Towards a Methodology for the Characterization of IoT Data Sets of the Smart Building Sector

Louis Closson, Christophe Cérin, Didier Donsez, Denis Trystram

Towards a Methodology for the Characterization of IoT Data Sets of the Smart Building Sector

Louis Closson*, Christophe Cérin†, Didier Donsez‡, and Denis Trystram‡
*Adeunis - University of Grenoble Alpes - l.closson@adeunis.com
† University of Paris 13 - LIPN, UMR CNRS 7030 - christophe.cerin@univ-paris13.fr
‡ University of Grenoble Alpes - LIG, Bâtiment IMAG - {didier.donsez,denis.trystram}@imag.fr

***Abstract*—The long-term objective of the paper aims to provide decision aid support to a technical smart buildings manager to potentially reduce the emission of data produced by sensors inside a building and, more generally, to acquire knowledge on the data produced in the facility. As the first step, the paper proposes to characterize the smart-building ecosystem's Internet-of-things (IoT) data sets. The description and the construction of learning models over data sets are crucial in engineering studies to advance critical analysis and serve diverse researchers' communities, such as architects or data scientists. We examine two data sets deployed in one location in the Grenoble area in France. We assume that the building is an autonomic computing system. Thus, the underlying model we deal with is the well-known MAPE-K methodology introduced by IBM. The paper mainly addresses the analysis component and the adjacent connector component of the MAPE-K model. The content of this layer, and its organization, constitutes the methodological point we put forward. Consequently, we automatically provide a complete set of practices and methods to pass to the planning component of the MAPE-K model. We also sketch a semi-automatic way of reducing the number of measures done by sensors. In the background of our study, we aim to reduce the operational cost of making measures with a much more sober approach than the current one. We also discuss in profound the main findings of our work. Finally, we provide insights and open questions for future outcomes based on our experience.**

***Index Terms*—Smart Building data sets analyses, IoT data sources characterization, Enabling technologies for the IoT, Building Information Modelling.**

I. Introduction

The definition of "Smart Building" has evolved [1], mainly to clarify what is meant by "smart." Historically, smart referred to a building that had deployed sensors to react in real-time to an event, for instance, lighting a hallway or controlling the heating or the air conditioning systems when a presence is detected. Then nowadays, the definition of smart building switches to a building that *addresses both intelligence and sustainability issues by utilizing computer and intelligent technologies to achieve the optimal combinations of overall comfort level and energy consumption*.

Authors in [2] introduce the vision of autonomic computing—computing systems that can manage themselves given high-level objectives from administrators. As noted in [3], the term autonomic computing is *symbolic of a vast and somewhat tangled hierarchy of natural self-governing systems. Many of which consist of myriad interacting, self-governing components that comprise large numbers of interacting, autonomous, self-governing components at the next level down*. We propose, in this paper, to consider a smart building (its digital manager) as an autonomic system because the concept corresponds to what the building manager wants. For instance, he does not plan to allocate people for manual or semi-automatic tasks in the building to regulate the temperature of the rooms.

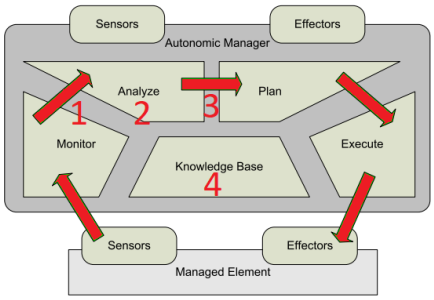


Fig. 1. The MAPE-K model

From an architectural point of view, see Figure 1, an autonomic element will typically consist of one or more managed elements coupled with a single autonomic manager that controls and represents them. By monitoring the managed element and its external environment and constructing and executing plans based on an analysis of this information, the autonomic manager will relieve humans of the responsibility of directly managing the managed element. This view is the so-called IBM MAPE-K model. It corresponds to the idea that we build knowledge over the managed element through monitoring, analyzing, planning, and executing sequential steps.

The scenario we investigate in this article is as follows. A particular actor, private or public, has the technical responsibility of managing a smart building. Moreover, he has to respect a corporate social responsibility policy, especially concerning the issues of digital sobriety. This actor uses a BIM-like tool for modeling. If the building is already operating, it is eventually controlled by a smart building manager.

A BIM (Building Information Modeling) designates the

This work has been co-funded by a CIFRE grant (reference 2021/1336) and partially supported by the Multi-disciplinary Institute on Artificial Intelligence MIAI at Grenoble Alpes (ANR-19-P3IA-0003). This work is conducted during the Délégation with Centre National de la Recherche Scientifique (CNRS) of Mr Cérin. Thanks to the institutional supports of the CNRS, university of Grenoble Alpes and university Sorbonne Paris Nord.

tools for modeling the construction information implemented by applications that allow modeling the data produced in the building, a structure, or an engineering structure, i.e., construction of great importance and size. It, therefore, has information about the locations and types of sensors deployed in the building, for example, temperature, humidity, and CO2 sensors. The data produced by the sensors can be sent to a cloud or kept at the edge of the network.

In this context and to go towards more digital sobriety, we aim to provide the building manager with a methodology to launch analysis on the sensor data and then build a smart building model. This objective serves as the problem we tackle in the paper. Finally, based on the previous automatic studies, we sketch some plans to reduce data emissions from the sensors while preserving the model. This part of the work is not deduced automatically in this paper for simplicity. The reader should notice that we are working on the analysis plan of the MAPE-K model (see symbol (2) on Figure 1) and also on the interfaces such as cleaning and planning (see symbol (1) on Figure 1) and predicting (see symbol (3) on Figure 1). We aim at generating knowledge for the element (see symbol (4) on Figure 1). But again, the core of the paper is on the analysis plan. Since a fully automatically managed building, in the sense of the MAPE-K model, is currently out of reach, the aim is to provide a decision aid, enabling the skilled expert to make informed choices about data production.

The Internet of Things (IoT) domain has been one of the fastest-growing areas in the computer industry for the last few years. The global sensor market is large and growing fast. By one estimate, it is projected to reach US $346 billion in sales by 2028, up from $167 billion in 2019 [4] . Consequently, IoT applications are becoming the dominant workload for many end-to-end systems such as clouds, fog, or edge computing systems.

The paper organization is as follows. In Section II, we make some parallels between our work and related works in the characterization of IoT data sets. Section III explains our methodology to analyze and characterize our data sets. We first explain the general principles, then give details about the different attributes of our studied data sets that we consider. We also explain the architecture of the IoT system producing the data and some physical characteristics of the system. In Section IV, we explore the data sets for analyzing purposes. We draw charts and figures to "summarize" the data sets. In Section V, we specifically interpret the charts and numbers to provide insights for future use regarding the data sets. Section VI concludes the paper.

II. Related work

In [5] authors deal with a generic but straightforward way to model the workload of typical IoT applications to obtain a practical and reproducible method to emulate loads for data centers. The authors combined a reference architecture and one IoT benchmark. The IoT benchmark populated the functional areas of the IoT reference architecture with benchmark components to emulate their typical behavior in terms of computing and communication requirements. In short, the paper's originality is in using a category of IoT benchmark, which is neither empirical nor mathematical. Contrary to this work, ours does not provide a realistic load generator, but we analyze real IoT traces. Our data could serve as a new user story under the paper's framework.

In [6] authors studied the electricity consumption in the Greener data set, which is a data set related to a tertiary building located in Grenoble, France. Authors pushed the idea that energy disaggregation methods, i.e., the Non-Intrusive Load Monitoring (NILM) methods [7], are powerful tools to get feedback on energy consumption. We also use the Greener data set but not the one related to the electricity consumption. We focus on the global quality attributes of the building and the CO2 metrics produced on multiple sensors located in various rooms. We provide orthogonal and complementary views of the building. Authors in [6] only consider the data quality aspect of the evaluated data set.

In [8] authors focused on the use of advanced artificial intelligence (AI) methods to optimize building energy usage while maintaining occupant thermal comfort. One key point is the definition of "comfort." The authors reviewed some known definitions and noticed that the question is still controversial somehow. In our work, we do not go in the direction of thermal comfort because some data that come with the definition are not available.

III. Data collection and analysis methodologies

A. General approach

The system we consider in this work has IoT devices that send data at irregular intervals, independent of each other. The nature of data can be binary data, strings, timestamps, reals, or integers. Some data sets present metadata. The methodology we developed in this paper, and the organization of the paper, follow the MAPE-K model for the production of knowledge (see symbol (4) in Figure 1). Thus, we first aggregate, clean, and filter the data (see symbol (1) in Figure 1). Then we put in place analyses (see symbol (2) in Figure 1). At last, we generate semi-automatic plans (see symbol (3) in Figure 1). All of this constitutes the methodology we promote in this paper.

In the case of time series, such as energy consumption data, completeness, outliers, timeliness, and accuracy are among the most critical dimensions of the problem. Completeness measures whether some data are missing, accuracy measures whether the samples are correct and reliable, and timeliness measures whether the information is up to date. At the same time, outliers are inconsistent data with the rest of the series, e.g., a statistical outlier. After cleaning the data, this paper mainly focuses on the outliers detection and classification problems. Indeed, we observed some messages in the raw data coming from the server used to query the data, hence the cleaning action. Moreover, the completeness analysis is complex because some data may be shifted in time and not missing due to the drift of the sensors' clock.

B. IoT data sets

All the two data sets we deal with in the paper come from a building in France. The building has more than 22,000 square meters of floor space divided over six floors and the roof. It is a massively monitored and controlled building with more than 1,500 sensors, including about 330 electrical energy meters. The measured data is used to monitor internal conditions and to track consumption. We use, for our experiments, a first sub-data set that corresponds to the monitoring of the building from Thursday, 16 December 2021, to Friday, 7 January 2022. This first sub-data set has 74 attributes (without the timestamp), ranging from the current electricity consumption to the wind direction. The second sub-data set, stored in CSV format, contains CO2 and temperature measures as well as the room's name where sensors are located and the timestamps for each measurement. The data set has 351 sensors for temperature and 76 sensors for CO2. All CO2 sensors are located in rooms with a temperature sensor. We found two CO2 sensors with outliers. The sensors have been canceled for the graphics. The date range for the data is also from Dec 16, 2021, to Jan 7, 2022. Each data set has approximately 38k lines. We can assert the data are spatiotemporal.

These two data sets have been built, by others, by concatenation of data with the granularity of 1 minute, using instant values for DS-2 and aggregated data for DS-1. We do not master this part of the production. For instance, we have no idea about the principle used for aggregation. At the moment, the two data sets are not publicly available for privacy concerns, but anyone can use a public API[1] to extract data and build his own data set.

C. Naming conventions

The data sets naming conventions are DS-1 for the dashboard CSV file and DS-2 for the comfort CSV file collected from the Greener opendata API.

IV. Data sets analysis and observations

Notice that a complete description of the data sets analyses is available as a technical report[2]. This paper synthesizes some relevant observations, and we assume the readers are familiar with machine learning algorithms. Otherwise, they will find more details in the technical report. We mainly use anomaly detection algorithms (Isolation Forest, Local Outlier Factor, and DBSCAN), clustering, or classification algorithms (Support Vector Machine).

A. Outline

The methodology we follow in this section for the analysis step is as follows. First, notice that we conducted different analyses to illustrate the directions of our work, and for the sake of conciseness, we did anomaly detection and classification. Second, we successfully applied a series of treatments on the CO2 and temperature attributes present in the two data sets to infer knowledge regarding the room with the most and the minor issues. Third, we examine the hierarchy empirically, given a certain amount of clusters.

B. Observations for DS-1

Our data sets are multi-dimensional; for instance, we have ten attributes for DS-1. In this case, we used the R tsne package for the visualization. The package implements a high-dimensional visualizing algorithm, which is called t-SNE. t-SNE stands for t-Distributed Stochastic Neighbor Embedding and its main aim is that of dimensionality reduction. It maps multi-dimensional data into 2D (or 3D) representation while preserving the 'structure' in the original data set space. In this way, we can detect patterns in the data set. Stochastic Neighbor Embedding (SNE) starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities representing similarities. Then the t-SNE non-linear dimensionality reduction algorithm finds patterns in the data by identifying observed clusters based on the similarity of data points with multiple features. It is a dimensionality reduction algorithm and not a clustering algorithm.

The number of anomalies given by the LOF algorithm is 91. The number of anomalies given by the HDBSCAN algorithm on DS-1 is 182, and the number of anomalies given by the iForest algorithm is 68. The main parameters of the LOF algorithm are 2 (minPts) and 10 (minPts) for the HDBSCAN algorithm, and 16000 (sample_size), 250 (num_trees), 0.65 (threshold) for the iForest algorithm. The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of the sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets; here are the three results of the anomaly detection algorithms.

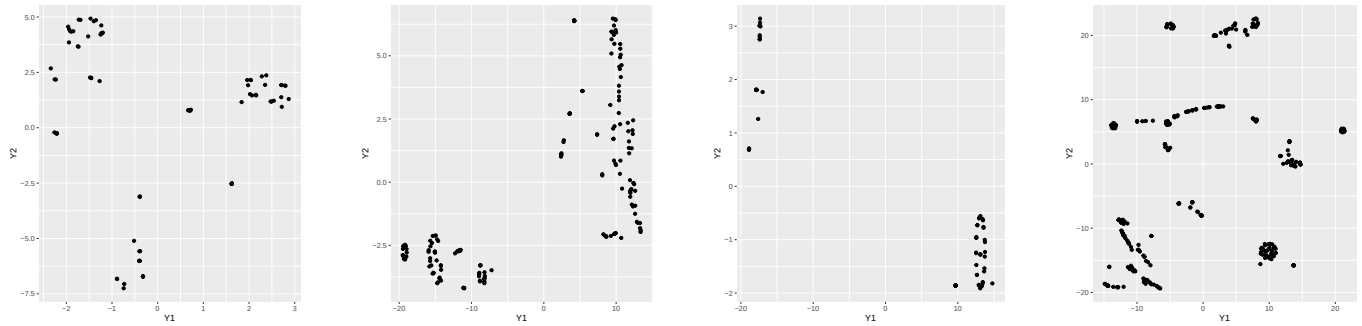
In Figure 2, according to the t-SNE algorithm, the anomalies for DS-1, with the three suggested algorithms (LOF, HDBSCAN, and FOREST), and we also gathered all the anomalies in one file, which allows for observing the sparsity of anomalies. We can say that the three algorithms provide different anomalies, which is confirmed by the measured Jaccard index, that is equal to 0.3%. Remind that the DS-1 has eleven attributes, hence the requirement for dimensionality reduction for the visualization.

Since the anomalies for DS-1 are different from one algorithm to another, we decided to classify DS-1 in two ways through the SVM algorithm. First, we used the SVM algorithm on the whole set of initial data. Second, we eliminated the union of anomalies from the initial data set; then, we used the SVM algorithm. The idea is to check if anomalies will or will not change the classification. We also decided to consider the union of the anomalies because they are pretty different. In the case of similarities in the anomalies, we suggest using the intersect of the anomalies, and those intersects will be removed from the initial data set.

As cross-validation, Tables I and II report the results for the inter-cluster analysis. Computing these indicators requires

[1] http://mhi-srv.g2elab.grenoble-inp.fr/django/API/

[2] https://gricad-gitlab.univ-grenoble-alpes.fr/batpred/ieee-isc2-2022



(a) LOF (b) HDBSCAN (c) IFOREST (d) All

Fig. 2. Anomaly detection for DS-1

TABLE I
COMPLETE DIAMETER (INTRA-CLUSTER METRIC) WITH DIFFERENT DISTANCE METRICS AND INTRA-CLUSTER DIAMETERS – DS-1' = DS-1 WITHOUT ANOMALIES –

Data set	Distance	c1	c2	c3	c4	c5	c6	c7
DS-1	Euclidean	586.4	201.7	399.1	768.2	408.4	226.8	81.1
DS-1'	Euclidean	161.6	41.0	174.0	226.5	129.9	51.6	36.9
DS-1	Manhattan	840.9	326.9	746.1	1526.0	900.4	481.9	189.9
DS-1'	Manhattan	585.4	200.9	388.5	664.7	406.6	219.9	80.9
DS-1	Correlation	0.2607	0.0157	0.2759	1.8792	1.3916	0.1086	0.0101
DS-1'	Correlation	0.2586	0.0152	0.2494	1.8863	1.4070	0.1105	0.0101

the `cls.scatt.data` function from the R `clv` package. Indeed, this function finds the most popular intercluster distances and intracluster diameters, i.e., it computes six intercluster distances and three intracluster diameters. For the sake of brevity, we present the *complete diameter* that represents the distance between the two most remote objects belonging to the same cluster, in Table I), and the *centroid linkage distance* (in Table II) that reflects the distance between the centers of two clusters $(v(i), v(j)$ clusters centers). The number of clusters was 7.

C. Observations for DS-2

The second dataset is relevant as the first one only provides aggregates. We lose spatial information, which is a critical feature in building management. As shown in Figure 3, information from DS-1 is not rich enough to explain such a wide distribution. Maxima, minima, and mean values are plotted with thick lines. However, outliers may only result from specific uses and cannot be considered abnormal at this point. The technical report provides the hierarchization made by the `scipy.cluster.hierarchy.linkage` algorithm [9] from the Python library `scipy`, with the linkage method ward, over CO2 sensors and weather data. Sensors indexes follow rooms' alphabetical and numerical ordering, so nearby sensors are consecutive. Similar spaces are clustered soon at the bottom of the graph, and very different behaviors lead to clusters being aggregated late at the top. We show that all weather data is clustered together and has little impact on CO2 clustering, even outdoor wind, while early clusters often aggregate pairs of consecutive rooms. We also provide a similar dendrogram resulting from temperature and weather data clustering. We identify four main clusters through data, showing how behaviors change from one floor to another. One smaller cluster links sunlight received to the temperatures in rooms. Our knowledge about the building makes us aware these rooms face the South-est and South directions.

V. FINDINGS AND COMING WORK

A. Analysis concerns - Findings for DS-1

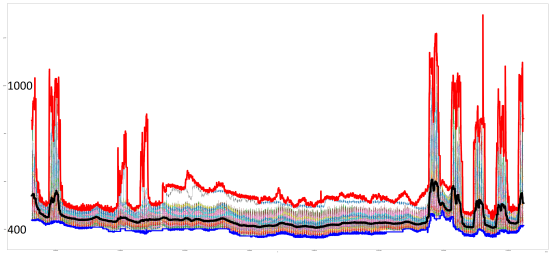
General conclusions from the initial observations and analyses over DS-1 include anomalies, classification, model fitting, and the potential for fewer measurements than are currently made. We noticed in Figure 2 that the different anomaly detection algorithms returned anomalies that were distinct from each other. We observed the phenomenon because the dots on the four sub-figures in Figure 2 appear in separate dials. The techniques used for the detection explain the different detections. For instance, the DBSCAN Algorithm is a density-based clustering non-parametric algorithm. In contrast, the iForest algorithm isolates abnormal points in the data set instead of a typical instance model. Thus, the result regarding the different outputs of the three algorithms was expected. However, we are faced with the question of exploiting the anomalies. In our case, we suggested taking the union of the sets of anomalies because the Jaccard index was tiny (about 0.3%).

The benefit of using the union of anomalies is demonstrated in Table I and Table II. The diameter of the clusters is much

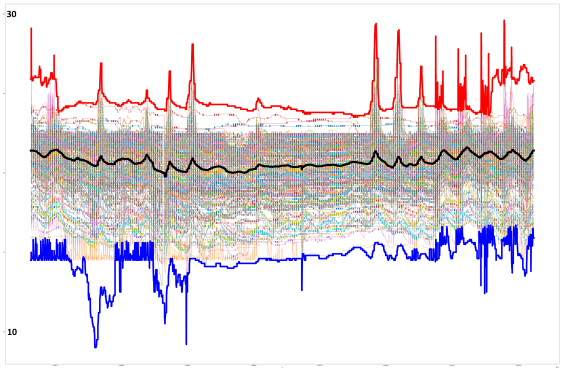
TABLE II
CENTROID LINKAGE DISTANCE FOR THE EUCLIDEAN DISTANCE METRICS

DS-1							
	c1	c2	c3	c4	c5	c6	c7
c1	0.0000	260.2037	342.6309	660.6085	947.0606	1170.3172	1232.5043
c2	260.2037	0.0000	340.4852	743.9295	971.6063	1197.1260	1253.7312
c3	342.6309	340.4852	0.0000	407.7996	638.1946	863.6430	940.8476
c4	660.6085	743.9295	407.7996	0.0000	323.1577	527.0483	657.6421
c5	947.0606	971.6063	638.1946	323.1577	0.0000	232.9979	453.5531
c6	1170.3172	1197.1260	863.6430	527.0483	232.9979	0.0000	367.6166
c7	1232.5043	1253.7312	940.8476	657.6421	453.5531	367.6166	0.0000

DS-1' = DS-1 without anomalies							
	c1	c2	c3	c4	c5	c6	c7
c1	0.0000	260.2371	340.7637	658.1367	944.5284	1167.5002	1230.9060
c2	260.2371	0.0000	340.7087	743.8000	971.0630	1196.9086	1253.7360
c3	340.7637	340.7087	0.0000	407.3638	637.0442	863.0272	940.5944
c4	658.1367	743.8000	407.3638	0.0000	323.0683	526.4675	657.6756
c5	944.5284	971.0630	637.0442	323.0683	0.0000	231.3793	453.0424
c6	1167.5002	1196.9086	863.0272	526.4675	231.3793	0.0000	367.4015
c7	1230.9060	1253.7360	940.5944	657.6756	453.0424	367.4015	0.0000



(a) Per room CO2 evolution (ppm)



(b) Per room Temperatures evolution (°C)

Fig. 3. Temperature and CO2 evolution over DS-2

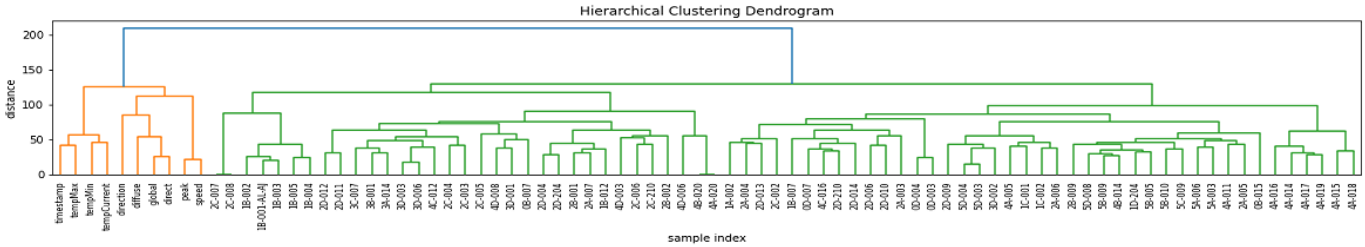
better when the anomalies are removed from the data set. They are smaller; hence the clusters are denser. Moreover, the intra-cluster distance varies little when working with the dataset without anomalies. We conclude that the centroids are not impacted by removing the anomalies. We also investigated finding the 'best' number of clusters. It is a critical issue to get representative subsets inside the data.

Next, we asked how the two datasets, DS-1 and DS-2, were related to the CO2 attributes. We indeed have both aggregate attributes (DS-1) and instantaneous data (DS-2), but we do not know how the aggregated data were produced nor if there is a link, for CO2, between the min, max and mean aggregated data. For this last problem, we performed a classification leading to regression, and we observed that we could not tell that the mean value of the aggregated CO2 was produced via an average computation over the aggregated min and max.

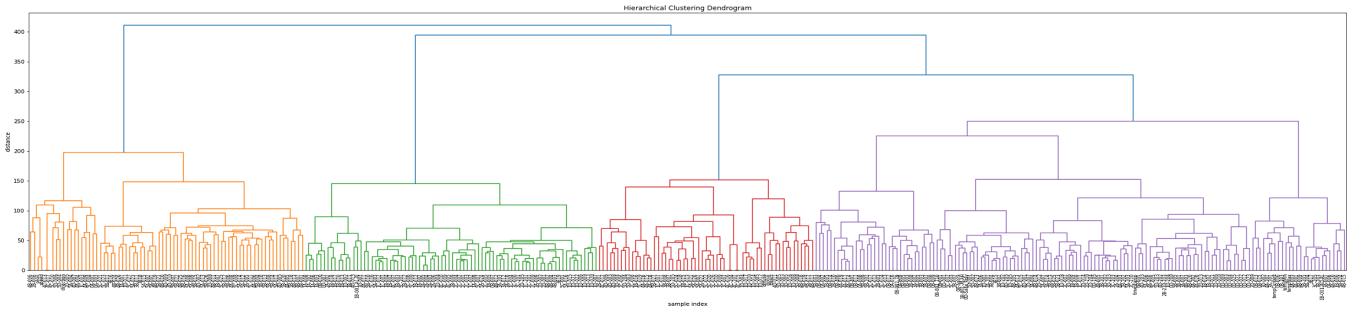
B. Analysis concerns - Findings for DS-2

General conclusions from the initial observations and analyses over DS-2 include the feasibility of behaviors clustering and inducted expectations. CO2 measures lead to clustering nearby rooms more often than temperature measures, in Figure 4. Temperature distribution over the building may vary more than CO2, which explains more temperature sensors. The pairing we notice still questions the use of 74 CO2 sensors and shows that information may be deducted from one sensor instead of two, allowing fewer sensors. For this purpose, we show weather data does not help and question the relevance of metadata to provide more information about expected CO2 and rooms' behavior.

The clustering of CO2 sensors displays 3 well separable behaviors while the temperature sensors clustering displays 4 clusters, one subdivided into 3. On-site metadata and further observations of the building could explain them. As figure 5 suggests, small clusters aggregate fast. They are all similar, so it would be hard to find why they are grouped, and bigger clusters would be even harder to explain. Figure 5(a) shows irregular plateaus that may hint at separable behaviors on smaller clusters when they are still 5, 13, 17, or even near 30 clusters left that new metadata as occupancy or architecture

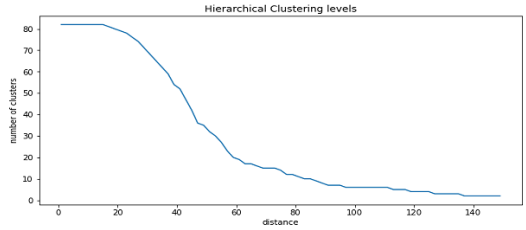


(a) CO2 clustering denrogram

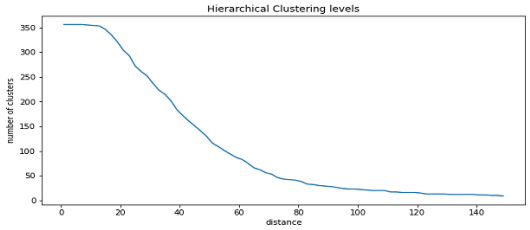


(b) Temperature clustering dendrogram

Fig. 4. Clustering dendrograms



(a) Number of clusters along CO2 clustering hierarchy



(b) Number of clusters along temperature clustering hierarchy

Fig. 5. Evolution of clustering along hierarchy

elements may explain. The smoothness of the curve on Figure 5(b) only suggests that explanations would be exponentially hard to define, as we need to combine more accurate expert knowledge and metadata.

C. Findings regarding correlation issues

To go in the direction of planning, we have made the following analyses. We extracted temperature and CO2 data from the building's rooms and meteorological data from a station on the roof. By type of expenditure, energy consumption data of the building are also accessible with an hourly step. The use of such data is described in the technical report.

The correlations obtained by comparing different parameters provide intelligence for decision support and planning. For example, they allow us to establish a clear link between the temperature of the rooms on the southeast façade and the exposure to the sun. We also find a relation between the outside temperature and the corridors. We observe a clear link between the energy dedicated to lighting and the power committed to ventilation, signifying that ventilation depends on presence. Groupings based on temperature data alone also allow us to identify classrooms and offices.

Based on these findings, we may next advise strategies to stop data production by sensors located in classrooms and offices with the same behavior regarding the temperature. Analyzing data from a regulated smart building is a complex problem because unknown algorithms already try to control rooms' behavior. Even more potential will lie in buildings without such fine control.

D. Future works

Based on the synthesis of our observations and analyses, explained in the two previous paragraphs, we propose exploring the following avenues. On the DS-2 data set, we have noticed similar rooms in the sense of clustering. We could undoubtedly forget some sensors or make them produce data less frequently than currently. We conducted, in parallel, an empirical study from the DS-1 data set, which, by varying the number of analyzed data, tried to preserve the clustering resulting from the complete raw data set. Regardless of the clusters' structural properties, random deletion does not allow us to conclude that

the approach is still valid. The properties of some clusters are preserved, but not for all. We propose understanding what maintains the properties of "same centroids" and the number of elements with the data forgetting rate.

We also measured that the number of anomalies and their management impacted the clustering and, thus, the data models resulting from the clustering. We propose observing the distribution of anomalies for the data set and interpreting them. For example, is attribute A abnormally high/constant/variable compared to the average behavior? Anomaly management and detection were done on the DS-1 data set, which does not contain a timestamp attribute. We propose to conduct a similar analysis on the DS-2 data set to observe whether the removal of temporal anomalies (marked on the heterogeneous data or not) affects the spatial clustering performed on only two attributes as we did. Perhaps it is appropriate to merge the DS-1 and DS-2 data sets, apply an anomaly detection algorithm on the flat spatial data set, and then apply spatial clustering. It seems to us that the general problem is to couple two algorithms on the same spatial data set.

On the other hand, perhaps we have too many approaches and algorithms to characterize smart building data, and the process should be refined to update a methodology with fewer tools. Finally, we can also imagine that the analysis and characterization are not done on a fixed data set but that the data set evolves. So, instead of investigating a massive data set in the Cloud, we could try to do it as close to the data as possible, for energy efficiency purposes, in the framework of Edge Computing. We can also imagine sharing the work between the Edge and the Cloud. In this case, the granularity of the calculations and data to be distributed in the Edge rather than in the Cloud arises. It also raises the question of designing online algorithms on low-cost devices or sensors performing online clustering and anomaly detection on time series.

The previous study exploits only the historical data, and the lack of contextual information hampers this approach. It is, therefore, necessary to add relevant data such as a room usage calendar, a building plan, or very short-term weather forecasts. Predicting a room's living conditions is possible based on those around it, and it will eventually compromise the distribution of sensors between rooms and the performance of the projection. Linking data from sensors and other easily accessible knowledge opens the way to automatic classification techniques that can be used without particular expertise. To obtain personalized management, they will allow adapting prediction algorithms to each building, regardless of the diversity of architectures, occupants, or the number of sensors installed.

VI. Conclusion

Experiments with massive deployment of smart sensors are becoming increasingly numerous, not only in research laboratories but also in cities. The COVID pandemic, for example, pushed the deployment of CO2 sensors in schools. Even if it is regrettable, this is a fact that a sober deployment and a reasonable a priori on the use of the data produced and the objective to be reached have not always been used to guide the deployments. Our work proposes a series of tests to characterize the data produced at the building level. The issues that we feel are important to address are related to the nature of the data, for example, the absence or presence of outliers and their management as they may impact the clustering and, thus, the nature of the data models resulting from the analysis. We propose in the article to build different views of the data. This process is a step forward so that the various stakeholders working on designing intelligent buildings can now rely on new business knowledge. We have proposed analyses and observations on actual data to illustrate the approach. We showed a profound potential reduction in data production due to specific sensors' "same" behavior. We conclude that we can indubitably go to a more sober smart building management.

References

[1] A.H. Buckman et al. "What is a smart building?" In: *Smart and Sustainable Built Environment* 3.2 (Sept. 2014). URL: https://eprints.whiterose.ac.uk/80714/.

[2] P Horn. *IBM, Autonomic Computing: IBM's Perspective on the State of Information Technology*. URL: https://people.scs.carleton.ca/~soma/biosec/readings/autonomic_computing.pdf.

[3] Jeffrey O. Kephart and David M. Chess. "The Vision of Autonomic Computing". In: *Computer* 36.1 (Jan. 2003), pp. 41–50. ISSN: 0018-9162. DOI: 10.1109/MC.2003.1160055. URL: https://doi.org/10.1109/MC.2003.1160055.

[4] Sri Chandrasekaran and Ravi Subramaniam. "Why IoT Sensors Need Standards - They could improve performance and spur development of new applications". In: *IEEE Spectrum* January 11 (2022). URL: https://spectrum.ieee.org/why-iot-sensors-need-standards.

[5] Summers J et al. Batz T Herzog R. *IoT Workload Emulation for Data Centers*. Tech. rep. Open Research Europe, 1:12, 2021.

[6] Benoit Delinchant et al. "GreEn-ER living lab: A green building with energy aware occupants". In: *2016 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*. 2016, pp. 1–8.

[7] G.W. Hart. "Nonintrusive appliance load monitoring". In: *Proceedings of the IEEE* 80.12 (1992), pp. 1870–1891. DOI: 10.1109/5.192069.

[8] Jack Ngarambe, Geun Young Yun, and Mat Santamouris. "The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls". In: *Energy and Buildings* 211 (2020), p. 109807. ISSN: 0378-7788. URL: https://www.sciencedirect.com/science/article/pii/S0378778819336527.

[9] Daniel Müllner. "Modern hierarchical, agglomerative clustering algorithms". In: *arXiv:1109.2378 [cs, stat]* (Sept. 2011). arXiv: 1109.2378 version: 1. URL: http://arxiv.org/abs/1109.2378 (visited on 01/28/2022).