DOI 10.5162/ettc2018/12.3

A Process Model for the Discovery of Knowledge in Sensor-Based Indoor Climate Data

Dominik Gruedl[1], *Thomas Wieland*[1]
[1] *Fraunhofer Anwendungszentrum Drahtlose Sensorik, Sonntagsanger 1, 96450 Coburg, Germany*
gruedldk@iis.fraunhofer.de, *thomas.wieland@iis.fraunhofer.de*

Abstract:
Sensors in office spaces collect data about the indoor climate in order to monitor and control HVAC-systems to ensure a constant air quality. The quality of indoor climate is dependent on various aspects, such as carbon dioxide, temperature, and humidity.

Approaches on searching patterns in this sensor data by using data mining techniques rarely explicitly apply established process models like KDD (knowledge discovery in databases) or CRISP-DM (cross industry standard process for data mining). These process models describe data mining as one of multiple steps in the whole process of extracting patterns from data. Other steps include the preprocessing of the data as well as the evaluation of the extracted patterns after the data mining.

This paper analyzes these aforementioned approaches and compares them to the established process models before deriving a process model that can be widely applied to data mining projects searching for patterns in sensor-based indoor climate data. The derived process model puts more emphasis on understanding the data and its context as preliminary steps to the extraction of patterns via data mining. In addition to facilitating new research on indoor climate data, the derived process model also allows to understand research in this field of study more easily.

Key words: CRISP-DM, Data mining, knowledge discovery in databases, process model, sensor-based indoor climate data

Introduction

The indoor climate in offices influences the health, productivity and comfort of the employees ([19], [20]). Indoor climate data collected by sensor networks can be used, among other things, to react to changes in the indoor climate. For example, if a threshold value of carbon dioxide is exceeded, a warning lamp can be switched on to indicate the increased value to the persons in the room.

In addition, various data mining methods have already been applied to detect patterns in sensory indoor climate data ([1], [4], [5], [6], [7], [11], [14], [17], [22]).

The research work mentioned above is mainly concerned with the testing of various data mining methods. However, this application is only one step towards the discovery of knowledge in large databases, for example described by the process models KDD (knowledge discovery in databases) [9] or CRISP-DM (cross industry standard process for data mining) [18]. The purpose of these process models is the standardized approach to discover novel, potentially useful, understandable and statistically valid patterns in the analyzed data [9]. In addition to data mining, this comprises, for example, cleaning the data of noise and inconsistencies and transforming the data for the respective data mining procedure by standardization or discretization.

Lack of reference to clearly defined process models makes it difficult to reproduce the results of research work in which data mining is used to determine knowledge in sensor-based indoor climate data. Furthermore, it is difficult to get an overview of research work with similar contents on the subject of data mining.

Methodology

This paper aims to define a process model that is particularly suitable for the discovery of knowledge in sensor-based indoor climate data.

First, the most commonly used process models for discovering knowledge in large data sets are described. These process models are examined to see to what extent they are suitable for describing research work that has already discovered knowledge in sensor-based indoor climate data.

Based on these investigations a process model is derived which takes the advantages and disadvantages of the established process models into consideration.

Subsequently, process models are discussed which were described and applied in two of the investigated research projects.

Established Process Models

Various process models exist which are intended to standardize the approach of data mining projects. The most commonly used non-proprietary process models are CRISP-DM and KDD [15] (Fig. 1).

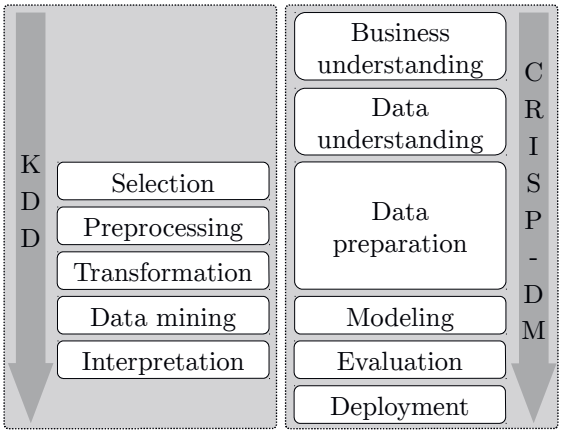

Fig. 1: Comparison of the process models KDD and CRISP-DM

These process models partly overlap. The KDD process model focuses mainly on data preparation and data analysis. The CRISP-DM process model, on the other hand, comprises steps for integrating data mining project into a business context. Nevertheless, the process models share several process steps ([3], [10], [21]).

The steps business understanding, data understanding and deployment of the process model CRISP-DM are not available in the process model KDD. These steps are very project- and company-dependent and therefore usually different for each data mining project. The step data preparation of the process model CRISP-DM comprises the steps selection, preprocessing, and transformation in the process model KDD [3].

Sensor-based Indoor Climate Data

Data mining as such is a very versatile toolset that can be applied to a plethora of situations. A typical use case is deducing information from data measured by distributed sensors, e.g. in buildings.

The presence of people in a room is an important information for the effective and efficient management of temperature and ventilation control devices [22]. Up to 16 % of energy is wasted due to increased energy requirements in unoccupied rooms [13]. Although the number of present persons could be determined through monitoring the persons in the room, this monitoring would be an intrusion into the privacy of the employees. Numerous research groups ([5], [6], [7], [11], [14], [17], [22]) therefore tried to determine with the help of data mining whether there are correlations between indoor climate data recorded by sensors and the presence of the persons in the room.

Opening and closing windows significantly influence the climate in a room ([2], [16]). Furthermore, natural ventilation is the most frequently used form for changing the room climate [8]. In [4], the authors investigated the causes of people's behavior with regard to opening and closing windows using data mining.

The climate and construction properties of a building have a major influence on its energy consumption and CO_2 emissions. For example, by using certain materials, the waste heat from direct sunlight could be used for more efficient temperature management [12]. So-called intelligent buildings are supposed to be able to be both comfortable and energy-efficient for the people present. In [1], the relationship between the optimization of energy consumption and the comfort of people in the room was scrutinized. According to [1], a room is perceived as comfortable if it is both thermally comfortable and optimally lit. Based on the data from these two aspects, classifiers were trained to identify energy-efficient and comfortable rooms.

Suitability of KDD

With the help of the highly detailed preparation of the data, consisting of the steps selection, preprocessing and transformation, corresponding modifications to the collected data can be traced very well (Fig. 2). In the research work examined, these steps of the KDD process model were applied in various degrees of detail.

Not all processing steps may be necessary for the discovery of knowledge in sensor-based indoor climate data. For example, if uniform sensors are used, the data selection step for this data can be omitted. Alas, not only specially collected sensor data, but also the data of third parties are used for the discovery of knowledge. Therefore an integration of the

sensor data and the third-party data is necessary. However, this integration was not described in the research work examined. Due to the lack of these explanations, it is difficult to understand the subsequent steps.

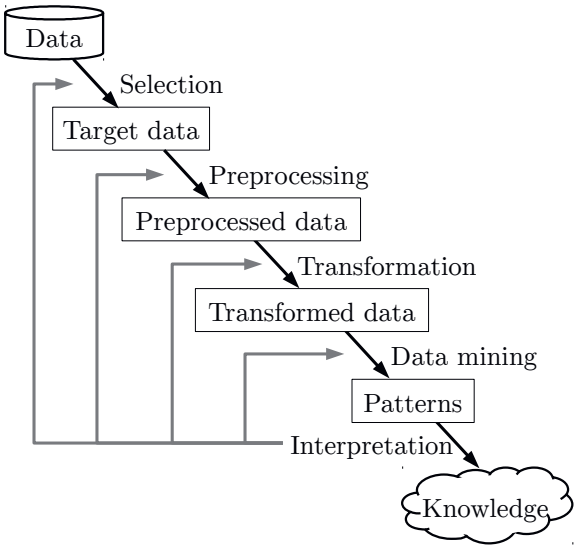

Fig. 2: Process model KDD [9]

Due to the technical conditions for wireless sensor networks, data collection errors may occur. If data is missing or faulty, e.g. due to defective sensors or problems in wireless communication, the data stock must be cleaned up before data mining. In addition, average values were often calculated for the indoor climate data to reduce the amount of data. These tasks are summarized in the data preprocessing step of the process model. This step of the KDD process model is particularly necessary when using feature selection and feature learning.

When selecting the data mining method, in most cases process-dependent data transformations must be carried out on the indoor climate data. Among other things, this process step includes the calculation of new attributes that are necessary for classification procedures. Smoothing can also be used as a method to remove peaks in the indoor climate data.

It may be very difficult to differentiate between the process steps data preprocessing and data transformation. For example, both process steps describe the aggregation as an instrument of the respective preparation. In addition, identical tasks are described with different names. This concerns, for example, the measures for removing noise during the data preprocessing and smoothing the data as part of the data transformation.

In the research work examined, the necessary domain knowledge is explained first. These preliminary considerations, which were made in almost all the research work examined, are not the subject of the process model KDD. In contrast to the CRISP-DM process model, business understanding, data understanding and deployment are not discussed in the KDD process model. The KDD process model therefore focuses on the purely technical aspects of data mining.

Suitability of CRISP-DM

The CRISP-DM process model is designed for data mining projects that take place in an industrial context. This means that the integration of the respective project into the operational processes is of particular importance (Fig 3.).

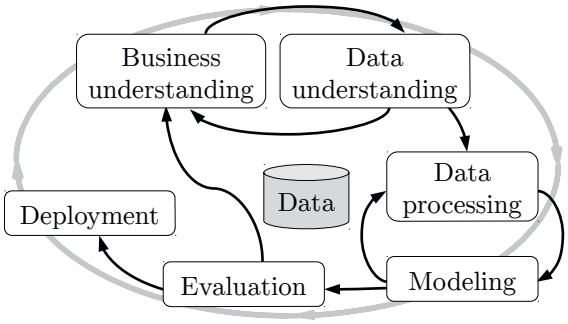

Fig. 3: Process model CRISP-DM [18]

For this reason, the facts and the existing data basis are first examined before working with the data. Furthermore, the deployment of the generated models after their evaluation is an integral part of this process model.

In the beginning of the research work examined, the context of the respective work was always described in great detail. The domain knowledge explained here is indispensable for understanding the following steps. The clarification of legal questions prior to data collection is also crucial before taking further action. If indoor climate data should be investigated in an operational context using data mining, financial or project planning questions could also be very important.

In the same way as the facts, the data which are to be examined are always thoroughly described. In addition to their description, the collection of the data was als discussed in detail.

According to the definition, data collection is not part of the CRISP-DM process model. The CRISP-DM process model is designed to examine data that already exists in the company. The purpose of data collection is therefore not primarily the discovery of knowledge, but the handling of operational processes. For example, delivery and invoice data is collected in order to track the flow of goods and money. Only during a data mining project this data is used for knowledge discovery. However, in the research work examined, the specified indoor climate data were collected exclusively for analysis by means of data mining.

The process step of data preparation combines the steps selection, preprocessing and transformation of the process model KDD. This summary makes it more difficult to trace the exact steps taken to prepare the data.

In the context of the research work under investigation, the deployment of the generated models was only discussed with the aim of optimizing them.

Defining the Ideal Process Model

Based on the investigations carried out, a process model can be derived, which is particularly suitable for the discovery of knowledge in sensor-based indoor climate data (Fig. 4).

The investigated process models KDD and CRISP-DM show specific weaknesses which exclude the general recommendation of one of these process models for application to the investigated research work.

However, in addition to these weaknesses, specific aspects of the process models could be identified, which very well describe the actual procedures of the investigated research work. In addition, some authors carried out measures, which are not intended as process steps in the process models examined.

By merging the benefits of the established process models and the best practices, almost all necessary process steps of the actual procedures can be represented in a unified and optimized process model.

This process model is ideal for understanding research work analyzing sensor-based indoor climate data as well as for planning and performing new research on this kind of data.

The steps of this process model can be summarized as follows:

1. Business understanding: The purpose of this process step of the CRISP-DM process model is to understand the task to be solved and the respective context ([3], [21]). The research work examined focuses exclusively on scientific issues that are almost exclusively located in the context of the energy efficiency of buildings. To deal with these questions, a high degree of domain knowledge from several areas of expertise is required, e. g. indoor climate, sensor technologies, human factors and architecture. In addition, there are relationships between the various domains. Changes in human behavior or the weather, for example, have an effect on the indoor climate.

Also part of this process step is the clarification of legal questions. E. g., the collection of the ground truth by means of cameras is prohibited under labor law.

2. Data understanding: This process step is used to determine the required data and to analyze the available data for its suitability for data mining ([3], [21]). For each context examined, different context data is required. For example, the context data for determining the presence differs greatly from the context data for determining the comfort of persons in the room.

3. Data collection: Data collection is not a process step of the examined process models. The required indoor climate and context data

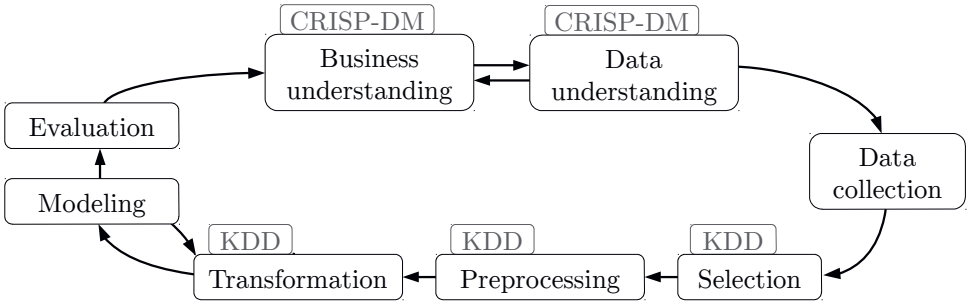

Fig. 4: Optimized process model for the discovery of knowledge in sensor-based indoor climate data

are collected based on the information from the previous process steps. This includes, among other things, the selection and installation of the sensors required for measuring the necessary room climate and context data, as well as the rooms provided for this purpose.

4. Selection: The aim of this step is to select and merge the data to be examined using all collected data ([3], [10]).

5. Preprocessing: During the preprocessing, the previously selected and integrated data is to be prepared for the data mining ([3], [10], [21]). This step is independent of the used data mining method. Part of this process step is the so-called data cleansing, i. e. the information-neutral removal of errors in the data. Another part of data preprocessing is the data reduction. In this process, subsets of the selected are formed in order to reduce the calculation effort during the data mining.

6. Transformation: The transformation step is used to prepare the data for a specific data mining method ([3], [10]). This step can be used, for example, to create new attributes that are required as class labels for classification methods.

7. Modeling: Various methods can be used to detect patterns in the processed indoor climate data. In most of the research work examined, classification was used to search for patterns in the respective indoor climate data. Although different forms of classification were used for this purpose, the basic procedure for generating classifiers was retained. This means that for training and testing the classifiers, corresponding training and test data was generated ([3], [10], [21]). When generating this training and test data, attention was also paid to ensure that the data was split according to both stratification and cross-validation. Only the authors of [4] performed both cluster and association analysis. The results of the cluster analysis were used for the creation of association rules.

8. Evaluation: Various metrics can be calculated to compare the results obtained from the previous data mining step ([3]. [10], [21]). In all research work that used classification to search for patterns in the indoor climate data collected, the accuracy of the respective results was calculated. [4] evaluated the cluster quality using the Davies-Bouldin index and the quality of the association rules using support and confidence.

9. Repetition: The examined research work concludes with the identification of open problems or the formulation of new research questions. I. e. it is not the immediate use of the results obtained in an industrial context that is proposed, but rather the optimisation of the respective model by means of further scientific studies.

Discussion

Two of the investigated research projects applied specially defined process models. This section discusses to what extent these process models are suitable for the discovery of knowledge in sensor-based indoor climate data.

The process model of [1] is based on the process model CRISP-DM. However, it does not cover all steps of the original process model (Fig. 5).

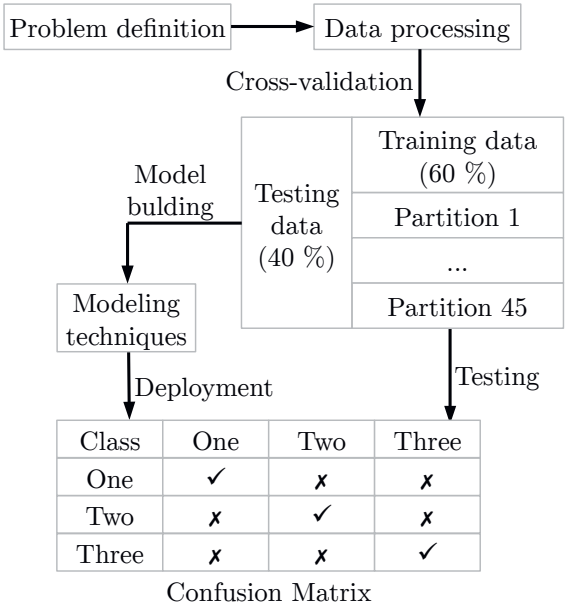

Fig. 5: Applied process model by [1]

For example, the steps data understanding and data preparation are combined in the step data processing.

As with all other research work, the description of the facts is given directly at the beginning of [1]. However, in the first step of its process model, [1] will limit itself to describing the problem to be solved. The process model proposed by [1] does not reflect the actual course of the data mining project.

[1] assigns the data collection to the second process step. This measure, though, is not part of the original process model CRISP-DM. Splitting the data records using cross-validation is part of the modeling according to the original definition of the process model and not to be seen as a separate step.

Furthermore, [1] is limited to classification procedures for the generation of models. Thus

the use of a confusion matrix is sufficient for the evaluation of the classifiers. If other data mining methods are used in future investigations, the process model must be adapted accordingly..

Furthermore, the process model does not indicate any jumpbacks to previous process steps.

The authors of [4] described in their research work a process model with three steps (Fig. 6).

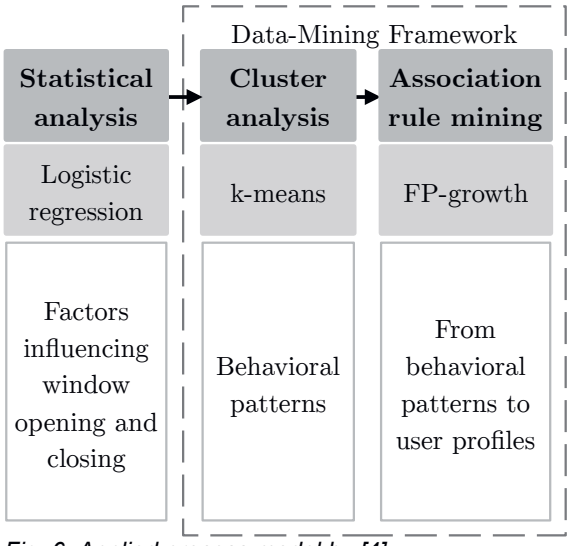

Fig. 6: Applied process model by [4]

Similar to the KDD process model, the process model in [4] mainly describes measures for data preparation and analysis. In contrast to the CRISP-DM process model, business understanding and data understanding deployment are not part of the process model.

The first step of the process model in [4] is to reduce the amount of data to be analyzed. Thereby, those indoor climate data sets are determined, which have the greatest influence on the investigated context. This reduction in data volume is independent of the subsequent data mining procedures.

The second and third steps of the process model of [4] are summarized as a "data mining framework". These process steps deal with the actual modeling by using clustering and association analysis. In the established process models, this procedure would be described as a return to a previous process step.

The process model of [4] does not provide for the evaluation of the generated model. Nevertheless, [4] conducted such an evaluation during their research.

Due to the fixed procedures for data preparation and analysis, the process model of [4] cannot easily be applied to other research work.

Conclusion

Within this paper, a process model was proposed that is suitable for the discovery of knowledge in sensor-based indoor climate data.

The established process models could only be mapped to the research work examined in accordance with their focal points. The process model KDD is particularly suitable for describing data preparation, whereas the process model CRISP-DM is particularly suitable for describing the situation.

The derived process model contains the process steps of the established process models most frequently used in the research work under investigation. In addition, data collection was added as a process step. This process model not only allows new data mining projects to be carried out transparently and comprehensivly in indoor climate data recorded by sensors, but also makes it easier to understand work already carried out in this area.

In the next step, the process model should be tested and refined by its practical use in data mining projects. Furthermore, the comparison of the process model with less known process models, which were not used for this study, would be conceivable.

References

[1] A. Ahmed, N. E. Korres, J. Ploennigs, H. Elhadi, K. Menzel, Mining building performance data for energy-efficient operation, Advanced Engineering Informatics 25(2), 341–354 (2011); doi: 10.1016/j.aei.2010.10.002

[2] R. Andersen, V. Fabi, J. Toftum, S. P. Corgnati, B. W. Olesen, Window opening behaviour modelled from measurements in Danish dwellings, Building and Environment 69, 101–113 (2013); doi: 10.1016/j.buildenv.2013.07.005

[3] J. Cleve, Data Mining, 2nd ed., Berlin, De Gruyter (2016)

[4] S. D'Oca, T. Hong, A data-mining approach to discover patterns of window opening and closing behavior in offices, Building and Environment 82, 726–739 (2014); doi: 10.1016/j.buildenv.2014.10.021

[5] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, Energy and Buildings 42(7), S. 1038–1046 (2010); doi: 10.1016/j.enbuild.2010.01.016.

[6] A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, H. Hjalmarsson, K. H. Johansson, Blind identification strategies for room occupancy estimation, IEEE, 1315–1320 (2015)

[7] T. Ekwevugbe, N. Brown, V. Pakka, D. Fan, Realtime building occupancy sensing using neural-network based sensor network, 2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST), IEEE, 114–119 (2013)

[8] V. Fabi, R. V. Andersen, S. Corgnati, B. W. Olesen, Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models, Building and Environment 58, 188–198 (2012); doi: 10.1016/j.buildenv.2012.07.009

[9] U. M. Fayyad, Advances in knowledge discovery and data mining, Menlo Park, AAAI Press (1996)

[10] J. Han, M. Kamber, J. Pei, Data mining: Concepts and techniques, 3. ed., Amsterdam, Elsevier/Morgan Kaufmann (2012)

[11] C. Jiang, M. K. Masood, Y. C. Soh, H. Li, Indoor occupancy estimation from carbon dioxide concentration, Energy and Buildings 131, 132–141 (2016); doi: 10.1016/j.enbuild.2016.09.002

[12] A. M. Khudhair, M. M. Farid, A review on energy conservation in building applications with thermal storage by latent heat using phase change materials, Energy Conversion and Management 45(2), 263–275 (2004); doi: 10.1016/S0196–8904(03)00131–6

[13] R. J. Meyers, E. D. Williams, H. S. Matthews, Scoping the potential of monitoring and control technologies to reduce energy use in homes, Energy and Buildings 42(5), 563–569 (2010); doi: 10.1016/j.enbuild.2009.10.026

[14] T. H. Pedersen, K. U. Nielsen, S. Petersen, Method for room occupancy detection based on trajectory of indoor climate sensor data, Building and Environment 115, 147–156 (2017); doi: 10.1016/j.buildenv.2017.01.023

[15] G. Piatetsky-Shapiro, CRISP-DM, still the top methodology for analytics, data mining, or data science projects (2014)

[16] I. A. Raja, J. F. Nicol, K. J. McCartney, M. A. Humphreys, Thermal comfort: Use of controls in naturally ventilated buildings, Energy and Buildings 33(3), 235–244 (2001); doi: 10.1016/S0378–7788(00)00087–6

[17] S. H. Ryu, H. J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques, Building and Environment 107, 1–9 (2016); doi: 10.1016/j.buildenv.2016.06.039

[18] C. Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing 5(4), 13–22 (2000)

[19] P. Wargocki, D. P. Wyon, J. Sundell, G. Clausen, P. O. Fanger, The Effects of Outdoor Air Supply Rate in an Office on Perceived Air Quality, Sick Building Syndrome (SBS) Symptoms and Productivity, Indoor Air, 222–236 (2000)

[20] P. Wargocki, D. P. Wyon, Ten questions concerning thermal and indoor air quality effects on the performance of office work and schoolwork, Building and Environment 112, 359–366 (2017); doi: 10.1016/j.buildenv.2016.11.020

[21] I. H. Witten, C. J. Pal, E. Frank, M. A. Hall, Data mining: Practical machine learning tools and techniques, 4. ed., Cambridge (2017)

[22] Q. Zhu, Z. Chen, M. K. Masood, Y. C. Soh, Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains, Energy and Buildings 141, 125–133 (2017); doi: 10.1016/j.enbuild.2017.01.057