# In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data

3

Ekaterina Petrova, Pieter Pauwels, Kjeld Svidt, and Rasmus Lund Jensen

**Abstract**

Cross-domain analytical techniques have made the prediction of outcomes in building design more accurate. Yet, many decisions are based on rules of thumb and previous experiences, and not on documented evidence. That results in inaccurate predictions and a difference between predicted and actual building performance. This article aims to reduce the occurrence of such errors using a combination of data mining and semantic modelling techniques, by deploying these technologies in a use case, for which sensor data is collected. The results present a semantic building data graph enriched with discovered motifs and association rules in observed properties. We conclude that the combination of semantic modelling and data mining techniques can contribute to creating a repository of building data for design decision support.

**Keywords**

BIM • Semantics • Data mining • Pattern recognition • Knowledge discovery

## 3.1 Introduction

Cross-domain analytical techniques such as Big Data analytics, machine learning, semantic query techniques and inference machines have made the prediction of outcomes in building design possible and much more accurate. Research has shown promising advances within the use of machine learning and data mining techniques for model predictive control, meta-modelling for design space exploration, grey box modelling and advanced control strategies related to building energy systems, etc. These approaches carry a powerful potential and can directly influence the decision-making process in the Architecture, Engineering and Construction (AEC) industry by infusing it with an evidence-based character. The latter is of direct relevance for high-performance building design, which employs strict performance criteria. Responding to these criteria ideally requires evidence-based multidisciplinary input. Nevertheless, many decisions are still based on rules of thumb and previous experiences, and not on documented evidence. This leads to inaccurate predictions and assumptions regarding input parameters (e.g. occupancy rate), rare revisiting of analytical and building models during operation, no modification of design assumptions based on actual performance and thus a difference between predicted and measured performance.

E. Petrova (✉) · K. Svidt · R. L. Jensen
Department of Civil Engineering, Aalborg University, Thomas Manns Vej 23, Aalborg, Denmark
e-mail: ep@civil.aau.dk

K. Svidt
e-mail: ks@civil.aau.dk

R. L. Jensen
e-mail: rlj@civil.aau.dk

P. Pauwels
Department of Architecture and Urban Planning, Ghent University, J. Plateaustraat 22, Ghent, Belgium
e-mail: pipauwel.pauwels@ugent.be



I. Mutis and T. Hartmann (eds.), *Advances in Informatics and Computing in Civil and Construction Engineering*,
https://doi.org/10.1007/978-3-030-00220-6_3

If knowledge discovered in building operation would be accessible, a design professional should be able to match the ongoing design with meaningful performance patterns. This article aims to investigate how data from buildings in operation can enable knowledge discovery and provide patterns that can be useful to inform future design processes. In particular, we consider available operational building data related to indoor space use, thermal performance and indoor climate collected from a culture and sports center. This use case is particularly interesting, as the building hosts different spaces such as conference and exhibition halls, ice hockey arenas, training facilities, swimming and wellness facilities, etc. The case provides operational building data captured through a sensor network and existing CAD drawings. From the collected datasets, we distil patterns and represent these so that they can be reusable by deploying the latest technological advances within Knowledge Discovery in Databases (KDD) [1] and semantic data modelling. The considered techniques are not often easily combined, especially not to inform future design decisions, which is the fundamental purpose of this study.

In this article, we first look into the diverse existing computational approaches for data analytics and knowledge discovery (Sect. 3.2), and semantic representation of building data (Sect. 3.3). In Sect. 3.4, we indicate how these data can be combined for knowledge discovery. We thereby suggest a system architecture aimed specifically at that purpose. Section 3.5 presents the use case we relied on for knowledge discovery, including the results obtained from that use case.

## 3.2 Data Analytics and Knowledge Discovery in the AEC Industry

The AEC industry nowadays generates large volumes of data associated with all stages of the building life-cycle. However, the traditional analytics can generate informative reports, but fail when it comes to content analysis [2]. As a result, data mining, pattern recognition and KDD have received major attention, as they can provide reliable results and effectively assist in analysis of data and extraction of knowledge. One definition of data mining is "the analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways so that data owners can fully understand and make use of the data" [3]. Furthermore, Bishop defines pattern recognition as "the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories" [4]. Finally, KDD represents the overall process of knowledge extraction, with knowledge being the end product of the data-driven discovery and data mining being the step in the process which employs specific algorithms to discover patterns in the given data [5]. Fayyad et al. [1] state that the fundamental objective is to discover high-level knowledge in low-level data and define the transformation steps of raw data into actionable knowledge, i.e. data selection, preprocessing, transformation, mining and interpretation/evaluation of the discovered knowledge.

Widely accepted data mining categories include classification, clustering, association rule mining, regression, summarization and anomaly detection, targeting either predictive (supervised, directed) or descriptive (unsupervised, undirected) analytics [1, 6]. Supervised approaches describe the qualitative or quantitative relationships between the input and output variables and rely on domain expertise and significant amounts of training data. As a result, discovery of novel knowledge is unlikely, due to the predefined inputs and outputs. Unsupervised approaches (e.g. clustering, association rule mining, etc.), however, excel in discovering the intrinsic structure, correlations and associations in data and do not rely on training data, as inputs and outputs are not predefined. While predictive techniques are backward oriented due to their predefined target, descriptive ones are forward oriented (no explicitly defined target) and make it possible to discover interesting patterns and relationships in the data [7].

Within the high-performance and sustainable building design domain, the use of predictive approaches is usually related to prediction of building energy use and demand [8–10]; prediction of building occupancy and occupant behaviour [11, 12]; and fault detection diagnostics [13, 14]. Unsupervised tasks usually complement and target framework development [15–17]; discovery of patterns in occupant behaviour for improvement of operational performance [18]; and extraction of energy use patterns [19, 20]. Of course, KDD applications in the AEC industry span over a much broader area than the main categories defined above. For instance, Jun and Cheng [21] target high-performance with classification models for sustainability certification evaluation and Peng et al. [22] propose the use of BIM-based data mining approaches for improvement of facility management, etc.

These studies all show promising results when it comes to improvement of the building operation and occupant comfort. However, using knowledge discovery in data to support future design decision-making is an area that is not explored in detail. Studies have explored pattern recognition in simulation data and information extraction from BIM design log files [23], data-driven approaches for energy-efficient design by BIM data mining [24], as well as use of data mining for extracting and recommending architectural concepts [25]. Even though these studies demonstrate promising results within the use of KDD for design decision support, they rely on patterns only in design data. The data analysis results coming from

existing buildings can rarely be linked to an early stage design, mainly because the data representations do not match. Thus, this study attempts to explore knowledge discovery in operational building data as a means to improve the decision-making in the performance oriented design process.

## 3.3 BIM and Semantic Representations of Building Data

The representation of building information nowadays typically happens using a BIM model, most commonly exchanged using the Industry Foundation Classes (IFC) data model, which captures building geometry, object properties, as well as semantics. The IFC schema is represented in the EXPRESS information modelling language. Any file exported to IFC is then typically an IFC STEP Physical File (IFC-SPF). Alternative formats for the IFC data model are available in XML, RDF and JSON. In all cases, however, the data model itself is derived directly from the EXPRESS or IFC-SPF format, making it the absolute reference.

Recent research and development initiatives have showed promising results using graph-based data modelling techniques, which are more common in a web environment (e.g. Neo4J, GraphDB). Such approaches are the preferred solution especially when a link needs to be made to outside data that is not typically captured in an EXPRESS-based format (e.g. sensor data, geospatial data). Typically, graph-based approaches focus entirely on the semantics and less on other specific data, such as geometry, large amounts of tabular data, etc. In such case, the semantic graph contains a direct link to the relevant information, which is kept in its original format. Both practice and research thus suggests the use of a graph-based format to capture building data, nevertheless keeping numeric data explicitly out of the semantic graph for computational performance reasons.

Representing semantic building data in a graph format can be done with the available ontologies by the W3C Linked Building Data (LBD) Community Group.[1] This includes a Building Topology Ontology (BOT) [26], a PRODUCT ontology, a PROPS ontology (properties), and an Ontology for Property Management (OPM). Using linked data technologies, links can then be maintained with other data [27], including operational data. For instance, device data can be captured using SAREF,[2] and sensor data can be represented using SSN[3] and/or SOSA.[4] For the building performance data, these ontologies do not serve well in case all operational data are targeted. In such case, a tabular format is still a lot more effective. The mentioned semantic ontologies can be used to capture static characteristics, such as averages, min-max values, features of interest, devices, and so forth.

## 3.4 Combining Semantics and KDD to Enhance High-performance Design: Proposed System Architecture

In this article, we consider the combination of KDD (Sect. 3.2) and building semantics (Sect. 3.3) for the purpose of design decision support. Most importantly, design decision support tools need to re-use the knowledge discovered in the available data through KDD and semantic data modelling. In this section, we focus entirely on discovering patterns using KDD and semantic data modelling, so that a repository of queryable design patterns can be built. Considering that the available data originate from multiple heterogeneous sources, a decentralized structure is preferred, which is most commonly realized using graph database approaches. Using these technologies, one can construct a web of semantic information in a decentralized manner, thereby allowing links between datasets, while respecting their original data structures. Transforming all data to a semantic format is possible and allows direct queries and applying semantic data mining techniques [28]. However, this approach may disallow many highly efficient data mining algorithms that can be used for retrieving useful knowledge. Instead, we propose to store the different kinds of data separately, thereby distinguishing between semantic data, geometric data and operational data (Fig. 3.1).

We additionally suggest a semantic data integration layer for linking the semantic data model of a building with its numeric representations and dynamic performance parameters. This layer serves as a reference model for the semantics of

[1] https://www.w3.org/community/lbd/.
[2] https://w3id.org/saref.
[3] https://www.w3.org/TR/vocab-ssn/.
[4] https://www.w3.org/ns/sosa/.

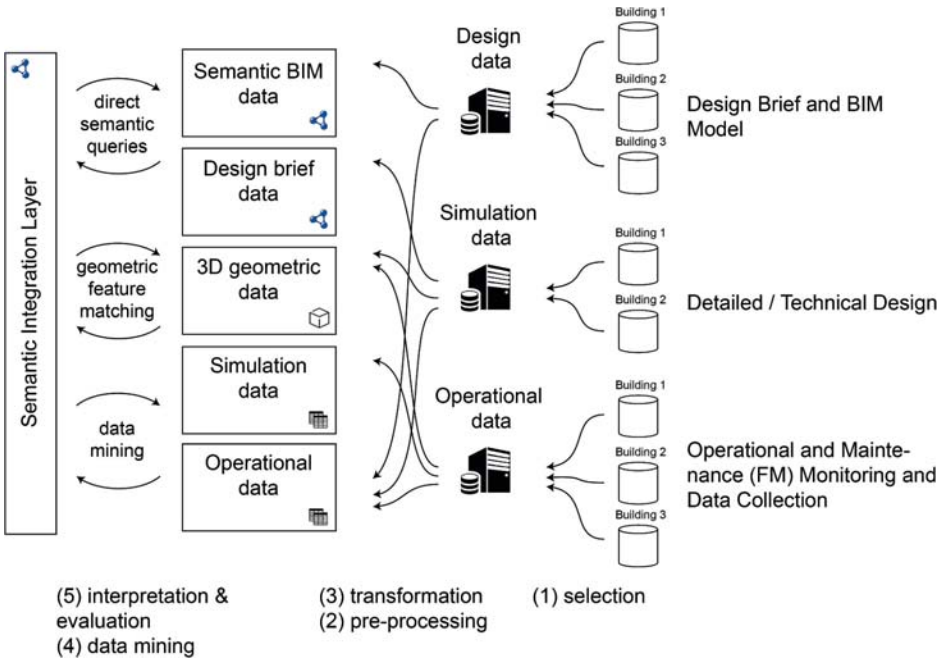


**Fig. 3.1** Proposed system architecture for the combination of semantics and KDD

the different data sources and makes integration possible by pointing from within the semantic graph to web server addresses for operational data streams and geometric data files. As a result, systems accessing this data can recognize the relevant associations.

## 3.5 Use Case: Gigantium Cultural and Sports Center

Gigantium is a large cultural and sports center in Aalborg, Denmark, which opened to the public in 1999. Initially, it housed a hall with indoor football and handball courts, a sports hall and meeting facilities. In 2007, two ice skating halls were added, followed by swimming facilities in 2011. Today, Gigantium hosts an ice skating arena and training facility, sports halls, a concert and exhibition hall, swimming and wellness facilities, athletics hall, meeting rooms, a conference room, a cafe, and a lobby. The total area of the center is about 34,000 $m^2$. The ice skating arena can host 5000 spectators and the main hall capacity during concerts is 8500.

Operational building data is being collected through a sensor network consisting of 35 nodes, divided in all spaces [29]. The nodes monitor Temperature (°C), Relative Humidity (%), Air Pressure (hPa), Indoor Air Quality [Total Volatile Organic Compounds ((TVOC), ppb) and $CO_2$ (ppm)], illuminance (lux) and motion. The purpose of the data collection spans from monitoring indoor climate and thermal comfort, to providing information on space use for maintenance of the facilities. Clearly, the diversity of facilities and activities will be reflected in the collected data. For instance, temperature and relative humidity for meeting rooms, ice hockey arenas, and swimming pool will clearly be different. As a result, this use case provides an ideal dataset that can be used to test the proposed knowledge discovery approach in diverse environments within the same building. Most importantly, the discovered patterns can then inform design decisions related to thermal comfort and indoor climate. For example, persisting issues have been experienced with overheating in the conference room, which has led to a decision to renovate the mechanical ventilation system. The discovered insights would be invaluable to the decision-making related to the system design, by preventing uninformed decisions or use of design parameters that previously led to these issues.

### 3.5.1 Capturing the Building Semantics Using a Semantic Graph

As the use case building was built in 1998, there was no BIM model or 3D geometry available as project data. Instead, access was only available to 2D CAD data in PDF format. In this research, we generated a semantic graph from the available data. The spaces are represented using the BOT ontology as *bot:Space* instances. Each of the spaces is linked to its corresponding sensor nodes. These are defined as *bot:Element* and *gig:SensorNode* class instances. The *gig:SensorNode* class is a direct subClass of the *sosa:Platform* class, which is defined by the SOSA ontology to "carry at least one Sensor, Actuator, or sampling device to produce observations, actuations, or samples". Each sensor node hosts sensors, tracking different observable properties (Sect. 3.5). The information is described in a graph, following a combination of the BOT and SOSA ontologies, including custom classes and properties (namespace "gig:").

Important to note is that the data values are not directly stored in the semantic graph. Instead, a custom *gig:values* datatype property points to a web address that returns the data values as requested using the HTTP protocol. One is able to add attributes to an HTTP request, thereby setting query parameters such as time frame and refresh rate (e.g. *from = now-30d&to = now&refresh = 30*s). The result includes the pointer to the data stream for a *sosa:Result* of a *sosa:Observation*. A full data sample is available[5], yet, access to the sensor data streams is obviously restricted.

```
inst:room_1
        rdf:type bot:Space ;
        rdfs:label "Main hall" ;
        bot:hasSpace inst:room_2 ;
        gig:hasSensorNode inst:sensorNode_00000097, inst:sensorNode_000000B0,
                inst:sensorNode_00000077 ;
        geom:hasGeometry "2000, 3000, 4000, 6000"^^wkt:linestring.

inst:sensorNode_00000097 rdf:type gig:SensorNode, bot:Element ;
        rdfs:label "00000097" ;
        gig:observation "Space use" ;
        sosa:hosts inst:sensor_00000097_1 ;
        gig:placement "Placed in the middle of the hall, 8m above the floor. "

inst:result_1 rdf:type sosa:Result ;
        rdfs:label "Result of observation of Relative Humidity" ;
        gig:values "https://gigantium.dk/Gigantium2018instances?orgId=1&datastream=true"
```

Although not in direct focus for this paper, geometry of spaces is also stored in this semantic graph (*geom:hasGeometry*). This representation relies on a Well-Known Text (WKT) and can be used for simple visualization of the relevant spaces in a web-based floor plan layout visualization.

### 3.5.2 Knowledge Discovery in Operational Building Data

According to Fan et al. [30], operational building data is essentially multivariate time series data, where each observation is a vector of multiple measurements, and time intervals between subsequent observations are fixed. In that case, knowledge discovery can help capture relationships between variables over particular time periods (frequent repetitive patterns (motifs) and association rules [31]). This article demonstrates the implementation of these approaches on the diverse data streams from the cafe in the lobby. The location is chosen for its varying number of visitors both on a daily basis and during events, thereby minimising the likelihood of discovery of patterns due to regularly scheduled events. The data is collected in the period 12.03–16.05.2018, which constitutes the full available dataset so far. The hourly observations are exported as CSV files and preprocessed to enable motif discovery. Missing data fields are treated with five iterations of multiple imputation by running the Expectation Maximisation bootstrap algorithm in R. Symbolic Approximate Aggregation (SAX) [32] is further applied for dimensionality reduction and transformation of the input time series into strings. The univariate motifs in the

[5]http://users.ugent.be/~pipauwel/CIBW78_additionaldata.html.

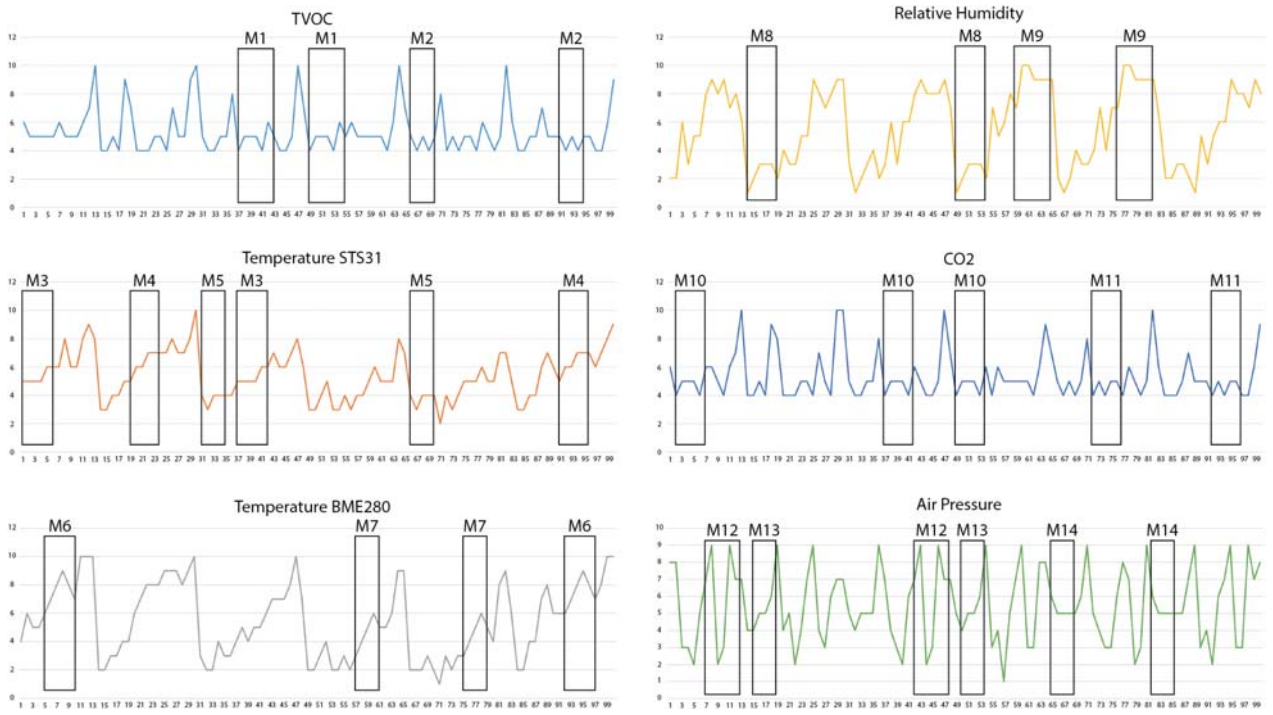


**Fig. 3.2** Discovered univariate motifs (M1–M14) in the observed variables

multivariate time series data are discovered by identifying Longest Repeated Substrings with Suffix Tree implementation [33]. All repeated instances in the symbolic representation of the time series were identified, as for this effort only disjoint and non-overlapping motifs were considered. Figure 3.2 shows a graphical representation of the labelled discovered motifs (M1, M2,…, M14) in the sequence of the six variables. Overlapping motifs, as well as motifs contained within other motifs were excluded from observation.

To enable association rule mining, the discovered motifs are further used to construct a co-occurrence matrix. The columns of the matrix correspond to the motif number and the values for each row (1 or 0) indicate whether an univariate motif occurs or not. For example, M3 co-occurs with M10 and M6. Using the co-occurrence matrix, we obtained 10 sets of co-occurring items for the considered space. Associations between the items of these 10 sets have then been identified by using the association rule mining algorithm defined in [34]. Setting the minimum support and confidence as 0.2 and 0.8 respectively, this results in 13 association rules with support equal to 0.2 and confidence 1. Nine association rules are related to the co-occurrence of M7, M9 and M14. Other association rules are M1 => M10, M3 => M10, M12 => M10, M13 => M8, the last of them being a bidirectional association rule. This means that, for instance, when M12 occurs, the probability of M10 co-occurring is 100%. In this case, the rule indicates an association between observation patterns related to air pressure and $CO_2$. Naturally, the meaning of the discovered rules needs to be interpreted relatively to the design purpose. To be able to use the discovered knowledge, it also has to be connected to the semantic graph in Sect. 3.5.1. This can be done by representing the rules in a semantic graph, and linking this graph to the representation of sensor node 00000014, to create a single motif-enriched graph.

## 3.6 Conclusion

Knowledge discovered in operational data can be linked directly to a semantic representation of the building and can also be used for retrieving and re-using patterns. In this work, we aimed at making high-performance design rely more explicitly on tangible evidence from operational building data. In order to untap as much knowledge as possible from available sources, data mining and semantic data modelling are used. The combination of these techniques is not often intensively deployed in an AEC context. Yet, this combination provides great advantages, as formal semantic query can be combined with flexible

and high-performing pattern recognition techniques. In this paper, we employ these techniques for the Gigantium Cultural and Sports Center in Aalborg. We hereby relied on the W3C ontologies for linked building data to model the building in direct connection to the available data streams. Furthermore, motif discovery and association rule mining were applied to the sensor data, thereby providing hidden knowledge through the semantic graph. This technique can in future work be used to build a repository that can inform any building designer of high-performing building design techniques.

**Acknowledgements** The authors would like to thank Dr. Mads Lauridsen and Aalborg Municipality for providing access to the sensor data used to perform the experiment.

## References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37–54 (1996)
2. Soibelman, L., Kim, H.: Data preparation process for construction knowledge generation through knowledge discovery in databases. J. Comput. Civil Eng. **16**(1), 39–48 (2002)
3. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press, Cambridge (2001)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, NY (2006)
5. Piatetsky-Shapiro, G.: Knowledge discovery in real databases: a report on the IJCAI-89 workshop. AI Mag. **11**(5), 68–70 (1991)
6. Han, J.W., Kamber, M., Pei, J.: Data Mining Concepts and Techniques, 3rd edn. Morgan Kaufmann, Waltham, US (2012)
7. Fan, C., Xiao, F., Li, Z., Wang, J.: Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review. Energy Build. **159**, 296–308 (2018)
8. Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H., Menzel, K.: Mining building performance data for energy-efficient operation. Adv. Eng. Inform. **25**, 341–354 (2011)
9. Wang, Z., Srinivasan, R.S.: A review of artificial intelligence based building energy use prediction: contrasting the capabilities of single and ensemble prediction models. Renew. Sustain. Energy Rev. **75**, 796–808 (2017)
10. Zhao, H., Magoulès, F.: A review on the prediction of building energy consumption. Renew. Sustain. Energy Rev. **16**, 3586–3592 (2012)
11. D'Oca, S., Hong, T.: A data-mining approach to discover patterns of window opening and closing behavior in offices. Build. Environ. **82**, 726–739 (2014)
12. Zhao, J., Lasternas, B., Lam, K.P., Yun, R., Loftness, V.: Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. Energy Build. **82**, 341–355 (2014)
13. Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y., Li, Y.: Case studies of fault diagnosis and energy saving in buildings using data mining techniques. In: IEEE International Conference on Automation Science and Engineering, pp. 646–651 (2016)
14. Pena, M., Biscarri, F., Guerrero, J.I., Monedero, I., León, C.: Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. Expert Syst. Appl. **56**, 242–255 (2016)
15. D'Oca, S., Hong, T.: Occupancy schedules learning process through a data mining framework. Energy Build. **88**, 395–408 (2015)
16. Fan, C., Xiao, F., Yan, C.: A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom. Constr. **50**, 81–90 (2015)
17. Yu, Z., Fung, B., Haghighat, F.: Extracting knowledge from building-related data—a data mining framework. Build. Simul. **6**(2), 207–222 (2013)
18. Xiao, F., Fan, C.: Data mining in building automation system for improving building operational performance. Energy Build. **75**, 109–118 (2014)
19. Miller, C., Nagy, Z., Schlueter, A.: Automated daily pattern filtering of measured building performance data. Autom. Constr. **49**, 1–17 (2015)
20. Wu, S., Clements-Croome, D.: Understanding the indoor environment through mining sensory data—a case study. Energy Build. **39**, 1183–1191 (2007)
21. Jun, M.A., Cheng, J.C.P.: Selection of target LEED credits based on project information and climatic factors using data mining techniques. Adv. Eng. Inform. **32**, 224–236 (2017)
22. Peng, Y., Lina, J.R., Zhang, J.P., Hu, Z.Z.: A hybrid data mining approach on BIM-based building operation and maintenance. Build. Environ. **126**, 483–495 (2017)
23. Yarmohammadi, S., Pourabolghasem, R., Shirazi, A., Ashuri, B.: A sequential pattern mining approach to extract information from BIM design log files. In: 33rd International Symposium on Automation and Robotics in Construction, pp. 174–181 (2016)
24. Liu, Y., Huang, Y.C., Stouffs, R.: Using a data-driven approach to support the design of energy-efficient buildings. ITCon **20**, 80–96 (2015)
25. Mirakhorli, M., Chen, H., Kazman, R.: Mining big data for detecting, extracting and recommending architectural design concepts. In: IEEE/ACM 1st International Workshop on Big Data Software Engineering, pp. 15–18 (2015)
26. Rasmussen, M.H., Pauwels, P., Hviid, C.A., Karlshøj, J.: Proposing a central AEC ontology that allows for domain specific extensions. In: Proceedings of the Joint Conference on Computing in Construction (JC3), pp. 237–244 (2017)
27. Pauwels, P., Zhang, S., Lee, Y.C.: Semantic web technologies in AEC industry: a literature overview. Autom. Constr. **73**, 145–165 (2017)
28. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery: a comprehensive survey. J. Web Semant. **36**, 1–22 (2016)
29. Rodriguez, I., Lauridsen, M., Vasluianu, G., Poulsen, A.N., Mogensen, P.: The Gigantium smart city living lab: a multi-arena LoRa-based testbed. In: 15th International Symposium on Wireless Communication Systems, Lisbon, Portugal (2018) (in press)

30. Fan, C., Xiao, F., Madsen, H., Wang, D.: Temporal knowledge discovery in big BAS data for building energy management. Energy Build. **109**, 75–89 (2015)
31. Fu, T.C.: A review on time series data mining. Eng. Appl. Artif. Intell. **17**, 164–181 (2011)
32. Patel, P., Keogh, E., Lin, J., Lonardi, S.: Mining motifs in massive time series databases. In: Proceedings of the 2002 IEEE International Conference on Data Mining. (2002)
33. Weiner, P.: Linear pattern matching algorithms. In: 14th Annual IEEE Symposium on Switching and Automata Theory, pp. 1–11 (1973)
34. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Disc. **8** (2004)