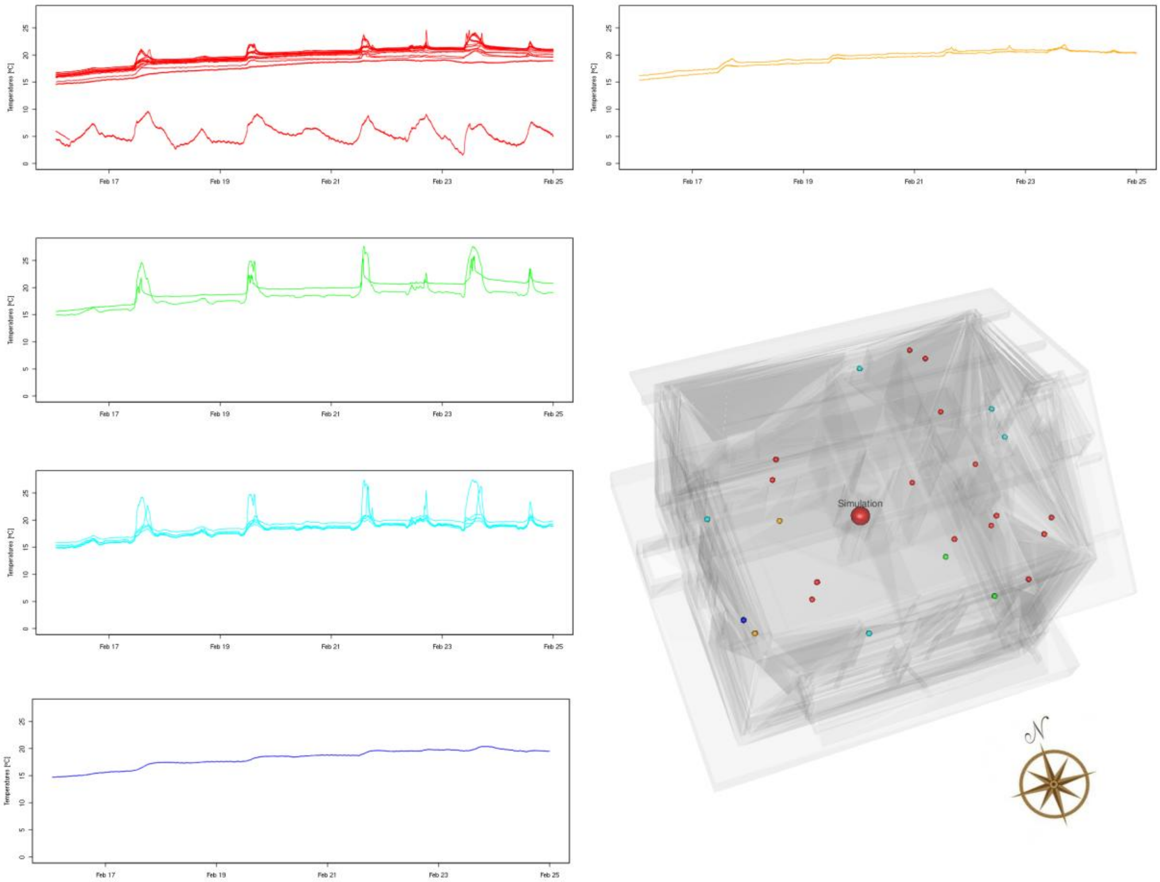

Figure 5: signals grouping using event-based distance.

Figure 6 shows the clustering results obtained using the CID and DTW algorithms. Clustering using the CID and DTW methods (figure 6) yields comparable results regarding internal sensors, grouped in a predominant cluster. Yet, we can observe that more discrimination is made between sensors inside the house with the CID method. This may result from relative exposition of these sensors to the sun illumination. Finally, an overview of the different results is depicted in table 2. This table highlights the fact that all air sensors are gathered in analogous clusters for all methods. Distinct properties of the time series independently appear in other clusters based on the distance chosen.

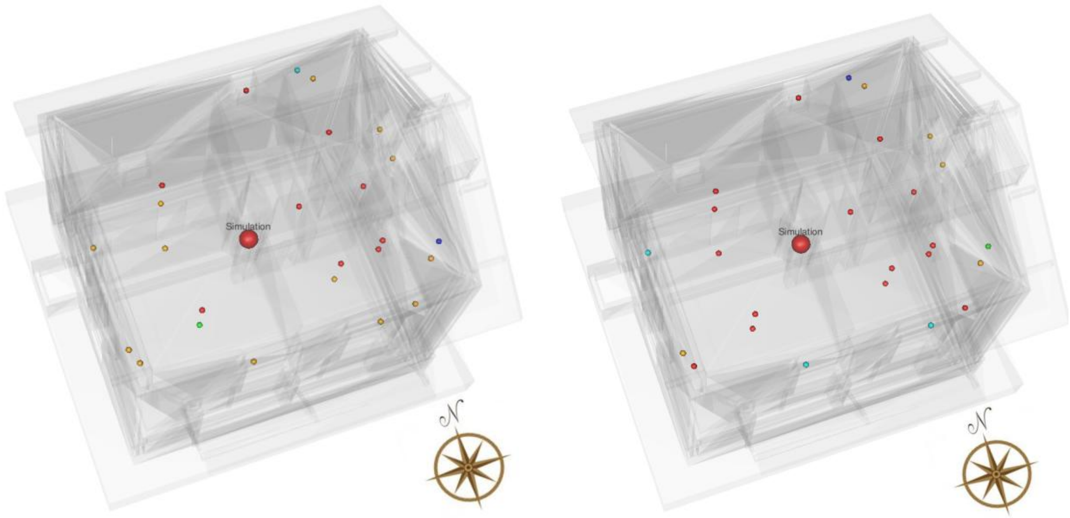

Figure 6: Sensors clustering based on two proposed distances. Left, CID. Right, DTW.

Table 2: Clustering results for the four presented distances. Euclidean Distance (ED), Event-based Distance (EV), Complexity Invariant Distance (CID) and Dynamic Time Warping (DTW). Cluster index are depicted for each sensor and each method.

Sensors names	Material	Room	ED	EV	CID	DTW
THC058	Wall	Room 3	1	1	2	1
THC059	Wall	Room 3	2	1	2	2
THC060	Wall	Room 2	1	2	2	1
THC061	Wall	Bathroom	2	1	2	2
THC063	Wall	Room 1	1	1	2	1
THC084	Wall	Room 1	1	1	2	1
THC085	Wall	Room 2	1	1	3	1
THC086	Wall	Room 3	1	3	2	1
THC112	Wall	Outside	3	1	5	3
THC130	Wall	Room 2	2	5	2	2
THC134	Wall	Bathroom	5	1	4	5
PTA010	Air	Bathroom	1	1	1	1
PTA015	Air	Room 1	1	1	1	1
PTA016	Air	Room 2	1	1	1	1
PTA017	Air	Room 3	1	1	1	1
PTA024	Air	Toilets	1	1	1	1
PTA025	Air	Landing	1	1	1	1
PTA035	Air	Room 3	1	1	1	1
PTA036	Air	Room 3	1	1	1	1
THCV007	Window	Bathroom	2	4	2	2
THCV008	Window	Toilets	2	4	2	2
THCV009	Window	Room 3	4	3	2	4
THCV010	Window	Room 2	4	4	2	4
THCV011	Window	Room 1	2	4	2	4
THCV012	Window	Landing	1	4	1	1

5. DISCUSSION AND CONCLUSION

Four methods have been presented to compute distances between signals recorded by temperature sensors. While such methods are still at their beginning for studying thermal processes in heavily instrumented buildings, many points appear encouraging. First sensors can be efficiently selected according to a wide variety of similarity measures. These distances can be chosen according to the goal pursued by the thermal expert. For example simple interpretations can be done about the presented distances. The Euclidean distance focuses on the relative average amplitudes between pairs of sensors. Previous results have shown that this distance can be useful to isolate sensors according to thermal position of the location (inside versus outside). Furthermore Euclidean Distance can be used to detect abnormal temperatures in a specific room of the house compared to the average one. While Euclidean Distance does not take into account the variance of the signals, the Complexity Invariance

Distance is able to discriminate sensors by taking into account their relative variance. It can thus be useful to detect high temperature variations inside rooms of the house. Dynamic Time Warping appears as a good candidate for identifying similar thermal responses, which are not synchronized. For example it can be useful to gather sensors located at similar positions inside walls (thus having comparable thermal responses). Lastly the event-based distance is able to focus on the synchronization of events. Its potential could thus be clearly profitable when short delay between sensors responses should be studied. Another advantage (not demonstrated in this article) is the possibility to compare inhomogeneous responses such as temperature and humidity. Indeed the pre-processing step consisting in transforming signals into events is applicable to any signal. Such a transformation always result in comparable spikes trains that can be evaluated through the spike-based distance.

Not surprisingly, the simulation signal is grouped with sensors located inside the building. As the simulation software is only designed to render average temperatures used to evaluate consumption, our result only confirm the coherent grouping strategies given by all methods. Clustering methods can further serve as tools to select sensors that can be compared with the simulation signal.

At its very beginning, the use of distance-based techniques to interpret thermal processes involves people from various fields. The continuous discussion between "data miners" able to highlight data dependences and thermal experts able to interpret results may improve our understanding of the thermal behavior of low-energy consumption houses, a compulsory step towards the drastic reduction of energy wastes.

REFERENCES

ADEME (2009) "Les chiffres clés du bâtiment, Energie, Environnement." *ADEME Edition 2009.*

Batista G. et al. (2011) "A Complexity-Invariant Distance Measure for Time Series." *SMD 2011.*

Crawley D.B., Lawrie L.K., Winkelmann F.C., Buhl W.F., Huang Y.J., Pedersen C.O., Strand R.K., Liesen R.J., Fisher D.E., Witte M.J., Glazer J. (2001) " EnergyPlus: creating a new-generation building energy simulation program " *Energy and Building,* vol. 33, p. 319-331

Huffmire T. and Sherwood T. (2006) "Wavelet-Based Phase Classification." *PACT'06.*

Keogh E. and Pazzani M. (1998) "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback." *In proceedings of the 4th Int'l Conference on Knowledge Discovery and Data Mining.*

Keogh E. and Kasetty S. (2002) "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration." *In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Korn F. et al. (1997) "Efficiently supporting ad hoc queries in large datasets of time sequences." *In proceedings of the ACM SIGMOD Int'l Conference on Management of Data.*

Kruskall J. B. and Liberman M. (1983) "The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules." Addison-Wesley.

Ratanamahatana C. and Keogh E. (2004) "Making Time-series Classification More Accurate Using Learned Constraints." *In proceedings of SIAM International Conference on Data Mining (SDM '04).*

Sankoff D. and Kruskal J. B. (1983) "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison." *Addison-Wesley, Reading*.

Victor J.D., Purpura K.P. (1997) "Metric-space analysis of spike trains: theory, algorithms and application." *Networks* 8, 127–164.

Victor J.D. (2005) "Spike train metrics" *Current Opinion in Neurobiology* , vol. 15 issue 5, p. 585 –592.

Wang C. and Wang X. (2000) "Supporting content-based searches on time series via Approximation." *In proceedings of the 12th Int'l Conference on Scientific and Statistical Database Management*.

Yi B. et al. (1998) "Efficient retrieval of similar time sequences under time warping." *IEEE International Conference on Data Engineering.*