Data analytics for smart buildings: a classification method for anomaly detection for measured data

Enguerrand de Rautlin de La Roy, Thomas Recht, Akka Zemmari, Pierre Bourreau, Laurent Mora

Data analytics for smart buildings: a classification method for anomaly detection for measured data

Data analytics for smart buildings: a classification method for anomaly detection for measured data

Enguerrand de Rautlin de la Roy[1], Thomas Recht[1], Akka Zemmari[2], Pierre Bourreau[3], Laurent Mora[1]

[1] Univ. Bordeaux, CNRS, Arts et Metiers Institute of Technology, Bordeaux INP, INRAE, I2M Bordeaux, F-33400 Talence, France

[2] Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR 5800, F-33400 Talence, France

[3] Nobatek/INEF4, 9 rue Jean-Paul Alaux, 33000 Bordeaux, France

E-mail : enguerrand.de-rautlin-de-la-roy@u-bordeaux.fr

Abstract. Data generated by the increasingly frequent use of sensors in housing provide the opportunity to monitor, manage and optimize the energy consumption of a building and the user comfort. These data are often strewn with rare or anomalous events, considered as anomalies (or outliers), that must be detected and ultimately corrected in order to improve the data quality. However, many approaches are used or might be used (for the most recent ones) to achieve this purpose. This paper proposes a classification methodology of anomaly detection techniques applied to building measurements. This classification methodology uses a well-suited anomaly typology and measurement typology in order to provide, in the future, a classification of the most adapted anomaly detection techniques for different types of building measurements, anomalies and needs.

1. Introduction

The most common use of Building Management System (BMS) and connected devices in the building sector has led to a democratization of edifices called smart buildings. A smart building can be seen as an association of multiple systems, software and sensors that aims to meet two main objectives: reducing both operational costs and environmental impact by managing and optimizing the energy use (1) and improve the comfort of the occupants (2). Smart buildings generate a massive amount of data from measurements (temperature, CO_2 concentration, occupation, *etc.*) that has to be analyzed in order to extract useful information for the user, the manager or the system itself. However, due to errors that may occur during the measurement phase (data transmission failure, accident, *etc.*) this collected data is rarely clean and straight usable. To address this problem, it is becoming essential to identify rare or anomalous events in datasets composed of varied data, to ultimately be able to correct them and to have data of sufficient quality to reach the objectives set.

Nowadays, anomaly detection is used in various domains for multiple purposes [1,2] such as financial with the fraud detection or in the security field to detect intrusion and even in the medical field to detect breast cancer. This approach has evolved from a simple use of expert-defined rules to advanced procedures that include statistical analyses and advanced machine learning techniques

[1,2,3]. However, even though several papers discuss about anomaly detection for different areas, those tackling the issue of building measurement flawed data are less common and often discuss about one method or family of anomaly detection for a specific measurement or problem [4,5,6].

Thus, this paper will attend to provide an overview and a global classification[1] of the most adapted anomaly detection techniques for different types of building measurements and anomalies. The goal pursed is to present and classify most efficient methods for every type of anomaly and measurement encountered and to highlight the added value (if any exists) of the methods that are currently, or might be, used.

The presented work is structured as follows: first, the construction of a classification method for anomaly detection techniques adapted to building measurements is described. Results and future work will be shown afterwards.

2. Methodology

2.1. Approach and organization

Although a lot of well-known and inspiring surveys on anomaly detection have been published over the past couple of years, this paper attend to focus solely on time series data from building measurement.

The following Figure 1, inspired and adapted from Raghavendra and Chandola's multi-disciplinary surveys [1,2], highlights the approach followed in order to build the classification and the selection methodology. First, an adapted typology of anomaly (1) and measurement (2) is built according to the characteristics of the data and anomalies studied, both directly depending on the field of application. These typologies are defined to include as many cases as possible while allowing new ones to be added without having to change their existing structure. Thus, the combination of these two specific typologies will allow us to provide an exhaustive classification method of anomaly detection techniques for different types of building measurements and anomalies (3). Those three parts will be discussed in the following section according to the structure described in Figure 1. The training method (supervised, semi-supervised or unsupervised) is a key feature of anomaly detection techniques as relevant as anomaly and data characteristics. However, this feature relies on the availability of labelled data to train a specific technique and does not intervene in the classification but later, while selecting and anomaly detection technique and thus, will not be discussed in this paper.

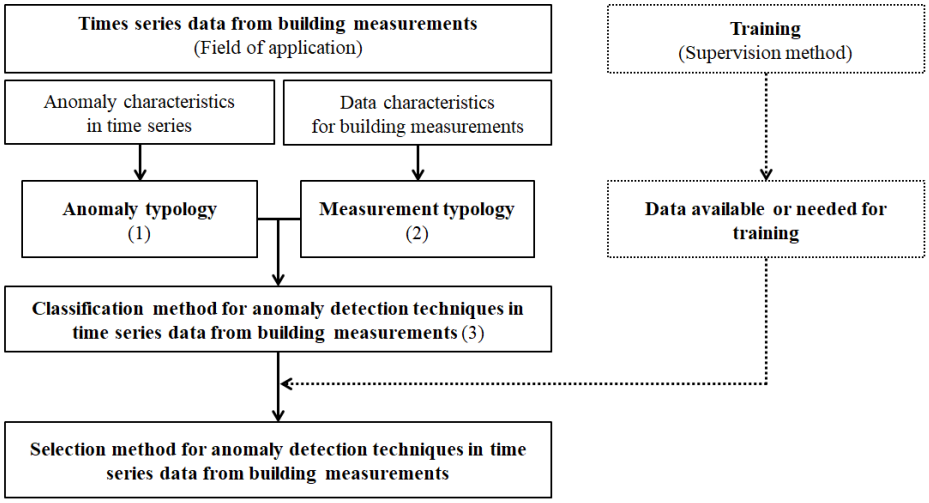

Figure 1: Key components of the classification and selection method for anomaly detection techniques in time series data

[1] Following the general sense of an arrangement in groups or categories according to specifics criteria

2.2. Anomaly typology for measured data

Literature usually classifies anomalies into three types: point, contextual and collective anomalies [1,2,3,7]. This classification has the advantage of being a multi-disciplinary approach but lacks some nuance when applied to times series alone. To address this issue, a typology of anomalies inspired by the work of [1,7,8] has been adapted and built to building measurements and usual anomaly encountered (see Table 1). This has resulted into a twelve elements classification depending on the type of the anomaly and its nature.

Table 1: Proposed anomaly typology for time series from building measurement

Nature			Type: Local	Type: Sequential	Type: Global
Independent	Value		**Single value anomaly**	**Successive value anomaly**	**Global value anomaly**
			Its own existence is sufficient to be detected		
	Timestamp		**Single timestamp anomaly**	**Successive timestamp anomaly**	**Global timestamp anomaly**
			Its own existence is sufficient to be detected		
Related	Univariate		**Single univariate anomaly**	**Successive univariate anomaly**	**Global univariate anomaly**
			One or more information from the same time series is required to detect it		
	Multivariate		**Single multivariate anomaly**	**Successive multivariate anomaly**	**Global multivariate anomaly**
			One or more information from other time series (one or more) is required to detect it		

The type of an anomaly refers to the range of the anomaly in the data: local if one element is involved, sequential for consecutives ones and global if all the series is affected.

On the other hand, the nature refers to the kind of attribute or combination of attributes (such as value, timestamp or both) for one or combined time series that enables to identify an anomaly.

For a specific time series, an anomaly is considered as independent (on a value or on a timestamp) when there is no need to collect any other information in order to be detected. For instance, an impossible high value, such as a relative humidity value of 120% presented in the Figure 2 (i), is considered as a single value anomaly. On the other hand, a lack of data during a time range as presented in Figure 2 (ii) represents a successive timestamp anomaly. For the first example, the anomaly occurs on the value axis while, in the second case, it appears on the time axis.

However, when one or more information is required to detect an anomaly within a time series, its nature is said to be related. If some information (or context) is provided by the same time series, the anomaly is considered as univariate. For instance, points highlighted in the Figure 2 (iii) have values which are found elsewhere in the time series but do not match with their close surroundings. In this case, time and value information must be both studied in order to detect the anomaly.

In a similar way, if information are provided from other time series (one or more), the anomaly is called multivariate. A successive multivariate anomaly is illustrated in Figure 2 (iv) by comparing two related times series: CO_2 concentration and occupation measurements for the same room.

This typology gives a brief overview of the difficulties, the required resources and the detection techniques that might be used or needed for detecting each anomaly presented. However, anomalies are not the only feature that needs to be considered while selecting anomaly detection techniques.

Depending of the nature of the data studied some detection techniques might not be optimal to use, and thus some information specific to the data studied must be added to complete this feature.

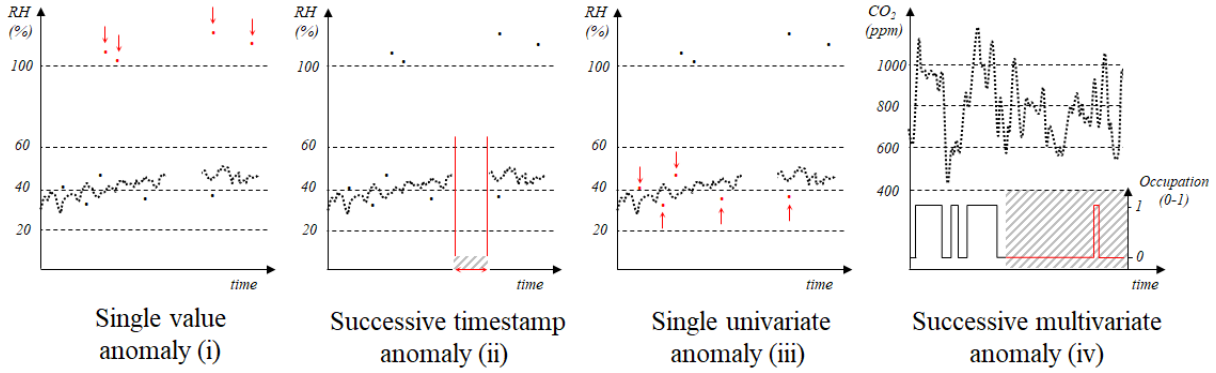

Figure 2: Examples of anomalies in time series

2.3. Measurement typology

As stated by Chandola [1,9], the nature of the data is a key aspect of anomaly detection techniques. This nature can be defined by several attributes (or variables) that are part of the data. By taking into account their number, their type (categorical, continuous, binary …) and connection to each other, the classification of anomaly detection technique can be refined.

In order to build a measurement typology and to prepare further tests, data reflecting the living and operating conditions of different types of buildings (residential or tertiary) have been collected and studied from several sources such as, BMS (Building Management System), connected devices, classical meters or other sources. Table 2 presents the categories studied and provides some examples.

Table 2: Studied categories of monitored data and examples

Categories of monitored data		Sub categories and examples
Indoor condition	Environment	Hygro-thermal (temperature, …), Air quality (CO_2, …)
	Occupants	Position (activity, …) , Control actions (opening/closing windows, …)
Outdoor condition	Environment	Hygro-thermal (temperature, …), Solar radiation, Pressure, Wind
Building management	Energy	Consumption (heating, cooling, …), Production (heating, cooling, …)
	Control system & devices	Status (ventilation, valve, …), Regulation (temperature, …)

The data has been separated according to two criteria that influence or might influence the nature of the data and thus the anomaly detection technique that can be used: the type of measurement and the type of measuring process. First, the type of measurement allows differentiating discrete or binary variables (valve opening, activity detection ...) from continuous variables (temperatures, radiations ...). On the other hand, the measurement process is considered in order to differentiate measurements from sensors with a constant acquisition frequency that are made on a regular and programmed timestamp (CO_2 concentration, relative humidity …) from those generated by an event (window opening, activity …). However, the impact of the type of measuring process on the nature of the data and therefore, on the anomaly detection technique, is still to be proven and remains as a conjecture.

In order to complete the measurement classification and to provide information for multivariate anomaly detection techniques, a last criteria, the connection between measurements, is considered and added to the classification. This connection is established through correlation techniques and domain expertise. As an example, the CO_2 concentration is classified as a continuous data with a regular timestamp which is linked to occupancy and activity measurements.

3. Results and investigations

Based on the proposed anomaly and measurement typology, we are able to present in Figure 3 (i) a classification method for anomaly detection techniques that best suit with the specificities of each dataset and anomaly encountered for building measured data. The next step, currently in progress, is to fill each sub-part of this classification by listing and testing anomaly detection techniques on collected measurements in order to classify them according to their use. For instance, Figure 3 (ii) presents the test of a STL technique (Seasonal and Trend decomposition using Loess) on a labelled indoor temperature in order to evaluate its capacity to detect successive univariate anomalies. Although this technique seems to work well by detecting most of the successive anomalies in the dataset (a), its inability to detect successive abnormal constant values (b) appears to restrict its use. By conducting similar tests on other anomaly detection techniques (Isolation Forest, Facebook Prophet, *etc.*), this classification may help to select an anomaly detection technique according to the needs of a project (expected accuracy, time limitation, *etc.*), its specificities (encountered data or anomaly for instance) or any eventual limitation (such as the availability of the data for training or to proceed multivariate anomaly detection techniques). Furthermore, particular attention will be devoted on multivariate anomalies detection techniques such as LSTM (Long Short-Term Memory) in future works. Indeed, despite a considerable work in the literature attending to detect independent and related univariate anomalies, there is a relative scarcity in anomaly detection techniques for related multivariate anomalies [2,10,11].

Although this classification method wishes to offer in the future a non-exhaustive and decomposed approach to select anomaly detection techniques for building measurements, some limitations have to be considered. An important part of these limitations is related to the data collected in order to train or use anomaly detection techniques. Depending on whether the data is labelled or not, its quality or quantity, it might be complicated to provide a satisfactory result. For instance, STL techniques cannot work if some data are missing in the dataset or a multivariate anomaly detection approach cannot be applied if there are no correlated measurements to be use with. Furthermore, an anomaly is supposed to be (and to remain) a rare event. Thus, it might be more efficient to focus solely on univariate anomalies than multivariate anomalies even if a few anomalies remain undetected. Indeed, in numerous cases, the increased accuracy provided by the use of multivariate anomaly detection techniques might not be significant enough to justify allocating the additional time and resources needed for them to run.

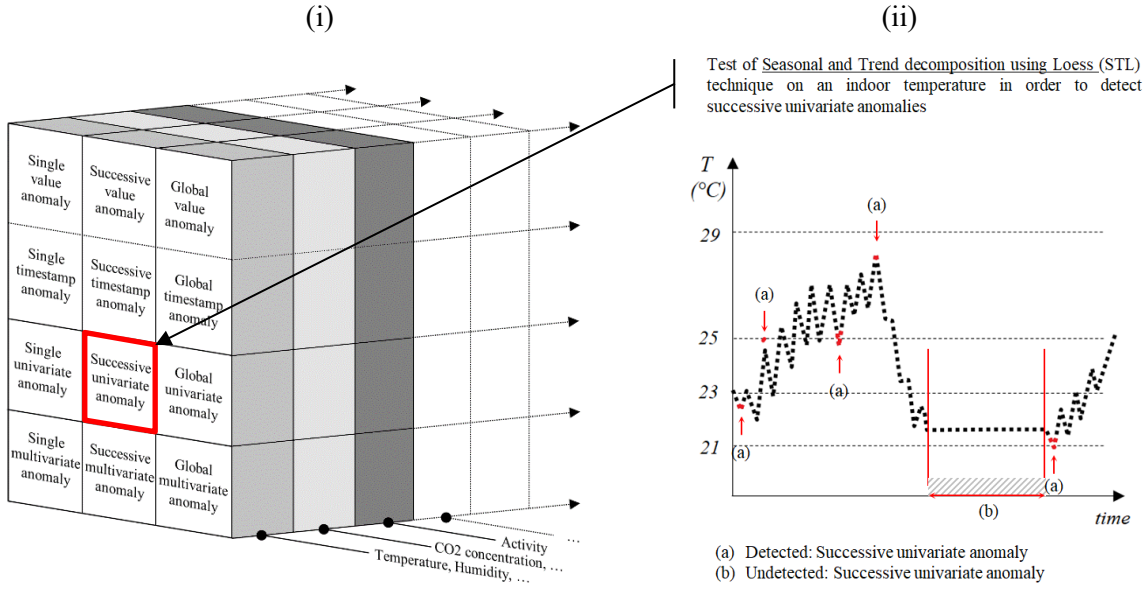

Figure 3 : Proposed classification method for anomaly detection techniques in time series data from building measurements

In addition to the classification and selection method for anomaly detection techniques, this work pursues two other aims. 1) The development of a library of measured data and labelled anomalies for the scientific community. Indeed no such open access labelled dataset exists to our knowledge, while it is necessary to evaluate results, and even train supervised machine learning models. 2) The evaluation of the reliability of the anomaly detection techniques in order to allow the qualification of the studied data throughout the processing chain (pre-processing, during processing and post-processing). This is in order to have an indication of the reliability (and use) of the data.

4. Conclusion and outlook

In this paper, we presented the construction and the implementation of a classification methodology of anomaly detection techniques adapted to most of the measurements performed in the building domain. The study of data and anomalies specific to time series allowed us to define two typologies that constitute the core of this methodology. Although still incomplete and to be tested, this classification methodology would allow, *in fine*, to select the most adequate anomaly detection techniques for each type of anomaly and data encountered (or available). This methodology might also offer a global overview of the techniques that are currently used or might be used, in the building sector and potentially optimize the choice of selected algorithms according to the needs, conditions or limitations of each study. Moreover, a modular system to deal with anomaly detection in buildings can potentially be derived from this work, considering a detection module per anomaly typology.

In future work we hope to automatize this selection and application process through an online platform in order to provide a quality check of the supplied data to define if a correction is doable.

References

[1] Chandola, V., Banerjee, A., and Kumar, V. 2009. “Anomaly detection: A survey”. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI=10.1145/1541880.1541882

[2] Raghavendra C., and Sanjay C. 2019. “Deep learning for anomaly detection: A survey”. arXiv preprint arXiv:1901.03407(2019).

[3] Zimek A, Filzmoser P. “There and back again: Outlier detection between statistical reasoning and data mining algorithms”. WIREs Data Mining Knowl Discov. 2018;8:e1280.

[4] Liu Y., Pang Z., Karlsson M., Gong S. Anomaly detection based on machine learning in iot-based vertical plant wall for indoor climate control Buil Environ, 183 (2020), p. 107212.

[5] Liguori, A., Markovic, R. & al. (2021). Indoor environment data time-series reconstruction using autoencoder neural networks. Science Direct. doi:https://doi.org/10.1016/j.buildenv.2021.107623.

[6] Yassine Himeur, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, Abbes Amira, (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Applied Energy. https://doi.org/10.1016/j.apenergy.2021.116601.

[7] Foorthuis, Ralph. (2020). On the Nature and Types of Anomalies: A Review.

[8] Foorthuis, R. “A Typology of Data Anomalies”. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, Cham, 2018. p. 26-38

[9] Tan, PN., Steinbach, M., and Kumar, V. 2005. “Introduction to Data Mining” (First Edition). Addison-Wesley.

[10] Teodora Sandra Buda, Bora Caglayan, and Haytham Assem. Deepad: A generic framework based on deep learning for time series anomaly detection. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 577–588. Springer, 2018.

[11] Guo, T., Lin, T. & Antulov-Fantulin, N.. (2019). Exploring interpretable LSTM neural networks over multi-variable data. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:2494-2504 Available from http://proceedings.mlr.press/v97/guo19b.html .