Smart buildings: a monitoring and data analysis methodological framework

Cristiana Bolchini[a], Angela Geronazzo[a], Elisa Quintarelli[a,*]

[a] *Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, via Ponzio 34/5, 20133 Milano, Italy*

Abstract

This paper proposes a methodology for monitoring and data analysis in buildings, leading the user in the design and implementation of smart environments able to collect information on spaces and the comfort perceived by their inhabitants. A systematic descriptive framework is defined to identify the most convenient strategies for monitoring and data interpretation, driven by the constraints and goals set by how data will be exploited. A wide body of field works has been analyzed and put in relation to the methodology, highlighting the difficulty to extract general guidelines with respect to the same goal and to reveal use case specificity, preventing the possibility to learn from previous solutions. We claim the need to overcome the specific features of each case study to elicit general knowledge, and thus the quest for a unified framework, useful in different scenarios.

Keywords: Monitoring, Smart building, User Comfort, User Behavior, Energy Efficiency, Sensor Data

1. Introduction

Energy efficiency and users' comfort are two of the main drivers for instrumenting buildings to make their management successful; indeed, second generation solutions are taking into account the former aspect with great emphasis, as a counterpart to the primary objective to reduce energy consumption. The final objective can be achieved by means of Home/Building Energy Management Systems (H/BEMSs) for a dynamic and adaptive control of the building, and/or by means of a refurbishment to improve the envelope should the thermal profile have critical performance. The way the building should be instrumented and how data should be collected, processed and analysed strongly depends on requirements (e.g., having a campaign with data for all seasons) and goals (e.g., assessing the users' comfort level on a monthly basis). Today it is easy to disseminate several sensors providing redundant information, since sensors and data storage have decreasing costs. Moreover, wireless technologies allow for a non-intrusive instrumentation of spaces, for short or long periods of time, in existing buildings. We refer to smart buildings as buildings empowered by ICT, with sensors, actuators and embedded systems that allow to collect, filter and produce information to be exploited to provide functions and services. However, the proposal applies to older buildings where the instrumentation is only temporarily applied to collect the needed information.

As a result, it is not so uncommon to produce large amounts of data; to be able to effectively handle them, it is necessary to face a number of challenges, such as i) selection of a sampling strategy, ii) necessity or not to clean data and correct errors, iii) summarization and aggregation strategies.

Indeed, some of these choices are strictly related to the main objective of the monitoring activity. However, since there is no de-facto standard approach, the space instrumentation/setup is usually carried out on the basis of an adapted subset of recommendations from the commonly adopted guidelines [1], personal experience and/or commercially available kits, so that each building setup seems to be a stand-alone, unique use-case. Thus, within the energy-efficient building scenarios, both methodological proposals and case study approaches rely on a customized (often omitted) solutions for monitoring the smart spaces and/or the user comfort.

In particular, given the large set of existing smart building monitoring experiences, it is timely to define a general methodological framework that identifies i) the main choices that are available in designing a monitoring campaign, ii) the alternatives for each identified choice, and iii) the effects on the final data. In fact, there are several works available in literature [2–4] presenting solutions that are possibly tailored for the specific application context, where choices are made based on the experience and rule-of-thumb estimations. Indeed, it is common to collect an excessive amount of data and/or adopt data analysis strategies that are oversimplified with respect to the final use of the collected data.

The main contribution of this paper is the introduction of a methodological and descriptive framework for the design and analysis of monitoring campaigns for buildings, within the context of energy efficiency, comfort and users' behavior analysis. We claim that the approach we

*Corresponding author

Email addresses: cristiana.bolchini@polimi.it (Cristiana Bolchini), angela.geronazzo@polimi.it (Angela Geronazzo), elisa.quintarelli@polimi.it (Elisa Quintarelli)

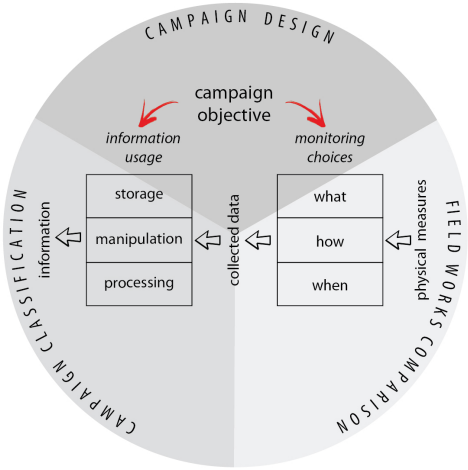

Figure 1: The conceptual schema implemented by the proposed framework.

have elicited starting from existing work, which systematically takes into account the relevant aspects determining the most appropriate monitoring strategy for buildings, is pivotal to an effective monitoring, control or assessment of the conditions of the spaces and their occupants. Moreover, it can also favour the re-use of existing solutions or a fair and useful comparison among different alternatives. In particular, Figure 1 depicts the main use in supporting the selection of the practical choices to be carried out in setting up the monitoring of variables and data analytics, as well as the framework exploitation for campaign design and analysis (classification and comparison).

To this aim, Section 2 introduces the goal of the work and the questions it aims at answering, providing the starting point for the proposed approach, discussed in Section 4. Section 3 outlines previous work and summarizes it in view of the proposed framework, for a classification of the existing solutions. Section 5 aims at showing how the proposed framework could be used to classify field works, for a more immediate analysis and comparison. The framework is then used to design the monitoring campaign in a local case study, presented in Section 6, along with the related analysis. Section 7 draws some conclusions.

2. Rationale and problem definition

The availability of possibly large quantities of data collected from sensors in instrumented ambients, to enhance energy efficiency and occupant comfort, poses some challenges related to the amount of collected data and their quality, when exploited to i) derive analyses related to energy usage and/or user comfort or ii) drive dynamic optimization algorithms, taking educated decisions to reduce consumption without compromising comfort.

Indeed, today it is reasonably easy and relatively cheap to instrument offices and rooms to collect ambient conditions (e.g., temperature, humidity, luminosity, ...) and energy consumption information. Based on the final goal and the duration of the monitoring, different solutions can be adopted for the sampling, either using a policy based on the change of value of the monitored measure above a chosen threshold, or by setting a sampling frequency. As an example, when deploying a monitoring solution for a home energy manager which is meant to operate for years, battery life is crucial and sensors transmit data when there is a significant event (the above-mentioned change of value above a certain threshold). Moreover, this solution is characterized by a possibly limited amount of data to be stored, corresponding to the changes of value. On the other hand, when performing a monitoring campaign to assess an apartment performance or to identify critical ambient conditions, battery life and data storage are not an issue, and regularly time-distributed series of data are produced, collected, post-processed and analyzed.

Another aspect that affects the choice of what data should be collected, events or data values, is the final goal of the sensing. If the information is used to trigger automatic actions (such as turning on/off the air conditioning when the temperature reaches a certain threshold), the event is relevant while the almost constant values monitored during the entire window of time are not. If the information is used to analyse an ambient, it is easier to extract trends and profiles, sampling at regular intervals straightforwardly provides the necessary information rather than having to reconstruct it from saved events.

Another interesting option when planning/designing the data monitoring strategy is the identification of how/when data will be exploited, either *real-time* or *batch*. A real-time exploitation of data, where "real-time" refers to the fact that data is used continuously, occurs when H/BEMSs use the monitored data to take educated decisions in the management of the building, or when data is used to feed an anomaly detection system that promptly reacts to specific identified critical situations. A batch exploitation of data occurs when data is collected for an a-posteriori assessment of the building/users' conditions, for profiling, or other data analytic activities that can be carried out off-line.

Monitoring aspect	Available alternatives	
Battery life	long	short
Amount of data	limited	high
Sampling strategy	change of value	fixed period
Collected data	events	values
Sensing goal	control	assessment/analysis

Table 1: Sampling strategies characteristics.

These are examples of the several aspects that need to be identified to determine the most effective (cost/time/effort) monitoring strategy, which are indeed dependent on the final goal of the activity, and Table 1 summarises these considerations. At present, as mentioned, there is no best-

practise methodology to guide a user in setting up the appropriate monitoring solution, but rather several reported case studies (e.g., [2; 4]).

Since not all of them tend to mention the adopted choices or provide a throughly explanation of the rationale behind such choices, it is sometimes hard to benefit from past experiences and to eventually compare different solutions. In this perspective, it is possible to formulate the problem we aim at tackling in the following way. In the building context, based on the final goal of the monitoring activity, what are the aspects that drive the design of the monitoring architecture and strategy, and what alternatives are available? Furthermore, given an application context, are there good-practise solutions that can be adopted?

The methodological framework we propose and discuss in the next section aims at supporting the user in taking into account all relevant aspects in a systematic way, to be able to design and implement an effective solution. Furthermore, it can be used as a way to systematically report and classify the adopted solutions for an immediate characterization and comparison of the relevant aspects in field works. Flexibility and extensibility are two key factors of the proposal, to allow the adoption of the methodology to as many application scenarios as possible.

3. Related work

Two bodies of works are reviewed in this section; field works used as a reference to design the monitoring campaign for improving users' comfort and feeding the H/BEMS presented in [5], and works presenting methodologies and frameworks to plan, deploy and analyse monitoring solutions.

Among the numerous case studies related to users' comfort and energy consumption monitoring, we here cite [2; 3], two surveys collecting solutions for monitoring i) users' activity and behavior in residential and non-residential sectors, and ii) users' comfort and behavior to put energy consumption in relation with indoor users' comfort, respectively. In these surveys a comparative analysis of the existing solutions is reported, although too many aspects are not specified and discussed, so that it is not straightforward to determine what approach would be best suited for a new field work. Therefore, other specific works have been taken into account, dealing with energy consumption monitoring and analysis, as well as users' comfort analysis with the aim of identifying a suitable monitoring infrastructure, data collection strategy and data analysis approach. More precisely, the authors in [6] present a middleware solution for the analysis of building energy usage in smart grid scenarios. The system analyzes and interprets energy usage in real-time, by collecting data from the underlying wireless sensor networks (WSNs). Moreover, it assesses the effectiveness of energy use by evaluating occupancy and occupants' comfort level. The work presented in [7] proposes a distributed energy monitoring system focusing on the IT infrastructure using WSNs, proposing a flexible solution to be easily extended. In [8], a strategy for occupancy estimation using a WSN based on temperature, humidity, light readings and audio levels is presented. Collected data are initially analyzed by using a set of statistical classifiers, then context information on room occupancy schedules and resource usages is integrated into the model. Results of this first analysis step are further forwarded to a new classification level that provides a fine-grained occupancy estimation and delivers it to support decision making processes. The adopted IT infrastructure and data analysis collection and analysis are the main contributions of the work described in [9]. The overall system is organized into five different layers devoted to energy measurement, data transmission through a dedicated bus, a data collector layer in charge of sending sensed data to a storage layer on the Internet, and a storage layer, in charge to provide easy access to the collected data for producing charts and summary statistics at different detailed levels. The solution presented in [10] consists of three modules comprising data acquisition and processing, a knowledge repository and a performance analysis system devoted to energy consumption prediction and modeling and decision support. Indeed the authors provide an overview of the proposed solution although some of the fundamental aspects characterizing the monitoring campaign are missing (e.g., missing and erroneous values detection and handling). Furthermore, the adopted solutions are hardly generalized so that it is hard to exploit the same setup in other field works or compare against.

When focusing on the collected data rather than on the IT infrastructure, the attention is on the interpretation of the data to to identify possible inefficiencies and recommend short and long-term strategies to improve energy usage. In particular, we report some approaches based on data mining solutions, where the quality of collected data plays a relevant role. The authors in [11] present an analysis of building environmental data, involving also buildings' physical features and climate conditions. The analysis refers to a university building and aims at effectively predicting rooms schedule by optimizing energy usage and users comfort. The work presented in [12] proposes a large scale analytics platform for real-time energy management within a university campus smart grid system. A multivariate predictive technique is employed to discover anomalous performance behavior; moreover the collected data is further analyzed to enable the creation of building consumption models. The solution presented in [13] aims at discovering useful knowledge in Building Management System (BMS) data. The analysis framework comprises four modules: a data pre-processing and partitioning module, to improve data quality and group data with relevant similarities using a clustering approach; a knowledge discovery module to extract recurrent consumption patterns and temporal associative rules; and a post-processing mod-

ule to improve results interpretability. In this perspective [14] presents data mining applications in the building domain, considering multiple purposes ranging from predictive analysis to descriptive tasks, and considers different applications comprising energy prediction and anomaly detection, building control and occupants behavior. The lack of an analysis framework makes it difficult to actually compare the different solutions, to determine what elements primarily affect the final results and what solutions could be reasonably applied to other contexts, a gap our proposals aims at filling by means of a systematic characterization approach.

With respect to such a methodological and systematic approach for supporting the design and deployment of the monitoring infrastructure, another group of solutions have been analysed. More precisely, the frameworks presented in [15] and in [16] are devoted to large sensor architectures management, investigate spatial sensors relations and analyze patterns in sensor metadata to effectively deploy a building monitoring campaign. The former investigates whether linear correlation or statistical dependency is better suited to infer spacial relations between pairs of temperature sensors in a commercial building. The analysis has been carried out in multiple test beds with data collected every minute; results show the efficacy of the linear correlation and stress the importance of the time series window size selection. The latter presents Zodiac, a framework able to name and classify sensor points in large building monitoring infrastructures by actively learning from sensor metadata thus requiring limited manual input data. The approach empowers a hierarchical agglomerative clustering method to group points on the basis of the intrinsic similarity in sensor metadata. The work presented in [17] focuses on energy-harvesting sensors, by addressing issues of scalability and deployment in indoor spaces of wide sensors networks to effectively support building monitoring tasks. Each device comprises an energy-harvesting power supply and a node trigger mechanism in charge of activating the device according to application-specific needs. The proposed architecture has been validated within multiple building specific applications. Common to these works is the definition of a practical framework for supporting the users in the efficient deployment of sensors for the monitoring campaigns, in a general way, since they seem to be independent of the final objective, which, in our experience, has a relevant impact on the way data should be collected and manipulated.

Finally, when considering data analysis and visualization, [18] and [19] present a comprehensive IT infrastructure devoted to data collection and storage for specific solutions for managing building information, respectively. BuildingDepot is an extensible and distributed architecture for storing and sharing building-related data. It provides standardized interfaces to easily access the stored data thus enabling end-users in the exploitation of fine-grained real-time data. SMAP is an IT infrastructure which allows to store and access building related time series data. It stores time series data streams with associated metadata and provides real-time querying and visualization tools for the collected data. Moreover, the infrastructure allows to push control strategies in the BMS. A step forward in building metadata representation is offered by [20], authors propose an ontology based representation for sensors and building's subsystems. The core of Brick is composed by an ontology, defining the main concepts and their relationships, specifically designed for the building domain concepts. The framework adopts the Resource Description Framework format to represent knowledge and SPARQL as query language. These solutions are mainly focused on the software infrastructure rather than on the monitoring issues.

The closest solution to the framework we propose is the one presented in [1], where guidelines are provided to determine, given the information the user is interested in (e.g., users' comfort), what data should be collected (e.g., indoor temperature and humidity) and with what frequency (e.g., once per hour). Indeed the reference provides a lot of information, mainly adopted in North America, according to regulations. However it focuses primarily on the definition of the data to be collected rather than the classification, characterization and documentation of the adopted choices. Moreover, no guidelines are provided for the manipulation of data (e.g., data cleaning, outlier detection, ...) that is eventually necessary based on the final campaign goal.

Table 2 offers a classification of the contributions previously described with respect to the main aspects they present, the IT infrastructure, the sensor network for collecting information or the data analysis/manipulation. A bullet in a column refers to the fact the main focus of the work is on that aspect. Those contributions that are actually reporting field works will be later used as case studies for the application of the proposed methodological framework in Section 5.

As stated, the goal of the proposed approach is to define a framework that, in as much as possible simple and comprehensive way, provides guidelines to design the monitoring campaign (e.g., by identifying the physical measures to be taken, such as indoor, envelope and outdoor temperature, and with what granularity) based on the final goal, also supporting a systematic classification and documentation of all the adopted choices.

4. The proposed methodological framework

As mentioned, there are several application scenarios for the exploitation of monitoring campaigns in buildings and environments, from temporary installations for profiling and assessing the environment and the user comfort, to permanent ones, to support H/BEMSs. The framework here proposed aims at supporting the user in the design and implementation of the monitoring infrastructure/campaign and data analytic, suitable for the widest

Work	IT infrastructure	Sensor network	Data analysis
[15]	•		
[16]	•		
[17]	•		
[18]		•	
[19]		•	
[20]		•	
[3]		•	•
[2]			•
[14]			•
[6]	•	•	•
[7]	•	•	
[8]		•	•
[9]	•	•	
[10]	•	•	•
[11]			•
[12]			•
[13]			•
[14]			•
[1]			•

Table 2: A summarization of the analyzed related work with respect to the main focus.

range of application scenarios and goals. Moreover, it elicits all relevant aspects for the definition of the monitoring campaign such that it can also be adopted as a systematic reference to classify and report field work. Since it refers to elements gathered from existing solutions and practical experiences, it is defined in a flexible way for an easy extension to other aspects not currently taken into account.

The idea stemmed from the challenge to setup a monitoring campaign and to compare previous field work in order to elicit useful knowledge from case studies described in literature. A monitoring and data analysis campaign has been carried out to perform a users' comfort analysis in the premises of a building of our university. The campaign has actually been planned referring to domain experts for the selection of the sensors to be used, for the sampling strategy to be adopted (and related parameters), and using a subset of all rooms to derive the overall comfort analysis. Starting from the initial setup planned by domain experts we present, through experimental simulation (e.g. by simulating different sampling strategies, by evaluating different pre-processing methods), an analysis of how to progressively extend it to comprise new dimensions at first neglected, and how to re-tune parameters which have been set according to common best practices.

4.1. Framework elements

The methodological framework identifies a set of so-called *dimensions*, that characterize the monitoring solution with respect to the given application scenario. As an example, the *objective* of the monitoring campaign is a characterizing aspect that drives some of the choices of the monitoring strategy: if the H/BEMS is the final goal, data should be collected in *real time*. The *campaign time span* is another characterizing aspect. We have called such characterizing aspects, *dimensions*. For each one of these dimensions, various alternatives exist, and we have tried to identify as many as possible of them, to provide a complete set of *values*. For instance, considering the objective dimension, possible values are an *energy audit*, *dynamic control* by means of a H/BEMS. When considering the campaign time span objective, possible values are *years*, *months* or *weeks*. Indeed, it is possible to introduce both new dimensions, and/or additional values, should they be deemed interesting, or remove those that are not relevant. For some values, such as *weeks*, a *parameter* might be appropriate/apply, such as the number of weeks the campaign is planned for.

The proposed methodology has been elicited on the basis of the monitoring campaign we are performing, and proposes some guidelines to resolve issues we have encountered during instrumentation and data analysis.

4.2. Dimensions, values and parameters

The dimensions currently included in the framework are the following ones:

- *objective*, that models the main goal(s) of the monitoring campaign;
- *campaign type*, used to differentiate situations where a single setup can be adopted or several ones are necessary based on different campaign phases;

- *data temporal usage*, whether collected data is used in "real-time" by a H/BEMS, or for off-line data analytic;
- *sampling strategy*, that specifies what kind of data sampling is adopted;
- *analysis granularity*, that refers to the possible time windows (e.g., hour, day, ...) adopted to analyze the data;
- *aggregation operator*, stating the aggregation operator adopted to summarize temporal (i.e. multiple values for a single time period) and spatial (i.e. multiple sensors record the same sensed variable) redundant values;
- *segmentation strategy*, related to the time window used to split the stream of data;
- *campaign time span*, that refers to the duration of the monitoring campaign;
- *instrumented spaces*, to specify whether the entire building / home is instrumented, or only some rooms;
- *redundancy*, stating if redundant information is collected to achieve a certain level of reliability, anomaly detection capability, ...;
- *data cleaning requirements*, related to the management of raw data and that can be further organized in
  - *outlier detection*, and
  - *missing values imputation*.

It is worth mentioning that dimensions are not orthogonal and influence each other. As an example, when the objective of a campaign is dynamic control, the sampling strategy is necessarily in real time, since decisions are to be taken on the latest values. This dependency between the two dimensions occurs in the specific case, but for other objectives the sampling strategy may not be determined a priori.

A detailed presentation of the dimensions and the alternative values follows, and the graphical view is presented in Figure 2.

4.2.1. Objective

This is the most important dimension, expressing the primary goal(s) of the monitoring infrastructure/campaign. Within the smart building domain in non-residential scenarios, it is possible to identify different use cases:

- Key Performance Indicators (KPIs) extraction with reference to users' comfort and energy usage,
- users' comfort analysis with reference to country-specific regulations,

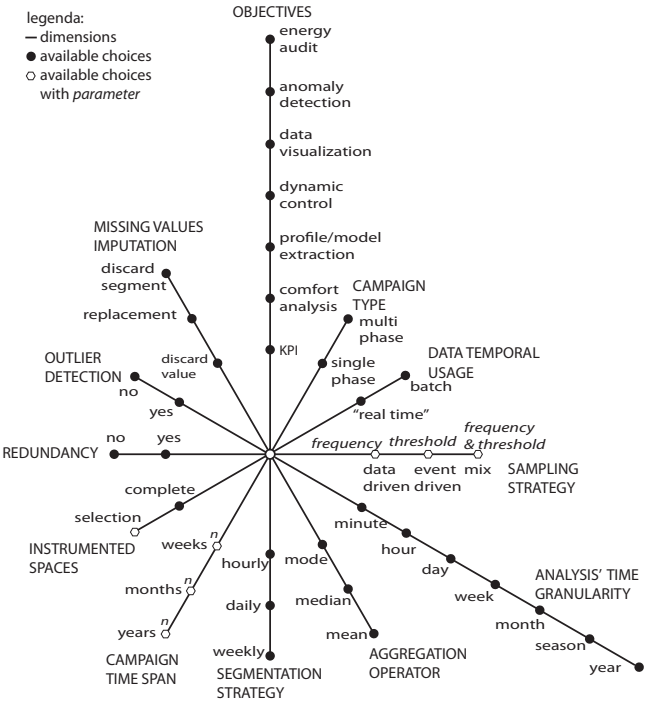

Figure 2: The monitoring campaign dimensions and alternatives identified by the proposed framework.

- thermal and energy modeling of the building and its subsystems, such as Heat Ventilation and Air Conditioning (HVAC), lighting, user devices,
- building envelope profile extraction, used to identify renovation/refurbishing actions aimed at improving energy usage and users' comfort,
- data visualization to enhance users' engagement to induce a sustainable behavior,
- information collection to develop a long-term decision support system for the H/BEMSs,
- data monitoring to collect the current status of the building(s) when H/BEMSs are exploited to optimize energy sources usage and/or interact in a demand-response scenario,
- energy audit processes, to investigate energy consumption trends and habits, to try to reduce energy costs and carbon footprint.

When considering different domains, such as residential or commercial ones, other objectives may come into play, and they can be added to this dimension.

In general, dimensions can be considered in an arbitrary order, however it is worth considering not only the single variables, but the set of significant ones, as a whole, to get the relation among different dimensions and evaluate the impact of each choice.

4.2.2. Campaign type

Most of the times, when performing a monitoring campaign, a single objective is pursued (e.g., users' comfort assessment) and the implementation choices are carried out at the beginning of the planning, and – provided they are wisely selected – the setup is not modified until the end of the campaign (except for occasional tuning). However, when considering a long-term project for energy consumption improvement, the monitoring campaign actually may go through different *phases*. More precisely, if a refurbishment is planned, the campaign is initially performed to gather the building current thermal profile and energy consumption information to identify what kind of intervention can be carried out to improve energy efficiency. Once the renovation takes place, the same monitoring can be exploited to assess the conditions of the building ex-post, and to feed a H/BEMS for an autonomous and optimized building management. Although the instrumentation may be the same, different sampling strategies and data analyses may be appropriate during the various phases of the monitoring campaign, and the framework models this aspect as well. A similar situation occurs in [21], where the monitoring campaign is divided into two different phases, characterized by different sampling frequencies and time spans; data is initially collected to train the model, and later sampled in a different way to validate it.

4.2.3. Data temporal usage

This dimension models the timing of the use of data, and we envisioned two possible alternatives, namely *batch* and "real-time", *dynamic*. The former scenario occurs when data is used off-line, for analyses and reports that are periodically performed, for instance, to assess KPIs or to evaluate comfort. In this scenario it is possible to deliver data to the final repository (in the cloud, on a server, ...) periodically, provided the latency is compatible with the exploitation of the data itself. Typically, data is consumed in batch, with a frequency that depends on the final campaign objectives. As an example, consider the field work reported in [5], where users' feedback on the perceived comfort collected by means of a mobile app is exploited at the end of the season, to statistically assess the average perceived comfort within the university classrooms. Data is processed in a batch, off-line mode. The latter alternative models the scenario where the monitored data is immediately and continuously used for decision making purposes, e.g., by a H/BEMS, anomaly detection or data visualization. In these scenarios data should be made available within a short time range to respect the constraints set by the application, e.g., H/BEMS systems in order to take timely decisions need update data. Still referring to the perceived comfort information collected from the users of a conditioned space, in [22] the information is dynamically exploited in real-time, to feed the HVAC control strategy in order to adjust the room conditions based on such feedback.

4.2.4. Sampling strategy

As already mentioned in Section 2, it is possible to adopt different sampling strategies. In a *data-driven* approach, sensor nodes periodically report their readings, in an *event-driven* one, sensor nodes only transmit data when a "relevant" change in the monitored value is detected. For a sensor monitoring a state (such as open/closed for a window), the change is implicitly defined, whether for sensors monitoring a discrete phenomenon (such as temperature), a threshold value has to be defined by the user, to characterize the gap between two consecutive values (e.g., 1 Celsius degree) to trigger the transmission of the data. The former strategy provides a continuous stream of samples and generates data that can be straightforwardly exploited (e.g., by means of charts, ...), although it usually creates huge amounts of data. The latter strategy requires less energy costs for the sensor (especially important in WSNs scenarios, since data transmission is the most expensive task) and usually a reduced amount of data to be stored. However, the critical aspect is the identification of a significant threshold, a parameter that is affected by the sensor accuracy and the accuracy required by data exploitation processes. For instance, in the experimental setup reported in [5], the threshold is twice the value of the sensor' precision threshold. Since for some sensors (such as the open/closed, on/off sensors) there is no use in collecting periodic samples, it is possible to adopt a mix of the two strategies, by using the periodic sampling for continuous variables and the event-driven one for sensors monitoring changes in state.

For the sake of providing some numerical information, with reference to the adopted field work, the sampling is performed every 6 minutes. Within a month, 7200 samples are collected and the temperature varies between 19.67°C and 30.96°C (Apr. 2016). By adopting an event-trigger sampling strategy, having a threshold set to 1°C, the number of samples in the examples is 322, amounting to less than 5% with respect to the complete data stream. This refers to a single measurement and has to be applied to all collected information in an instrumented building.

Associated with the sampling strategy, the user has also to determine the specific *frequency*, the *threshold*, or both (in case of a mix of strategies), to be used, as parameters for this dimension values.

4.2.5. Analysis' time granularity

The choice of a reference time interval for the analysis (i.e., daily, weekly, monthly and seasonal) is of paramount importance. On different time spans it is possible to experimentally observe different correlations among variables, e.g., the correlation between internal and external temperatures is evident on a monthly basis, but is much less marked on a daily one. The granularity deeply impacts other dimensions as well, like outlier detection and missing value analysis. As an example, consider a web application visualizing the energy footprint of every employee. The visualization of daily figures requires a sampling frequency

as high as 1 sample per minute, and the application of processing procedures to identify and replace missing values and outliers, to deliver an accurate daily picture; when considering monthly statistics, samples could be collected once every 5 minutes, ignoring missing and outlier values, still giving users an accurate view.

4.2.6. Aggregation Operator

To summarize the huge amount of sensed data it is necessary to adopt aggregation strategies by means of aggregation operators. The aggregation could concern values sensed by redundant sensors, and in that case it should provide a single consistent value. It can be performed by using an operator such as the average, the median, the mode applied to the original sensed values. On the other hand, it could be possible to use one of the available values. In the setup reported in [5], the air temperature provided by the sensors co-located with the CO2 level is inaccurate with respect to the value sensed by the correspondent one co-located with relative humidity, therefore a single value is used. Aggregation may involve more complex computation if performed as the feature extraction step before machine learning and data analysis tasks, as proposed in [23]. Moreover, an aggregation could be applied also to distill groups of values in a single one, e.g., to summarize values of a long time period in a single significant one, by using the mentioned operators.

4.2.7. Segmentation Strategy

Data collected from a single sensor across time represents a time series, i.e. $T = t_1, ..., t_m$ is an ordered set of m values sampled by a sensor. Time series segmentation is the procedure necessary to split the series into a set of disjoint segments [24]. In general, the problem of segmenting time series data is very complex, and related work in the domain of building monitoring typically segment data on a daily basis [25]. Indeed, it is common to segment on a daily basis, discard non-occupied hours and non-working days, and then perform the analysis. To provide a more general framework, it is interesting to evaluate different segmentation intervals, e.g., weekly basis, as suggested by the characteristics related to a working week, to detect useful knowledge such as frequent patterns or relevant changes. Furthermore, an hourly segmentation is significant in high-resolution analysis of H/BEMSs setpoints or to highlight and isolate system anomalies. Indeed, the monitoring strategy should be design taking into account the target segmentation window in order to assure the availability of a sufficient amount of sampled data for the analysis. Real-time applications don't consider this dimension as the information is continuously consumed.

4.2.8. Campaign time span

The selection of the appropriate duration of the campaign is influenced by the two identified skew factors: users' behavior and weather conditions. Indeed, the time span should allow the identification of existing dependencies between the monitored variables if any, and should be, as much as possible, representative of the climatic zone, including the variety of seasonal conditions. For instance, if the goal of the monitoring campaign is the efficiency of a cooling system in wintertime, it could be appropriate to monitor the spaces for a two to three weeks in winter, to limit resources. On the other hand, when considering the general performance of a H/BEMS, the time span should span through different seasons. In general, it is not easy to identify adequate periods to perform short-term monitoring campaigns; however, from our monitoring experiences we have noticed that a two-week monitoring campaign is often sufficient to highlight data features and issues and to suggest general considerations on the problem in the considered season, e.g., temperature perceived by users, main climatic impacts as orientation, outside temperature ranges etc. Since it is necessary to specify, for the selected value (e.g., years, months and weeks) the desired granularity (e.g., 2 weeks or 6 months), the values of this dimension have a parameter to be specified.

4.2.9. Instrumented spaces

Based on the overall goal of the campaign, it can be necessary to select only a subset of the building spaces to be instrumented for the monitoring activity. In particular, when the objective is not the dynamic control of the building (which requires the *complete* instrumentation of all spaces), it is possible to monitor a representative set of spaces, a *selection*, for gathering significant data and profiles. With respect to this choice, it is important to make sure that spaces with different exposures (e.g., North and South), operational patterns, destinations of use and occupants' schedules are included. On the other hand, not relevant spaces should be excluded from the analysis, e.g., it is not significant in a wall thermal mass performance analysis to monitor spaces actively heated/cooled if the HVAC does not play a relevant role in the analysis objectives or accurate information regarding operational patterns is not clearly available/collected.

4.2.10. Redundancy

Sensors, thanks to their low cost, are usually placed in a redundant way to increase data availability needed to achieve a certain level of reliability. Moreover, several variables are often involved in the monitoring campaign thus generating a multi-dimensional stream of information. Adopting a redundant strategy introduces the need to aggregate redundant values in a single consolidated value, thus designing an IT infrastructure able to gather and summarize data, possibly on the fly, to make them available to stakeholders. It is worth noting that placing multiple sensors to monitor the same variable and to ensure data availability is not always an effective strategy, since it prevents data losses caused by single points failure but does not prevent data leaks caused by network issues or central infrastructure problems, thus a different

strategy should be designed. Moreover, the correlation between subsets of monitored data need to be then investigated to reduce the number of sensors involved in the monitoring. For instance, the relationship between temperatures measured at different heights in the same room is clearly linear; this behavior suggests that after a first period of monitoring, having understood the process dynamics, it is sufficient to monitor only the temperature at a certain height and then to infer the others (e.g., by using a linear regression method).

The data collected from Wireless Sensor Networks (WSNs) are often incomplete, in particular sensors can run out of batteries or network issues can prevent sensors from communicating with the IT infrastructure in charge of data storage. It is worth noting that the adopted sensor redundancy does not improve measurements availability when connectivity problems occur, since they typically involve the whole sensor network. In our specific test bed redundancy is ineffective and carries little benefit, while it introduces the need for an aggregation strategy to deal with redundant sensed data. For instance, in our campaign, which was run for a month, redundancy has been effective only for one day, when a sensor provided only 5% of the expected data thus actually benefiting from the second sensor's values (ref.: May 2016).

4.2.11. Outlier detection

Sensed values gathered through WSNs are often inaccurate with respect to two main types of outliers, erroneous values and inaccurate values, that might need to be identified and eventually treated with different strategies. The former are values that deviate significantly from the surrounding values (e.g., a negative value of relative humidity or a steep temperature variation) and are occasionally sensed when the sensor batteries are exhausting or due to the presence of external disturbance sources. These values need to be first identified by using an inspection method and then removed with an ad-hoc procedure. In this domain, it is vital to compare sensed values with summarized historical records, that can be used as a representative reference to identify erroneous values in an automatic way. It is important to highlight that the presence of outliers impacts markedly on analysis' results by introducing substantial distortions. Inaccurate values often occur in data sensed through WSNs: it may happen that values fluctuate among a set of previously measured values and after some unstable measures the value stabilizes. This behavior introduces a weak distortion because sensed values usually have a small deviation from real values and are easily cleaned through filtering methods or can be simply ignored.

Outlier analysis in time series data, e.g. sensor readings, focuses on the identification of unexpected behaviors across time. We tested various outlier detection methods including static and dynamic techniques [26], [27]. It should be stressed that static techniques should be adapted to the streaming case by segmenting in advance the time series, otherwise the removal of outliers becomes ineffective. We employed different filters, such as the moving median filter and Fast Fourier Transform filter, as well as static methods, among which the generalized extreme Studentized deviate and inter-quartile range filter on a daily basis. These two techniques suffer from the following drawback: they should be applied on normal distributed data, which generally does not apply to environmental data or energy usage data. However, incorrect values are properly identified but they do not detect inaccurate values. Moreover, it is necessary to segment the data and apply the procedure on a daily basis, thus introducing a computational overhead and leading, in general, to poor results if the subsequent analysis adopts a different segmentation interval. Conversely, applying a non-linear filter is effective for inaccurate as well as erroneous values, at the cost of evaluating the trade-off between different choices of parameters. On the collected data set, a simple moving median filter produces excellent results: the computational complexity is reduced and thus it can be managed even by sensors or gateways; indeed, they can filter values in real-time or with a distributed procedure.

To improve data accuracy we applied a moving median filter. As previously mentioned, a median filter is a filtering technique often used to remove noise: the main idea is to run through the data entry by entry, replacing each entry with the median of neighboring entries after setting an appropriate window.

It is worth mentioning that the window size should be chosen with reference to the sampling frequency and data segmentation, i.e., for the *real-time* scenario a smaller window might be adopted to improve computational efficiency.

The use of a moving median filter allows, in this particular domain, to achieve satisfactory results by using a computationally efficient method, which might be applicable in real-time and directly on sensors or on gateways, i.e., it will be applicable to every scenario and also to the real-time one.

Outliers arise when the sensed value changes and tends to take the previous or the subsequent detected value, for this reason the above detailed filtering strategy is very effective with high sampling frequencies.

These considerations are supported experimentally by the collected data. As it can be seen in Figure 3 showing sensed temperatures of a day segment, there is noise in the collected data since the temperature changes abruptly too often, therefore sensors sometimes provide erroneous readings. In particular, several spikes can be noticed. When applying the adopted filtering solution (Figure 4), some minor spikes are still present but with a limited impact. Thus, with reference to the stated goal we consider this method an effective solution, characterized by a low computational complexity.

4.2.12. Missing value imputation

In sensed data, missing values are of two different types: occasional and burst.

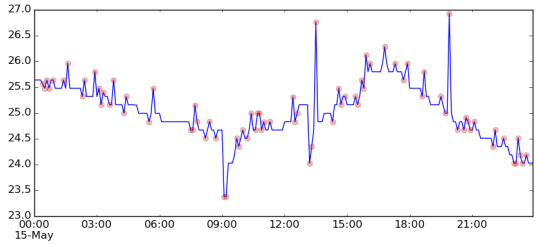

Figure 3: Single day segment: raw temperature values.

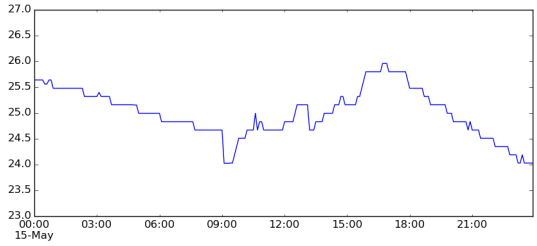

Figure 4: Single day segment: filtered temperature values.

Erroneous single values, e.g., values produced by the outlier detection process, can be suitably replaced, with one of the adjacent values considered as correct, or ignored. In the same way, random missing values are due to temporary interference in sensors or networks and can be replaced with a neighboring value (or the average of a group of values) or, alternatively, can be ignored. It should be noted that this type of missing values is irrelevant in many exploitation scenarios, and could be easily overlooked.

On the other hand, missing bursts of data are difficult to handle because they introduce a significant distortion in the subsequent analysis. In this case, two alternatives are feasible: data segments can be discarded or fitting data can be artificially introduced. The first approach can be adopted when the number of samples is below a certain availability threshold. For instance, in comfort evaluation, when a day has less than 200 samples out of the expected 240 (collected with a 6 minutes frequency) the segment is discarded, because it has been experimentally observed that missing data introduce a skewness. When it is critical to discard data, trying to keep at least partial data segments, it is necessary to identify the most appropriate method among the available ones, to infer the missing values. This can be done starting from similar historical series, or from the phenomenon model, if available. It should be emphasized that this task is computationally complex, but generally shows good results using methods as k-Nearest Neighbor regression, when considering highly incomplete segments.

5. Exploiting the framework to classify existing field works

To validate the proposed methodology and framework and to evaluate its usefulness from the point of view of the analysis and comparison of existing field works, we here use as a common descriptive framework to characterize and classify the subset of field works discussed in Section 3. The analysis has been carried for a larger number of existing solutions, however, due to space limitations, we here report only a subset of them, with the goal of evaluating different strategies adopted to pursue the same objective (some solutions that presented synthetic datasets or datasets with too few details have been omitted). More precisely, we analyzed each work with respect to the framework dimensions and values, in order to gather the relevant information of the monitoring campaign. Table 3 summarizes the results of the analysis; the columns of the table list the framework dimensions and for each field work a • appears if it was possible to gather the corresponding information from the description. The works are clustered by objective and the table does not report the specific values for the dimensions for clarity – except for the objective, however the description of the various works in the remainder of the section presents such information.

In [11], [28] and [29] the authors present a comfort analysis under different perspectives. More precisely, [11] investigates the relationship between building characteristics and its performance by applying data mining methods; the specific target is to predict building performance indicators related to thermal comfort and indoor daylight with reference to climate conditions and building features. The monitoring involves ten rooms with different intended uses, including seminar rooms, offices and laboratories. The analysis adopts different sampling frequencies on the basis of sensors technology, i.e., 15 minutes sampling for wired sensors and 1 minutes sensing for wireless sensors. Moreover, data is collected only during working hours, i.e., from 7:30 am to 10:00 pm on weekdays, and from 9:00 am to 7:00 pm on weekends. Missing values are filled by interpolation and a processing step to remove noise and outliers is performed. A feature extraction procedure is performed in order to compute different indicators and use them within classification models. The monitoring has been carried out for 41 months.

In [28] a thermal preference assessment method using adaptive thermal comfort is presented. The main aim of the research is to compare and predict the thermal performance of different buildings, taking into consideration building materials, orientation, shading, occupant behavior and weather and environment surrounding the building. Each module shows different building features and has been monitored for 12 months with over 100 sensors with a data driven approach using a granularity of 5 minutes. Sensed data is segmented on a monthly basis and an annual analysis is performed.

The authors in [29] evaluate indoor air quality and

	Framework dimensions											
Field work	Objective	Campaign type	Data temporal usage	Sampling strategies	Analysis time granularity	Aggregation operator	Segmentation strategy	Campaign time span	Instrumented spaces	Redundancy	Outlier detection	Missing values imputation
[11]	comfort analysis	•	•	•		•		•	•		•	•
[28]	comfort analysis	•	•	•	•		•	•				
[29]	comfort analysis	•	•	•				•				
[30]	data visualization	•	•	•	•		•	•	•			
[31]	data visualization	•	•	•								
[10]	data visualization	•	•	•								
[9]	energy audit	•	•	•	•	•	•	•	•			
[32]	energy audit	•	•	•	•	•	•	•			•	
[33]	energy audit	•	•	•	•	•			•			•
[34]	occupants' behavior	•	•	•		•	•	•	•			•
[35]	occupants' behavior	•	•	•			•		•			
[36]	occupants' behavior	•	•	•	•		•		•			•
[37]	data mining	•	•	•		•	•	•	•		•	
[38]	data mining	•	•	•		•	•	•	•			

Table 3: Use of the proposed framework for the characterization and classification of field works.

occupants' thermal comfort in hot humid environments, assessing thermal and air movement acceptability levels through environmental monitoring and users' opinion collected by means of extensive surveys. The paper adopts a measurement protocol oriented to users based on a survey filled every 30 minutes. A 5 minutes interval is adopted for the micro-climatic station installed within each building, excepts for the air velocity which, due to its changeable nature, is recorded when the survey is filled. In this use case, the data sampling strategy for environmental sensors is data driven, while it is event driven for hot-wire anemometer due to air velocity tendency to vary rapidly through time and space. Each time a user completes a survey, 30 instantaneous air speed samples are collected and an average value is then computed. This work can be considered an example of a complex monitoring strategy, with multiple issues that need to be tackled and where multiple parameters can be adopted for the same dimension. Overall, it can be observed that this objective requires a single phase campaign and is typically exploited in batch mode. The sampling strategy is data driven with a sampling frequency of 5 minutes in [11] and [28], whereas it is based on sensor technology in [29]. The second and the third work provide an analysis with annual and seasonal granularities, the campaigns span for 4 years in the first case and for 2 months in summer and 2 months in winter for the third case. [28] runs a campaign for 1 year considering the complete monitoring of a single building. The authors of [11] mention an outlier detection and missing value replacement procedure, however they do not provide details.

The authors in [30], [31] and [10] focus on data visualization by performing a real-time usage of the data.

In [30] a visualization tool is presented: it provides tenants with a meaningful and timely feedback of energy usage in order to engage them in the energy efficiency process of a university campus. A mobile app provides real-time visualization of data collected every 30 minutes through meters; the analysis is conducted on data segmented on a weekly basis. The subsequent analysis extracts the consumption model for each monitored building on an annual basis.

In [31] the authors propose a framework for energy monitoring and management in factories, focusing on energy used by each productive asset. The framework incorporates standards for online visualization and performance analysis. Moreover, KPIs are extracted in real-time allowing more timely decisions. The framework performs an event driven analysis by detecting significant events within data streams. The adopted sampling rate is between 2 ms and 1 minute. The peculiarity of this work is the application to the industrial scenario and the online exploitation of sensed data by using a complex event processing analysis.

The work proposed in [10] presents an integrated performance monitoring system useful to promote energy aware-

ness by visualizing electricity data and proving information to enable occupants to adopt preventive or corrective actions thus optimizing energy usage. The monitoring campaign is organized into two different phases with different exploitation strategies. During the first step data is collected every hour to train prediction models, whereas during the second one data is collected to validate the models, each step lasts one year and involves six buildings devoted to research and educational purposes. The analysis considers a segmentation on a daily basis and provides prediction on a hourly basis.

The adopted sampling strategy for these works is data driven, respectively every 30 minutes, every 1 minute and every 2 milliseconds. Furthermore, in [30] the analysis granularity is weekly, the campaign involves every building within the university campus for twelve months.

In [9], [32] and [33] the authors focus on the problem of energy audit.

[9] presents an internet-based energy monitoring system for large public buildings, comprising supermarket, hospital, university campus, offices buildings and hotels. The platform releases in real-time energy consumption information by means of charts or tables; data is sensed every 5 minutes and hourly, daily, monthly and yearly aggregations are provided. Further processing steps are not detailed, nor mentioned.

The authors in [32] propose a method for modeling and predicting daily energy consumption to support building management systems in energy usage forecast and anomaly detection. Building data is sampled every 30 minutes for one year, a processing step involves feature extraction from the sensed time series followed by an outlier detection procedure. The analysis comprises different segmentation strategies such as daily, weekly and monthly, adopting different analysis granularities.

The work presented in [33] details a method to predict building energy demand based on the decision of the tree approach. The work monitors energy consumption in 80 residential building in six Japanese districts; data are sampled every 5 minutes (energy data) and 15 minutes (environmental parameters). Datasets containing missing values have been discarded and a pre-processing step computes aggregated values for hourly, daily, monthly and annual consumption before the analysis.

In [9], [32] and [33] energy consumption is evaluated, in either real-time and batch mode. All the approaches consider a data driven solution with a sampling frequency of 5 minutes, 30 minutes and 5 or 15 minutes, respectively. The analysis is carried out by considering hourly, daily, monthly and annually aggregation in [9] and in [33], daily, weekly and monthly aggregation in [32]. In [9] the monitoring involves more than fifty buildings for a two-years period. In [32], data is segmented on a daily basis, then aggregated to extract mean, peak values and auto-regressive parameters; after feature extraction an outlier detection procedure is applied using GESD and Q-test rules. In [33] data are aggregated over 5 minutes and missing values handling is mentioned without further details.

Works outlined in [34], [35] and [36] deal with the problem of occupants' behavior by performing the analysis in a batch fashion. More precisely, the authors in [34] focus on energy waste pattern discovery, by analyzing rooms occupancy and energy consumption due to lighting, in a secondary educational building. The monitoring has been carried out for a full academic year in three classrooms, different in size and schedule. Data on power consumption and occupancy are sampled every minute, then aggregated every 15 minutes to allow feature extraction, which comprises average, median, mode and boolean labels computation, e.g., on/off for lighting usage. Missing values are replaced with the last recorded value. Furthermore, data is segmented on the basis of different daily windows (morning, evening and night). The analysis is performed on an annual basis.

In [35] the relation between occupants' behavior and energy consumption is investigated; the study employs statistical analysis with data-mining techniques to identify operational patterns in measured data. The monitoring spans across two years and involves sixteen offices within a naturally ventilated and cooled building. The monitoring adopts a mixed sampling strategy; energy and environmental parameters are sampled every 15 minutes in a data driven mode, whereas users' behavior is tracked in an event driven mode. The analysis is performed at various granularities including hourly, daily and seasonly. The paper does not report details regarding data pre-processing excepts for a normalization task.

[36] develops an analysis approach using office appliances power consumption data to learn the occupants' behavior. The method is tested in a medium office building by recording every 5 minutes the energy consumption caused by each occupant, then metered data are used to derive occupants' behavior models. The analysis is carried out on a daily basis, after segmentation and discarding incomplete segments.

All these works perform a data driven sampling using 15 minutes frequencies in [34] and in [35] and 5 minutes frequency in [36], respectively. No details are provided regarding outlier detection and missing values handling.

Data mining is the focus of the monitoring campaign performed in [37] and [38]. The work presented in [37] applies mining methods to BAS data acquired within a commercial building, the data have been collected for 8 months every 15 minutes; the purpose of the mining is to discover useful hidden knowledge to enhance building operational performance. The work mentions outlier detection using inter-quartile range filtering and extract features based on different daytime ranges, after segmentation on a daily basis. In [38] data mining techniques are applied for detecting abnormal lighting energy consumption. The case studies analyze an office building for two months, where environmental parameters are sampled every one hour while energy consumption is aggregated to compute an hourly average and a daily peak. In both

cases, the monitoring involves a single complete building and the analysis is performed in batch mode. In the former data are segmented in day, evening and night ranges and a feature extraction (minimum, maximum, standard deviation and average) procedure is applied. Moreover, pre-processing to filter outlier has been performed through inter-quartile rule, and missing values are replaced using a moving average filter with a window of size 5. The latter work mentions the removal of incomplete series.

From our experience, most approaches tend to focus on the most immediate dimensions (e.g., sampling strategy), neglecting the discussion of other relevant aspects (e.g., missing values handling, redundancy) such that it is rather difficult to deploy a similar solution or to find answers to common problems. In general, the attention is on the peculiarities and innovative aspects of the field work, omitting elements of the monitoring campaign that actually have an impact on the overall cost and quality of the adopted solution. The proposed framework thus aims at guiding the users in reporting all relevant aspects in a systematic, shared way.

6. Exploiting the framework to design a monitoring campaign

In this section we discuss how the framework has been exploited to design and tune the monitoring campaign presented in [5]. More precisely, we here analyse the effectiveness of parameters chosen during the campaign setup with reference to the impact on results, with respect to three dimensions: sensors *redundancy*, *sampling strategy* and *outlier detection*. The initial choices were driven by domain experts' best practice and experience, and no detailed information on the best practises in similar cases was systematically available from field works.

The campaign focused on the users' comfort analysis in one of the university premises: it has targeted a five-stories building, occupied by faculty and administrative offices (140 rooms), dated 1961. Four contiguous offices located on the first floor have been instrumented: they are similar in size (3.50 x 6.27 x 3.30 meters), two of them facing South, two facing North. Each office is equipped with the same set of sensors, monitoring external and internal conditions, among which it is worth mentioning three co-located air temperatures and relative humidity points. The *objective* dimension is the user' comfort analysis, specifically a fine-grained comfort analysis on a daily basis is performed. In line with experts' best practice, a *single-phase* campaign has been planned, by adopting a *data-driven* sampling strategy (a sample every 6 minutes). Sensors are placed using a *redundancy* policy, i.e., two temperature sensors situated at 1 m from the floor in each room. The *campaign time span* is set to one year and involved four offices (i.e., thus corresponding to a *selection* of *instrumented spaces*).

To infer perceived comfort we refer to the current national regulation ISO 7730/2005 [39], briefly summarized in Table 4; it should be noted that a different norm, based on discrete temperature ranges (e.g. ASHRAE 55-2007 [40]), could have been applied. The daily percentages have been computed by adding, for sake of representation, a fourth class which contains samples outside the intervals defined by the normative. The operative temperature is approximated by the air temperature, as common practice recommends for an approximate thermal comfort analysis. To determine the level of comfort (i.e., class) based on the room air temperature, each class is derived by computing the percentage of samples included in each temperature range within the referenced time period.

Comfort class	Summer	Winter
A	24.5 ± 1.0 °C	22.0 ± 1.0 °C
B	24.5 ± 1.5 °C	22.0 ± 2.0 °C
C	24.5 ± 2.5 °C	22.0 ± 3.0 °C

Table 4: Comfort classes from ISO 7730/2005.

Sensors redundancy is adopted in WSNs to improve data accuracy as collected data is often incomplete (sensors can run out of batteries or network issues can prevent sensors from communicating with the IT infrastructure in charge of data storage). It is worth noting that the adopted sensor redundancy does not improve data availability when connectivity problems occur. For instance, in our campaign, redundancy has been effective only for one day in a month, when a sensor provided only 5% of the expected data thus actually benefiting from the second sensor's values (ref.: May 2016). It can be conclude that in our specific test bed with reference to the outlined objective, *redundancy* is ineffective and offers little benefit, while it introduces the need for an aggregation strategy to deal with redundant sensed data.

Collected data is pre-processed by: i) aggregating and averaging multiple values, ii) discarding days with a limited amount of samples – we set a threshold to 200 samples out of the total 240 –, and iii) ignoring occasional missing data.

A second analysis has been carried out to determine whether the initially sampling frequency and the need to remove inaccurate values were appropriate with respect to the campaign objective. To this end, raw sensed data has been resampled using different strategies, also applying a filter to clean data:

- raw data collected every 6 minutes, without pre-processing, is labeled `raw`
- raw data collected every 6 minutes, filtered using a median filter with a window of size 5, is labeled `flt`
- raw data resampled every 18 minutes, filtered using a median filter with a window of size 5, is labeled `s18flt`

- raw data resampled every 60 minutes (i.e. same sampling detailed above), filtered using a median filter with a window of size 3, is labeled `s60flt`
- raw data resampled every 180 minutes, without preprocessing, is labeled `s180`

Tables 5 and 6 report, as examples, data collected in one office for a month in summertime and in wintertime, respectively. Here resampled data are compared to the raw data and among themselves with respect to the results of the classification (e.g., classes of comforts) and the incurred errors, and the amount of collected data. As it can be seen, the use of a sampling interval equal to 18 minutes allows us to obtain excellent results form the comfort analysis results, however the reduction in the number of samples is not relevant. More interesting results correspond to sampling every 60 minutes, since it produces accurate results while significantly reducing the amount of collected data (1/10 the initial number of samples). Moreover, it can be noted that the resampling is representative for the problem as sampled values shift inside the three provided classes, so generally the comfort level derived when reducing sampling time is consistent with the original one. With reference to the *outlier detection* dimension, the applied filtering strategy is considered an effective solution, characterized by a low computational complexity, as it allows an improvement in results accuracy.

To summarize, the values adopted for three analysed framework dimensions have been re-tuned, considering the experimental analysis here outlined, and the following values have been finally chosen: a single-phase *campaign type* with no *redundancy* policy; the *campaign spans* across one year and involves a subset of *instrumented spaces*. Given the technological setup, a data driven *sampling strategy* has been adopted. Moreover, for a fine-grained comfort level analysis, the *sampling frequency* has been set to 60 minutes, the *analysis' time granularity* to a month, and a daily *segmentation* solution is used. *Outlier detection* is performed using a median filter, incomplete segments are discarded and single *missing values* are neglected.

7. Lessons learned and conclusions

The availability of a broad body of work consisting of field studies that have their own peculiarities and often omit aspects that are considered of limited relevance, makes it particularly difficult to setup an efficient monitoring campaign, unless referring to domain experts' experience and best practice. Yet, even when general guidelines are available, such directives might be not optimized with respect to the specific task at hand. As an example, it is common practise to sample indoor conditions every 6 minutes, however such a frequency is not high enough when envelope performance models need to be extracted for the ambient (a sample per minute is appropriate in the investigated spaces), and it is too high when the monitoring is performed to estimate the users' comfort (a sample every hour is appropriate). The optimal frequency actually depends from the several other aspects of the monitoring campaign and a systematic way to report the adopted choices and their impact would better provide support to other researchers dealing with similar issues. In fact, domain experts' knowledge is fundamental, and it is paramount to be able to model it in order to actually exploit it.

In this perspective, the use of the proposed framework allows for an as much as possible complete modeling of the relevant aspects of a monitoring campaign, so that it can lead the design phase, as well as the characterization/classification phase, providing a common methodology to reference field work and adopted solutions.

In this paper we have analyzed the issue of ambient monitoring, aiming at defining some methodological guidelines for collecting and managing information. The outcome is a methodological descriptive framework allowing one to systematic design and analyze monitoring campaigns and collected data, towards a systematic, unified approach in modeling, characterizing and evaluating possible solutions. We have applied the proposed framework to existing field works presented in literature, showing the comprehensiveness and flexibility of the modeling approach, as well as its usefulness towards a more systematic characterization of the monitoring and data analysis approaches. Moreover, we also report the application of the framework in the design phase of a campaign, to identify and define the relevant aspects with respect to the campaign main objective. The framework integrates domain experts' best practises and experiences, putting them into relation with the actual exploitation of the collected data, to optimize efforts and costs with respect to the campaign ultimate goal.

Acknowledgments

This work has been partially funded by Regione Lombardia under grant no. 40545387 "SCUOLA - Smart Campus as Urban Open LAbs" and "EEB - Edifici a Zero Consumo Energetico in Distretti Urbani Intelligenti" (Italian Technology Cluster For Smart Communities) - CTN01_00034_5940 The authors thank prof. Giuliana Iannaccone and Marta Sesana for the fruitful discussions on data sensing solutions.

[1] ASHRAE. Performance_Measurement_Protocols_for_Commercial Buildings. ASHRAE, 2010.

[2] Tuan Anh Nguyen and Marco Aiello. Energy intelligent buildings based on user activity: A survey. Energy_and_buildings, 56:244–257, 2013.

[3] Olivia Guerra-Santin and Christopher Aidan Tweed. In-use monitoring of buildings: An overview of data collection methods. Energy_and_Buildings, 93:189–207, 2015.

[4] Zhun Jerry Yu, Fariborz Haghighat, and Benjamin CM Fung. Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities_and_Society, 2015.

Methods	Number of samples	Average error [%]				Maximum error				Averaged
		Class A	**Class B**	**Class C**	**Other**	**Class A**	**Class B**	**Class C**	**Other**	**Cumulative**
raw vs flt	6,877 / 6,853	1.35	0.98	0.93	1.12	3.75	4.58	4.16	2.92	4.38
s18flt vs flt	2,297 / 6,877	0.94	1.28	1.61	1.21	4.36	5	7.92	7.92	5.04
s60flt vs flt	688 / 6,877	1.98	2.35	3.57	2.45	7.04	11.66	12.09	8.34	10.35
s60flt vs flt	687 / 6,877	2.42	3.29	3.56	2.04	7.04	11.66	13.75	8.34	11.31
s180 vs flt	230 / 6,877	4.3	5.39	5.16	4.47	13.75	15.83	22.08	16.67	19.32

Table 5: Comfort analysis May 2015.

Methods	Number of samples	Average error [%]				Maximum error				Averaged
		Class A	**Class B**	**Class C**	**Other**	**Class A**	**Class B**	**Class C**	**Other**	**Cumulative**
raw vs flt	4,967 / 5,142	1.61	1.55	0.93	0.78	4.56	5.12	3.49	3.2	4.87
s18flt vs flt	1,724 / 5,142	1.48	1.24	1.69	1.24	3.84	3.37	5.45	4.15	5.65
s60flt vs flt	506 / 5,142	3.8	5.57	3.45	4.8	20.43	18.38	11.09	15.21	17.62
s60 vs flt	488 / 5,142	3.28	4.73	2.43	3.92	15.66	12.45	10.51	15.71	14.36
s180 vs flt	164 / 5,142	8.62	10.79	6.7	5.43	20.06	25.96	22.37	15.9	31.54

Table 6: Comfort analysis November 2015.

[5] Antimo Barbato, Cristiana Bolchini, Angela Geronazzo, Elisa Quintarelli, Andrei Palamarciuc, Alessandro Pitì, Cristina Rottondi, and Giacomo Verticale. Energy optimization and management of demand response interactions in a smart campus. Energies, 6(9):398 (1–20), 2016. ISSN 1996-1073. doi: 10.3390/en9060398.

[6] Deokwoo Jung, Varun Badrinath Krishna, Ngo Quang Minh Khiem, Hoang Hai Nguyen, and David KY Yau. Energy-track: Sensor-driven energy use analysis system. In Proc. ACM Workshop on Embedded Systems For Energy-Efficient Buildings, pages 1–8, 2013.

[7] Neil Klingensmith, Dale Willis, and Suman Banerjee. A distributed energy monitoring and analytics platform and its use cases. In Proc. ACM Workshop Embedded Systems For Energy-Efficient Buildings, pages 1–8, 2013.

[8] Aftab Khan, James Nicholson, Sebastian Mellor, Daniel Jackson, Karim Ladha, Cassim Ladha, Jon Hand, Joseph Clarke, Patrick Olivier, and Thomas Plötz. Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In BuildSys@ SenSys, pages 90–99, 2014.

[9] Liang Zhao, Ji-li Zhang, and Ruo-bing Liang. Development of an energy monitoring system for large public buildings. Energy and Buildings, 66:41–48, 2013.

[10] B. Sučić, A. Anđelković, and Ž. Tomšić. The concept of an integrated performance monitoring system for promotion of energy awareness in buildings. Energy and Buildings, 98:82–91, 2015.

[11] Ammar Ahmed, Nicholas E Korres, Joern Ploennigs, Haithum Elhadi, and Karsten Menzel. Mining building performance data for energy-efficient operation. Advanced Engineering Informatics, 25(2):341–354, 2011.

[12] Natasha Balac, Tamara Sipes, Nicole Wolter, Kenneth Nunes, Bob Sinkovits, and Homa Karimabadi. Large scale predictive analytics for real-time energy management. In Proc. IEEE Int. Conf. Big Data, pages 657–664, 2013.

[13] Cheng Fan, Fu Xiao, Henrik Madsen, and Dan Wang. Tem-poral knowledge discovery in big bas data for building energy management. Energy and Buildings, 109:75–89, 2015.

[14] Zhun Jerry Yu, Fariborz Haghighat, and Benjamin CM Fung. Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities and Society, 25:33–38, 2016.

[15] Merthan Koc, Burcu Akinci, and Mario Bergés. Comparison of linear correlation and a statistical dependency measure for inferring spatial relation of temperature sensors in buildings. In Proc. ACM Conf. Embedded Systems for Energy-Efficient Buildings, pages 152–155, 2014.

[16] Bharathan Balaji, Chetan Verma, Balakrishnan Narayanaswamy, and Yuvraj Agarwal. Zodiac: Organiz-ing large deployment of sensors to create reusable applications for buildings. In Proc. ACM Int. Conf. Embedded Systems for Energy-Efficient Built Environments, pages 13–22, 2015.

[17] Bradford Campbell and Prabal Dutta. An energy-harvesting sensor architecture and toolkit for building monitoring and event detection. In Proc. ACM Conf. Embedded Systems for Energy-Efficient Buildings, pages 100–109, 2014.

[18] Yuvraj Agarwal, Rajesh Gupta, Daisuke Komaki, and Thomas Weng. Buildingdepot: an extensible and distributed architecture for building data storage, access and sharing. In Proc. ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, pages 64–71, 2012.

[19] Stephen Dawson-Haggerty, Xiaofan Jiang, Gilman Tolle, Jorge Ortiz, and David Culler. smap: a simple measurement and actuation profile for physical information. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, pages 197–210. ACM, 2010.

[20] Bharathan Balaji, Arka Bhattacharya, Gabriel Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, Jason Koh, Joern Ploennigs, Yuvraj Agarwal, et al. Brick: Towards a unified metadata schema for buildings. In Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, BuildSys '16. ACM, 2016.

[21] Boris Sučić, Aleksandar S Anđelković, and Željko Tomšić. The concept of an integrated performance monitoring system for promotion of energy awareness in buildings. Energy and Buildings, 98:82–91, 2015.

[22] Soazig Kaam, Paul Raftery, Hwakong Chen, and Gwelen Paliaga. Time-averaged ventilation for optimized control of variable-air-volume systems. Energy and Buildings, 2016.

[23] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of pre-diction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy, 127:1–10, 2014.

[24] Tak-chung Fu. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1):164–181, 2011.

[25] Jarke J Van Wijk and Edward R Van Selow. Cluster and cal-endar based visualization of time series data. In Proc. IEEE

Symp. Information Visualization, pages 4–9, 1999.
[26] Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han. Outlier detection for temporal data. Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1):1–129, 2014.
[27] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials, 12(2):159–170, 2010.
[28] Aiman Albatayneh, Dariusz Alterman, Adrian Page, and Behdad Moghtaderi. Assessment of the thermal performance of complete buildings using adaptive thermal comfort. Procedia-Social and Behavioral Sciences, 216:655–661, 2016.
[29] Christhina Cândido, Richard de Dear, and Roberto Lamberts. Combined thermal acceptability and air movement assessments in a hot humid climate. Building and Environment, 46(2):379–385, 2011.
[30] Richard Bull, Dave Everitt, Graeme Stuart, and Martin Rieser. The gorilla in the library: lessons in using ict to engage building users in energy reduction. 2012.
[31] Konstantin Vikhorev, Richard Greenough, and Neil Brown. An advanced energy management framework to promote energy awareness. Journal of Cleaner Production, 43:103–112, 2013.
[32] Xiaoli Li, Chris P Bowers, and Thorsten Schnier. Classification of energy consumption in buildings with outlier detection. IEEE Transactions on Industrial Electronics, 57(11):3639–3644, 2010.
[33] Zhun Yu, Fariborz Haghighat, Benjamin CM Fung, and Hiroshi Yoshino. A decision tree method for building energy demand modeling. Energy and Buildings, 42(10):1637–1646, 2010.
[34] David F Motta Cabrera and Hamidreza Zareipour. Data association mining for identifying lighting energy waste patterns in educational institutes. Energy and Buildings, 62:210–216, 2013.
[35] Simona D'Oca and Tianzhen Hong. A data-mining approach to discover patterns of window opening and closing behavior in offices. Building and Environment, 82:726–739, 2014.
[36] Jie Zhao, Bertrand Lasternas, Khee Poh Lam, Ray Yun, and Vivian Loftness. Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. Energy and Buildings, 82:341–355, 2014.
[37] Fu Xiao and Cheng Fan. Data mining in building automation system for improving building operational performance. Energy and buildings, 75:109–118, 2014.
[38] Imran Khan, Alfonso Capozzoli, Stefano Paolo Corgnati, and Tania Cerquitelli. Fault detection analysis of building energy consumption using data mining techniques. Energy Procedia, 42:557–566, 2013.
[39] CEN. Analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort. Standard ISO EN 7730, Brussels, 2005.
[40] ASHRAE. Thermal environment conditions for human occupancy. Ashrae standard 55-2007, Atlanta, 2007.