Data Modelling and Time Series Classification of Building Energy System Datapoints (Which May Also Be Quite Long And Stretch Several Lines) Here

Master Thesis

Your Firstname(s) Your Lastname(s)

The present work was submitted to the

Chair of Communication and Distributed Systems

RWTH Aachen University, Germany

Advisor(s):

Dipl.-Inform. Erika Mustermann
Max Mustermann, M. Sc.

Examiners:

Prof. Dr.-Ing. Klaus Wehrle
Prof. Dr. rer.nat. Albert Einstein

Registration date: January 1, 2019
Submission date: August 30, 2019

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

Your Lastname(s), Your Firstname(s)
Name, Vorname/Last Name, First Name

XXX XXX
Matrikelnummer (freiwillige Angabe)
Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende ~~Arbeit~~/~~Bachelorarbeit~~/ Masterarbeit* mit dem Titel
I hereby declare in lieu of an oath that I have completed the present ~~paper~~/~~Bachelor thesis~~/Master thesis* entitled

Data Modelling and Time Series Classification of Building Energy System Datapoints
(Which May Also Be Quite Long
And Stretch Several Lines) Here

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

Aachen, 30.08.2019
Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen
*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
Para. 156 StGB (German Criminal Code): False Statutory Declarations
Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence
(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.
(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Aachen, 30.08.2019
Ort, Datum/City, Date

Unterschrift/Signature

Abstract

Außerhalb der Umgebung hingegen erfolgt die Trennung des Donaudampfschifffahrtskapitäns nicht korrekt.

Kurzfassung

In der `otherlanguage*`-Umgebung erfolgt die Trennung des Donaudampfschifffahrtskapitäns korrekt.

Contents

1

Related Work

1.1 Data Modelling in Building Management Systems

[ ]BRICK schema, its key features (ontology, based on semantic web)

[ ]Smart Data Models, its key features

1.2 Time Series Classification

[ ]Bode et al. (EBC paper). Unsupervised time series clustering. Why poor results?

[ ]Chen at al. (A Metadata Inference Method for Building Automation Systems With Limited Semantic Information)

Automated detection of BMS datapoint metadata has been addressed in recent work using time series data and datapoint labels for detection.

El mokhtari et al. [MEm21] propose classification of sensor time series using a CNN that receives datapoint time series and its derivative as inputs. They further propose a grey-box model to calculate correlation scores between estimated control time-series and unknown control time-series datapoints.

Mertens et al. [MW23] evaluate several traditional machine learning algorithms to classify ventilation system time series and achieve best results *(mention accuracy value)* with XGBoost, ExtraTrees. For the classification of a test datapoint, they classify each 24-hour window of its time series individually, and then assign the majority predicted class over all windows as the final prediction of a datapoint.

2

Problem Statement

The previous chapters introduced the required domain knowledge about Building Energy Systems (BES) and their role in enabling smooth and efficient functioning of building infrastructure. They account for a large proportion (mention percentage here as a citation) of the energy use in the real-estate industry. BES operation optimization is a multidimensional problem including energy consumption, comfort, and control requirements. In this chapter, we elaborate on a specific facet of the BES operation optimization problem, namely the structuring and classification of BES time series data. We begin by motivating our thesis in section 2.1. We list the challenges of structuring and ordering BES data and the origins of these challenges in section 2.2, followed by a summary of the research gaps that remain unaddressed in section 2.3. Finally, we derive a problem statement by enumerating the problems this thesis aims to address and defining the scope of our work in section 2.4.

2.1 Motivation

[x]Give outline of the subsections and their focus

[x]Outline motivation for thesis (Why is my thesis important? What is the long-term goal? Why will the world end without my thesis?)

Building Energy Systems have become a commonplace feature in the building sector. The backbone of BES are sensors and actuators, which monitor, collect, and transmit measurements and event data from distinct, remote points in a building. Currently, the installation of sensors, actuators, building monitoring equipment and the integration of datapoints into BES are two separate tasks performed distinct from each other. This separation of concern creates a challenge for datapoint integration into BES since intergation requires knowledge of the building topology and location of datapoints. Overcoming this challenge requires manual effort and expertise to first ascertain the building topology and later match datapoints to the

topological representations in BES. Avoiding this manual effort would save time as well as costs, which can then be assigned to more critical needs.

Accompanying the advent of BES is the emergence of large corpora of BES data resulting from the storage of BES time series datapoints. BES data is used for analysis and modeling tasks such as efficiency monitoring and fault detection. Raw, unstructured time series data generated by BES datapoints must be structured and complimented with requisite metadata to be fit for consumption by analysis tasks. Data models provide a solution to represent information of buildings and building-related equipment in a structured manner. In addition, depending on the use-case motivating a data model, it may also capture semantic relationships between datapoints, introduce human-readable labels,and provide support for translation of building data to other machine-readable formats. Datapoints often have constraints on the underlying values they track, owing to the physical nature of the measured metric. Value range (minimum, maximum), datatype, unit of measure are examples of such constraints. In order for building datapoints to be structured using a data model, building topology as well as behaviour of building-related equipment needs to be known. The identification and association of raw time series data to their representation in BES forms the background for the motivation of this thesis.

As mentioned earlier, BES comprise of several sub-systems. For example, an HVAC system is made up of a thermostat, heat exchanger, combustion chamber, fan(s), condensing unit, and damper. Optionally, (mention optional HVAC components here) may also be included in a HVAC system. Therefore on an abstracted level, different BES consist of a constant set of components, differentiated only by optional components appropriate to regional or temporal climatic conditions. Subsequently, the general relational and semantic information of BES sub-systems remains (to a large extent) unchanged across distinct installations. This information may contain valuable knowledge of building topology and utilizing this largely constant semantic information may aid the classification of BES datapoints. Thus, the primary motivation of this thesis originates from the potential of using generic semantic information of BES to enhance classification of BES datapoints.

Encoding this high-level semantic information could be achieved with a data model. One of the possible categroizations of data models is into three types: conceptual, logical, and physical data models. These three types differ in the level of data abstraction they offer, with conceptual models being the most abstract and physical models the least. A conceptual data model represents the overall structure and establishes the scope of the data model from a organisation standpoint. A logical data model adds more detail to the conceptual model by establishing entities, their attributes, and relationships. However, it does not include specifications of the underlying hardware and software technology. A physical data model takes into account technical factors such as the DBMS, and data storage platform to detail a database-oriented implementation of the data model. For this thesis, the motivation to apply generic semantic information of BES for classification of datapoints guides us to the choice of a logical data model to collect relationships within BES. Our final motivation, therefore, is to adequately represent generic semantic information of BES with a data model to subsequently classify BES datapoints.

Automating the classification of BES datapoints and ensuring their conformation to a data model is interconnected with several other objectives in BES operation

optimization. For example, Digital Twins in the building sector are developed to simulate the introduction of new equipment in a BES without the need to conduct a full-scale physical installation of equipment. However, knowledge of building topology and metadata of building equipment is a pre-requisite for Digital Twin development. An approach for identifying datapoint metadata from raw time series data will fasten construction of Digital Twins. Consequently, fault detection through tracking and visualization of structured BES data becomes easier since logical and locational relationships are on hand to pinpoint root causes of malfunctioning equipment. Adhering BES data to a data model enables smooth exchange of data for inter-building comparison of BES operation. Retrofitting buildings can be expedited since time to associate digital datapoints with their physical entities in the building is shortened. Lastly, by assigning datapoints to their data model representations, we build a bridge towards implementing a physical data model from a logical data model. Thus, the broader motivation of this thesis is manifold with the long-term goal of facilitating the optimization of other BES end goals.

2.2 Challenges

[x]For data modelling, challenges of unstructured datapoint names, no standarised data format for capturing building context information

[x]For TSC, challenge of not knowing building datapoint metadata. Which tasks does this affect?

[x]Why does faulty data need to be handled? (sensor error, missing data)

Above, we laid out the premise and the current circumstances that motivate our thesis. In this section, we list the challenges that arise from these circumstances. These challenges majorly stem from the unstructured and unannotated nature of BES data.

C1: Datapoint labels A key challenge in the building sector is poorly annotated BES datapoint labels. The use of numeric identifiers and acronyms in datapoint labels causes them to not be human-readable. In most existing buildings, datapoint labels do not follow a standard semantic model and are not well-documented [CGS+20]. As highlighted previously, labels with limited semantic information lead to challenges in several analysis tasks such as development of Digital Twins, efficiency monitoring, fault detection, as well as detection of datapoint type which we expand on in the next section.

C2: Unsupervised time series classification A direct consequence of the previous challenge leads to the difficulty of unsupervised time series classification. Due to the absence of usable datapoint labels, automated detection of datapoint types needs to be performed in an unsupervised fashion by relying on the numerical time series data of BES datapoints. [SNCG19] attributes the challenge of unsupervised clustering to time series features being less valuable compared to descriptive metadata labels.

C3: Structuring datapoints Currently, the building sector faces the absence of a standard sector-wide data model for capturing context and objective information

within buildings. The choice of a data model differs with respect to each distinct BES vendor. The challenge lies in the design of a data model that contains valuable semantic logic for improving time series classification of BES datapoints. In addition, the data model also needs to be serializable into a machine-readable format (the choice of which also depends on the underlying DBMS or data storage provider). Fulfilling this two-fold objective would ensure that our data model is capable of representing datapoint metadata as well as compatible with BMS software.

C4: Erroneous time series data Erroneous data hinders the use of raw time series data for analysis tasks, and mandates data preprocessing techniques to be applied prior to analysis tasks. Teh et al. [TKLW20] conduct a systematic review of 57 publications to ascertain the different types of physical sensor data errors. They conclude that outliers, missing data, noise, constant value, and stuck-at-zero are common error types addressed in research. Outliers are values that largely deviate from the normal behaviour provided by the model. They greatly influence statistical values of time series data such as mean, median, variance, and therefore provide an inaccurate summary of the data. Li et al. [LP14] attribute missing data to several causes such as network congestion, sensor device outages, and environmental interferences. Missing data over longer time periods can negatively affect both size and continuity of available data. Noise are small variations in the dataset due to degradation in sensing quality over time, or imprecise measuring of values. Stuck-at-zero errors correspond to high dead-times in sensor readings resulting in zero values. Constant values, similar to stuck-at-zero errors, are generally a result of transmission issues or malfunctioning sensors.

Having listed the challenges surrounding our research area, in the following section, we analyse how existing research overcomes these challenges and their shortfalls, thereby identifying research gaps that remain.

2.3 Existing Work

[ ]This section is incomplete. Mention more papers and their research gaps to make it comprehensive

[ ]Summarize all research gaps into final paragraph

This section mentions the contributions of existing research towards automated inference of metadata for building system datapoints. We summarize these works in the context of our challenges. Finally, we observe why the existing literature does not yet meet all of our challenges.

El mokhtari et al. [MEm21] propose a convolutional neural network (CNN) architecture to classify sensor time series points. Their approach, relies first on classifying sensor time series, then estimating control time series behavior to classify control time series and finally classifying subsequent response time series. This study however focuses only on sensor and control time series and does not consider boolean, and setpoint time series. They also consider only 4 classes of sensor time series. Mertens et al. [MW23] perform a comparison of non-deep learning techniques and feature selection methods for time series classification of BAS points. However, this work

evaluates supervised classification techniques, thereby relying on well-annotated datapoint labels. This assumption, as we highlighted above, does not apply in the real-world to the majority of buidings. Secondly, since only ventilation system datapoints are considered, reproducing their results on other BES sub-systems might involve additional effort. For the unsupervised time series classification problem, Bode et al. [BSBM19] employed unsupervised clustering methods and compared their results with supervised techniques. As features for clustering, both statistical features as well as an auto-encoder for unsupervised feature extraction were tested. The study concluded that unsupervised clustering results were below par with supervised approaches. Details about hyperparameters, design, or analysis of the auto-encoder features could not be found. The clustering algorithms were trained on 20-day and 1-day time series datasets. While these small lengths might be beneficial in scenarios facing a shortage of BAS point data, they are insufficient to extract long-term seasonal or temporal features in the data. The work that comes closest to tackling our set of challenges is Chen et al. [CGS+20], who develop a method to infer metadata of Building Automation System (BAS) points using numerical time series data only. Their method relies on different feature domains (statistical, time, frequency) to extract robust features for classification. Further, they propose an association method to discover functional relationships between datapoints at the zone level.

To the best of our knowledge, no existing article has considered both unsupervised classification of BES datapoints and use of a data model for BES datapoint classification.

2.4 Problem Statement

[x]How does a good data model look like? What goals should it achieve?

[x]How do you plan to use the data model for time series classification? (mention anecdotes, conceptual sketch)

[x]Outline scenario (application to one or multiple AHU)

[x]Mention rough measures for evaluation

[x]What is the challenge? What steps are involved in resolving it? Under what conditions must a solution work?

We divide our problem into four parts: development of a generic data model for BES datapoints, unsupervised time series classification of BES datapoints, exploiting the generic data model logic as features for time series classification of BES datapoints, and assigning classified datapoints to our generic data model to obtain a specific data model. The distinction between a generic and a specific data model is made in the following paragraphs.

At the time of conceptualizing this thesis, there does not exist a defined standard for modeling BES datapoints. A data model must be established during the installation and integration phases of BES so that associated BAS components (database, data broker) have a data schema to store, query, and modify data by. As mentioned in

2.1, design of data models are influenced by the underlying type of data usage. Since there is no single correct choice for a data model, we consider a set of criteria that our data model must fulfill. Datapoints, such as those in the ERC industrial test hall, are used as the working set around which we construct our data model. Our data model aims to encode context information for datapoints. As an example, for a sensor installed within a BES system measuring a physical metric, relevant metadata may be a human-readable label for the sensor datapoint, location of sensor with respect to the building topology, highest measurement resolution offered by the sensor device. Our data model may also contain constraints for datapoints, examples of which are provided in section 2.1. Moreover, logical relationships between entities are key to understand the complete building topology and functioning of building energy systems. An entity-relationship model suits our scenario to capture logical relationships between BES entities, and constraints applicable to BES components/equipment. An entity-relationship model (ER model), as the name suggests, aims to describe relationships between interrelated things of interest (entities). To recall, one of our motivations is to represent generic semantic information of BES with a data model to subsequently classify BES datapoints. Considering this, our data model must be broad in its design so as to encapsulate semantics, and non-specific without containing details explicit to any particular installation of a BES. At the same time, the data model must contain valuable information to distinguish one datapoint from the other. This presents an inherently hard problem for certain types of datapoints that seldom differ in their time series data, such as boolean (on/off) datapoints, error sensor points (error/no error), and control commands. Therefore, the difficulty of our data modeling problem exists in obtaining a balance between generalizability and sophistication.

While data modeling infers metadata of BES datapoints, we also need to identify the point type (class) of BES datapoints without depending on datapoint labels. This leads us to the unsupervised classification problem. Although existing research has investigated approaches for supervised classification of BES time series using semantic labels with success, unsupervised approaches relying solely on time series data have been less addressed. In 2.2, we highlighted why unsupervised classification is a hard problem. Addressing the unsupervised setting is an important research goal because datapoint labels are often unavailable or poorly annotated. In such cases, a supervised classifier is inapplicable due to the absence of groundtruth labels for training a classifier. Therefore, unsupervised classifiers are trained using only the features of time series data. Computing discriminative features for time series classification is a research direction in itself. In the context of feature quality, 'discriminative' refers to the ability of a feature to assign datapoints to separate classes based on distinct feature values for the datapoints. [RBLG23] conducts a review of feature engineering methods for time series classification and their predictive capabilities. To capture various properties of time series, it is desirable to have diverse features, e.g., seasonality, trends, autocorrelation. [FSZ+19] investigate deep learning-based feature engineering methods for building energy prediction and conclude that it can be very difficult to construct features from the original data for building energy predictions, as building operation data are highly correlated and noisy. While feature engineering is a crucial ingredient in classification, the other major decision influencing performance is the choice of a classifier. Recent work in the field has tapped into the ability of deep neural networks (DNNs) to learn hierarchical feature represen-

tations for classification problems. DNNs are effective at learning features through a series of non-linear transformations referred to as layers. Keeping in mind our overarching goal of utilizing the data model for datapoint classification, the format in which the classifier expects features must be compatible with the format(s) in which the data model's information can be expressed in. Since DNNs take numerical data as input vectors and automatically learn features, they are less suitable for our idea of extracting features from a data model. Instead, non deep classifiers, e.g., rule-based classifiers use features encoded as rules. Each time series instance input through such a classifier is classified on the basis of its satisfiability of the rules.

This bring us to the problem connecting data modeling and unsupervised time series classification in our work, summarized by the following question: how do we convert semantic information from a data model into features utilizable by a time series classifier? BES datapoint metadata is often embedded as text and number attributes or string-type description fields in a data model. This metadata needs to be firstly parsed into a classifier-friendly format. An algorithmic parser could utilize regular expressions to identify recurring/common relationships in the data model. Additionally, human expert knowledge might be required to transform unique metadata belonging to specific BES components. Secondly, parsed features must be filtered based on their classifying merit. Example of metrics to evaluate the "goodness" of features are information gain and Gini coefficient. An ideal classifier would combine features from both the time series as well as features extracted from the data model to obtain a robust feature set to classify BES datapoint into their point types. Mean accuracy, precision, recall, and F1-score are popular evaluation measures for the classification problem. By dividing our dataset (datapoints from the ERC industrial test-hall) into a train and test set, the train set would serve the data modeling and classifier training tasks, and the classifier would be evaluated on the test set containing data previously unseen by the classifier. Optionally, datapoints from a secondary BES could also be used for classfier evaluation, thereby assessing the approach's generalizability to more than one BES.

Once we have the classified datapoints for our test-hall, we want to assemble it into their appropriate representations in the data model. Although not one of the mainstream motivations for our thesis, linking the physical datapoints with their digital representations is a logical extension of our classification task to facilitate the use of a specific data model as a data schema for a DBMS or a BAS data aggregation software. We use the terms 'generic data model' to refer to a high-level logical data model representing BES datapoint semantics, and 'specific data model' to refer to a low-level physical data model instantiating each physical datapoint as an instance/object of its class. Since the EBC team at E.ON ERC employs FIWARE as the central data monitoring platform for its building management needs, we consider FIWARE-NGSIv2[FIW] as the serialization format for our specific data model.

To summarize, our problem statement considers the full lifecycle of structuring BES datapoints from raw time series data to assigning them to a physical data model through the unsupervised classification of BES datapoints. Having defined our problem statement, we outline the design of our method in the following chapters.

Bibliography

[BSBM19] Gerrit Bode, Thomas Schreiber, Marc Baranski, and Dirk Müller. A time series clustering approach for Building Automation and Control Systems. *Applied Energy*, 238:1337–1345, March 2019.

[CGS+20] Long Chen, H. Burak Gunay, Zixiao Shi, Weiming Shen, and Xiaoping Li. A Metadata Inference Method for Building Automation Systems With Limited Semantic Information. *IEEE Transactions on Automation Science and Engineering*, 17(4):2107–2119, October 2020. Conference Name: IEEE Transactions on Automation Science and Engineering.

[FIW] Fiware-ngsi v2 specification. `https://fiware.github.io/specifications/ngsiv2/stable/`. Accessed: 2023-08-26.

[FSZ+19] Cheng Fan, Yongjun Sun, Yang Zhao, Mengjie Song, and Jiayuan Wang. Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy*, 240:35–45, 2019.

[LP14] YuanYuan Li and Lynne E. Parker. Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks. *Information Fusion*, 15:64–79, 2014. Special Issue: Resource Constrained Networks.

[MEm21] J.J. McArthur and Karim El mokhtari. A data-driven approach to automatically label BAS points, 2021.

[MW23] Noah Mertens and Andreas Wilde. Automated Classification of Datapoint Types in Building Automation Systems Using Time Series. In Frédéric Noël, Felix Nyffenegger, Louis Rivest, and Abdelaziz Bouras, editors, *Product Lifecycle Management. PLM in Transition Times: The Place of Humans and Transformative Technologies*, IFIP Advances in Information and Communication Technology, pages 495–505, Cham, 2023. Springer Nature Switzerland.

[RBLG23] Aurélien Renault, Alexis Bondu, Vincent Lemaire, and Dominique Gay. Automatic feature engineering for time series classification: Evaluation and discussion. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2023.

[SNCG19] Zixiao Shi, Guy R. Newsham, Long Chen, and H. Burak Gunay. Evaluation of clustering and time series features for point type inference in smart building retrofit. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '19, page 111–120, New York, NY, USA, 2019. Association for Computing Machinery.

[TKLW20] Hui Yie Teh, Andreas W. Kempa-Liehr, and Kevin I-Kai Wang. Sensor data quality: a systematic review. *Journal of Big Data*, 7(1):11, Feb 2020.

A

Appendix

A.1 List of Abbreviations