POLITECNICO DI TORINO

Corso di Laurea Magistrale

in Ingegneria Energetica e Nucleare

Tesi di Laurea Magistrale

Mining operational data for anomaly detection in buildings

Relatore

prof. Alfonso Capozzoli

Candidato

Daniele Mauro Mazzarelli

Correlatori

Ing. Marco Savino Piscitelli

Ing. Silvio Brandi

Aprile 2018

To Antonella and my family,

that always support me.

Summary

1. Introduction and objectives 1
2. Data Analytics in energy and buildings: Literature review 3
   - 2.1. Prediction of energy consumption 5
   - 2.2. Profiling 6
   - 2.3. Benchmarking 6
   - 2.4. Fault detection and diagnosis 7
   - 2.5. Occupant behaviour 8
   - 2.6. Methods 9
     - 2.6.1. Cluster 14
     - 2.6.2. Classification tree 17
     - 2.6.3. Association Rules Mining 20
       - 2.6.3.1. Definition 20
       - 2.6.3.2. Types of Association Rules Mining 22
       - 2.6.3.3. Temporal Association Rules Mining 24
       - 2.6.3.4. Algorithms 25
       - 2.6.3.5. Discretization 33
3. AFDD methods' application 36
   - 3.1.1. Fault Detection and Diagnosis in AHU 37
   - 3.1.2. FDD methods applied to other building energy cases 48
4. Methodology 55
   - 4.1. Description 55
   - 4.2. Intermediate developing phases 65
5. Case study 69
   - 5.1. Dataset description 69
   - 5.2. Selection of the variables 77
   - 5.3. Application of the methodology 82
6. Results 89
   - 6.1. Extracted rules interpretation 89
   - 6.2. Diagnosis results interpretation 113
7. Discussion 119
8. Conclusions 121
9. Acknowledgement 123
10. References 124

11. Glossary 127
12. List of figures 129
Appendix 131
A. Faults coding 131
B. Variables coding 132
C. Distribution and discretization 133
D. Extracted rules resume 140
E. Faults implication of extracted rules violation 145
F. Faults-rules matrices 149
G. Score calculation 151

1. Introduction and objectives

The employment of ICT is becoming more and more penetrating in everyday life environment, improving the quality of life and bringing automation at every level of the surrounding world. Smart devices exploiting ICT allow actions previously impossible to do because of the lack of tools permitting these actions or the necessary information.

The fortune of the "Smart" approach lies in the possibility to be applied to several field and several devices, providing immediate and sensible improvements to users' life. The enabling of previously "unthinkable" operations, it is permitted by the detailed data collection conducted by Smart devices, which gather more information helpful in enhancing the grade of the description of the environment which they are integrated.

Supposing a total penetration of Smart devices in the environment, the awareness of the system's functioning is complete for the user, which is able to manage the system on the go, eventually preventing an expert's inspection of the system, since these devices can carry out monitoring, storing and elaboration of data regarding activities or system's parameters. The great potentiality of this approach is the retrieving of helpful information for a deep understanding of the environment in an easy to understand and accessible way, enhancing the awareness and control of the user.

In any case, the higher detail level of description of the environment brings necessarily to a higher dimensionality of data, since the number of quantities monitored to increase the detail of description should be increased. The higher dimensionality of the data handled represents an obstacle hard to overcome for the exploitation of the collected data. For this reason, the extraction of knowledge from this data is prerogative of automated Data Analytics methods that can handle this huge amount of data.

This kind of procedure is becoming more and more popular in various fields, due to its straightforward approach and immediate improvements obtained. Building's energy is one of the fields which this approach has been applied too and its study results very active in recent years.

The interest in the study of energy applications of Data Analytics lies in the great possibility of energy saving by the reduction of wastes, pointing to a more sustainable environment. It is worth noting that the rationale of the energy saving by a better management of an energy system is completely inserted in a sustainability perspective for a global reduction of the anthropogenic greenhouse effect, thanks to the reduction of primary energy requirement and emissions of GHG.

Considering that buildings' energy for non-industrial usage accounts for a total consumption of 56.9% (U.S. EIA, 2013), which is responsible for the 24% of the whole world 's anthropogenic CO_2 emission (Day, et al., 2013), the margin of improvement in this sense is wide, supporting the effort in studying building's energy.

In building's energy field, the monitoring of a Building Management System (BMS) can interest HVAC systems as well as electrical power related systems (e.g. electrical appliances), in order to keep track of the functioning and the consumption of the energy systems of the building.

The majority of recent studies in this field concerns HVAC systems, mainly because this kind of systems represents the 50% of energy consumed by a commercial building, thus its influence on building energy balance is high. (DOE, 2016)

What legitimizes even more the study of the employment of Data Analytics for this kind of application is the fact that AHU's energy consumption reaches the 40% of the total energy consumed by industrial sites, because of inefficiencies. (HVACSWGSpin12007, 2011)

In fact, due to lack of proper maintenance, failure of components or incorrect installation, AHUs are often run in inappropriate operational conditions, therefore the investigation on the improving of this kind of equipment's management is legitimate. Considering how common is this kind of equipment, strategies to reduce energy waste for these systems can drive to a great saving in the global building energy usage. Actually, the portion of total energy consumption of developed country related to HVAC systems is 10-20% (Pérez-Lombard, 2008), therefore the energy saving in this fields can bring to great advantages in a global sense. In fact, it has been estimated that the potential energy saving due to Fault Detection and Diagnosis strategies is 10-40% according to age and condition of the equipment, maintenance practices, climate and building use. (Kang & Golay, 1999)

Given that, it is clear how the developing of Data Analytics methods for building's energy applications, represents a great way to reach significant improvements in the rational use of energy.

The objective of this dissertation is to portray the typical Smart approaches of Data Mining to energy building applications, dealing with the common Data Analytics methodologies, particularly focusing on the Fault Detection and Diagnosis of energy systems. Fault detection is one of the scopes of Data Mining for building's energy applications and it has a great potentiality concerning the achievement of energy saving by the enhancement of the management of the system and the improvement of performance of the system. In particular, this dissertation is intended to provide a framework of the common methods used for this sake and to propose a novel methodology to perform Automated Fault Detection and Diagnosis. The methodology proposed has been developed to fulfil the requirement of being eventually applicable in a real-time implementation and support the management of the system by automatically spot the presence of anomalies and provide a guess of their causes. Furthermore, this novel methodology aims to provide easy to understand results, so to be exploitable even by non-expert users. Guaranteeing an easy comprehension of the results found is a crucial point since for a better management of the system is necessary the comprehension of the condition of the system to all players, at the same time making the results exploitable without banalizing them.

In the following sections, a review of main Data Mining applications and methods is reported, in order to provide an overview of the problem and the techniques employable for the purposes. After that, a novel methodology to perform Automated Fault Detection and Diagnosis for building's energy systems is proposed with a description of the whole procedure. Finally, the test and validation of the methodology proposed are reported, presenting and discussing the results obtained.

2. Data Analytics in energy and buildings: Literature review

The approach considered in this dissertation is the one typical of the Data Analytics, more specifically of Data Mining, which finds great interest in literature but still low employment in real building energy applications, although the advantages are many.

Data Analytics approach considers the possibility to extract helpful knowledge from the data derived from the monitoring of the system, so the analysis of data leads to the coming out of information that unless are ignored or considered impossible to obtain.

Data Mining is a branch of the Data Analytics which gathers all the methods able to "mine" the hidden knowledge from the data available, providing information impossible to obtain differently.

Considering the unconventional Data Mining nature for energy application, the expert intervention is mandatory for the full understanding and exploitation of the results obtained by means of these promising techniques. The knowledge of the field of application and the framework of the problem, allow the expert to transform sterile results in actual intervention for optimization of the energy system and improving the management and performance of the system itself.

Data Mining (DM) methods serve as a tool for extraction of hidden knowledge from the huge amount of data provided by the BMS, but the system's expert can interpret the results and guide the process of knowledge extraction in order to achieve the goal of the analysis.

The structure of these methods can be summed up in four main phases (Figure 1), applied in cascade, but each of them is fundamental for the analysis.

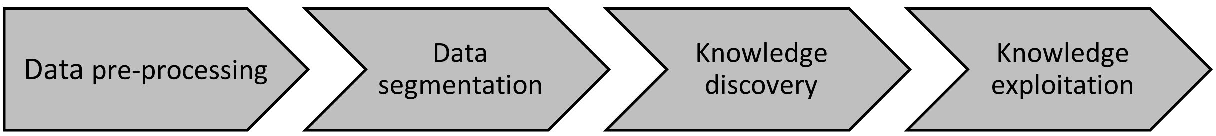


Figure 1 General Data Mining framework

Raw data from the monitoring system must firstly be manipulated in order to prepare and characterize them during the data pre-processing phase. Data preparation consists in the data cleaning, data transformation and data reduction which are three crucial steps in order to firstly manage inconsistent values, outliers or missing values, then to scale data to treat contemporary variables with different values' ranges and finally reduce the dimensionality of the data to enhance computational efficiency. After the data preparation, data characterization ends the data pre-processing phase with the visualization of the data just manipulated. The visualization of the data results essential for the first comprehension of the data processed and the problem facing with.

Data segmentation phase results helpful in the improving the knowledge acquisition since in this phase the correlation among variables contained in the data are used to subset it and treat separate groups of data, rather than the whole dataset. In this phase, the domain expertise plays an essential role in the choice of the best way to split data, without degradation of information provided, unless significant test or clustering analysis are adopted to split data according to the intrinsic nature of the variables themselves.

The third phase can be considered as the effective knowledge mining phase, matter of fact that in this phase are applied the methods supplying the hidden knowledge that DM process is looking for. In this phase the approach is twofold, as supervised or unsupervised techniques can be used according to the kind of analysis is intended to perform or the quality of data available. Unsupervised methods are able to provide results without the intervention of the system expert in the formulation of the model or inserting specific input data coming from his knowledge, unlike for the supervised methods. Through years several methods have been developed to fulfil to the necessity coming from the available knowledge of the system or kind of data, for both the approaches.

The last phase represents the interpretation and selection of the knowledge extracted in the previous phase. In this phase, the role of the system expert is fundamental in order to transform the results of DM process in interventions to perform for energy efficiency improvements. (Capozzoli, et al., 2016)

The knowledge extracted from data by mining methods can be exploited only if are functional for supporting the decision making, therefore if they are relevant for the scope which they are generated for. For this sake, the knowledge extracted should accomplish the goal of being informative and easily understandable. These requirements are necessary due to the fact that an energy miner system is composed by four roles which all should be able to manage, without any problem, the information provided by the analysis and to exploit the knowledge provided.

The process of DM can be conducted on different levels; since the system's energy miner roles are the Energy Manager, the Energy Analyst, the Consumer and the User.

The first one is in charge of the providing of energy services, so this player treats high-level information analyzing data at a scale much large than the building one, as can be the city district scale.

The second player is the consumption expert and has complete awareness the system at the component level. This player treats data with a further level of detail, in order to comprehend the functioning of the system and extract the required knowledge from the huge amount of data to analyze.

The Consumer is interested in managing the system for efficiency optimization and energy saving preserving the desired level of comfort. The information provided to this player should be as intuitive and clear as possible, since his awareness of the system is limited, despite the fact he plays an important role in the decision-making process.

The last player is the person who uses the system and is interested in the reduction of energy expenditure keeping the desired comfort parameters constant. The information at this level should regard only the practical actions and behaviour that can be done to achieve the goal. (Capozzoli, et al., 2016)

It is worth noting that a single person can impersonate one or more than one role, even though the level of information treated is different for each role.

In addition to the different levels of investigation of DM process, the process is made more complex by the different nature of data treated.

In fact, energy-related data collected by the BMS, describe several internal and external aspects of the building, matter of fact that the description of energy processes involves different aspects, such as weather, hardware information, operational parameters, user utilization of the building or physical parameters. The kind and the amount of data available represents a fundamental point to consider during the study of a system, as long as it influences the kind of implementation applicable and the depth of investigation permitted.

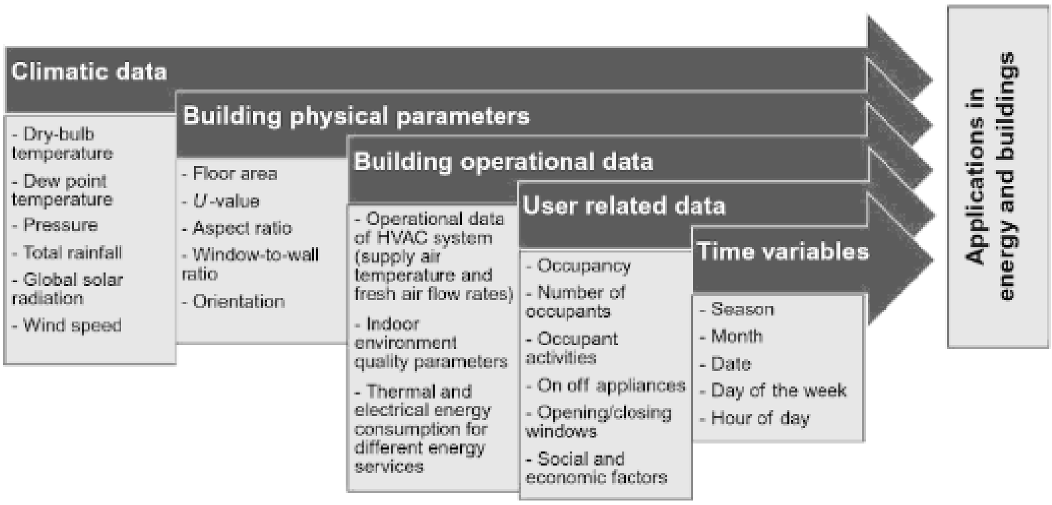


Figure 2 Classification of energy applications-related BMS data (Capozzoli, et al., 2016)

The amount of data available can be enormous, hence hard to use without any automatic method able to extract knowledge from this, alternatively sterile, data. Data mining processes fit this purpose and demonstrate itself as an effective way to obtain otherwise unknown knowledge.

The data mining process is guided by the goal of the analysis at every step of the procedure, from the setting of methodology's parameters to the knowledge extraction from the results. For this reason, it should be done a distinction among the main scope of investigation for energy applications of data mining, as they can change the way the methodology is applied, and results are interpreted. (Capozzoli, et al., 2016)

2.1. Prediction of energy consumption

One of the most studied application of data mining in building energy field is the prediction of energy consumption. The prediction of the behavior of the system and its energy consumption results very helpful in several applications. In fact, the accomplishment of this goal represents an important tool for the performance evaluation, operational optimization, evaluation of Demand Side Management strategies or Fault Detection and Diagnosis, as well as for energy saving purposes.

Even though the implementation of the predicting model does not represent a great concern, the accuracy of input parameters and the retrieval of needed data and necessary information for the construction of a robust model represent a delicate point. In fact, monitored quantities sometimes cannot be sufficient for the construction of a good predictive model, since several are the aspects to consider in order to predict the behaviour of system getting more and more complex. Those aspects influencing the system, conditioning the prediction of the energy consumption, can be as internal (i.e. appliances usage, indoor comfort parameters) as external (i.e. climatic condition), which are indirectly implied in the system's behaviour. (Capozzoli, et al., 2016)

In addition to this problem, the accuracy of the sensors' data has to be evaluated too, due to the poor maintenance of these devices, as well as bad calibration or installation. Sensors' expenditure for building monitoring is often not a priority, matter of fact the application is not critical, therefore the low accuracy of sensors' data can derive from low resolution or low-quality measurement devices.

For this reason, the construction of prediction models should be conducted on the track of a rigorous methodological process to obtain the best accuracy and robustness possible from the available data. (Capozzoli, et al., 2016)

2.2. Profiling

Another building energy application of data mining is the energy profiling, which provides typical patterns of the energy consumption at whole building level or component by a component on different time scales. This information is useful to detect the deviation from an expected trend of a particular quantity, so to find anomalies in functioning, rather than that it is possible to find typical load profile. The detection of the anomalous profile with respect to the typical one helps to reduce energy waste in case of the anomalous functioning of the system, while the typical load profile identification itself can be exploited by energy provider to create specific tariffs for the class of customers with similar load profile.

As a result, the load profile mining provides a piece of time sequences that describes that recursive and typical functioning of the system. These pieces are extracted from time indexed sequences, namely time series, which are characterized by large dimension and constant updating during the monitoring. Time series data are as complex as interesting because the knowledge extracted with patterns mining techniques results very helpful and with straightforward utilization. (Capozzoli, et al., 2016)

One of the main applications of temporal data mining is, indeed, the extraction of motifs and discords from the time series, namely the identification of common and uncommon patterns within the temporal sequence. The study of these patterns raised a particular interest for load profiling, matter of fact they describe an approximation of the behaviour of the quantity which they are referring to, therefore the research along the time series is not constrained by the exact value, so referring to only a particular condition. For instance, the same motif can be identified in high consumption condition as well as low consumption condition. In this way, the behaviour of the system is independent of the seasonal or calendar effect, since it is generalized and scaled to different levels of consumption. (Capozzoli, et al., 2016)

2.3. Benchmarking

Another possible application of data mining to energy building is the benchmarking of the performance of the building itself, by means of the calculation of an unequivocal parameter so to assess the bias from a reference value considered as optimum or desired value. In order to get a reference value a set of similar buildings is used, so to have a statistical relevant measure of the value of the benchmark parameter considered.

For the sake of generalization of the benchmarking parameter, it is important to take into account internal and external factors influencing the performance of the building, so to avoid having values specific to particular conditions, eventually anomalous or rare. Possible internal factors influencing the performance indicator can be the number of occupants in the building, which, for instance, can influence the internal load and changing the heating requirement. On the other hand, external factors that make fluctuate performance indicators can be weather condition, especially if exceptional or severe. (Capozzoli, et al., 2016)

The performance indicator of a benchmark analysis can be exploited for comparing the building with other similar ones with public benchmark methods, rather than to compare the building's performance value with previous or expected ones with internal benchmark methods.

Public methods are helpful for municipalities that are intended to track the energy performance of their pool of building, so to schedule interventions for improving energy performance of buildings showing poorer performance with respect to the others.

Once a building has obtained an unequivocal parameter's value, by means of a public benchmark method, its energy performance is quantified, so it can be classified according to the obtained value.

In this kind of application of data mining, the results are often exploited for funds allocation or interventions schedule, so the results must be easily understandable even without any expertise of the system. Therefore, the indicator should facilitate the comprehension of the results, in favour of a complete awareness of the condition of the building. (Capozzoli, et al., 2016)

2.4. Fault detection and diagnosis

Starting from the point that the system is not always running under the expected conditions, anomalies spotting becomes a particular topic of investigation in building energy field.

Anomalies in a building energy building are represented by unexpected operational condition, which can be caused by improper installation, bad operation or the occurrence of faults.

Unattended values of energy-related data, whatever is the cause, drive to performance degradation or, in case of high deviance from the expected condition, possible damage to the system itself. For this reason, fault detection and diagnosis should always be correlated to an indication of following intervention schedule or best practice suggestion, so to fix the cause of the anomaly. (Capozzoli, et al., 2016)

In this sense, data mining can provide an automatic method of detection and identification of faults causing the anomalous operation of the system. The growing diffusion of BMS's data offers the opportunity to exploit analytical methods to assess every observance and determine if it is whether fault related or not and, eventually, identify the cause.

The most important advantage of the application of data mining in this field is represented, indeed, by the possibility to spot the cause of the energy waste without any intervention in the diagnosing procedure. (Capozzoli, et al., 2016)

In literature there are many examples demonstrating how data mining for this field of application can be focused at different level of investigation, since the analysis can be conducted at whole building level or component level. Component level assessment, namely bottom-up approach, guarantee the description of the system's behaviour at a grade of detail that fault isolation can be performed, while building level one, namely top-down approach, is uncapable to investigate the causes of the anomalies.

The detection of faults is strictly related to the corresponding energy performance degradation; therefore, it is fundamental associate this kind of analysis to the evaluation of energy indicators. The best way to evaluate the performance degradation is the calculation of indicator, which are commonly used in benchmark application. This approach guarantees a full understanding of the condition in which the system is running. (Capozzoli, et al., 2016)

Despite the advantages in improving the awareness and the understanding of the results, the performance assessment is often skipped in the actual proposed methodologies, similarly for the fault evaluation.

The diagnosing phase is still representing a promising procedure, but with poor utilization, matter of fact that the actual implementation requires a high distribution of sensors to monitor all system's component, which means higher costs.

In the direction of overcome this issue, recent studies are reduced the number of sensors needed through the use of low-cost algorithms, which create analytical redundancy and virtual sensing.

Actual methods can even handle multiple faults at the same time linked by cause-effect logic even on different system's level. (Katipamula & Kim, 2017)

The implementation of methods applied to achieve the fault detection and diagnosis goal can be considered a great way to reduce energy waste and rationalize the use of the system. (Capozzoli, et al., 2016)

2.5. Occupant behaviour

One of the main unknown in the energy building data utilization is the consideration of the occupants' behavior. The occupants' behavior represents the way the system is used; therefore, this factor has a dramatic impact on energy consumption.

The way the occupants of a building act, influencing the buildings' energy related quantities, is completely unpredictable and hard to model, since it is impossible to schedule exactly the way people should use the building or impose to people good practice actions to unconditionally stick to. Moreover, the diversity of human nature provides an additional grade of uncertainty to the reaction of occupants to indoor conditions or system' s operation changes.

It is worth pointing that occupants' behavior can be completely independent from optimization of system's performance, since occupants can set the operating parameters differently with respect to the expected case, for the sake of preserving their comfort or other reason not strictly related to the energy field. (Capozzoli, et al., 2016)

Factors influencing the occupants' behavior are several, but they can be grouped in five categories considering internal and external parameters. These categories are physical, contextual, psychological, physiological and social. The first comprises physical parameters such as internal and external temperature or humidity, wind speed, illumination or odor. The second refers to the building characteristics such as its dimension, insulation or type of HVAC system employed. The third refers to the drive to satisfy needs. The fourth takes into account age, activity level and health of the occupants. The last one considers possible interaction among occupants. (Capozzoli, et al., 2016)

For this reason, data mining technique have been employed to extract useful information to predict or model the way people use the assessed system and apply modification to the system that can result in improving of the performance of system during its running.

Once the occupants' behaviour has been understood, a list of good practice action can be provided to occupants, so to correct the bad behaviour had, in addition to possible adjustment to the system's parameters in order to encounter occupants' needs and reduce the energy waste. (Capozzoli, et al., 2016)

2.6. Methods

More than the other applications of DM in building energy field, Automated Fault Detection and Diagnosis (AFDD) gathers particular attention of worldwide researchers, thanks to its easy applicability and great potentiality in energy saving rather than in management improving.

For this reason, many methods have been implemented to achieve this goal, with different grade of knowledge of the system and different level of computational effort required, in order to adapt the method to the user's demand and available expertise.

In the direction of making the description of the common methods used for AFDD in building energy field straightforward, Katipamula and Brambley provided a classification that gathers all the examples present in literature. (Figure 3)

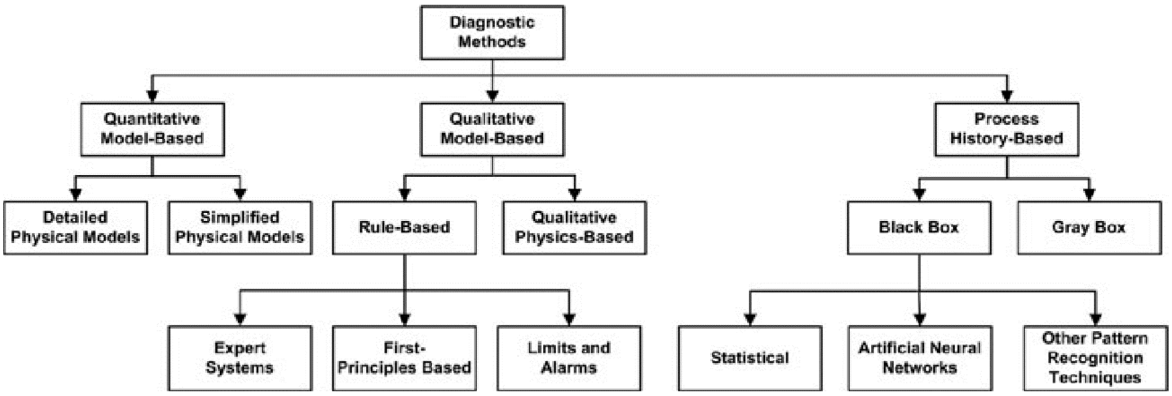


Figure 3 Classification of AFDD methods (Katipamula & Brambley, 2005)

According to this classification (Figure 3), the whole range of methods is divided in three main categories, which are quantitative model-based, qualitative model-based and process history-based, according to the way the model used for the investigation is constructed and how it exploits available data.

Most of these methods are able to perform fault detection by means of residual analysis, which needs an expected value to compare to measured observance. Therefore, the methods exploiting the residual analysis to detect faults differ for the estimation of the expected value, which can be conducted by means of models of the system analysed. Models constructed through these methods can have different level of accuracy as well as implementation cost. (Katipamula & Kim, 2017)

Starting with the methods with the lowest implementation effort requirement, process history-based methods come out, demonstrating the most used type of methods used at the moment. The reason why these methods are preferred is that they are easy to implement and do not require a deep knowledge of the system which they are applied to.

These methods consider data-drive approach, since the knowledge is extracted from the data available, without any exhaustive model or prior system 's deep knowledge requirement. Mainly these methods can be classified in grey and black box, according to the kind of parameter analysed by the method. Grey box methods are able to estimate parameters that can express a linkage with physical quantities driving the system, but they are not capable to model the system behaviour. Black box ones consider mathematical expressions treating only inputs and outputs of the system, but the parameters estimated are not attributable to the actual physical behaviour. Due to the low requirement and low cost of black box models, they are the most studied, since they represent a valid option for AFDD in real time cases. Typical black box methods are statistical ones, Artificial Neural Networks (ANN) and pattern recognition techniques. Statistical methods are several, since this category embraces polynomial regression, autoregressive (AR), principal component analysis (PCA) and partial least squares methods. (Katipamula & Kim, 2017)

Polynomial regression is a technique founding on the construction of a linear mathematical expression to describe the relationship between parameters' expected values and the actual values coming from the BMS. For the application of this technique the expected values are considered constant and independent from the operating condition. As a consequence, the method exploits residual analysis to detect anomalies in the functioning of the system, making comparisons between the actual and the expected values of the considered parameters.

Autoregressive methods have been demonstrated helpful in time series data forecasting and modelling, since it makes prediction of the parameter's present values using a stochastic difference equation applied to previous data instances of the same time series. The results of the AR model are corrected by means of a optimization stage that minimize the overall error between the model output and the actual input-output. Even though this technique provides quite accurate results even in non-linear cases, without any deep knowledge or accurate model implementation, it requires a great amount of data for training, in order to reach a good level of accuracy.

Another common statistical method is the PCA, which bases its widespread use in the capability of reducing the dimensionality of space of system's variables, making it helpful in complex system's fault detection cases. It is able to detect faults due to the fact the anomalous data are usually very different from the median of the normal data, therefore the great distance of the values discriminates the faulty data from the faultless data. This technique's criterion is as simple as limiting, matter of fact it is unable to take into consideration the causality of the relationship among faulty data. As a result, the lack of cause-effect recognition, prevents the use of PCA for diagnosis.

ANN represents an alternative to statistical black box models, which results good for identifying even a complex relationship between a set of network inputs and outputs.

By means of a network training procedure the functional relationships are extracted from data, providing patterns of whether normal or faulty condition.

For this purpose, this technique exploits sets of nodes distributed in multiple layers, passing data each other as computational elements. As a consequence, it can find the functional map of inputs and outputs using input/output training pairs.

The main issue to face with is the necessity of enough data for training, so to enable the classification of faulty and normal data once the corresponding patterns are identified. Although the data requirement is high, ANN is a largely diffused technique thanks to the fact it can treat even non-linear system and classify raw BMS data directly, making this technique a good candidate for on line AFDD.

The last category of black box methods is the pattern recognition, which mainly classifies the observed condition comparing the observed pattern with the ones associated the faulty conditions. Techniques as Support Vector Machine (SVM), Gaussian mixture or Bayesian network can be used to generate a set of patterns embracing the various faulty and normal condition and successively compare the observed behaviour of the system with the ones defined by the patterns in the set generated in the training phase. (Katipamula & Kim, 2017)

The comparison of the patterns allows the identification of the state of the system, so to detect and isolate the eventual fault occurring.

Association rules mining (ARM) can be useful in the pattern extraction phase, since it represents an effective unsupervised way to find correlation among variables of dataset. The implication of ARM can be helpful in finding the common patterns related to the normal operation in a fault-free dataset or common patterns of faulty conditions in dataset containing data related only a specific fault. Perpetrating the rules extraction in a series of dataset containing a single fault data a time, it is possible to map the whole range fault considered and obtain the reference pattern to compare in an on line AFDD application.

Due to its unsupervised nature, ARM is appealing for cases of poor knowledge of the system to analyse, therefore its implication as pattern recognition technique for AFDD is largely studied.

Instead of the employment of the black box methods, another way to exploit process history data is the implication of grey box models. These models are able to combine historical data with system's physics principles to construct a model that lower the data requirement of black box models. In fact, the grey box models found the estimation of parameters necessary for the analysis on the formulation of mathematical expression which takes into account first principles or physical knowledge of the system, but the parameters are estimated using measured data training. In this way the model is able to predict the expected normal behaviour of the system, so to provide reference values in a residual analysis involving measured data. This characteristic makes the grey box models employable in real time AFDD applications. (Katipamula & Kim, 2017)

The grey box models, therefore, are able to operate even outside the limits of the training data, unlike the black box ones which are obliged to stick to the condition considered in the training data. As a result, the black box models require a great amount of data for training to cover a large range of fault conditions to detect, even though the range of recognizable faults is limit to the already encountered ones. The computational cost, for this reason, is high, even if still manageable for black box ones. Grey box models overcome the issue of high data requirement, lowering even the computational cost, by using physical principles, enabling even the recognition of cause-effect logic, so the possibility of conduct diagnosis after the detection of the faults. On the other hand, the implication of physical principles for the construction of the model, implies a high level of expertise to formulate the model parameters and the form of the model itself, which it is not needed for black box models.

If a higher level of accuracy is prescribed or a deeper knowledge of the system is available, the system can be modelled using qualitative models. (Figure 3) These models are constructed basing on *a priori* knowledge of the system without the implication of historical data for training. The mainly used kind of qualitative model-based method is the rule-based, while the other kind exploits qualitative physics to construct the model. Qualitative physics-based models use a qualitative description of the system, so to avoid the implication of time-consuming mathematical formulation or the requirement of accurate measurement or description of the system. These models are simple and have a low computational cost, but the accuracy of results is incapable to reach high levels and their implication is limited to the possibility to model in this simple way the system.

Rule-based approaches need an *a priori* knowledge of the system, in order to construct a set of IF-THEN rules, which test the faults' symptoms and classify the operational condition. In this way an inference mechanism is used to verify whether the system is running in normal condition or in fault condition on the basis of the symptoms observed. The application of the rules can drive to the isolation of the fault if the rules are constructed in a way such that different faults can be separately diagnosed. (Katipamula & Kim, 2017)

Historical data are useful for this method, in order to set properly, or fine tune, thresholds so to avoid false alarms and detect faults with the best accuracy possible. The rules can be constructed to detect faults at component level or at system level, as well as tailored for whether detection or diagnosis. As regards diagnosis performed by means of this method, the set of rules must be constructed at component's level, so to push deeper the analysis. Even though this method can classify data on the basis of prior knowledge, anomalies related to not yet encountered faults can be detected as well, even though diagnosis cannot be performed.

The source from which the rules come from is the main discriminant of this methods, since all of them are characterised by the inference structure of testing with a rule a time. In fact, the rules can derive from expert system or first principles related to the functioning of the system. In the first case, they come from a deep knowledge of the system as expert of it, while the other case considers the functioning of the system in steady state condition to apply rules based on physical laws driving the system itself.

As a consequence, qualitative model-based methods are useful in non-critical applications, due to the fact that they do not treat precise numerical parameters or expression, so with a low computational cost. On the other hand, it is worth considering that the model is strictly interlinked to the system which it is applied fitted to, as well as the available knowledge of the developer. For this reason, as the system gets complex, adding extra rules to enlarge the range of faults considered can complicate the model, jeopardizing the effectiveness of the model itself.

The kind of methods described above can reach a high accuracy but cannot describe deeply the behaviour of the system. The most accurate model type is the one constructed with quantitative methods, since they use mathematical models for each component of the system to detect and diagnose faults and their causes during both transient and steady-state behaviour. (Figure 3)

This kind of model must be validated with experimental data in order to guarantee effectiveness and accuracy of faulty and fault-free identification. Hence, the model construction cost is affected by the implementation cost and the validation cost, which can be as high as the first one if the amount of data used for validation is large. In order to construct these models, the system must be known in depth, since models take into account the mathematical equations governing the behaviour of the system. (Katipamula & Kim, 2017)

According to the kind of equations used for the model implementation, these methods can be subdivided into detailed or simplified physical models. In fact, the first class of quantitative models are constructed by means of the most accurate expression possible for each component of the system, so to obtain expression able to describe the behaviour of the system in steady state or in time depending condition. In the simplified physical models, physical quantities are calculated with lumped parameters approach and assumptions are made, therefore a certain grade of approximation is introduced, even though the accuracy of the model remains high. For this kind of quantitative models, the partial differential equation governing the system, employed in the detailed approach, are substituted by ordinary differential equation and algebraic equations.

As can be easily guessed, the computational cost is high for the implementation and the development of these models is a not negligible commitment, due to the high level of detail and the high amount of data to treat for formulation of the model and following validation. It is worth noting that, even though the model constructed is able to describe accurately the outputs of the system in both steady state and transient condition for faulty and fault-free conditions, the model is specific to the system. In fact, the generalization of the model constructed is hard to perform, since the knowledge used for the model construction is specific to the system itself, therefore for different system is necessary implement almost-brand-new models.

In view of this wide selection of methods, the choice of the most appropriate one represents a point of particular investigation. In fact, several are the aspects to contemplate for the method's choice, such as the type of system, the necessary degree of knowledge for the model implementation, the cost-to-benefit ratio, the degree of automation in the application of the model, the tolerance to false alarms and the input data required. All these aspects represent constraints driving the selection of the method towards the most feasible one. Choice usually is driven by availability of tools or interesting of the developer, but in real life implementation often this approach leads to problems. Diagnosis is not always conducted due to low resolution of data available or the enough isolation of the fault provided by detection method. (Katipamula & Kim, 2017)

The most stringent aspect among all is the data requirement of the method, since on the amount and the kind of data available limit the choice of the method, as well as the depth of the investigation. In order to perform fault detection, large data are necessary, because in cases of limited data the results' sensitivity is very poor. As regards fault diagnosis, limited data cannot embrace the condition of the system for all faults considered, therefore isolation can be hard to perform. For this sake, rule-based or knowledge-based approach can be used to obtain good results.

A particular aspect to consider is the false alarms rate, which should be as low as possible even in non- critical application as can be commercial building HVAC. It should be considered that, if the false alarm rate is high, especially in non-critical applications, users can be driven to disable the FDD system, nullifying the modelling work. Hence, thresholds should be high enough to avoid frequent false alarms, but tight enough to detect promptly the faults.

The cost feature of the methods relies only on the computational one, since the sensors required for the monitoring of the system to study do not represent a great expense, due to the non-critical application considered.

Even though these methods can represent an expense for the system manager, the benefits that they can provide are huge, matter of fact they are able to avoid energy waste, which can be easily transformed in money waste.

Any fault detection and diagnosis method should include energy and cost assessment to evaluate the quantify the loss attributable to the fault detected during the analysis. This phase is useful to study the severity of the fault, since each fault isolated is associated to the level of degradation of system's performance. In this way it is possible to prioritize the repairing intervention.

In this phase performance is taken into account comparing expected values with current ones, in order to identify how a fault affects the whole system's efficiency. Unfortunately, this phase is often skipped, so the application of these methods drives to blind schedule, not influenced by the severity of the fault found. (Katipamula & Kim, 2017)

Successively, in this section, a particular focus on the most common and promising techniques is reported, especially dwelling on cluster analysis, classification tree and association rules. These three techniques belong to the category of unsupervised methods for the knowledge extraction since the first two are able to group data just by analysing the similarity of the values included in the data, while the third technique is a pattern recognition one already mentioned above.

2.6.1. Cluster

The cluster analysis is way to divide the data into groups of similar data, with an unsupervised approach, exploiting a similarity criterion to gather observances from a dataset. Most of the cases which this technique is applied, the results obtained represents a starting point for further application of other Data Mining techniques. Even though this technique is not employed for its own sake, the information that provides is fundamental for the whole Data Mining procedure.

The cluster analysis is able to associate a class to each observance of the dataset, only basing on the information found in the data itself.

According to the kind of approach used for finding the clusters from data, the methods used for clustering analysis can be classified in hierarchical or partitional, exclusive or overlapping fuzzy, complete or partial.

Partitional versus hierarchical is the most common labelling used to classify these methods, since it defines the main approach used to construct the clusters with the data. In fact, the first approach consists in the division all observances in non-overlapping groups, ensuring that each object is in only one group. The other one constructs the clusters starting from one for each data object, merging successively all the clusters until all the objects are included in the same cluster. Considering the way this approach operates, they may be called un-nested or nested respectively. (Han, et al., 2011)

The other classification interests the way the objects are located in the clusters in terms of belonging to a single or multiple group. In some applications can be helpful to have a less stringent classification, permitting multiple simultaneous labelling. Exclusive methods do not allow the assignment of more than one label per object, while overlapping ones consider the case which the classes can refer to different characteristics, so the multi labelling can be possible. Fuzzy logic can be applied to clustering as well, assigning a membership weight for every label to each object of the data. In this way, a value from 0 to 1 gives an indication of the probability that an object may belong to each cluster. (Han, et al., 2011)

The last classification considers the distinction between methods able to use up all the objects or not in the labelling process. In fact, some methods can detect objects hard to assign to a cluster, matter of fact these objects may be dissimilar to all the cluster created, therefore they are excluded from the classification. Partial methods represent the kind of methods useful for the outlier detection, since outliers are data very dissimilar to the rest of the data, therefore its identification can be performed by excluding them from a classification of the rest of data objects. (Han, et al., 2011)

The clusters obtained from the different kind of clustering analysis methods are not all of the same kind, according to the kind of method used and the kind of data handled. For this reason, clusters can be distinguished in well-separated, prototype-based, graph-based, density-based or shared-property ones.

The first case gathers the kind of clusters which the intra-cluster similarity is high and the inter-cluster similarity is low, making clear the distinction among the clusters.

Prototype-based ones refers to the generation of clusters around prototypes, which are the reference for the similarity computation for the cluster which they belong to, such as centroids, i.e. the average of the objects in the cluster, or medoids, i.e. the most representative objects in the clusters.

Graph-based clusters are constructed on the basis of a graph representation of the data objects, therefore the connections among the objects represents a criterion of grouping.

The density-based ones take into account the location of the data objects, since the criterion of grouping upstream the generation of these clusters is the density of the positioning of the objects. According to this criterion, the cluster is a region of high density of objects, therefore the shape of the clusters can be any, since the idea of similarity is the free from the concept of actual distance among objects.

The last kind of clusters is the shared-property one, which considers the clusters as an ensemble of objects sharing some property. This definition is as general to gather all the previously mentioned categories but takes into account even other more abstracted criterion. (Miller, et al., 2018)

In order to better understand the way cluster analysis operates, some examples of common methods employed for this purpose are briefly described below.

A typical density-based method is the DBSCAN, with a center-based approach, that is an example of simple to implement density-based method. The effectiveness of this method basis on the computation of the counting of the number of points within a specified radius from a particular point. As a result, the data objects can be classified as core points, if within a dense region, border points, if on the neighborhood of the edge of a dense region, or noise points if located in a dispersed region. (Han, et al., 2011)

A region is considered whether dense not sparse according to a user-defined parameter stating the minimum number of points necessary to detect a dense region. Other than the radius length and the minimum number of points for dense regions, the user is in charge of setting the number of cluster too, since the number of centers from which compute the density evaluation is fixed during the procedure.

This method is able to perform the cluster analysis even in cases the shape of the cluster is not globular or the quantity of noise of data is high, but it has some problems in the cases the dimensionality of data is high, or the density of the clusters is various.

Another kind of method very diffused, thanks to its simple implementation, is the hierarchical clustering, which can be performed with opposite approach: agglomerative or divisive. The divisive one starts with a unique cluster containing all the objects and then recursively split it up to the creation of a cluster for each data object. The agglomerative one proceeds in the reversed sense, merging similar clusters from single objects ones to a unique cluster. (Han, et al., 2011)

This kind of method is typically represented by a tree-like scheme, namely dendrogram, reporting the order of split/merge and the level of similarity reached between clusters at the moment of splitting/merging. Alternatively, the relationship among the cluster generated may be represented by a nested cluster diagram representation.

The most used technique of this kind is the agglomerative one, exploiting different criterion of similarity for the computation of the merging of the clusters. In fact, since the calculation of the similarity of the clusters affects the way the clusters are generated, the choice of the criterion should be performed on the basis of the goal of the analysis.

The main criteria used for the hierarchical clustering are the single link, complete link, group average and the Ward's method. The first one computes the similarity on the basis of the minimum distance between the points of 2 clusters. The second one considers the maximum distance between the points of 2 clusters. The third computes the similarity as the average of the distances between the points of 2 clusters. The last criterion computes the similarity as the distance between the centroids of 2 clusters. In order to do so, the last method calculates the increasing of the sum of squared error (SSE) (Equation 1) associated to the merging of 2 clusters and minimizes it.

The SSE calculation is perfumed by the Equation 1, considering the distance between the centroids of all the k clusters and the points of the points of the other clusters.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2$$

Equation 1

Choosing the similarity criterion means drive the analysis to a certain type of cluster, since the way the similarity is computed shapes the clusters obtained. In fact, the single criterion can handle non-elliptical shapes, the complete one prefers globular shapes, even breaking large clusters. Group average and Ward's ones tend to prefer globular shapes too.

It is worth noting that an approach like this is lacking a global view of the data objects distribution, matter of fact that the merging is operated locally without a general consideration of the effects of that action. On the other hand, this way of proceeding guarantees the possibility to perform the analysis without the issue of the selection of starting points, since every point is considered equally important. Another issue related to this method is the final definition of the clusters. In fact, the merging or the splitting of the clusters is irreversible, so the distance of the clusters merging is monotonically increasing. This last feature is not respect for the only case of Ward's method, which has a global overview of the data condition by the calculation of the SSE. For this similarity criterion, it can happen that the merging at a certain level interests 2 clusters with a distance lower than the one of the clusters merged previously, if the operation brings to the minimum increasing of SSE at that step. (Han, et al., 2011)

Alternatively to hierarchical and density-based methods, another widespread approach is the partitional one of the K-means. This method is a prototype-based one, that elects as prototype the mean of the points in the cluster, with a user-defined number of clusters.

Unlike the hierarchical methods, K-means is an iterative method, which reconsider the position of the centroid at every step, in order to minimize the overall SSE. In fact, after an initialization of the position of the centroids of the clusters and the assignment of data objects to the cluster of the closest centroid, the SSE is computed so to update the position of the centroids and minimize the overall SSE (Equation 1). The procedure goes on until the position of the centroids is no more updated.

The update of the positions of the centroids is conducted in direction of the finding of the local minima of SSE, but these cannot be found if the positions of the centroids at step 0 is far from the one that brings to a local minimum. Therefore, the initialization of the centroids represents an issue to face with if those positions are chosen randomly.

Although the selection of initial position of centroids may represent a problem, the method results quite effective and can handle a variety of type of data. Another weak point of this method is the shape of the clusters which it is intended to detect, since the non-globular shapes are hardly identified by K-means, in particular way if outliers are present, since they affect the updating of centroids positions making them less representative of the cluster which they are referring to. (Han, et al., 2011)

In any case clustering analysis results one of the most used technique to extract class from a set of data objects, limiting the intervention of the user in setting constraints in the implementation, matter of fact it is a purely unsupervised and effective method.

2.6.2. Classification tree

A classification technique, or classifier, is an approach that makes possible the construction of a classification model from an input dataset. Common techniques of this kind are decision tree and rule-based classifiers, neural networks, support vector machines and naïve Bayes classifiers. A learning algorithm is employed for this sake, creating a model which has to be able to fit the input data 's attribute and class label. The model must fit the input data and well predict class labels of records never seen before as well, therefore a general approach of this technique is to induce a model by means of a learning algorithm applied on a training data set, then use the learn model to classify elements from a test data set. Training data set contains already the classification information, so it is possible to evaluate the performance of the model counting the correct and incorrect classification.

The evaluation of the performance of the model can be represented in a table known as confusion matrix, reporting class by class how the model classified the elements. From this kind of table can be extracted even performance metrics, as accuracy or error rate, that sum up the goodness of the model in single values. Accuracy is the ratio of correct prediction over the total number of predictions, while error rate is the number of wrong predictions divided by the total number of predictions of the model.

Classification and regression tree (CART) is a kind of decision tree, which strength is its easy implementation. The decision tree is usually fully grown initially and pruned successively in order to reduce the number of branches, making its interpretation easier and avoid over-fitting.

Over-fitting is an issue to face with in a decision tree classification, since it consists in a well performing model with training data, but scarcely performing in the testing phase.

Since the model is used in 2 different phases, error associated to a classification operation can be of two kinds: training error or generalization error. The first is related to the error in the classification of training data set, while the latter is related to the error in classification of observations never seen before.

Training error is reduced as the number of nodes decreases, monotonically, while generalization error does not follow the same trend. While the tree is small, both errors are large, so this case is named model underfitting. Enlarging the tree, the errors will decrease both since the model learns more and more the true structure of the data. As the number of nodes becomes large, the generalization error starts increasing, unlikely the training error which keeps decreasing. The situation is known as model overfitting and it can be caused by presence of noise or lack of representative samples. Pruning is a strategy that can be adopted to avoid this phenomenon. (Tan, et al., 2005)

CART is a decision tree method that recursively partitions data in order to find a prediction model. Partitions are binary splitting which are represented as decision tree. This kind of representation allows an easy interpretation and identification of the rules which drives to data splitting.

The splitting of the data is driven by an impurity measure, in particular the Gini index for CART method, in the sense that the difference between before and after impurity is maximised, in order to find the best split. The method keeps splitting until a stopping criterion is reached, such as the number of observations in the node is less than a threshold value.

The Gini index is defined as

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

Equation 2

where $[p(i|t)]^2$is the portion of the elements in class i at a given node t, with c the total number of classes. (Yan, et al., 2016)

The impurity difference, at a splitting node t, for a particular split s, is defined as

$$\Delta I(s,t) = I(t) - (P_L(I(t_L)) + P_R(I(t_R)))$$

Equation 3

and t_L are left and right child nodes of the splitting and P_L and P_R are the proportions of the observations going into left or right child nodes.

As metric to evaluate the quality of the pruned tree, a cost-complexity measure is used, so to find the pruned tree which has an error comparable with the fully grown one. A common index used is the F-measure, defined as

$$F - measure = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Equation 4

having TP is true positive (correctly classified elements), FP is for false positive (elements labeled with a class label, even belonging to other classes), FN is for false negative (elements labeled with different class label, even belonging to the positive class).

The cost from the number of leaf nodes of a fully grown tree is calculated by means of a non-negative parameter α, defined as

$$\alpha = \frac{R(t) - R(T)}{|\check{T}| - 1}$$

Equation 5

having as denominator the number of leaf nodes minus one and at denominator the difference between a single node misclassification rate and the tree misclassification rate, which are defined as

$$R(T) = \sum\nolimits_{t \in \check{T}} R(t)$$
$$R(t) = r(t)p(t)$$
$$r(t) = 1 - \max_i \ \mathrm{p}(i|t)$$

Equation 6

where p(t) is the probability of the observations in the node t.

A value of α is calculated iteratively, stopping the growth of the tree each time at a node of higher level, proceding backward along the tree. Relative error associated to the pruning can be implement using the k-fold validation. This method divide the training data in k groups using the alternatively each group once for training and k-1 times for validation. The realtive error versus the complexity parameter α is used to find the deepness of the best pruned tree. In fact the optimal pruned tree is the one which relative error does not have any significant decrease if cost complexity is decreased further more. Relative error is defined as

$$RE = 1 - \frac{n_{correct}}{n_{total}}$$

Equation 7

where $n_{correct}$ is the number of correctly classified observations and n_{total} is the total number of observations.

Once the tree is pruned, F-measure metric can be used to evaluate the accuracy of the model for each class identified. (Yan, et al., 2016)

The classification technique results very helpful, as well as the cluster one, in data segmentation phase. Data segmentation is often used to contextualise data to treat. In fact, it is often used in the preliminary phase of an Association Rules Mining methodology, so to carry out the rules extraction only from data belonging to the same context. The context can be found segmenting data according to temporal information, for instance can be used seasons,month,day of the week or hours in order to group data. Association Rules that can be contextualised are more interesting, since can refer to specific situations, therefore the FDD can be conducted for specific cases or specific regime of operation of the system, obtaining a deep understanding of the system. (Fan, et al., 2015)

2.6.3. Association Rules Mining

2.6.3.1. Definition

Data is getting more and more penetrating and big sized for every field of application of data mining. The huge amount of data available drives easily to the necessity to simplify the results obtained and get them more practical and easy to interpret. Given that, a particular data mining technique may help going towards this direction, since it can find a correlation among different operations or data or variables, providing information easy to interpret. This technique is Association Rules Mining, which provides rules that describe the way data are presenting and interacting. This technique has been largely used for commercial purposes, since can find the most frequent items present in the same purchase, that means that retailers may improve marketing promotions, inventory management and relationship with customers using this information.

The associations coming out with this technique are hidden in the data collected, but due to the size of data sets that usually are available these are hard to find differently. Therefore, this technique can be used in every field in which the amount of data available is huge, in order to extract hidden knowledge.

Data are organized as an ensemble of transactions, namely instances, numbered by transactions ID. Each transaction is a group of items, formally item sets. Items may be organized even according to the type of information reporting for each transaction, therefore each item may be referred to an attribute of the data set. This way of representing data is typical of time series in which instances are time instants and attributes are the variables included in data. In this case item sets are the composition of items belonging each to a different attribute.

Attributes may express information in different ways, matter of fact they can be binary, categorical or continuous, according to the values they can assume.

An association rule is an expression defining a logic implication between the presence of an itemset, having stated the presence of another itemset, assuming the form X→Y, with X and Y as disjointed item sets, thus their intersection is void.

According to this form, X may be called the left-hand side (LHS) or antecedent, while Y may be named right-hand side (RHS) or consequent. (Li, et al., 2017)

The rules found can represent the association between frequently occurring item sets or only casual co-occurrence of item sets. As a consequence, it has been introduced some parameters that evaluate the strength of the rules obtained. These parameters are Support and Confidence, which are helpful to discriminate frequent to infrequent rules and weak to strong rules. More specifically, the Support is the percentage of occurrence of a rule in the dataset, while the Confidence is the probability to find Y after encountering X in a transaction. If the Support is high, the rule is widespread in the dataset, that means that the rule is very common. On the contrary, a low Support is describing a rule found only by chance.

If the Confidence is high, X and Y are often present in the same transaction, thus the rule represents a strong relation between the item sets. On the other hand, a low Confidence represents a weak implication between the item sets.

Another parameter is used sometimes to better understand the validity of the rule found, that is the Lift. This parameter measures the dependency between antecedent and consequent. In fact, if its value is higher than 1 states a positive correlation between antecedent and consequent, therefore the antecedent's presence leads to the presence of the consequent. On the other hand, if Lift is minor than 1 it means that consequent is negatively affected by the presence of the antecedent. In the case Lift is exactly equal to 1, the two item sets are completely independent.

Formally the parameters are defined as:

$$\text{Support}(\text{X} \rightarrow \text{Y}) = P(X, Y) = \frac{N(X, Y)}{N}$$

Equation 8

Support, defined such that, is the joint probability of the two item sets. $N(X, Y)$ is the number of transactions in which the rule is present, N is the total number of transactions.

$$\text{Confidence}(\text{X} \rightarrow \text{Y}) = P(X|Y) = \frac{P(X, Y)}{P(X)}$$

Equation 9

Confidence results the conditional probability of the consequent, given the antecedent of the rule.

$$\text{Lift}(\text{X} \rightarrow \text{Y}) = \frac{Confidence(\text{X} \rightarrow \text{Y})}{Support(Y)} = \frac{P(X, Y)}{P(X)P(Y)}$$

Equation 10

This being so, the association rules mining is conducted imposing a minimum threshold of confidence and support in order to focus the rules research on the kind of rules most relevant for the user's purposes. (Li, et al., 2017)

It is worth noting that the rules determined by this technique, must be considered in their inferential nature. Therefore, the association stated by the rules must be considered not as logical correlation, but resulted from inference, thus a causality's implication is not always definable. Due to its inferential nature, even strong rules, with high support and high confidence, can be meaningless for the user, therefore the rules must be evaluated by the user to filter out the uninteresting rules, despite of parameters values.

2.6.3.2. Types of Association Rules Mining

Starting from the market basket analysis, the association rule mining can be extended to many applications implying the research of typical patterns inside a dataset, according to the kinds of data or the need of the users. Then, from mining item sets of user's interests, association rules can be extracted. The pattern mining, namely item sets mining can be classified, basing on the pattern diversity to mine. Such classification is resumed in the Table 1.

Classification by pattern diversity	
Basic patterns	• Frequent pattern • Closed pattern • Max pattern
Abstraction levels	• Multilevel • Single-level
Number of dimensions	• Single-dimensional • Multidimensional
Types of values	• Boolean • Categorical • Quantitative
Constraints used	• Constraint- based • Approximate • Compressed • Near-match • Top-k • Redundancy-aware top-k

Table 1 Classification of ARM by pattern diversity (Han, et al., 2011)

Frequent pattern research can present itself in many forms, since the kind of pattern appearing in the frequent item set pool can be analysed more deeply or not. In fact, pattern can be defined as closed or max if in frequent item sets sub pattern are excluded. A sub pattern is an item set contained in at least another pattern of same or higher length. (Table 1)

In the first case the sub patterns with the same support are excluded, while the other case refers to the exclusion of all the sub patterns of each frequent item set regardless the value of support related to them.

The kind of item included in the pattern can be another criterion of classification, due to the fact that the abstraction level of the item in the pattern can be multiple or single. According to this classification, multilevel patterns include item belonging to categories that comprise one in the other, since the higher-level category can be more general than the lower level one. For instance, in a market basket case the attribute can be "buy", while the different level of abstraction can arise in item such "fruit" and "apple".

In the case which the items involved in the pattern are belonging to different attributes, the pattern can be considered as multidimensional, since each attribute is a dimensionality for the pattern. (Table 1)

The kind of data of the attributes involved can determine the classification of the patterns, since the data can be quantitative, namely continuous variables, categorical, namely discrete variables, or Boolean, thus defining the presence or not of an item, as it can be found in a typical market basket case. (Table 1)

In the frequent pattern searching some constraints can be applied to focus the research to a certain type of pattern. In fact, patterns discovered can respect a generic constraint set by the user, be approximate, compressed or near-match, if the support is counted considered similar or almost matching patterns, or, furthermore, top-k, extracting only the k most frequent item sets, eventually excluding redundant patterns. (Table 1) (Han, et al., 2011)

The patterns may be classified even according to the kinds of data involved, therefore the Table 2 collects the cases considered by this classification.

Classification by the kinds of data involved	
Kinds of data and features	• Frequent item sets • Sequential patterns and time series • Structural patterns
Application domain-specific semantics	• Spatial • Temporal • Spatiotemporal • Image or video • Software programs • DNA sequences, etc.
Data analysis usage	• Pattern-based classification • Pattern-based clustering

Table 2 Classification of ARM by the kinds of data involved (Han, et al., 2011)

The nature of the data considered in the patterns is the basis of this classification, in fact frequent patterns can be named frequent item sets, sequential patterns, temporal patterns or structural patterns, according to the kind of pattern that can extracted. The first case refers to the pattern made up of a set of items, that is common in situations which there is not an order in the data. The second case considers as input data a sequence of ordered events, eventually with temporality. Temporal frequent pattern extraction is very helpful, for example, if the data comes from a monitoring system of a building. Therefore, frequent patterns for this category are pieces taken from the input sequence. The latter is a general definition, because structural patterns are formally any minor element of graphs, lattices, trees, sequences, sets, single items or combinations of these structures. The classification of the frequent patterns can consider the domain semantic, that means that a pattern is considered as specific for the kind of domain of the mining application. In fact, patterns can be spatial, temporal or spatiotemporal basing on the nature of the data, otherwise they can be pieces extracted from images, videos, software programs or DNA sequences. This categorization is very helpful in the individuation of the most efficient approach to use, since the nature of data affects the way it should be treated. (Table 2) (Han, et al., 2011)

As far as concerns the data analysis usage, it is worth noting that the approach at the frequent patterns mining is varies with respect to the goal of the methodology. Many are the chances to use this technique in a Data Mining procedure, nevertheless it is mostly employed in classification, for feature extraction, or clustering, applied to high-dimensional data. (Han, et al., 2011) (Tan, et al., 2005)

2.6.3.3. Temporal Association Rules Mining

As regards energy building application, particular interest has been directed on the development of multidimensional temporal patterns, because many variables taken into consideration during the analysis of patterns are extracted from time series, which comes from the monitoring along a specific period of time. Data of this kind is organized in a dataset with all the monitored quantities as attributes and transactions as temporal log identified by timestamps. Every piece of the sequence that can be considered relevant for the analysis is an event Temporal association rules are able to express a relationship between sequences taken from a time series, preserving the consequentiality of the happening of the events within the implication.

Formally temporal association rules mining (TARM) represents the problem of finding the relationship between events occurring with a certain time difference in a time series, thus

$$X \xrightarrow{t} Y,$$

Equation 11

meaning that the occurrence of X implies the occurrence of Y after a time *t*. The temporal information, therefore, can be resumed as both the respect of temporal order of happening, i.e. the consequent can never anticipate the antecedent, and the respect of the time window constraint, i.e. the temporal distance *t* has a limit. The latter defines the difference between sequence pattern mining and temporal pattern mining, since even the sequence patterns respects the first condition. (Martínez-de-Pisón Ascacíbar, et al., 2009) (Martínez-de-Pisón Ascacíbar, et al., 2012)

The definition of an appropriate time window lies with the user, since it depends on the requirements of the problem to analyse. The time window width can even let vary, repeating the operation of rules mining for different width, therefore for each rule can be associated a minimum time window size in which it is present. This information, for instance, may be exploited to find the different response time of the variables affected by the same event. The power of this kind of analysis is not only the deeper understanding of the propagation of events' effects in the system, but also the chance to predict the behaviour of the system, considering eventually dynamics without carried on a deep study about it. (Fan, et al., 2015)

Formally, rules mining that does not consider temporal information is of the type of intra-transactional, since the item sets composing the rule are taken from the same transaction, while the approach considering temporality is named inter-transactional.

In literature there are examples of both types regarding rules mining applied to time series, but the second kind results more useful, due to the prediction power resulting from the different information provided.

In the inter-transactional approach, item sets composing the rule are taken from a portion of time series, appearing respecting the order they have in the time series. The portion of the time series is an ensemble of instances with consecutive timestamps, thus it is nothing but a series of consecutive transactions. The time window considered is sliding along the time series maintaining fixed the width, that means that it can be positioned starting from every timestamp. In this way there are not any constraints in the positioning in time, thus any event in the time series may be part of a time window in every location inside the window. (Figure 4)

The computational effort is increased considering the time window, thus inter-transactional techniques require a higher cost with respect to the intra-transactional ones. This cost is related to the generation of all the portion of time series sequences, namely time windows, which are partially overlap if consecutive. In fact, strictly consecutive time windows, are almost identical, differing only for a number of instances equal to the step of the sliding. For instance, if the sliding step is equal to one, thus the time series is scanned without holes in the analysis, a window is identical to the previous apart from a single instance. Basically, given a portion of the time series of a time window width long, cutting out the first instance and appending at its end the immediately consecutive instance observed in the time series, it is obtained the following portion. (Figure 4) (Fan, et al., 2015) (Martínez-de-Pisón Ascacíbar, et al., 2009) (Martínez-de-Pisón Ascacíbar, et al., 2012)

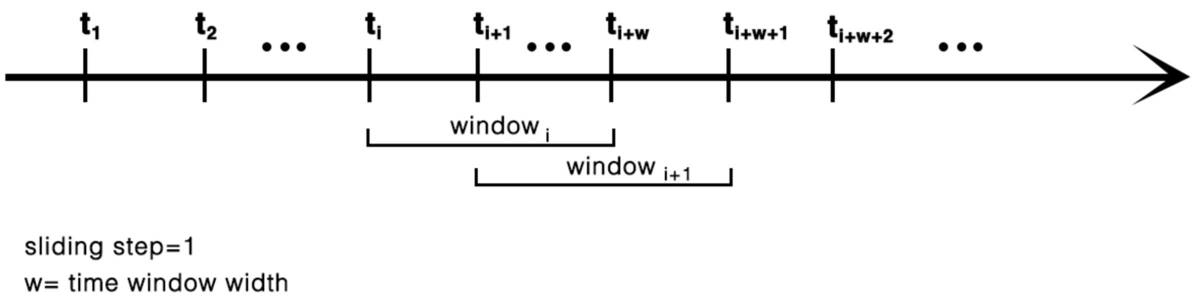


Figure 4 Sliding window schematic

2.6.3.4. Algorithms

Association rules mining can be summed up as a two-steps procedure, since most of the algorithms starts finding frequent item sets first, then find the association rules.

In the first phase the item sets selected must respect the minimum Support threshold to be considered frequent, successively, the Confidence is taken into account. In fact, only in the second phase the minimum Confidence threshold is used to generate rules according to this restriction. Considering the nature of the steps, the attention of the algorithms' developer has been focused on the first one, since the other one requires minor computational effort. In fact, the second step is nothing but generation of possible combination of items taken from the frequent item sets, with a consequent filtering according to the minimum threshold of Confidence.

The first step can be different basing on the approach that the algorithms have, matter of fact that the process can be characterized by a candidate generation or not, prior the actual frequent item sets generation. The candidate large item sets generation is a sub-step according to which, the dataset is scanned generating all the possible combinations that can exist with the item sets belonging to the dataset. After that, the candidates are filtered according to the minimum Support value. The candidates are generated considering item sets ordered by size, therefore first item sets to consider are the single item's ones, then the 2 items' ones and so on and so forth, up the size of the longest item sets possible. After the candidates' generation, the candidates are evaluated in terms of Support, so to eliminate the ones that do not respect the minimum threshold. (Kaur, 2014)

Several algorithms have been developed through years, according to the need and the applications. The most used is the highly widespread Apriori algorithm, which has been developed in 1994 by R. Agrawal and R. Srikant for Boolean association rules mining. A typical application of Apriori is the market basket analysis, since it analyzes the presence or not of items in market baskets.

This algorithm is a candidate generation one and it is named such that due to the criterion used to eliminate the uninteresting candidates. In fact, the algorithm bases its effectiveness in the fact that it considers as frequent all the nonempty subsets of a frequent item set, as an *a priori*. As a consequence, it is not necessary to verify the respect of the minimum Support condition for all the item sets present, because every time the size of item sets to generate is increased, frequent item sets generated in previous iterations are taken into account as the bases for the candidates to generate. Basically, candidates are generated starting from the frequent item sets already found. After the candidates' generation, the new candidates found are evaluated in terms of minimum Support and discarded if necessary. (Figure 5) (Agrawal & Srikant, 1994)

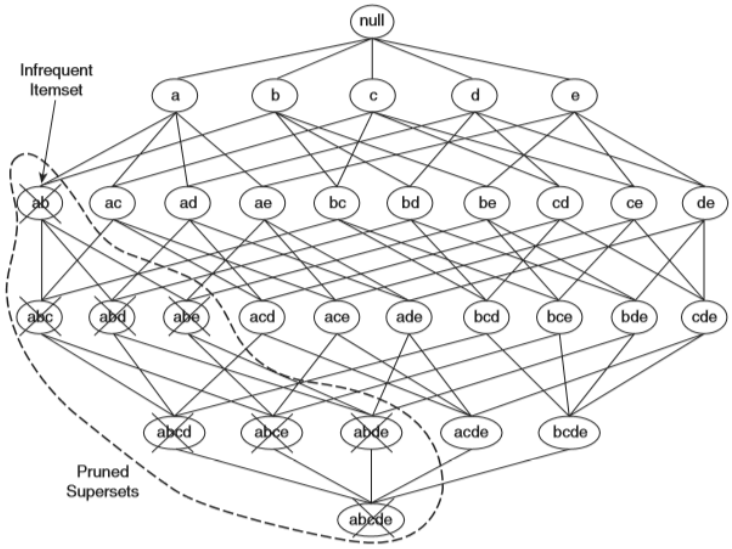


Figure 5 Apriori pruning from infrequent itemset (Tan, et al., 2005)

Apriori algorithm main disadvantages is the compulsory scan of the database in order to check the eventually huge set of candidates, but most of all the computational cost is not irrelevant. In fact, although the Apriori principle prevents the complete generation of candidates that can possibly found in the dataset, handling the whole candidate set represents a not negligible computational effort.

Frequent item sets can be extracted from a dataset with other similar and equally diffused algorithm as ECLAT or FP-Growth. The latter has particular fortune because of its approach, which has been demonstrate as faster than Apriori's one, matter of fact it avoids the candidates' generation in the frequent item sets generation, which represents the main bottleneck of the algorithm. (Han, et al., 2004) (Han, et al., 2011)

According to FP- Growth algorithm, firstly data are represented in a compact form, namely frequent pattern tree (FP-tree). Data organized such that, are distributed in a tree-like graph, with a single item per node. Then the dataset is divided in pieces according to the FP- tree organization, that are mined separately. Constructing the tree-like graph, dataset is scanned in order to count the items' occurrences. In this way, only frequent items can be easily spotted in the graph, respecting the support limitation. After the dataset is scanned the items are located in the graph so to allow to read each transaction along the branches. In this way the dataset is scanned only twice. Each node represents an item that belongs to the item sets, reporting the counts of occurrences of the item in item sets with the prefix that may be composed going backward along the branch which is located. Each node is linked to others so to construct item sets. For tree-like form construction it is necessary to order items inside transactions in support decreasing order, so to find the most frequent items in upward.

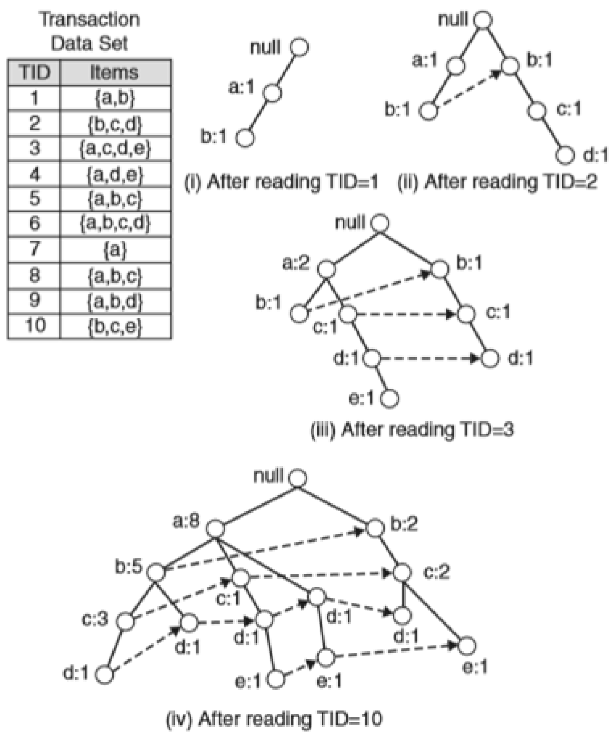


Figure 6 Construction of FP-tree (Tan, et al., 2005)

As a result, each node is shared by the item sets containing it. (Figure 6) After the FP-tree construction, the actual frequent pattern mining phase comes along. In this phase the frequent patterns are found with suffix-based approach. (Han, et al., 2004) (Han, et al., 2011)

Frequent patterns are searched starting from the last items as ending items of the path. In fact, an ending item is chosen, then the paths upstream the item along the branches are considered as prefix. A tree graph that does not contain the suffix item is constructed, reconsidering the counts for each node, such that only transactions with the suffix item are taken into account. Successively branches representing item sets that do not respect the minimum Support threshold are pruned. As a consequence, the frequent item sets ending with the suffix item have been found. After that, the algorithm proceeds shortening more the branches of the original FP-tree, so to consider as suffix item the one located a level immediately above the previous. Therefore, the algorithm considers as suffix items each in order of increasing counts.

The operations are perpetrated until the root node is reached. As a result, the frequent item sets are found in order of suffix item, from the least occurring to the highest. This divide-and-conquer approach results particularly efficient, since it is able to divide the problem into sub-problems to be faced one by time.

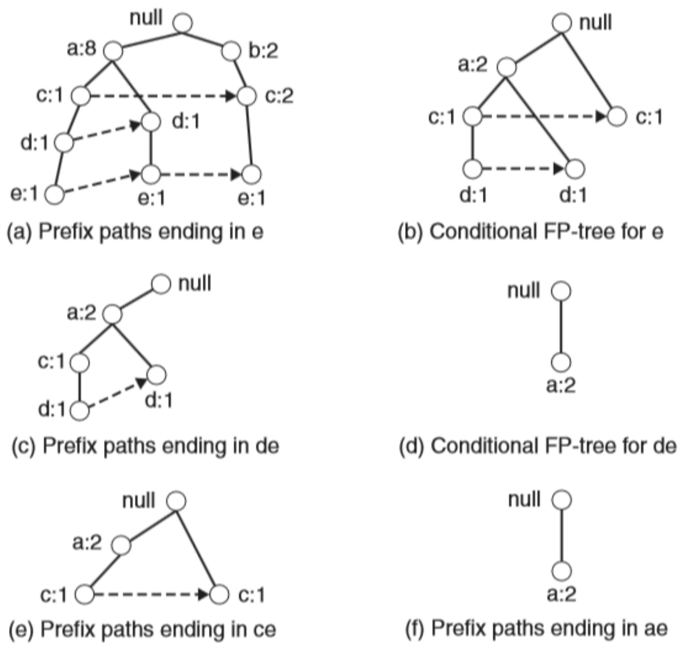


Figure 7 Application of FP-growth algorithm to find frequent item sets ending in "e" (Tan, et al., 2005)

Thanks to this, FP-Growth results significantly faster than Apriori, unless the FP-tree constructed is very complex. In that case the effectiveness of the FP-Growth algorithm is put on a strain. (Han, et al., 2004) (Han, et al., 2011) (Tan, et al., 2005)

The choice of the algorithm is strictly related to the kind of data available. In fact, the algorithms described above can extract association rules easily from categorical data, but they ca not handle continuous variables. Facing this issue, the solution commonly used is the discretization of the continuous variable, to transform it in categorical. That means that data must be manipulated, eventually with loss of information or with the generation of a huge number of rules, most of them meaningless.

Another approach is to use quantitative association rules mining, that allows the extraction of rules, without the prior application of discretization methods. In fact, this kind of algorithms generates rules providing the intervals of values that the variables involved assume.

The typical format of quantitative association rules is of the kind:

$$\{X \in [x_1, x_2]\} \rightarrow \{Y \in [y_1, y_2]\},$$

Equation 12

with $[x_1, x_2]$ and $[y_1, y_2]$ the intervals in which respectively X and Y are defined.

Quantitative association rules mining is characterised by 3 main approach: data cube-based, clustering-based and statistical methods-based (to find exceptional behaviour).

The first type of method requires the use of a cube-like graph to organize data of multi-dimensional dataset. This kind of method is very helpful in simplifying the approach of mining multi-dimensional association rules, so it is convenient in quantitative rules mining, if the attributes have been priorly discretized. In this way, any of the frequent pattern generating algorithm that can handle only categorical attributes can be applied. In fact, the algorithms recognize the pairs attribute-value as items, as far as they treat to single attribute items.

The Apriori principle can help in the computation of the data-cube exploration. The second one uses cluster analysis to find more dense area in each attribute, that corresponds to more frequent occurrences. Therefore, clusters that do not satisfy the minimum Support condition are discarded. This operation is perpetrated for each attribute present in the dataset.

After that, the remaining clusters of an attribute are combined with the ones belonging to another attribute, applying the minimum Support criterion at the resulting item sets. The process continues adding a dimensionality a time, until the dataset is exhausted. The Apriori principle can be applied in this approach too, in order to speed up the computation.

The last approach provides information on the bases of statistical analysis. Therefore, an interesting rule that this method can find may state the unlikeness of the values of an attribute's subset, with respect to the overall mean. For the extraction of this kind of rule, a statistical test is required, such that it is possible to identify an exceptional behaviour. For instance, if the mean salary of females is significantly different from the overall mean, calculated considering both females and their complement, then the rule associating the "sex"'s attribute value "female" to the corresponding value of the attribute "mean salary", represents an exceptional behaviour's rule. (Han, et al., 2011)

An alternative and very common approach of quantitative association mining is the optimization-based one, that is adopted by Quantminer algorithm. This method comprises the application of a genetic algorithm to find the interval of attributes involved in the rule, by means of the optimization of additional parameters related to the rules. The algorithm basis its power on the optimization of intervals of attribute already inserted in rules templates. Basically, Quantminer generates the whole range of possible combinations of item sets, in order to create all the possible rules that can be extracted.

Since attributes can be categorical or quantitative, the approach is two-fold. In this case of categorical attributes, the values assumed can be set by the user or, using an Apriori-like frequent pattern generation, defined by the algorithm. (Fan, et al., 2015) (Fan & Xiao, 2016)

In the case of quantitative attributes, for each rule template a Genetic algorithm is used to find the best interval in which the attributes' values rely.

A rule templates set can be also provided by the user, if the research should be focused on specific rules. The genetic algorithm considers for each rule template, the whole set of intervals for every attribute involved in the rule template, that have support beyond the minimum support threshold. This set of intervals is used as a pool from which intervals are randomly taken to create the initial population. After that, mutation and crossover operation generate the following population's generation. Crossover combines 2 intervals, while mutation modifies the amplitude of a single one. After that, the best interval for each numeric variable involved in each rule template is searched maximising the fitness function associated to the rule template. The fitness function is calculated by means of the Gain parameter, with the expression below.

$$Gain(X \rightarrow Y) = Support(X, Y) - MinConf \times Support(A)$$

Equation 13

$$Fitness(X \rightarrow Y) = Gain(X \rightarrow Y) \times \coprod_{A_i \in A_{num}} \left[1 - \frac{size(I_{A_i})}{size(A_i)}\right],$$

Equation 14

with A_{num} is the number of numeric variables in the rule, I_{A_i} is the interval of the variable A_i, $size(A_i)$ is the total range of A_i. This phase is carried on trading-off the length of the interval and the fitness function related to the rule that the algorithm is analysing. As a result, small intervals and high Gain rules are preferred. (Fan, et al., 2015) (Fan & Xiao, 2016) (Han, et al., 2011)

As concerns temporal association rules mining, the most common algorithms come from the adaptation to the problem, of algorithms used for other kind of association rules mining. That is the case of T-Apriori, which follows basically the approach of Apriori, but considers the temporal information in the candidates' generation, so to take into account only candidate that respect a maximum time span constraint and temporal consequentiality as well. (Zhai, et al., 2018)

Alternatively, Apriori can be used after a segmentation of the sequence in micro subsets that are already expressing the time span information, since the data contained belongs to different timestamps temporally spaced according to the width of the time span. (Bhuvaneswari & Umajothy, 2013)

Another example of temporal association rules mining derived from Apriori is the Mining Frequent Item sets within a Transaction-sensitive Sliding Window (MFI-TransW). This algorithm implements a sliding window to generate frequent item sets according to a time window width that embraces more transactions. Each transaction considered by the algorithm represents an instance of the dataset related to a single time stamp. (Li & Lee, 2009) (Martínez-de-Pisón Ascacíbar, et al., 2009)

Often the methods are borrowed from sequential rules mining, in which the consequentiality is considered, so only the time span constraint must be added in the computation. In fact, the sequential rules mining problem is the association rules extraction from an ordered list of item sets, namely events.

This is the case of TRuleGrowth, that is the temporal transposition of RuleGrowth, which is used for sequential rules mining. The temporal information is added simply inserting a sliding window implementation in the algorithm.

Dataset in input is in vertical format, that means that each row refers to a transaction, namely an ordered sequence.

This algorithm uses a pattern growing approach in order to consider only database occurring rules as potential valid rules.

Pattern growing approach starts with a 1*1 pattern, composed by an item set for each side of the rule, that respects the minimum Support requirement, then recursively add an item a time to expand right or left side. (Fournier-Viger, et al., 2012) (Fournier-Viger, et al., 2015)

Every time a new pattern is generated, its validity is evaluated in terms of minimum Support and minimum Confidence.

Consequentiality in the pattern is considered checking if the item inserted in the pattern respects the order of occurrence in the sequences, so to avoid temporality's inversions.

A sliding window is implemented in order to grow the pattern taking items from a limited time interval, respecting the temporal constraint.

The sequences width affects the sliding window passage, since the sliding window is forced to pass along the sequence. In this way it is contemplated a continuity between transactions. (Fan, et al., 2015) (Fournier-Viger, et al., 2012) (Fournier-Viger, et al., 2015)

As regards temporal association rules mining, a very effective approach is the one proposed by Zaki with the cSpade algorithm. This algorithm extracts sequential rules considering some user-defined constraints. These constraints may drive the mining by controlling length, width, gap or time window of the rules. Setting these parameters, the rules extracted are filtered out, in order to respect the restrictions, therefore only useful rules are provided to the user. Parameters controlling length and width of the rules modify the shape of them, since control how many events should be considered in the whole pattern or per each side of the rule respectively. Parameters controlling gap and time window drive the choice of the events involved in the rule, since the gap parameter limits the interval of time that must elapse from the LHS and RHS, while the time window is the dimension of the time interval in which the whole sequence occurs. (Zaki, 2001)

The frequent pattern search is perpetrated by means of the construction of a frequent sequence lattice, starting from a vertical formatted database. A frequent sequence lattice is constructed so to take into account all the possible combinations of the events in the database, therefore the exclusion from the lattice of the nodes that do not respect the temporal consequentiality, is carried out secondly.

In this way the cases of temporal inversion of the patterns are discarded and the support counting is conducted only for the patterns that can be considered as sequences, namely that respects order of occurring. For this sake, it is fundamental the specification of sequence ID and event ID for every event, which identify the temporal position of the events in the whole database, by means of the sequence ID, and the temporal position inside its sequence, with event ID.

In literature cSpade is mostly used for applications in medical field, such as data analysis of medical treatments of admitted patients, or DNA sequence analysis, but its characteristics made it helpful even on building energy applications. (Zaki, 2001)

Another interesting approach is the one proposed by the EDMANS research team and implemented in the CONOTOOL software. This algorithm is able to extract inter-transactional association rules from multivariate time series, starting from a specific episode of particular interest. In fact, the research is focused on the patterns that drive to an episode already known as characteristic of a fault. The user, therefore, knows the parameters that represents a fault for the system and search the causes of the fault in terms of sequence of episodes happening prior the faulty event. Running the CONOTOOL software, the time series are transformed in series of transitions, sequences of events, which are then grouped into episodes, according to expert criteria, before the application of the algorithm itself. After this a database of frequent item sets appearing into a sliding window pre-set by the user is created. In fact, the time windows' patterns that contains episodes occurring before the faulty episode are taken into consideration and used to construct a database of transactions. Once the database has been made, a frequent item set algorithm of intra-transactional type is used, namely ECLAT, that with Apriori-like approach finds the patterns that most frequently drive to the fault pattern. The construction of such database guarantees the respect of temporal constraints, so the application of intra-transactional algorithm does not affect the "inter-transactionality" of the whole method. Supporting this, it can be demonstrated that inter-transactional item sets are a composition of intra-transactional item sets.

Handling the database of this kind allows to find, by means of intra-transactional methods, temporal association rules involving episodes with prefixed consequent, antecedent occurring logically and temporally before, with a time gap fixed by the user and within a time window fixed by the user as well. (Martínez-de-Pisón Ascacíbar, et al., 2009) (Martínez-de-Pisón Ascacíbar, et al., 2012)

As a result, the CONOTOOL software allows the user to focus the research exactly on what is already known as interesting or useful for the analysis, by setting temporal constraints and providing to the software the consequent of the rules to extract. The fact that it permits the use of intra-transactional algorithms to extract inter-transactional association rules gives a great freedom to the user in the rules mining, therefore the methodology can be adapted to the kind of data on which it is applied to. In this way, the user can choose the algorithm that better fit the data to analyse. (Martínez-de-Pisón Ascacíbar, et al., 2009) (Martínez-de-Pisón Ascacíbar, et al., 2012)

2.6.3.5. Discretization

Discretization is a fundamental and necessary preliminary phase in an association rules mining methodology, since most of the algorithm can handle only categorical variables, therefore continuous ones must be transformed in discrete ones. In temporal DM, most of the time series treated refers to continuous variables, therefore the discretization step is mandatory for this kind of application.

The choice of the discretization method represents a great issue to overcome since the quality and the number of rules extracted at the end of a DM process are seriously affected by how the discretization is conducted. (Fan & Xiao, 2017)

A mis-discretization can bring to a poor description of the possible states of a variable, as the number of interval can be too high or too low, in addition to the possibility to have a wrong positioning of breakpoints. If the number of interval is too low, information can be lost, since data can be mis-categorized, eventually including anomalous values in a normal state category. On the other hand, if the number of intervals is too high, the relevance of the discrimination between a state and next one is jeopardized, besides the fact that the computational cost is increased, not necessary resulting to a better quality of the rules extracted. (Fan & Xiao, 2017)

Several techniques have been developed, so to find the best way to transform a continuous variable in a discrete one. The most common and easy to use methods are equal frequency, equal interval and expert discretization.

The first kind of discretization 's criterion is that each interval must have same number of values. Since the number of interval is fixed and provided by the user, the discretization algorithm must count the elements in order to equally distribute them among the intervals. The second one splits the elements in a fixed number of interval provided by the user as well, regardless the number of elements that can be included in each interval or the shape of the distribution of the variable which it is discretizing. The latter is an absolute supervised technique, in fact according to this technique the user defines the number of interval and the breakpoints' positions. Therefore, a great knowledge of the variable is required in order to split data in the most convenient and useful way. This kind of discretization may be very useful in the case the user knows exactly the threshold within which data can be considered normal. Since the user has not always a great knowledge of the system or the results of previous methods may be unsatisfying, several other methods, in particular unsupervised ones, have been developed.

K-means discretization is one of them and basis its effectiveness on the possibility to find intervals as clusters in a monovariate environment, eventually optimizing the number of interval, optimizing the number of cluster found by K-means algorithm.

A very efficient method to transform a continuous variable in a symbols series of reduced dimension is symbolic aggregation approximation (SAX). This technique is able to compact data preventing the information loss.

First of all, the time series that has to be processed by SAX algorithm, is standardized so to have mean value equal to 0, standard deviation equal to 1 and fragmented into pieces of the original time series of the same temporal width. SAX is characterized by two parameters, that are Word size and Alphabet size. (Figure 8)

The time series with this method is transformed in a stream of alphabets that identifies uniquely the values assumed by the processed variable. Therefore, alphabet size defines the number of interval used to discretize the variable. Furthermore, the variable is considered normally distributed and the positioning of breakpoints is carried out in such a way that the area below the curve of the variable's distribution is equally divided among the intervals.

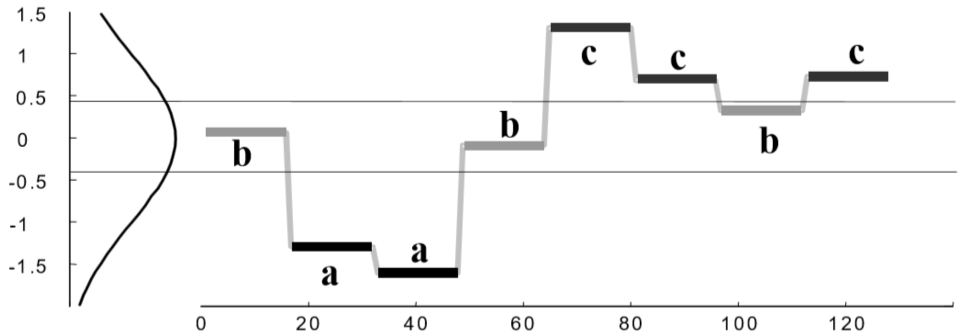


Figure 8 Example of SAX application with Word=8 and Alphabet size=3 (Lin, et al., 2003)

As a consequence, the area below the Gaussian curve of the variable distribution is sliced such that each slice has an area equal to 1 over the alphabet size. The original time series is organized in equally sized fragments, each represented by a Word. A Word is an ensemble of alphabets that sets out the level assumed by the variable into the time interval of the fragment which the Word is referring to. In fact, the fragment is divided in sub intervals of the same length and the number of intervals coincides with the Word size. Therefore, each alphabet defines the level assumed by the variable in a specific time interval. For instance, given a 1-year data stream of a continuous variable, assuming to fragment it into 1-day chunks and that the Word size is equal to 3, it means that alphabets refer to the value assumed by the variable in 8 hours. The higher the values of the two parameters, the higher the grade of detail of information provided, but the higher the computational cost as well. This technique is very efficient and straightforward, matter of fact there is not the need of high domain expertise or pre-processing. (Fan, et al., 2015)

Despite the time series fragmentation and word sizing, the interval identification can be used as a discretization technique.

A variation of SAX technique has been proposed in literature in order to enhance the quality of the discretization in the cases which the variable is not properly Gaussian-like, that is the adaptive SAX (aSAX). This technique starts from aSAX discretized variable and optimizes the positions of breakpoints using an algorithm derived from k-means clustering, with k equal to the alphabet size. In fact, it is an iterative process that calculate at each step the centre of mass of each interval, then computes the new breakpoints position as the middle of the distance between two adjacent centres of mass. If the total representation error variation is within a defined threshold, the process is stopped, otherwise the process starts over.

This method can be applied to a training set of the original time series but using a piecewise aggregate approximation (PAA) representation of the time series drives to a faster convergence. (Pham, et al., 2010)

According to this method the only parameter set by the user is the number of intervals, namely the k value. This process can be turned in a completely unsupervised one if the initialization is conducted with k-means discretization algorithm with k optimized by means of some internal validation indices (e.g. Dunn index, Davies-Bouldin index, etc.). (Pham, et al., 2010)

3. AFDD methods' application

In Data Analytics, several are the techniques that can be used for Automated Fault Detection and Diagnosis (AFDD). Some of them have been already mentioned above, but many others populate the AFDD literature with good results.

Different approach can be used in order to accomplish the objective of AFDD, as quantitative, qualitative or process history-based. The choice of the method to use is influenced by several factors and it implies the acceptance of drawbacks related to the limits of the method chosen.

The techniques mentioned above are rarely applied on their own, especially for AFDD purposes. This is basically caused by the limitations imposed by single methods, which prevent the employment of the methods as they are, unless the user is confident with the acceptance of the constraints.

Through years, worldwide researchers attempt to overcome techniques limits, especially using different methods in combination, so to take the best part of each of the methods used. This kind of approach, in AFDD problem solving, is more and more diffused, since the such methodologies results quite effective. (Katipamula & Kim, 2017)

For this reason, in this section it is presented a series of methodologies implying the combination of different techniques that singularly cannot reach the quality of the results that are able to get within a combination methodology.

For the purpose of overcoming the methods' limits or to push forward the efficiency of the methodology, several combinations of different kinds of methods have been implemented through years. One of the main goal of the combination of different methods has been the enlargement of the analysis to simultaneous fault detection. In this way a AFDD technique is able to distinguish faults even occurring contemporary, identified from a list of possible faults, with improved accuracy, robustness and reliability with respect to non-combined methods.

A grey box analysis can be associated to a black box statistical polynomial regression model in order to use two cost-effective methods finding multiple faults and isolate each of them. A quantitative model-based method can be applied together with a black box model increasing robustness and reduces noise, disturbances and uncertainty.

A quantitative model based on first principles can be used to model the system and then a SVM acts as a faults classifier in order to detect and diagnose basing on statistical learning theory. Alternatively, it is possible to use a PCA method instead of a SVM in order to reduce data volume and computational cost. (Katipamula & Kim, 2017)

Following, in this section, it is reported an insight of the major methods employed for AFDD in AHU systems, which are the most common ones, and for the other energy building 's systems. AHU application of AFDD represents the most interesting case, since the potentiality of energy saving is enormous if it is considered the high diffusion of this kind of system.

3.1.1. Fault Detection and Diagnosis in AHU

HVAC systems represents the principal cause of energy consumption in majority of the buildings in the world, therefore the preference in the investigation of preventing energy waste in this kind of equipment is widely justified.

Considering the large diffusion and the great potentiality of operation efficiency improving, many FDD methods have been developed to spot the cause of inefficiency and schedule countermeasures to solve it.

Yu et al. reviewed several FDD methodologies applied to the AHU case, defining 10 characteristics that an effective FDD methodology should have to get the best results in this application. These characteristics are: quick detection and diagnosis, isolability, robustness, novelty identifiability, classification error estimate, adaptability, explanation facility, modelling requirements, storage and computational requirements, multiple fault identifiability (Figure 9). These characteristics cannot be found contemporary in any of the common technique used for FDD in AHUs, therefore the choice of the technique implies the acceptance of a compromise between the presence of some of the characteristics and the absence of the others. (Yu, et al., 2014)

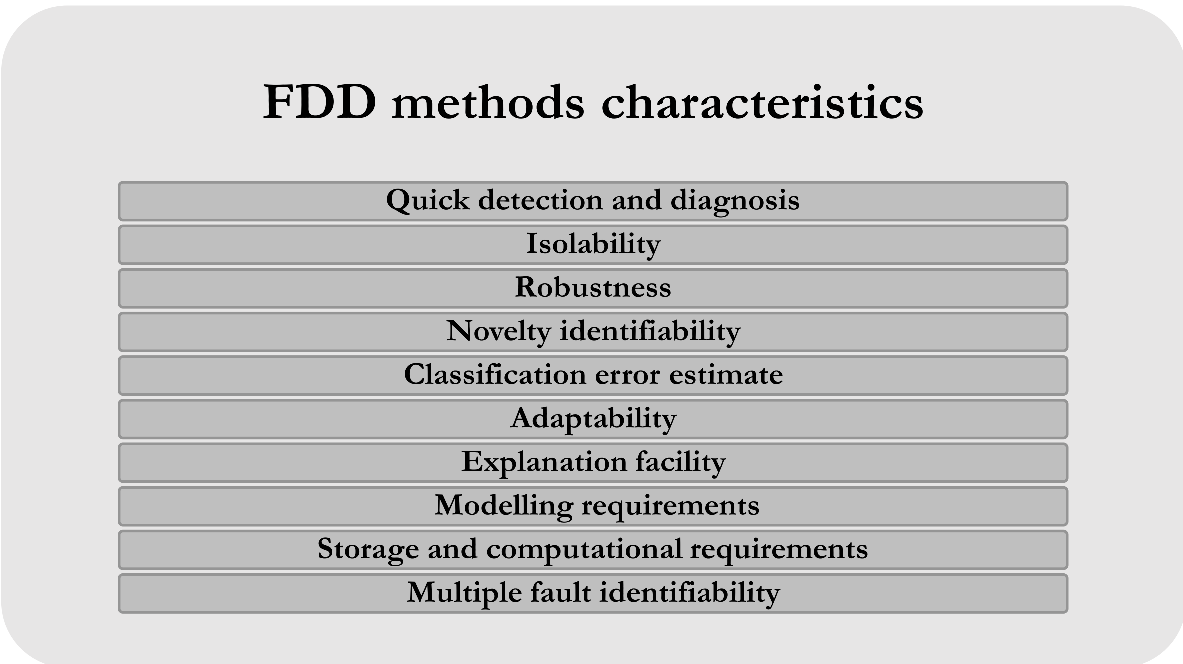


Figure 9 FDD method main characteristics

The quick detection and diagnosis is the main requirement for a diagnostic system, matter of fact the prompt identification of the faults occurring is at the basis of the procedure. The detection of faults is conducted in order to find them quickly and schedule as soon as possible countermeasures to fix the problems. Unfortunately, a too tight threshold can lead to a high rate of false alarms, since even though in normal operation noise of data can be placed beyond the threshold limits. This property is common to most of the categories of FDD, that are model-based, knowledge-based or data driven. . (Figure 9) (Yu, et al., 2014)

The isolability represents the possibility to separate and diagnose contemporaneous faults. The issue about the multiple failure isolation is represented by the summation of the uncertainty of the model developed, with the failure data related to a different fault. This summation effect emerges in particular in the residuals analysis. (Figure 9) (Yu, et al., 2014)

Robustness is the property of being not affected by noise or uncertainties introduced in the model. This feature should face with the quick detection one, since in most cases the higher the robustness, the higher the lower the false alarm rate. As a consequence, a robust model is not sensitive to small variation of data, so to absorb disturbance introduced by input variables. (Figure 9) (Yu, et al., 2014)

Novelty and identifiability feature indicates the capability of identification of fault not already encountered or taken into account in the model construction. Therefore, approaches like historical-based are not suitable in cases which unknown faults it is required to spot, due to the fact that they are able to identify only faults considered in the data used for the construction of the model, unlike statistical methods as PCA. (Figure 9) (Yu, et al., 2014)

Classification error estimate refers to evaluation of the error committed in the diagnostic decision. This feature is a measurement of accuracy and reliability of the method, so to provide to the users the elements for the assessment of the methods and stimulate them to be confident with them. This analysis can be performed on every kind of approach, even on model-based ones, which cannot avoid tolerating a residual error, as low as it is. (Figure 9) (Yu, et al., 2014)

The adaptability property to exploit the same model on systems differing by some structural changes, namely equipment changings. These methods can be applied to different systems, since they are not constructed on the basis of information specific of the system which they are applied to. Historical- based methods, as well as knowledge-based ones implies the construction of the model tailored to the system which is applied to. On the contrary, model-based approach can permit a certain grade of adaptability according to the way the model has been constructed. (Figure 9) (Yu, et al., 2014)

Explanation facility represents the capability of a model to detect a fault and provide at the same time the location and the mode of failure associated to it. This feature is particularly aspired in cases of real-time application, matter of fact the intervention should be scheduled without further inspection or investigation. Models that are constructed on the basis of first-principles equations or a priori knowledge can provide this information. (Figure 9) (Yu, et al., 2014)

For real-time application it is worth to consider the requirements of model, storage and computational effort. All of these three features should be minimized to guarantee an effective implementation of real-time FDD. In particular, the model requirement refers to the fact that the lower the number of models used, the faster the implementation. Fast implementation allows the method to be used in real-time applications. As a consequence, rules-based and historical-based approach fulfil this requirement, since there is no need of any models to use them. Generally model-based approach cannot satisfy this condition, but the novel approach called decoupling-based feature can provide fast implantation as well, because it is the result of manufacturer's data and virtual sensing. (Figure 9) (Yu, et al., 2014)

The last characteristic refers to the power of isolation of different simultaneous faults, thus referring to a condition common in large and complex systems. Unfortunately, there not any examples of implemented methods, which can fulfil this requirement. (Figure 9) (Yu, et al., 2014)

As regards AFDD methods applied to AHU case, a specific classification has been introduced by Yu et al. (Figure 10), partly overlapping the Katipamula and Brambley classification of AFDD for HVAC systems application. (Katipamula & Brambley, 2005)This classification takes into account the approach and the exploitation of data, so to group methods in an analytical-based, knowledge-based and data-driven ones.

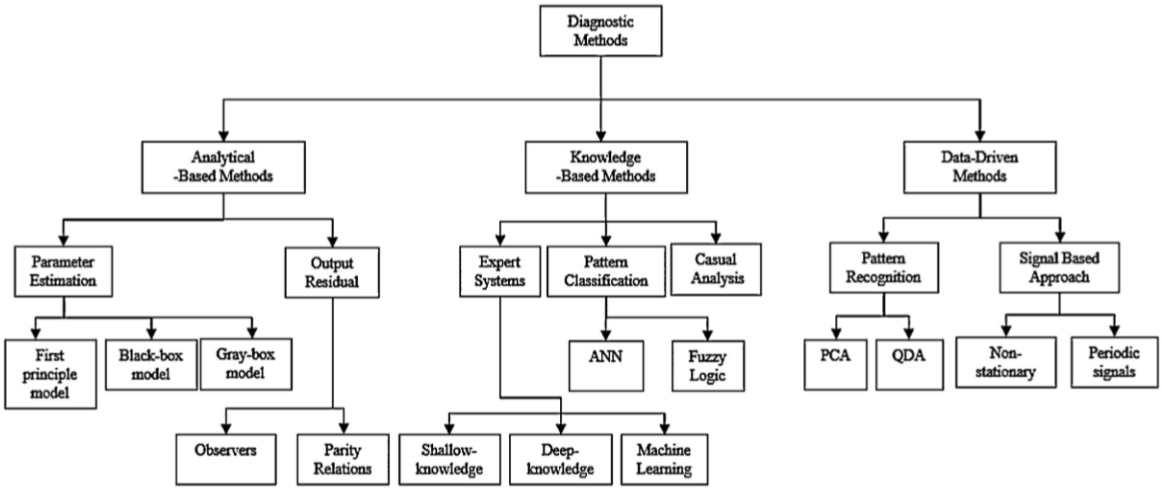


Figure 10 AFDD methods for AHU classification (Yu, et al., 2014)

In the first class all the methods which model construction is involved are placed in. Basically, it gathers all the methods that provides expected values on the basis of a model and, by means of residual analysis, then use it for detection and diagnosis of the faults. Approaches for this category are twofold, since the models can be used to generate directly expected values to compare with real input-output data or can be used to generate parameters' estimation to use in residual analysis. The parameters' estimation approach exploits the system's model constructed in order to create an estimator able to generate values starting from measured inputs and outputs of the system. These values are compared with the ones taken from the measured outputs of the system. The parameters' estimation can be conducted with different kinds of models, since the investigation of the system can consider dynamic effects or only steady-state conditions. The models that considers dynamics of the system are constructed on the basis of first principle, therefore physical relationships are considered and modelized, providing accurate results, but computationally demanding. Due to the high computational-cost, this model is hardly used for real-time applications, thus grey-box and black box are preferred in this case. Black box models are largely used in AHU application, especially for real-time implementations, thanks to its fast computation, although parameters used by these models have not a physical meaning by themselves.

The other type of analytical models is output residual models, which perform residual analysis on real input-output data and estimated states. Observer methods construct linear or non-linear models from input signals and measured signals of state variables. In this way output state are estimated and compared with actual output of the system. Parity relations methods are an alternative to observer methods, with simpler approach, matter of fact they develop linear-algebraic models. Even though does not require deep mathematical knowledge of the system functioning, it is still suitable for FDD of AHU.

Alternatively to analytical-based methods, knowledge-based methods can be used for the scope as well, in particular in the cases which the construction of a mathematical model cannot be implemented for the lack of information or to avoid high-costly methods. Common FDD for AHU methods of this class are expert systems, pattern classification and casual analysis. Casual analysis investigates the bias of variables from expected values, therefore every difference from expected value is considered as a symptom. Symptoms are taken into account with their sign, so to identify positive or negative correlation among co-occurrence of the symptoms. FDD is conducted analysing the symptoms of cause and effect identified. As concerns expert knowledge, this group of knowledge-based methods gathers the approaches that may exploit first-principle knowledge and systems description of its composition, to construct rules. This kind of methods can be shallow-knowledge expert systems if the rules are of the IF-THEN type, deep-knowledge expert systems if the rules come from first-principle laws, or machine learning methods. The latter type is most promising one, matter of fact considers those methods that are able to extract autonomously from system's data the knowledge required to construct the rules necessary for the implementation of the method itself. After the rules generation, they are tested by means of real building data for rules tuning. Another kind of knowledge-based methods implies the definition of non-linear correlation between data patterns and fault classes. Typical examples of this group of methods are ANN, fuzzy logic, SVM classifier, Bayes classifier. These methods are able to perform AFDD without model construction or expert knowledge, but at the cost of the requirement of robustness and accuracy of data used as input for the patterns generation.

A particular example of knowledge-based method for AFDD applied even in AHU cases is ARM, because this method can extract rules necessary in an unsupervised way, so to provide to the user a set of indication that can report the behaviour of the system, even though the system knowledge is not available priorly to the application of the method. The interpretation of the rules by the system's expert permits to perform FDD.

Data-driven approach can be used instead of the other two aforementioned ones. This approach exploits statistical methods to reduce dimensionality, therefore it results helpful in AFDD of AHUs, since these are systems considerably complex and with a large number of variables treated in some cases. At this class of methods belong signal processing and pattern recognition methods. As regards the first category of methods, they are rarely adopted in AHU cases, since only non-stationary signals methods have been used for this purpose. Wavelet transformation is the only method of this kind employed, but it was used in combination with a common multi-variable statistical method, that is PCA. PCA is suitable to fault detection thanks to its property of dimensionality reduction, but it cannot provide any information about cause-effect logic, therefore it cannot be implied in a fault diagnosis phase. To overcome this issue, PCA should be combined with other pattern recognition methods, allowing even an on-line implementation.

It is worth to consider that the technique described above are characterised by different advantages and drawbacks, which can drive the choice process, accepting to tolerate related disadvantages. In order to overcome the disadvantages, several combinations of different methods have been used. Main goals of combinations' development are the enhancement of accuracy, robustness and reliability.

Analytical methods have to face with their weakness in modelling requirement, novelty identifiability and computational requirement, therefore they can be combined with methods that are able to lower the cost of implementation and requirements. In particular, many examples imply the combination of analytical-based methods with knowledge-based to perform fault detection and fault diagnosis with two different approaches. The model constructed is used only for fault detection, exploiting the rules-based method to speed up the computation of diagnosis and enhance the explanation facility as well.

Alternatively, analytical methods can be combined with data-driven methods, such as PCA, which can reduce the computation time since the noise and disturbances are reduced as a result of the reduced dimensionality. The novelty identifiability is permitted as well, in addition to the reduction of model requirement, although the memory and the computation requirement for training phase is increased.

Data-driven approach can be helpful even in the acceleration of knowledge-based methods, lowering the dimension of input data of knowledge-based methods. Particularly used is the combination of data-driven methods with ANN, obtaining a method with better robustness and precision.

In order to overcome to the disadvantages encountered and acquire the lacking characteristics combinations of all the three categories of approaches should be used. (Yu, et al., 2014)

Dey and Dong proposed a methodology involving Bayesian Belief Network (BBN) to perform FDD in AHU cases. This methodology guarantees good results with a poor training requirement, since exploits a set of expert rules in combination with Air handling unit Performance Assessment Rules (APAR). The latter are able to identify the mode of functioning of the AHU by the analysis of data of control and occupancy. The modes of operation considered are heating, cooling with outdoor air, mechanical cooling with 100% outdoor air, mechanical cooling with minimum outdoor air and simultaneous heating and cooling.

These rules have been proposed by the NIST in the formulation and thresholds. In order to use these thresholds according to NIST procedure, it is necessary to set some parameters that can be estimated with a model-based approach. As a result, a little training is required for the use of the thresholds.

APAR are used as symptoms, while expert knowledge for cause-effect correlation for the construction of BBN. Once the BBN has been constructed, it must be trained by means of the exploitation of data from case studies taken from literature. At this stage the training is used to learn some parameters useful for BBN, that are the prior and conditional probabilities. The first kind of probability is related to the occurrence of a typical fault, while the other type states the probability of happening of a certain fault, if a rule is verified. Following this phase, the monitoring data are processed and used in the APAR rule-based analysis, so to create symptoms for BBN. In particular, data related to dampers and coils signals are used to define the mode of operation, according to APAR indications, then the set of rules referring to the mode identified are applied to the data. The output of the application of the rules are considered as symptoms helpful for BBN. The symptoms passing through the BBN provide, in case of single or co-occurring faults, a list of faults with their probability of occurrence. In this way diagnosis is conducted and allows the schedule of intervention operation to fix the problem related to the faults identified. (Dey & Dong, 2016)

Bayesian Belief Network is a probabilistic graphical model that shows the relationship among variables reporting their conditional dependence. Can be used as a knowledge discovery method due to the fact that it can perform diagnosis of systems with uncertain or discordant information. As regards AHU's systems, training with faulty dataset is required to exploit the diagnosing power of BBN, unless user expertise is used to skip the training phase as in the methodology proposed by Dey and Dong. Faulty data are rarely available in real building dataset, since a faults' mapping is required. As a consequence, expert knowledge can make BBN applicable even in poor data cases. (Dey & Dong, 2016)

The BBN shows the connection among variables by means of nodes' connections made with an arc representing the "direction" of the cause-effect relationship. In fact, the two correlated nodes are connected by an arc which points from the node representing the cause to the effect one. The parent nodes that are not implied in any connection as child node, namely they do not have any arc pointing to them, are called root nodes.

Two parameters characterize the nodes basing on the fact that they are root or child nodes. At each root nodes are associated prior probabilities for the states of nodes, while for the states of the child nodes are used conditional probabilities. Node' states are events related to the variables of the node considered, which can be used for the purpose of the BBN analysis. Therefore, the exploitation of the probabilities of happening of the events is at the basis of the BBN. Conditional probabilities are used to express the probability of happening the events of child nodes considering the happening of the events of parent nodes. As a consequence, it is necessary to calculate all the conditional probabilities related to every combination of the available relationship for each child node. The higher the number of parent nodes associated to a child node, the higher the number of parameters required.

The conditional probabilities can be calculated exploiting the Bayesian inference theory, which allows to extract these probabilities from the qualitative structure of the BBN model and the prior probabilities. The Bayesian theory is express by the formula

$$P(A|B) = \frac{P(AB)}{P(A)} = \frac{P(A)P(B|A)}{P(B)}$$

Equation 15

considering P(AB) as the joint probability of events A and B. Furthermore, it is possible to calculate the marginal probability for every event as

$$\mathrm{P(A)} = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$

Equation 16

Therefore, Bayesian theory can be expressed with conditional and marginal probabilities as

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

Equation 17

This expression permits the calculation of posterior probabilities using prior ones.

The structure is the result of qualitative interpretation of the relationships among the systems' variables constructed by the expert. In FDD cases, the model can be constructed in 2 possible ways, which are the manual construction of the expert or the learned construction. The first is the mostly used since it is the easiest to implement, even though a high expertise of the system is required. The latter implies the use of learning dataset containing fault and fault free data, in order to report the whole range of combinations of relationships among nodes. In FDD cases, the model gives a representation of the relationships among symptoms and faults, in order to perform diagnosis. The symptoms are considered as child nodes, since the faults are the causes of the events belonging to the child nodes. It is worth noting that it is possible to encounter even single symptoms related to an only fault, that means that the diagnosis of that specific fault can be conducted with the only verification of the symptom, making that fault's diagnosis straightforward. Other relationships make the diagnosis less immediate, since several faults can imply the same symptom, eventually they can be related to different components. To overcome this issue, Dey and Dong proposed a hybrid approach, which implies, after the manual construction of the BBN with expert's rules, a training phase with actual data, eventually from different AHUs if datasets are universal. As a consequence, FDD conducted with this approach of FDD results easy to implement, requires a less effort in data retrieving and gives good results in diagnosing. (Dey & Dong, 2016)

Zhao et. al. proposed a method exploiting the BBN manually constructed as well but exploiting symptoms' rules constructed as residual analysis. In fact, the BBN comprises a set of fault nodes and evidence nodes, such that each fault is connected to a set of rules that can describe the behavior of the system in that fault condition. The rules in evidence nodes are mainly constructed as the difference between the measured value and the estimated one on the basis of a polynomial regression model, trained on fault free data, that takes as input a value of a correlated variable from BMS data. For instance, an evidence node may consider the supply fan power consumption from BMS and the value of the same parameter estimated with a polynomial function of supply air flow rate, since it is a strictly correlated variable. The thresholds, within which the difference between these values is considered normal, are calculated with the t-student method. This approach considers the variable as normally distributed, in this case the variable is represented by the deviation of the BMS value from the expected one, so setting a threshold of 3 times the standard deviation of the variable guarantees a confidence level of 99.73%. As the confidence level is such high, it is licit to consider values beyond this threshold as anomalous. The standard deviation of the

Otherwise, the residual analysis can be conducted on the measured data and a known set point value, since the expected value in normal condition is a fixed value. In this study, the prior probability related to the faults' state are taken from previous surveys on a several VAV systems of the same kind or estimated by the authors if lacking. Conditional probabilities necessary to calculate the posterior probabilities are estimated by the authors on the basis of the data available on the system used for the validation of the method. In fact, the data used are from the ASHRAE RP-1312, which provide a map of VAV common faults, since a set of faults was artificially implemented in the facility which the monitoring data comes from. The experimental facility used for data collection has been used for other similar studies in the years, therefore the amount of data available on that specific VAV system allows the inference of probabilities of fault's occurring. (Zhao, et al., 2017) (Zhao, et al., 2015)

The faults nodes present several states, so to consider several faces of the same kind of fault, as the bias from the fault free condition can be positive or negative, as well as slight or large. The states considered are strictly related to the fault which they belong, so to cover the range of the main fault conditions for each fault node.

Fault detection is conducted on the basis of the residual analysis at evidence nodes levels, matter of fact that the violation of the rules in the evidence nodes provides an error signal that is exploited for the isolation of the fault successively. The evidence nodes' rules violated are used as input of BBN in the diagnosis phase, so to obtain beliefs of the faults states related to the evidence nodes. Beliefs are analyzed according to some main rules constructed with the goal of sifting the posterior probabilities got from the BBN-based analysis.

Firstly, two diagnosing rules are applied to check if a fault can be isolated, as these rules provide a criterion of selection of the most probable belief among the all found. The rule 1 states that the fault state to consider as the cause of the evidence nodes errors signals is the one with the belief value higher than 0.7. The rule 2 is applied in cascade with respect to the other one if the rule 1 cannot isolate a fault and states that the belief to choose is the highest one if it is higher of more than 0.3 with respect to the second highest one. (Zhao, et al., 2017) (Zhao, et al., 2015)

According to this methodology, these rules determine the only conditions able to isolate a fault, therefore the rest of the cases have to be investigated deeper in order to determine if a fault is actually present or the fault detection phase produced a false alarm. For this sake a set of diagnosing rules are applied at this stage, so to evaluate if the system's performance is affected by the violation of the evidence nodes' rules. As a consequence, if any of the rules of this stage is violated it means that a fault occurs, but the BBN is not able to diagnose it, therefore it should be diagnosed manually with the help of a checklist constructed sorting the beliefs found at the previous step. Once the checklist is completed the actual state is evaluated to create an "additional information node" that can take into consideration that condition too, in order to fine tune the method and enlarge the range of possible situation that it is able to detect. In the case which the performance evaluating rules are respected, it means that the fault detection stage generates a false alarm, therefore the system is considered fault free.

The results obtained shows that faults can be isolated by means of this method in most cases, since small faults are resulted hard to diagnose, matter of fact that they have small effect on the functioning of the system. Even though the methodology is as easy to implement as effective, the definition of parameters represents a great issue to face with, due to the fact that an high number of parent node related to the same child node can make the conditional probabilities calculation too complex or the finding of the necessary parameters is strictly related to the knowledge of the system and prior expertise. (Zhao, et al., 2017) (Zhao, et al., 2015)

A method proposed by Yu et al. implies the use of FP-Growth algorithm in order to find correlations between different variables and looking for hidden information about the actual operation of the system. FP-Growth is an algorithm able to find correlations between contemporaneous events belonging to different variables, in a faster way with respect to the most used Apriori algorithm. Several software can allow the implementation of FP-Growth, such as SPMF or RapidMiner. This algorithm can manipulate only categorical data, as the majority of association rules algorithms, so a prior data pre-processing is necessary.

Therefore, the methodology proposed implies the collection of data from the monitoring part of the system to study, then a data pre-processing phase to eliminate outliers (i.e. remove noise in generated data) and to discretize data is carried out. Successively association rules are found and analysed to extract hidden knowledge. The data used to test this methodology is 2-years monitoring dataset of 2 building with same location but different features and utilization. The spaces to study are AHU served and their utilization varies from office to chemical lab.

The methodology proposed includes the association rules mining (ARM) in 4 different cases per each building's dataset. In fact, the ARM is conducted for a typical day in 1st year, a typical day in 2nd year and the whole 1st and 2nd years. In this way, 4 rule sets are found, then they are compared coupling the rules sets of the same kind. (Yu, et al., 2012)

The comparison phase belongs to the knowledge extraction's phase, since the presence of concordant or discordant rules may be studied at component level and system level to search an anomalous behaviour. The rules are studied differently in the cases they are referred to a typical day or the whole year. In the first case rules are compared with the information known about the system from the expert of the system itself, so to optimize the control and the use of the system. In the other case, the rules found are used to perform fault detection, since more than one day data are considered, which embrace faultless and eventually faulty data. In the yearly data rules, 2 categories are identified from the rules sets comparison.

If the same rules are found in both the dataset compared, it means the that rules can be used as reference for normal operation of the system, therefore the violation of that rule represents an anomalous behaviour. This kind of rules can be used for on line fault detection as they represent a necessary condition to respect in normal operation state.

If there are rules discordant in premises or conclusions, it means that further investigation is required, in particular at component level, so to understand which is the rule that describes the expected behaviour and identify the faulty operational data. The deeper analysis of the discordant rules can involve even the analysis of rules found for different but analogous components, in order to get aware of the real functioning of the system. (Yu, et al., 2012)

Fontugne et al. proposed the Strip, Bind and Search method, enabling the possibility to detect multiple faults with the use of data-driven and qualitative approaches. In particular the method in the strip and bind phase identifies inter-device patterns isolating occupancy-induced trends and finding the correlation between the devices. The search phase keeps monitoring the devices over time, reporting deviations in their behaviour.

A great advantage of this technique is that it is not required any prior knowledge of the system to filter dominant trend, such as occupancy or weather influence. Data derived from devices' monitoring are considered as raw signals which are treated as summation of sets of components called Intrinsic Mode Function (IMF). These components are revealed by means of a sifting process carried on by the Empirical Mode Decomposition (EMD). The sifting process isolates components by frequency from the highest to the lowest in the signal, iteratively isolating the component with the highest frequency in the signal and repeat the operation to the residual of the signal.

In the successive step the IMF are aggregated, building clusters of components of near frequency, so to divide components of raw signals in clusters of different time scale ranges. The choice of time scales and thresholds is a qualitative matter, so that high frequencies are classified as noise, medium ones as devices usage and low ones as devices patterns. Time scales are defined by the expertise of the user, the knowledge of the system and the goal to reach. (Fontugne, et al., 2013)

Once IMF have been aggregated, a pairwise comparison is conducted in order to find correlation between devices at different time scales. In this way the correlation between two devices is evaluated related to the behaviour in the time scale of our interest, eliminating the influence of dominant factors. From this pairwise comparison, a correlation matrix is constructed helping the identification of misbehaving.

The search phase points to exploit the correlation information in a unsupervised way, keeping monitoring the devices and comparing data collected with a normal inter-device usage pattern, reporting deviation from expected behaviour. Medians of observations are considered as normal patterns, since anomalies, faulty deviations, are rare by definition. (Fontugne, et al., 2013)

Yang et al. combined a statistical residual-based technique to a Fractal correlation dimension-based one, in order to improve the sensitivity of the analysis. In fact, the first is used to detect large bias, while FCD-based is used for small bias since it is more sensitive. Statistical residual-based uses Support Vector Regression to compute predicted values to compare to measured values. If a large bias is not found, analysis is pushed forward to find potential small biases.

Fractal dimension is a statistical quantity indicating how completely a fractal appears to fill the space. Fractal dimension measures the dimensionality of the space occupied by a set of random points. Fractal dimension is calculated for measured data and predicted data both, in order to compute the difference and identify small biases. (Yang, et al., 2011)

Wu and Sun proposed a methodology of FDD with a training phase carried out off- line and a real-time phase. The methodology is applied to AHU case, analysing data in terms of space and time, so to partition data according to them and contextualise the statistical analysis conducted after the partitioning. The goal of the statistical analysis is the threshold identification for FDD. In this phase, belonging to the off-line part, PCA and correlation analysis are used to find normal patterns, therefore thresholds of normal operation come as a consequent. Even though the analysis is focused on the energy consumption, the other variables are taken in consideration as well by the techniques used. Once the thresholds are defined, the real-time procedure can be applied, since it includes a sensor FDD phase, then a cross-level FDD with faults identification and finally diagnosis of faults analysing at component level. The thresholds are updated considering real-time data at the end of the process. (Wu & Sun, 2011)

Sensor FDD is necessary in real-time procedure, so to process only reasonable data. This phase is conducted by means of a trend analysis of data, identifying thresholds for each sensor data stream. Data from components are analysed to check abnormal values in absolute terms if values assumed are beyond thresholds or in relative terms if the values are abnormal compared to the ones from similar component of the system. Comparison among similar components is used to have a confirmation of the presence of a fault for a system's component. The power of this method is the use of only a parameter to check for fault detection, namely the flow of energy consumption of a single component, and the contextualization of the analysis, that is the partitioning before the thresholds' identification. (Wu & Sun, 2011)

Qin and Wang proposed a methodology that uses several approaches to perform FDD in VAV units. According to this method, a set of faults for VAV units are identified, then a set of causes are found, so to direct the research towards them. The investigation on the presence of the faults is performed cascade alike, guaranteeing a multiple faults identification as well. In this way, interaction among faults is taken into account, as well as it is prevented the interaction among faults that have to be studied in different moments. Faults that are correlated are studied in parallel, while the others are studied at sequential steps. (Qin & Wang, 2005)

Expert rules, performance indexes and PCA are employed along the procedure's steps in order to approach differently and in the most convenient way to the different faults. In the cases which performance indexes are used, the 3-sigma method is used, so a fault is detected if the value assumed is beyond the threshold of 3 times the standard deviation calculated in normal condition. If expert's rules are adopted the value provided by the sensors are checked by the rules and considered faulty if exceed the user defined threshold on the parameter's error. PCA is used only in the case of flow sensor bias, since the bias in flow values can drive to modifications in the behaviour of a large number of variables in the system, therefore this kind of multivariate analysis technique results very helpful in this case. A sensor bias can be considered harmless if small and within the range, but if because of it the values are close to the threshold the control chain can be affected to it. Other methods can fail in sensor bias detection due to the large number of variables related to the flow measurement's value. (Qin & Wang, 2005)

Liang and Du proposed a hybrid method that merges a model-based approach with SVM, in order to exploit the high accuracy of the system's model and the power of classification of faults of SVM. The model used in the methodology is a lumped-parameter one, so to guarantee a high level of accuracy without a high-cost implementation and computation.

The model has been developed off-line on the basis of normal operation data, so to exploit it to generate values employable in statistical methods and create residuals. During on-line phase, by means of the residuals analysis, comparing residuals with pre-set thresholds, faults can be detected. During this phase the model values are compared with real plant data so to check if the error is within thresholds limits. Therefore, the values provided by the model are considered as the expected values for the parameters taken into account by the monitoring of the system. The choice of thresholds represents an issue to face to, since should be find a trad-off between the increase of the probability to detect faults and the preventing of false alarms. (Liang & Du, 2007)

Successively SVM can be used for diagnosis. The SVM requires a little effort on training before its application, but the diagnosis accuracy obtained is high diagnosis. The SVM results a cost-effective technique for handling the faults diagnosis, considering non-linear effects too. For this kind of application SVM used is a layered one, so to allow it to classify all the faults considered, one for each layer. The checks of the presence of the faults are conducted in cascade, with most frequent faults first and then the others in frequency decreasing order, to speed up the computation. (Liang & Du, 2007)

3.1.2. FDD methods applied to other building energy cases

Even though AHU represents the highest share in the total energy consumption of a building, other energy systems of building has been taken into consideration too, in order to apply the FDD approach in the energy saving for these cases.

Considering the diversity of the nature of the rest of the systems, the methods applied are set properly for the system taken into account. In this sense, the work of the system expert points to the choice of the most appropriate and effective method for the specific kind of application.

Following it is reported a selection of the most interesting methods applied to various building energy cases such as district heating or chillers.

One of the most employed technique is ARM, due to its great potentiality and easy comprehension, therefore many are the examples of application of this technique in FDD.

ARM for FDD can be applied to several kinds of systems, in fact in literature there are examples even of application of this technique to District Heating cases. A particular methodology, proposed by Xue et al., implies the use of partitioning techniques before the ARM phase, such that the mining is contextualized. Data cleaning and transformation remain crucial phases to perform before the Association analysis, but prior the ARM is conducted data are treated as to find clusters that highlight different operating patterns.

In the methodology proposed the data transformation phase implies the reduction of the dimensionality of data, selecting a lower number attribute for the analysis, then the discretization of variables is conducted to obtain categorical data, which ARM algorithm is able to treat. The discretization issue is overcome, in this case, using an unsupervised technique to group data, that is K-means. It is a partitive cluster algorithm that can find the best interval for discretization in unsupervised way, therefore the user it is not involved in the selection of the boundaries of interval used for data grouping.

The K-means algorithm can be used even for the following phase which is the cluster analysis applied to the whole dataset. This phase is conducted to find different operational patterns and conduct the ARM only in defined condition, thus the rules found are strictly related to the operational condition of the cluster which they were found. Other algorithm can be used alternatively to K-means, such that partitioning around medoids (PAM) or agglomerative hierarchical clustering. The first is a partitive algorithm, as K-means, that divides data in a pre-defined number of cluster and then optimize the position of the reference point of clusters. The latter is a hierarchical algorithm, which creates clusters with a bottom-up logic, starting from a number of cluster equal to the number of elements and reaching the situation which all elements are included in the same cluster.

The methodology in the successive step uses a quite common ARM algorithm to perform the analysis, that is Apriori. This algorithm is quite efficient in finding association between events of different variables happening in the same time. The last phase is the interpretation of the rules found, since not all rules are useful for the analysis, therefore knowledge of the system and the comparison of the rules found can bring to the extraction of information from rules. (Xue, et al., 2017)

Cluster analysis in the specific case study treated by the authors of the methodology, identifies 3 different operational condition, named slightly, moderate or severely cold, according to the quantity of power required by the user, namely the shape of the profile of power itself. This 3 clusters defines 3 different regimes that can be found along the heating season. Analysis of rules has been driven to classify them in 3 categories according to the goal that can be achieved by means of them, that is identification of different regulation strategies, fault detection or identification of energy inefficient operation.

In particular the first category exploits the possibility to have a deeper understanding of the behaviour of the system studied, since it is possible to find association between variables and components which can be related to regulation strategies (which are unknown to the user) that can influence the following control strategy. The second category leverages on the chance to find rules showing anomalous association of well-known components, therefore a fault can be detected. A fault can be detected even comparing rules with data related to it, so looking for anomalies in time series or illogical operations. The last category refers to the cases which anomalous operation is carried on and energy waste is performed as out of schedule functioning of certain components. (Xue, et al., 2017)

Bynum et al. developed an Automated Building Commissioning Analysis Tool (ABCAT) which merges user expertise with quantitative methods in FDD problems. ABCAT is a white box model, a calibrated, simplified and first principles based mathematical model, which is able to predict energy consumption under given weather conditions. Since it is proposed as a simplified and cost-effective method, it only requires 3 sensors, which are in charge of collecting whole building electricity consumption, heating and cooling. (Bynum, et al., 2012)

The model is used to produce a normal condition reference of the behaviour of the system, considering actual weather condition, in this way data collected are identified as anomalous or not based on expected values calculated by the model. Deviation from expected values is monitored and presented to the user in several graphical representations in order to let the user judge if there is a fault or not and which kind of intervention is necessary. If a deviation is found, but there are not any faults to report, the model is recalibrated using the data just collected. Diagnosis is carried on by the user after a fault is detected, basing on his expertise, since the model is not accurate in the system description to diagnose automatically the system. (Bynum, et al., 2012)

Li and Braun have been using a decoupling feature analysis combined with qualitative approach, so to find multiple simultaneous faults. The technique of fault detection is able to find individual faults using a normalized distance fault detection classifier

$$(Y - M_{normal})^T \Sigma^{-1} (Y - M_{normal}) \quad \underset{\omega_2:Faulty}{\overset{\omega_1:Normal}{\lesseqgtr}} \quad (\chi^2)^{-1}\{(1-\alpha), m\},$$

Equation 18

with $(Y - M_{normal})^T \Sigma^{-1} (Y - M_{normal})$ as the normalized distance, $(\chi^2)^{-1}\{(1-\alpha), m\}$ as the threshold of the normalized distance for normal operation, $(\chi^2)^{-1}\{,\}$ as the inverse of the Chi-square cumulative distribution function. A is the false alarm rate, *m* is the degree of freedom or dimension that is equal to the number of chosen state variables. Class ω_1,normal operation, is selected if the left-hand side is less than the right-hand side, and class ω_2, faulty operation, is selected otherwise. Because of the modeling error, M_{normal} is not exactly zero, so the equation takes modeling error into account to statistically evaluate whether *Y* is zero or not.

Using only normal operation data, it identifies a normal condition area outside which there are faults. Diagnosis is made finding the location of the fault, so x and y of the points outside the normal operation area. The coordinate Y is defined as $Y = F(X)$. If F(x) is known, diagnosis is easy to implement, but if it is hard to have a formulation for F(x), a McLaurin's series is used to express F(x) in an analytical form. Therefore

$$Y = F(0) + \frac{\partial F}{\partial X}(0)(X - 0) = JX,$$

Equation 19

where $F(0) = 0$ and

$$J = \frac{\partial F}{\partial X}(0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Equation 20

Therefore $X = J^{-1}Y$, that is equivalent to write the signum function of each member of the equation, using a rule-based method proposed by Rossi and Braun (1997) , so

$$Y_{sign} = J_{sign} X_{sign} = sign\left(\left[\frac{df_1}{dx_i}, \frac{df_2}{dx_i}, \dots, \frac{df_n}{dx_i}\right]^T\right)$$

Equation 21

$$X_{sign} = sign\left({J^T}_{sign} Y_{sign} - [n, n, \dots, n]^T\right).$$

Equation 22

As a result a rule-based approach supports and simplifies the identification of the coordinates of the fault making easier the diagnosis. J_{sign} is generic for the same kind of system so the model is not strictly related to the system which is applied to, furthermore change pattern directions converts an infinite classification into a multiple classification. The issue of this formulation is the incapability to handle multiple faults.

In order to enable multiple simultaneous faults identification, a decoupling-based fault diagnosis method has been implemented. Therefore the previous formulation has been modified as

$$PY = PJX$$

Equation 23

$$Z = \Lambda X = |\lambda_1 x_1, \lambda_2 x_2, \dots, \lambda_n x_n|^T,$$

Equation 24

where $\Lambda = PJ = Diag(|\lambda_1, \lambda_2, \dots, \lambda_n|)$, $Z = PJ$ is the transformed feature vector, and $P = \Lambda J^{-1}$ is the transformation matrix to make Λ diagonal. There exists an infinite number of transformation combinations of Λ, *P* and *Z* that can be obtained by arbitrarily choosing a diagonal Λ if matrix *J* is non-singular (this can be guaranteed by proper choice of *Y* physically). This transformation decouples interactions among the different faults and makes each entry of the feature vector *Z* only correspond to a unique fault entry vector *X* and vice versa.

$$X = \Lambda^{-1}\mathrm{Z} = \left[\frac{z_1}{\lambda_1}, \frac{z_2}{\lambda_2}, \dots, \frac{z_n}{\lambda_n}\right]^T$$

Equation 25

The signum function has been used at this point to eliminte impacts of linearization operation and driving-condition-independence assumption on diagnosis.

$$sign(X) = sign(\Lambda^{-1})sign\_stat(Z)$$

Equation 26

Where *sign_stat(Z)* is a signum operation of each entry (z_i) of matrix *Z* in a statistical sense, such that

$$sign\_stat(z_i) = \begin{cases} -1, if \ \dfrac{(z_i - \mu_{i,normal})}{\sigma_{i,normal}} < -\sqrt{(\chi^2)^{-1}\{(1-\alpha), 1\}} \\ 0, if \ \dfrac{(z_i - \mu_{i,normal})}{\sigma_{i,normal}} \leq -\sqrt{(\chi^2)^{-1}\{(1-\alpha), 1\}} \\ 1, if \ \dfrac{(z_i - \mu_{i,normal})}{\sigma_{i,normal}} > -\sqrt{(\chi^2)^{-1}\{(1-\alpha), 1\}} \end{cases}$$

Equation 27

It can be seen that the signum operation of each entry (z_i) of matrix *Z* is equivalent to evaluating the one-dimensional fault detection classifier of normalized distance inequation.

$$X_{sign} = \left[\frac{sign\_stat(z_1)}{sign(\lambda_1)}, \frac{sign\_stat(z_2)}{sign(\lambda_2)}, \dots, \frac{sign\_stat(z_n)}{sign(\lambda_n)}\right]^T$$

Equation 28

This formulation simplifies fault detection from high-dimensional problem to n one-dimensional problems. It makes the method more generic and system-independent, avoiding rules strictly related to the system. (Li & Braun, 2007)

Lin and Claridge developed a method to identify faults analysing consumption data and outdoor temperature, in combination with the use of a quantitative method.

A fault is found if the difference between simulated and measured consumption data is greater than a standard deviation of the residuals between measured and simulated consumption in the baseline period and persists for at least 20 days, which are consecutive when ordered according to increasing or decreasing outside air temperature within data being analysed. This statement has meaning assuming that there are not any faults in the baseline. For this sake a simulation model is needed in order to predict normal building energy performance, making the reference data for the fault detection investigation. (Lin & Claridge, 2015)

The procedure involves the creation of a fault detection matrix, in which the fault detection criterion is applied day by day, associating a fault index to each day. The detection matrix contains date, average outdoor temperature and fault index for each day to investigate. The fault index is defined as

$$Fault\ Index = \begin{cases} 1 & E_{mea} - E_{sim} > SD\_baseline \\ -1 & E_{mea} - E_{sim} < SD\_baseline \\ 0 & otherwise \end{cases}$$

Equation 29

With this matrix mean outdoor temperature is correlated to daily energy consumption and the fault index.

Successively a transition matrix is built as a portion of the detection matrix updating iteratively the matrix, adding one day more each step if no faults are detected.

The transition matrix is used in the fault identification phase, which consists in the examination of fault the index day by day. If the fault index is 1 or -1 for a day, rows of transition matrix greater than or equal the date of fault are imported in a sub transition matrix. A sort in ascending way of sub transition matrix is made. The sub transition matrix has the same information of the previous matrices but reports the sum of fault indexes too, since the last column contains the sum of indexes of following 20 rows. A fault is identified if the sum of the indexes is 20 or -20.

Severity of fault is evaluated with energy consumption indexes as Energy Impact of Fault, that is the ratio of cumulative energy consumption variation from fault occurrence to fault identification day to annual energy consumption when there is no fault, or Annual Energy Impact, that is the ratio between the total energy consumption variation in one year due to a synthetic control change and annual energy consumption with no fault. (Lin & Claridge, 2015)

Wang and Cui combined a black box model-based approach to a Principal Component Analysis, so to implement an on-line method able to identify sensors' faults before the data from monitoring is used for fault detection of the system itself. This method improves the robustness of the residual analysis based on a black box model, by means of a PCA analysis of the quality of sensors' data.

Only if a measurement is considered valid, it is tested for system's fault detection. Both the schemes are based on steady-state, so data from system's panel and BMS goes through data pre-processor, to filter out transient data, then anomalies in sensors' operation is investigated. Q-statistics using PCA is applied so to detect faults in sensors. If Q-statistics is higher than a defined threshold, Q-contribution are analysed so to detect fault affecting the sensor. Otherwise data are passed to the system's fault detection procedure.

With PCA analysis only variables of great concern are kept and used for the analysis, transforming the set of correlated variables in a set of uncorrelated variables. In the system's fault detection phase, the method is considerable as a residual analysis of performance indices, which uses as reference model for normal operation data a black box model. (Wang & Cui, 2006)

Wang et al. developed a method to improve the robustness of a simulation model in order to fit better to the real system. This method uses an EnergyPlus simulation model based on monitoring data, so to combine real data and first principles. In this way the objective of this methodology is to overcome the limits of the software in modelling completely the plant, if it has configurations different from the ones pre-defined in the software itself. Therefore, the model is calibrated using monitoring data, so input parameters of the simulation model are adjusted according to the deviation of energy consumptions, both components' level and whole building level. As a result, prediction is improved thanks to monitoring data. Once the model is implemented and fine-tuned, a series of energy conservation measures are identified, using the model to evaluate the effectiveness. (Wang, et al., 2013)

Zhao et al developed a method merging knowledge of the system and data-driven techniques, by means of a decoupling-based method, so to deal with multiple simultaneous faults. In the application of this method is crucial the identification of specifically pertinent decoupling features for each subsystem. In fact, decoupling-features are tailored to the faults of the systems studied, taking into account the performance sheets produced after manufacturer's factory tests. (Zhao, et al., 2014)

The method consists in defining common faults for the system and associate them virtual data points coming from monitoring data. Virtual data points are calculated extracting information from correlation between actual measurements and theoretical data points described in manufacturers' data sheets. Therefore, virtual data points are indexes indicating how the system is working related to the prescribed operation or normal operation. For this sake data must be filtered in order to eliminate the transient condition and evaluate only steady-state condition.

Once a set of indices for virtual data points calculation is identified and modelled, the FDD methodology is applied, reporting a fault every time the values associated to a virtual data point' index is beyond a defined threshold. These thresholds must be tuned in order to accomplish both the objective of sensitivity and robustness of the methodology. (Zhao, et al., 2014)

Arseniev et al. proposed a methodology that exploits ANN to accomplish FDD, but on the basis of a rule-based training of the system. In fact, a rule-based system is implemented to provide expert knowledge to the procedure and considerably speed up the process of FDD. The training of the ANN system is executed by means of a supervisor constructed on a rule-based approach, then the learning phase using the ANN is carried out. (Arseniev, et al., 2009)

Kocyigit proposed a methodology exploiting Fuzzy Inference System (FIS) together with Artificial Neural Network (ANN) to conduct diagnose faults in a vapor compression refrigeration system. The two techniques are used to classify system faults or sensors' error from sensor data. FIS is able to classify data by means of fuzzy logic, by the application of Membership functions (MF) and rules. These rules are linguistic-type, namely if-then form, so to assign a membership degree to data, therefore each input data get a value for every class defining the degree of belonging to that class. FIS is used for sensor error detection and diagnosis, since it can categorize data basing on the values collected by the sensors, so can spot anomalous values for one or two sensors at the same time. ANN is used to conduct fault diagnosis on the system, therefore output data of FIS fault diagnosis is then processed by ANN for training and testing. In this way the methodology can conduct fault diagnosis for sensor data fist, then for system data. (Kocyigit, 2015)

4. Methodology

In this section a novel methodology to perform Fault Detection and Diagnosis is proposed, by means of a general description of the approach and the discussion of the different techniques used. Furthermore, a general discussion about how the method has been developed is provided, even though the detailed description of intermediate results is omitted, since they were not relevant for the purpose, but useful for the choice of the best techniques to employ in each phase of the methodology.

The development of the methodology pointed towards a most unsupervised and automatic approach possible, in order to receive as input a time series from BMS data and provide as output the indication of the cause of an eventually occurring fault. The process aims to be unsupervised in order to be generalizable and scalable to every kind of system, without the intervention of the operator or the system expert in the input information compiling. In this way, the results are not influenced by the prior knowledge of the system of expert, which can insert specificity and uncertainty to the results. The methodology thought as unsupervised, can adapt itself to the kind of data which it is applied to, therefore automatically recognize the best way to treat that data.

Furthermore, the methodology has been intended to be employed even on-line application, in order to be a support for FDD to the management of the system right during the control of the system itself.

4.1. Description

The methodology proposed has been tested on ASHRAE RP-1312 dataset and has been thought on the basis of different methodologies retrieved in literature. The methodology bases itself on the idea of applying the association rules mining with a temporal relationship among the events involved in the rule. The rules, which this methodology is looking for, describe a cause-effect relationship among events occurring in a time interval. In this way not only instantaneous co-occurrence of events of different variables, but even occurrence of events with a temporal dependency are taken into account, so to consider even a sort of inertial effect in the occurring events or a temporal lag in the reaction of the different variables involved in the system's analysis.

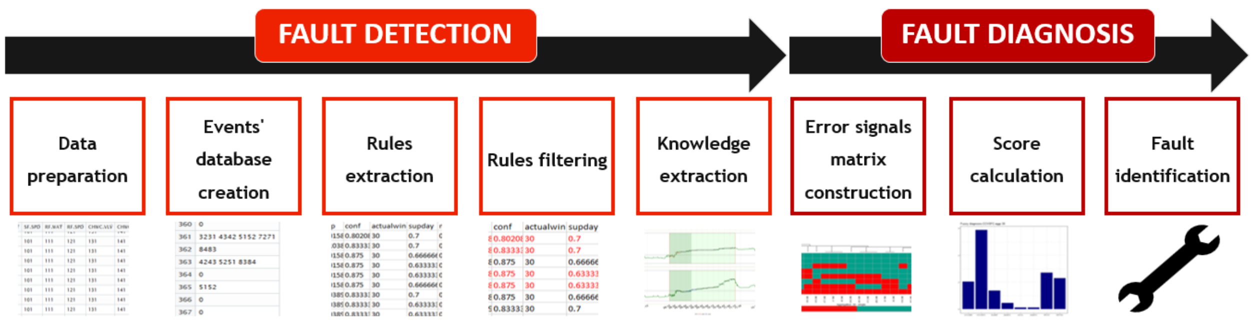


Figure 11 Framework of the Fault Detection and Diagnosis methodology

Temporality is considered in the rules mining as the methodology takes into account the consequentiality of the happening of events and a time interval constraint, which limits the temporal distance between the antecedent and the consequent in the rules.

The implementation of the whole methodology has been carried out on R, using several packages of functions. The whole methodology comprises two main stages: the fault detection and the fault diagnosis.

The fault detection stage of the methodology is made up of 5 phases: data preparation, database creation, rules extraction, rules filtering and knowledge extraction. (Figure 11)

The fault diagnosis stage comprises the error signals matrix construction, the score calculation and the fault identification at the end. (Figure 11)

In first phase of fault detection, namely the Data Preparation phase, raw data are aggregated, discretized and encoded, so to have low dimensional time series. Original dataset which the methodology can be applied to, may have a frequency of sampling higher than necessary, that means that the amount of data to treat may be high, with a large computational cost and unnecessary sensitivity in the time series considered. To overcome these issues, aggregation can be carried out. The most used aggregation technique is the piecewise aggregation approximation (PAA), which approximates the time series to the mean value calculated in fixed width time interval. As a consequence, the PAA time series is a piecewise constant function of the same length of the original time series, therefore the size of the time series can be reduced substituting each interval with a single point assuming the value assumed along the interval which is referring to. The choice of the width of these intervals represents the issue to face to in this step. In fact, the higher the aggregation time interval, the lower the computational cost of the rest of the procedure, but the lower the grade of the quality of information carrying in the PAA series too. Therefore, the choice of the interval of aggregation depends on the sensitivity and the level of knowledge the user is interested to.

Discretization is an obliged step, since most association rules algorithms use categorical variables. This represents a great issue to overcome since the quality and number of rules extracted at the end, are seriously affected by this step. A mis-discretization can bring to a poor description of the possible states of a variable, for instance because data belonging to anomalous states can be included in a state that describes a normal operation. (Fan & Xiao, 2017)

Once data has been aggregated and discretized, data is encoded in order to identify uniquely the variables which rules are referring to. In this way variables are numbered as well as states of the variables, therefore for each value the rightmost digits identify the state, while the rest of the digits identify the variable (Figure 12). In the application of the methodology the code can accept a number of states, i.e. intervals of discretization, for a maximum of 5, matter of fact it is not needed to have a higher number of states, so only one digit has been allocated for level identification and even more than one for variable identification, but it only affects R code writing.

RA.HUMD	RA.m3m	OA.TEMP	SF.WAT	SF.SPD	RF.WAT	RF.SPD	CHWC.VLV	CHWC.LM	CHWC.DAT	CHWC.LWT	E_ccoil
64	71	81	91	101	111	121	131	141	152	162	171
64	71	81	91	101	111	121	131	141	152	162	171
64	71	81	91	101	111	121	131	141	152	162	171
64	71	81	91	101	111	121	131	141	152	162	171
64	71	81	91	101	111	121	131	141	152	162	171
63	71	81	91	101	111	121	131	141	152	162	171
63	71	81	91	101	111	121	131	141	152	162	171
63	71	81	91	101	111	121	131	141	152	162	171

Figure 12 Dataset of discrete encoded variables

After Data Preparation, dimensionality of data has been reduced further, transforming the several discrete multivariate time series into a single multivariate events' timeseries organized in an inter-transactional database, in the Events Database Creation phase.

In this phase, first of all, the multivariate time series has to be transformed in events' times series. In order to do so the dataset has been manipulated to construct a multivariate time series containing only the transitions for the variables considered. In this way each instance can assume the value 0 if there are not any changes of state in that instant of time or the code identifying the transition. For each variable a run length encoding is carried out, by means of the function "rle" of R's base package. This function locates the changes in time series allowing to store only the passage from a level to another for each variable. Consequently, transitions are encoded such that it is possible to merely understand the variable involved and the kind of transition occurring. The transition code is simply the concatenation of the codes of the starting state and the ending state of the transition. Formally it consists in a transformation of an array of numeric or factor, to an array of character, thus starting from this point each instance of the time series is handled as a string.

At this point each time series is composed of a bunch of zeros interrupted by transitions' codes. The following step is the concatenation of all the variables at every instance, so to compact all the time series into a unique one. This means that item sets are created for every instance of the database, keeping non-transitions' instances as 0. As a consequence, the instances in which there are many events occurring are represented as a series of spaced transitions' codes. (Figure 13)

4243 7172
0
3231 4342 5152 7271
8483
4243 5251 8384
0
5152
0
0
0
-

Figure 13 Transitions' time series

This approach is similar to one used in the CONOTOOL software developed by EDMANS research team, except for the kind of events taken into consideration, since in this methodology only incremental and decremental events are considered, while the CONOTOOL software can recognize and store incremental, decremental, horizontal or according to a threshold events. (Martínez-de-Pisón Ascacíbar, et al., 2009) This approach tries to add to the transition's direction information, the severity information, similarly to Xiao's work about temporal rule mining, in which at each event is associated a trend that identifies the grade of change of it, i.e. distinguishing large, moderate and slight changes. (Fan, et al., 2015)

Once the transitions' time series has been created, the following steps have the goal to obtain the inter-transactional database of events, so to create the input for an association rules mining algorithm which can look for rules present into sequences, i.e. transactions.

For the sake of generality, the inter-transactional approach is used, but in contrast to the CONOTOOL software, the whole dataset is organized in sequences, not only the temporal windows involving the events of our interest. In this way it is not necessary to know priorly the kind of faulty event to look for, therefore the symptom to search can be any. The CONOTOOL approach basis on the knowledge of the symptom related to the fault to investigate, so it does not result completely unsupervised, unlike the methodology proposed in this dissertation.

For the purpose of creating an unsupervised data base of transactions, a sliding window is implemented for constructing a database of transactions composed by chunks of the times series, resulting from the overlap of the time windows. The user can define the width of the time windows and eventually repeat the procedure for different windows width, such that it is possible to relate association rules to time windows' width.

This technique can be considered as an unsupervised sliding window, since there are not any constraints related to events present in the time series. The only constraint that can be possibly set is the interruption of the sliding of windows at the end of the day or not. This constraint is useful if the dataset, which the methodology is applied to, contains consequential day or not. By means of it, the case of positioning a time window between two days consecutive in the dataset, but not temporally consecutive can be avoided. Considering that the time windows are sliding a timestamp by time, two consecutive sequences differs only for a single item set, in addition to the fact that the database constructed has a number of sequences, i.e. transactions, equal to the number of instances of the time series. Basically, each sequence of the database represents the portion of the time series starting from the corresponding instance and ends a time window width after. (Figure 14)

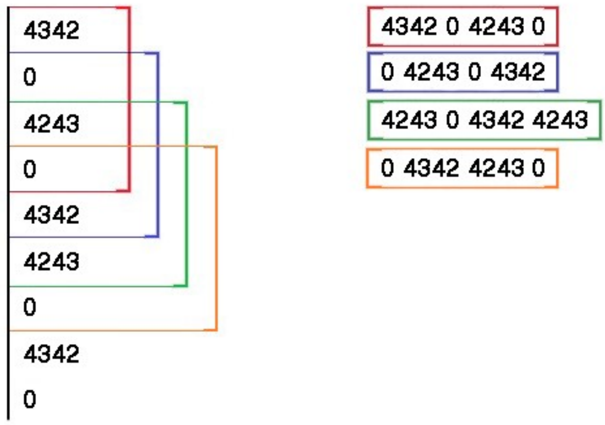


Figure 14 Inter-transactional item sets database creation

The implementation of the sliding window such that, allows the possibility to use sequential rules mining algorithms or intra-transactional algorithms, maintaining an inter-transactional approach.

The Rules Extraction phase consists in the applying of association rules mining algorithms in order to retrieve rules from the database. Several algorithms have been applied during the methodology test, to find the most convenient and effective one. The cSpade algorithm results the optimal choice in terms of results and convenience, since it is included in a function belonging to the "arules" R package, so it is not necessary to use other tools to apply it, provides many rules with high number of variables involved, it is quite computationally low-cost, it can use a "0-less" database as input (i.e. in favor of computational efficiency), it provides the pattern of the rules in temporal order in contrast with lexicographic order of other ones and has a set of parameters that can be used to enhance the focus of the rules searching (e.g. setting a minimum or maximum gap between antecedent and consequent pattern or limiting the length of the antecedent or the consequent pattern).

The rules extracted has to be filtered in different ways basing on the kind of algorithm used, therefore the Rules Filtering phase is conducted at this point. Since 0s can be present in the rules, vitiating the corresponding values of support and confidence or given non-sense to rules, these rules have to be removed from the rules set. Furthermore, some algorithms can generate multiple version of the same rule, i.e. permutation of the events involved in the rule, even swapping antecedent events with consequent ones, therefore these cases have to be removed.

Only one kind of filter is common to every algorithm, that is the sub-pattern identification. This kind of filter removes the rules that express the relationship between events already described in other rules with longer pattern. Basically, subpattern's rules are portions of bigger rules, so they are removed since the situation which they are describing is already considered differently.

Another aspect to take into account is that the Rules Extraction phase is strictly related to the parameter set for the rules mining, such as minimum support and minimum confidence of the rules to find. That is an important point since, as concerns support, it is worth noting that it is calculated on the basis of the inter-transactional sequences, therefore it should be kept in mind that the number of sequences is huge, since there is a sequence for each instance of the time series. As a consequence, values of support found by the algorithms is very low, thus the introduction of some extra parameter is necessary for the sake of better comprehension of the quality of the rule found. For this purpose, it has been implemented the support daily based, i.e. SUPP.DAY, and the mean occurrence per day, i.e. MOD. The first it is nothing else the ratio of the number of days which the rule is present over the total number of days. (Equation 30) Essentially, it is the percentage of days which the rule is occurring. The latter is the total number of occurrences divided by the number of days, namely the mean number of times which the rule occurs in a day. (Equation 31)

$$SUPP.DAY = \frac{|rule\ occurring\ day\ |}{|day\ of\ dataset|}$$

Equation 30

$$MOD = \frac{|rule\ occurring|}{|day\ of\ dataset|}$$

Equation 31

These two parameters are not provided by any algorithm, therefore the calculation of these has been implemented separately, by searching the pattern along the whole time series, bearing in mind the time window constraints and the consequentiality of the occurrence of the events present in the pattern. Since the support calculated by the algorithm itself is vitiated by the huge number of transactions generated and treated, these two parameters have to be recalculated with different formulations. According to the formulation of CONOTOOL, the support the number of the transaction is reduced to the number of the transactions containing the consequent of the rule. In this way the denominator is dramatically lowered obtaining relevant values of support. This different formulation results effective in the situation like this which the number of transactions is very large. (Martínez-de-Pisón Ascacíbar, et al., 2009) After the calculation of support with this different formulation, a filtering of the rules if performed exploiting this parameter, setting a significant threshold to filter out uninteresting for the purposes rules.

The analysis on multiple time window includes the filtering of the same rules showing themselves on different time windows width, therefore another stage of filtering takes into account this information. In fact, after this type of filtering, to each rule is assigned the lowest value of time window width related to them, matter of fact wider time span simply contains the shorter ones. In this way, considering a real time application of this method, it is not necessary to allocate the space for a wide time window if the rule extinguishes itself in a shorter time lapse.

During the search of the pattern in the time series, even other information is stored, such as the time of the day which the rule is present or the actual time window. The time of the day related to a rule allows distinguishing rules related to the startup, the shutdown or the operation along the day. This information is fundamental, since the violation of a rule can be a symptom of a fault at a time of a day, while can be absolutely normal in another time of the day. To understand the time of the day effect in the categorization of the rules, it is enough thinking about how the variables act differently at the startup and the shutdown.

The actual time window gives an information about the distribution of the temporal length of the patterns of the rules found. Basically, the methodology associates to each occurrence of the rules along the data set, the temporal distance between the first event of the antecedent and the last of its consequent. In this way, even though a time window is set, the user can have an indication of the most common interval of time related to the extinguishing of the rule when it is occurring. For instance, in real time application, if the user observes the violation of a rule after the most actual time window, he receives a first alert which can be useful to detect and eventually diagnose the fault before the time window is elapsed. This parameter comes from the observance of the rules extracted by the ASHRAE RP-1312 dataset, which have showed a very shorter actual time window for the majority of the occurrences with respect to the formal time window.

Both of this information can be helpful in supporting in the last phase of fault detection, i.e. the knowledge extraction, which is strictly dependent on the goal of the analysis, the kind of data available and the grade of knowledge of the system analyzed.

Finally, Knowledge Extraction phase can be carried out according to the need and kind of analysis intended to perform. For instance, the Knowledge Extraction phase can be oriented to normal operation identification by means of dominant rules, as well as to fault spotting, looking for low support rules. Normal operation rules can be used for fault detection, matter of fact they may represent a reference behavior of the system, which violation is considerable as faulty. Direct fault spotting can be conducted by means of the identification of rarely occurring rules, therefore with low support but high confidence. In this way, the anomalous correlation is identified directly in the ARM, but the interpretation of the anomalous rules can be hard to perform, since these rules may be highly influenced by the noise generated by outliers.

The fault diagnosis stage can be conducted after the fault detection stage is terminated and a fault map is available. In fact, the fault detection stage can be performed only if further knowledge is available, matter of fact that the violation of some rules has to be related to the possible presence of a certain fault, so to schedule a specific intervention to solve the problem causing the fault.

The diagnosis stage can provide an indication of the possible cause of the anomaly on the basis of the rules that have been violated.

Exploiting the information of the fault map, the rules can be related to the faults in terms of violation or not of them. In fact, the fault map is nothing else the census of the occurrences of the rules, associating them with moments tagged whether as faulty or not. By the knowledge of the presence of not of the rules' patterns in faulty moments can be constructed a contingency matrix reporting the relationship between the presence/absence of a rule with the faults comprises in the fault map.

Therefore, at this point, the rules are organized in a matrix as function of the faults that can be implied by their violation. A matrix constructed in this way is able to express the relationship between the rules and the cause, in terms of reaction of the system to a specific fault.

It is worth noting that the diagnosis phase can be perpetrated in the only case a map of faults is present. In fact, the association between rules and faults is obtained by inference from a dataset containing recognized faults and associated to the monitored data. There is no way to construct the fault-rules matrix if there is not a fault mapped dataset supporting the process.

By means of the inference analysis it is possible to enhance the detail of information provided by the matrix of rule-fault linkage. In fact, each cell of the matrix reporting the violation of the rule can provide deeper information about the kind of violation that characterise that specific linkage rule-fault. In Figure 15 is provided an example of rules-faults matrix with a deeper level of information regarding the violation of the rules. In this kind of matrix, the violation cells contain a code which states if the violation, in that case, interests the absence of the only consequent, the absence of the only antecedent or the absence of both sides of the pattern.

In fact, the code "A" means that the only antecedent is respected, the code "C" means that the only consequent is respected, while the code "X" states that neither the antecedent nor the consequent are present. It should be pointed out that, since the rules are strictly related the time of the day in which they are occurring, the inference analysis of their violation research should be conducted in the same time frame so to obtain relevant results regarding the absence of the rule considered.

	r1	r2	r3	r4	r5	r6	r7	r6	r7	r8	r9	r10
CCVS65	X		CX		X			X		X	CX	AX
CCVSFC	X	AX	X	X	X	X	X	X	AX	X	X	X
DLBSF	X									X		AX
EASFC		X										
HCVL												
OASFC												
RFCF			ACX	X		X	X		CX		ACX	ACX
RFF30	CX	CX								AX		AX
wind	180	180	180	180	180	180	180	240	240	240	240	240
actualwin	120	90	120	90	60	90	90	90	90	120	150	120

Figure 15 Rule-fault matrix with violation characterisation

The information of the kind of violation referred to a fault comes in handy in the case of real time implementation of this method, since the kind of violation of this rule can speed up the diagnosis and the understanding of the behaviour of the system. In fact, it can be that a kind of violation for a rule is characterising for a specific fault, therefore the occurrence of that particular violation can diagnose merely the cause, without the application of the rest of the diagnosis process.

Although the characterisation of the violation provides a deep understanding of the linkage between rules and faults, it should be pointed out that, for the sake of diagnosis, it is sufficient to analyze the linkage between rules and fault as binary, since the simple violation information is enough to establish a linkage between rules and faults. The violation characterization information can be used to split the rules and considered the violation mode as separate, but the complexity of the analysis increases dramatically affecting the computation and the results.

As a consequence, for the diagnosis phase the matrix of rules- fault linkage considered is the contingency matrix constructed only with 0 corresponding to violations of the rules and 1 for non-violations.

Showing the whole pool of rules together in the representation of a contingency matrix allows identifying similar rules, not in terms of similar pattern, but in terms of similar implication. In fact, as can be seen in Figure 16, there are cases in which the same set of faults drives to the violation of the same set of rules. In these cases, the rules can be considered similar, since provides the same information about the implication of faults. The identification of groups of rules characterized by the same kind of faults implication, can be easily carried out by eye if the number of rules is small, but as the number arises the process becomes unbearable. For this reason, a cluster analysis is conducted, so to keep the whole method the most unsupervised and automatic possible.

The matrix constructed has the form of a contingency matrix, which reports only binary values to state the presence or not of the fault for that rule.

Since the matrix's values can be considered as binary, with 0 in violation cells and 1 in respect cells, the function for the calculation of the distance for the cluster analysis should be suitable for this kind of values. For this reason, the function of the distance considered is the Jaccard one which computes the distance between binary quantities as:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Equation 32

with A and B two ensemble of binary values and J (A, B) the Jaccard index. In this way the distance is computed by counting the number of 1 in the ensembles.

After the employment of different hierarchical clustering methods, such as Ward, Complete, Average and Single. The choice of the most suitable method, for the kind of problem considered, results the Single method. This method takes as similarity criterion among the elements to include in the same cluster, the minimum distance. Therefore, the distance between the elements in the same cluster is the lowest possible. As a result, the rules grouped in the same cluster at distance 0, are the ones with the exact equal faults implication.

Figure 16 Fault-rules matrix of violation

In the representation of the matrix, the red cells refer to the violation of the rule, while the greenish ones refer to the respect of the rule, therefore this representation provides an easy to understand and eye-catching portray of the linkage between rules describing the system's behavior and the possible faults related to their violation.

Both the rules and the faults in the matrix are re-arranged according to two different logics. The rules are ordered according to the cluster found so to have on the left rules with a high number of faults implicated and, on the right, a low number of faults implicated. As concerns the faults, they are ordered in ascending order on the basis of the number of rules implying the fault.

This arrangement provides information about the discrimination power of the rules and the diagnosability of the fault. On one hand, rules implying a single fault have the best discrimination power, since they are able to detect only a fault by their violation, making easy the diagnosis. On the other hand, faults implied by a small number of rules makes the isolation of them easy since they specifically implied by those rules.

As a consequence, red cells located in the upper right part of the matrix are considerable the best linkage between rules and faults to obtain good results in diagnosis.

The matrix in Figure 16 is an indication of the description of the whole pool of rules extracted from the dataset, but a similar matrix can be constructed with the set of rules violated in a real-time application, reporting the misbehaviour of the system.

This kind of matrix reports, therefore, only the rules violated, which can be considered as system's error signals.

Once the Error Signals Matrix has been constructed, the isolation of the fault responsible for the anomalous functioning of the system can be conducted in the Score Calculation phase. In fact, for each fault it is computed a score which gives an indication of the violations related to that specific fault. The comparison of the adimensional values obtained by the score analysis gives the indication of the most probable faults present in the system. This score can assume values higher than one, hence they are not an indication of the absolute value of the fault's presence probability, although it gives an indication of isolability of the fault in relative terms.

Figure 17 reports an example of score analysis for an error signal matrix, which is able to identify the fault by means of the large difference from the score of the other faults.

The calculation of the score must take into account the discrimination power of the rules considered, therefore it is a weighted count of the 0s of the matrix, perpetrated row by row. In this way the score is an indication of the number of rules of error signals matrix implies the fault, but the weights bias the calculation towards the rules with a high discrimination power. The weights are calculated as

$$w = \frac{1}{|Faults|} + \frac{|non - violation|}{|Faults|}$$

Equation 33

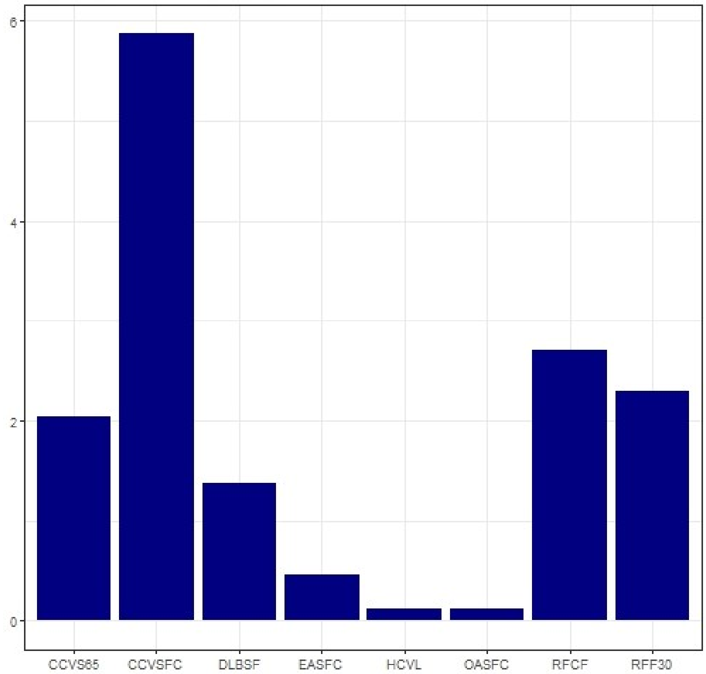


Figure 17 Score visualization for fault diagnosis

Since the weight is associated to the rule, in this way the rules with a single violation gain the highest value. Due to the shifting term summed to the fraction of non-violation, the weight can assume values from $\frac{1}{|Faults|}$ to 1. The shifting term is necessary to avoid assigning weight equal to 0 to rules having a violation per each fault. These rules are useless for diagnosis, but are indicator of the health of the system, matter of fact they give an error signal in every fault condition. As a consequence, they are fundamental for fault detection, but neutral to diagnosis. These rules are not filtered out since they do not affect the diagnosability of the faults, given that they provide the same step to all the faults' score.

In the case reported in Figure 17 the fault is easily isolated, but in other cases the difference can be very low or even null, thus the isolation of the fault in those case can be conducted by setting a threshold as a percentage of the maximum level.

If there are more than one fault inside the threshold set, it means that it is not possible to isolate a single, due to the contemporary presence of more than one fault or the presence of symptoms shared by different faults. Different faults can drive the system to react in similar ways, since the faults can be related to the same component or components regarding the same sub-operation.

The scores calculated in the cases of multiple faults presence within the threshold, can be used as a priority indication for a manual diagnosis. At this point, a fault is detected but the cause is not sure, thus the operator should check the most probable causes by his own using the diagnosis results as a support for the investigation.

In the case which a fault is isolated without any doubt, the operator can schedule an intervention to fix the fault and bring the system back to its normal condition, since the cause of the malfunctioning has been identified.

4.2. Intermediate developing phases

The process of development of the proposing methodology, comprises different intermediate phases that represent the process of optimization of methodology's steps. Each attempt's result has been analyzed critically, so to find the modification that can be applied to obtain better results.

The first phase starts from the consideration made on the work of Yan et al. (Yan, et al., 2016) The methodology proposed in this work comprises the construction of a decision tree by means of CART technique. Input variables are attributes of the dataset used, while output variables are faults' tags, since the dataset used provides data classified related to different faults' situations. Results were interesting, since the decision tree found is completely data-driven, easy to implement, quite robust and it provides rules which define exactly the condition that is related to each fault. Knowing that, the user can immediately diagnose a fault following the conditions along the branches of the tree constructed. Following the description of the authors of the methodology, it has been replicated using the same dataset, obtaining similar results with respect to the ones obtained by the research team.

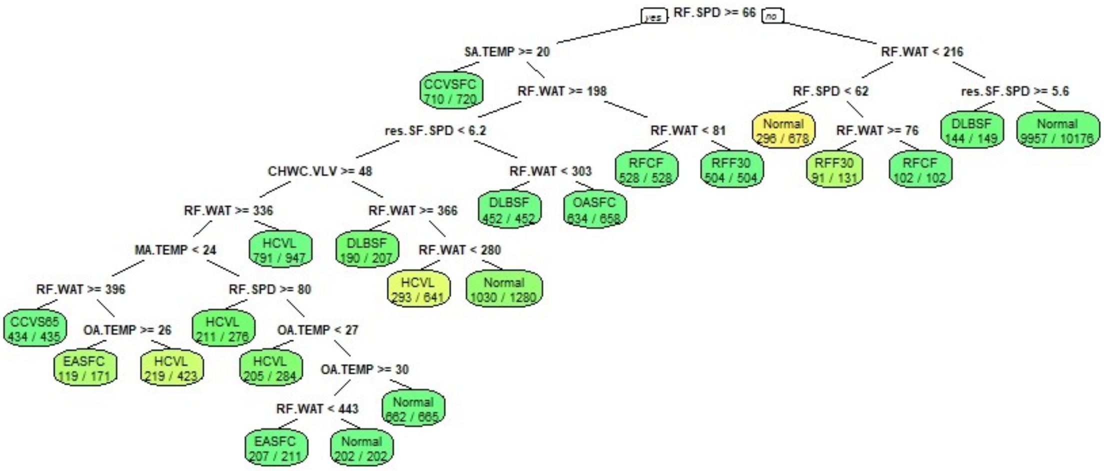


Figure 18 Decision-tree obtained with Ma et al. approach

That brought me to considering even another kind of approach, that is the association rules mining one, eventually with temporal information. In fact, the decision tree described above, it is not able to consider temporal effects, since it analyses data instance by instance, therefore the conditions stated by the tree must be respected in the same moment to define a related fault tag.

Temporal lag between variables can be even enough significantly large, even in a system like an AHU in which inertial effect is very low, to justify the use of temporal methods.The effectiveness of association rules has been tested with different algorithms and approach, inspired by literature already present methodologies.

Data used for the following attempts are the same of the decision tree, still exploiting the decision tree information as reference for discretizing the variables. In fact, since a discretization must be carried on prior the use of association rules mining techniques, the splits of the decision tree has been used as discriminant in the interval of discretization definition. In this way instantaneous information is still taken into consideration. As a consequence, the number of intervals for each variable is the number of splits related to that variable, considering the whole decision tree, plus one.

The approaches used in this phase were two, that involved inter-transactional association rules mining or data segmentation. In the data segmentation approach, data are divided according to the schedule of operation, that means that data referring to periods which the system is turned off are separated from the one referring to the "on" operation. This criterion for the split came out from a series of CART analysis conducted on data, considering the time variable as input and one attribute a time as output variable. Only time has been considered as input variable, due to the nature of data considered, since data are not affected to calendar effect being data of simulated schedule.

As a consequence, the association rule mining is conducted in parallel in the two datasets resulting from the split. The analysis at this stage has been conducted using cSpade, Quantminer and Trule-Growth algorithms, so to make a comparison among them.

The idea of using Quantminer and Trule-Growth of segmented data has been inspired by the works of Xiao et al. (Fan, et al., 2015) (Fan, et al., 2015) (Fan & Xiao, 2016) (Fan & Xiao, 2017)

In this case, the only algorithm that relies its work on temporal information is Trule-Growth, but the kind of association rules found with this algorithm has been of poor interest, as the ones resulting from cSpade's algorithm. The Quantminer results were interesting, matter of fact there were no need of prior discretization, but the logic of the algorithm does not consider temporal effects, so correlations are limited.

In the other approach, a sliding window has been implemented in order take in consideration temporal effect in the association rules mining. This approach is inspired by the CONOTOOL algorithm, proposed by the EDMANS team. (Martínez-de-Pisón Ascacíbar, et al., 2009) The association rules mining has been conducted with Trule-Growth, cSpade and FP-Growth. The latter is not a sequential algorithm, therefore considers the entire time window as contemporary, regardless the order of occurrence inside the time window. This point is reflected in the results as low confidence and low occurrence parameters' values. The other two algorithms generated accordance results, such that the results were completely interchangeable.

This being the following attempts has been focused on the use of the inter-transactional approach, with cSpade, due to the fact that it provides similar results of Trule-Growth, but it opens to large possibilities of customizing the research. The cSpade algorithm allows the use of constraints, which are able to drive the results, such that the user can filter out easily uninteresting association rules or focus the research on a certain type of rules. For instance, it is possible to extract rules that implies a specific event, rather than rules with a user-defined gap between antecedent and consequent or with a maximum number of events involved. (Fan, et al., 2015)

Analyzing deeply the results obtained, it came out that the rules extracted are very specific to the data used, due to the choice of variables made. In fact, the presence of variables that can be present in a dataset of experimental data such the one used in this case but is unlikely to find in data from real building's systems. Therefore, using the same dataset, the choice of the variables involved in the analysis has been made again, using the generality as main criterion for the choice, in addition to the fault oriented one.

Basically, the variables chosen are related to the components which the dataset mapped the fault, but at the same time, it is a variable that can be easily encountered in a common dataset of real systems' monitoring.

Standing back from the decision tree's variables, the technique used for discretization has been changed as well, preferring an unsupervised method. The discretization method chosen has been the k-means optimized with aSAX algorithm. (Pham, et al., 2010)

5. Case study

In this section it is described how the methodology has been applied to the ASHRAE RP-1312 dataset in order to test and validate it. The description of the application of the methodology lets show up even the way the methodology has been adapted to the dataset and the system which refers to, so to extract hidden knowledge and exploit in the best way possible the data available.

The kind of dataset can be considered unusual, matter of fact it does not derive from a real monitoring, but it is constructed artificially with the intention of gather data related to several AHU common faults. Even though the dataset cannot emulate completely the actual functioning of an AHU system it provides precious information about the operation of the system in fault condition, tagging the moments of faultless and faulty running.

The methodology described has been constructed thinking on the possibility of employment of this dataset, therefore some steps have been made real only thanks to the nature of the dataset used, especially as regards the diagnosis phase, which basis itself on the information provided by the fault tagged moments. Nevertheless, this methodology is not specific to this kind of dataset, since a real monitoring one can report tagged faults as well, as additional information provided by the system manager.

Following, in this section, it is provided an insight on the dataset used, in order to have complete awareness of the information available, then the application to the dataset is described, so to explicit the details related to each step of the procedure.

5.1. Dataset description

In 2011 the project ASHRAE RP-1312 with the scope of building an Air Handling Unit's model has been concluded, publishing results and data used for the study.

Particularly interesting are data collected for the validation of the model, in fact it has been used a monitoring system on a real AHU in order to identify the behaviour of the system in several conditions. The dataset used for the scope of this project contains operational data of the system in faulty and fault-free condition, so to study the behaviour of the system even in unexpected conditions.

The site where monitoring has been carried out is the Iowa Energy Center Energy Resource Station (ERS). ERS is a test facility built in order to make comparisons of different energy efficiency solutions and record energy consumption. As regards the project ASHRAE RP-1312, since data has been collected for the sake of validation of the AHU model developed, the system of ERS is operated recreating the particular situations needed for the study.

The facility is equipped with 3 AHUs in order to serve common areas with one of these (AHU-1) and 4 zones for each of the remaining (AHU-A and B).

The AHU-A and B are perfectly identical and serve specular zones, thus the 2 AHUs can be considered as identical in technical features and functioning. The opposite zones served by AHU-A and B are facing east and west, so be completely comparable even under load aspects.

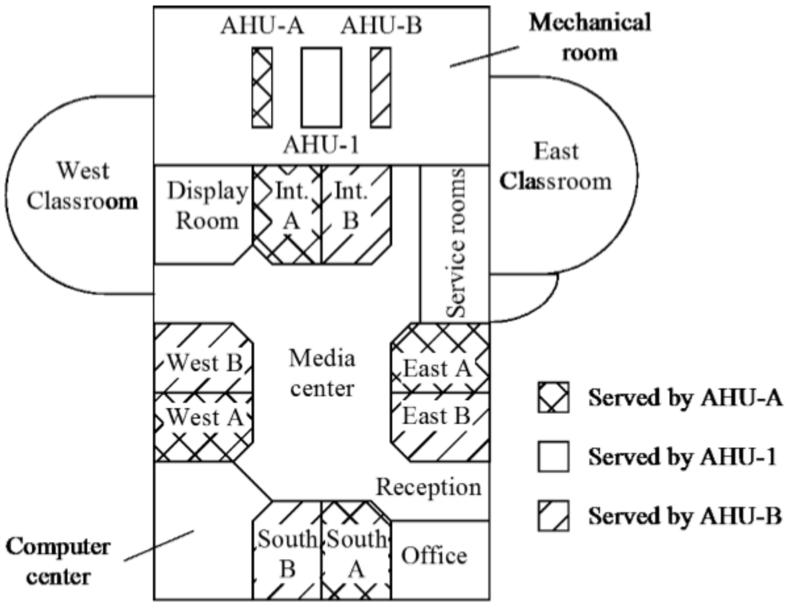


Figure 19 ERS facility (Wen & Li, 2011)

The system considered is a Variable Air Volume AHU with a mixing chamber, to mix return air with outdoor air with dampers controlling the percentage of mixing. There are 2 coils, one for heating and the other for cooling the mixed air. In the site there are VAV devices to locally adjust temperature of the supply air, so there is only a coil in the AHU.

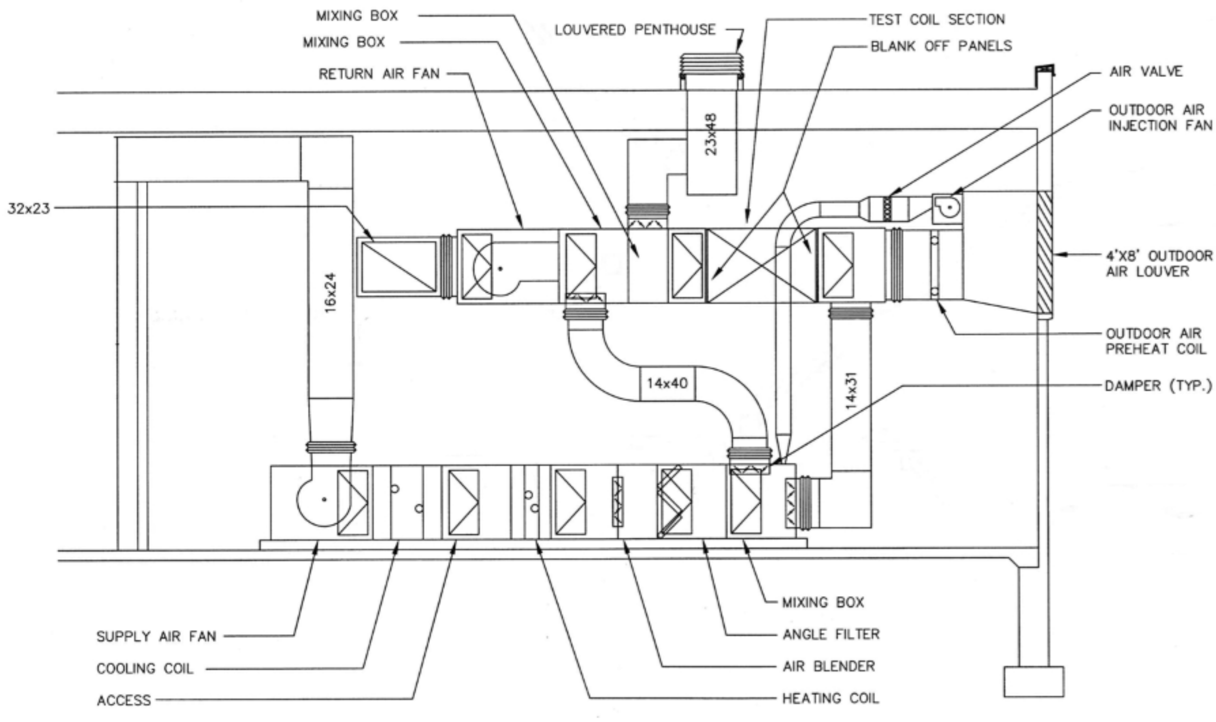


Figure 20 AHU- A and B schematic (Wen & Li, 2011)

Monitoring has been carried out for 24 hours per day, recording data every minute, scheduling experiments all day long. Since different situations have been recreated through the implementation of artificial faults along the day, data related the behaviour of the system in different operational regimes with different faults condition are available. As a consequence, daily data are representative of the behaviour of the system in a particular fault condition or in fault free condition.

Since AHU-1 serves only common areas, while the other two AHUs serve twin and specular areas, AHU-A and -B are monitored in order to build the dataset useful for model validation. A set of common faults are artificially generated singularly only on AHU-A, while AHU-B is always carried out in fault free operation to have a reference for the faulty unit's data. The 2 AHUs run in the same operational condition, controlled in the same way. Artificial faults experiments' solution has been chosen to validate the models in order to avoid simultaneous faults effect, which may be present in real data utilisation.

Tests on system have been carried out in three different periods, therefore scheduling experiments in different seasons, thus they were conducted in winter, spring and summer. The object of this choice is to study the behaviour of the same fault in different weather, configuration and strategy condition, since some faults can present different symptoms depending on these factors. Faults presenting same symptoms has been conducted in only one period. Faults were introduced manually only on one of the 2 AHUs; therefore, it can be possible to analyse symptoms. Even though they are identically constructed, differences may exist, so fault free days were scheduled during tests periods to understand the difference between the AHUs.

The Table 3 reports all the experiments conducted and how they were artificially implemented.

Category	**Fault description**	**Magnitude or location**	**Summer**	**Winter**	**Spring**	**Implementation**
Controlled Device	OA Damper Stuck	Fully Closed	X	X	X	Manually control OA damper at faulty positions
		40% Open			X	
	OA Damper Leak	45% Open	X			
		55% Open	X			
		52% Open		X		
		62% Open		X		
	EA Damper Stuck	Fully Open	X	X	X	Manually control EA damper at faulty positions
		Fully Close	X	X	X	
		40% Open			X	
	Cooling Coil Valve Stuck	Fully Closed	X		X	Manually control valve at faulty positions
		Fully Open	X	X	X	
		Partially Open - 15%	X			
		Partially Open - 20%		X		
		Partially Open - 50%			X	
		Partially Open - 65%	X			
	Heating Coil Valve Leaking	Stage 1 - 0.4GPM	X			**Manually** open heating coil bypass valve
		Stage 2 – 1.0GPM	X			
		Stage 3 – 2.0GPM	X			

Equipment	AHU Duct Leaking	after SF	X			Remove sealing from one access door
		before SF	X			
	Heating Coil Fouling	Stage 1		X		Partially block heating coil using a piece of cardboard
		Stage 2		X		
	Heating Coil Reduced Capacity	Stage 1		X		Manually throttle heating coil balancing valve
		Stage 2		X		
		Stage 3		X		
	Return Fan complete failure		X		X	Manually stop return fan
	Air filter blockage fault	10%				Partially block air filter using a piece of cardboard
	Air filter blockage fault	25%				Partially block air filter using a piece of cardboard
Controller	Return Fan at fixed speed	30%spd	X			Manually maintain return fan speed at faulty speed
		20%spd			X	
		80%spd			X	
	Cooling Coil Valve Control unstable		X			Reduce the PB value for supply air temperature PI control algorithm to be 50% of its original value
	Cooling Coil Valve Reverse Action		X			Change cooling coil valve scaling factor
	Mixed air damper unstable				X	Change the PB value for supply air temperature PI control algorithm from -45.7 to -10
	Mixed air damper unstable/Cooling Coil Control Unstable				X	Change the PB value for supply air temperature PI control algorithm from -10 to -5
	Sequence of Heating and cooling unstable				X	Increase the PB value for supply air temperature PI control algorithm
Sensor	OA temperature sensor bias	+3 F			X	Manually change the sensor calibration equation in the control system
	OA temperature sensor bias	-3 F			X	Manually change the sensor calibration equation in the control system

Table 3 Experiment conducted for ASHRAE RP-1312 (Wen & Li, 2011)

As concerning the system settings, following are reported the main setup information for AHU settings, zone settings, and heating and cooling plants settings.

The site's usage is simulated as a commercial building, so it is occupied only from 6:00 to 18:00. Lighting and baseboard heaters are scheduled to simulate the normal activities in a commercial building. Both window and door side light fixtures turn on from 8:00 to 18:00. There are no windows coverings, so sun load influenced the control of the chillers utilization. In fact, when load is light, only chiller provides chilled water, while when load is heavy, chilled water supply temperature setpoint is 3,3°C, obtained by mixing chiller leaving water with thermo storage tank 0°C water. The boiler outlet temperature set point is 60°C, in order to provide hot water to the coil. The heating and cooling set-points are 21°C and 22°C respectively.

Even a minimum ventilation requirement is fulfilled, limiting the outdoor air damper position at least at 40 %. Economizer is enabled when outside temperature is less than 18,3°C, while supply air temperature set point is 12,8°C.

Even though AHU-A and B are identically constructed and operated, some differences can emerge, therefore fault free tests were scheduled during test periods, so to study the match of the 2 systems. During these tests the 2 AHUs are controlled and operated using the same conditions, in order to report the deviation of a set of performance indexes defined as the percentage of difference or the difference between the values of the 2 AHUs. Energy performance indexes are compared by means of percentage of difference, defined as

$$Percentage\_of\_difference = \frac{Measurement\_A - Measurement\_B}{Measurement\_B},$$

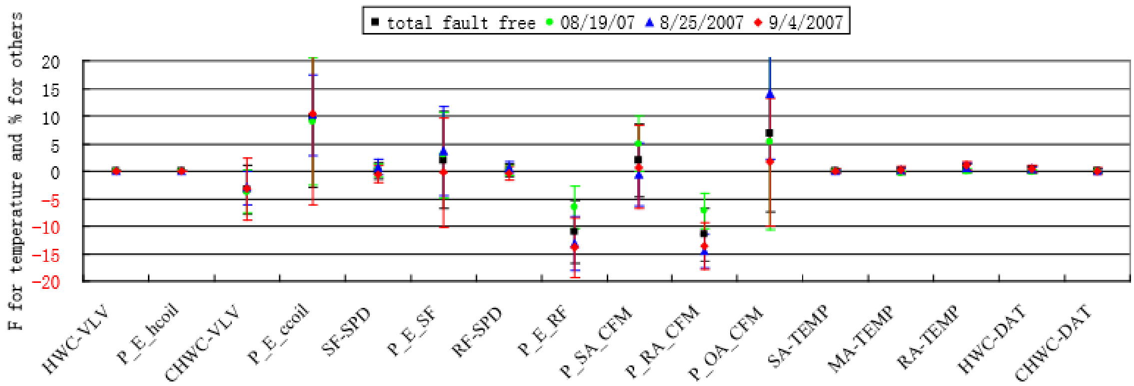
Equation 34

while measurement perfomance indeces are compared by means of real difference values.

For this analysis sixteen indices has been used, including twelve direct measurements, namely, supply air temperature (SA-TEMP), mixed air temperature (MA-TEMP), return air temperature (RA-TEMP), heating coil discharge air temperature (HWC-DAT), cooling coil discharge air temperature (CHWC-DAT), heating coil valve position (HWCVLV, controller output signal), cooling coil valve position (CHWC-VLV, controller output signal), supply fan speed (SF-SPD, controller output signal), return fan speed (RFSPD, controller output signal), supply airflow rate (SA-CFM), return airflow rate (RACFM), outdoor airflow rate (OA-CFM). The sixteen indices also included four energy performance indices, namely supply fan electricity consumption, return fan electricity consumption, chilled water heat transfer rate, and heating water heat transfer rate.

In this way error in calculating those indices is evaluated for each period of testing, so critical parameters are identified for each period.

The results of these tests are reported below.



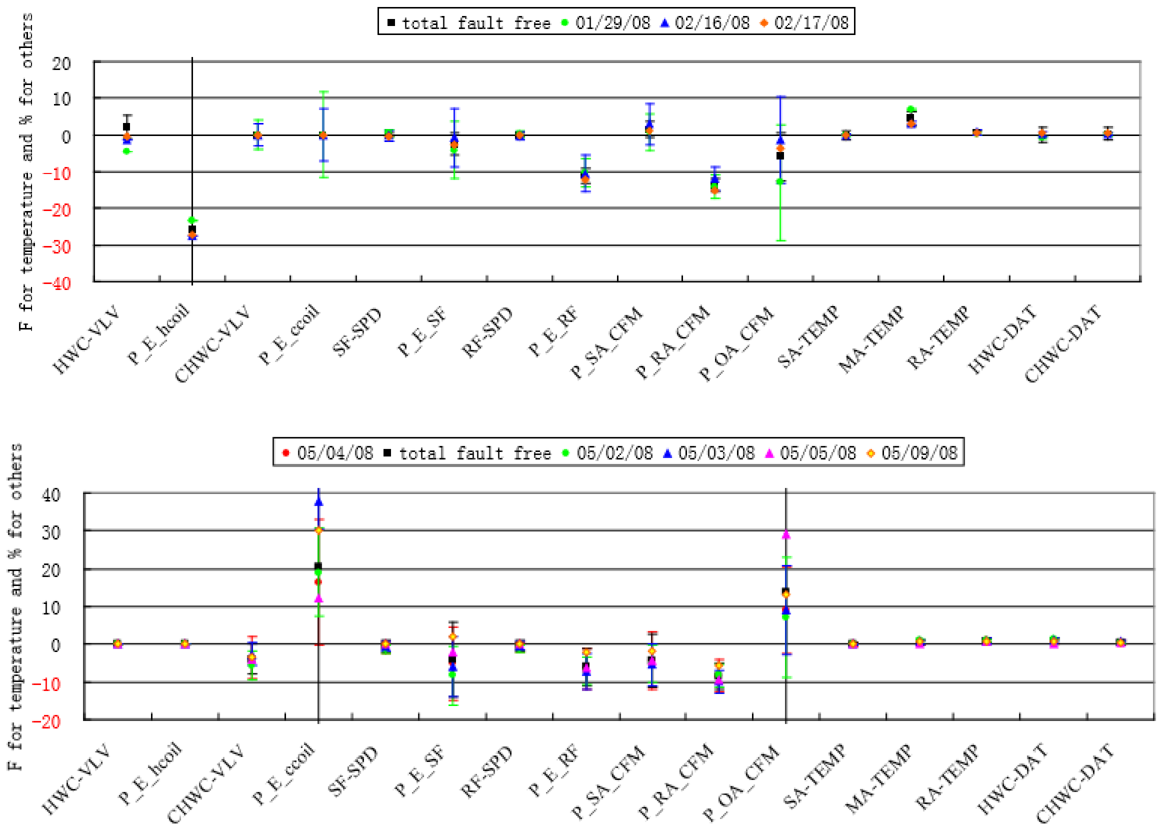


Figure 21 Match test of AHU-A and B for summer, winter and spring cases (Wen & Li, 2011)

The same kind of study has been conducted for each fault implemented during tests period, therefore it has been found correlations between variables considered contextually to each fault. Thanks to this information, a symptoms matrix has been constructed, reporting symptoms of each fault in terms of sign and magnitude of variation for each variable considered, for each test period.

The resulting symptoms matrix are reported in Table 4 for the 3 test periods.

For summer period, the table is reported.

Fault description	Date	HWC-VLV	P_E_hcoil	CHWC-VLV	P_E_ccoil	SF-SPD	P_E_SF	RF-SPD	P_E_RF	P_SA_CFM	P_RA_CFM	P_OA_CFM	SA-TEMP	MA-TEMP	RA-TEMP	DAMPERS	SA-SP
EA Damper Stuck (Fully Open)	8/20/2007	0	0	0	0	+	+	+	+	0	+	+	0	0	0	0	0
EA Damper Stuck (Fully Close)	8/21/2007	0	0	0	0	-	-	-	-	0	-	-	0	0	0	0	0
Return Fan at fixed speed (30%spd)	8/22/2007	0	0	0	0	++	++	--	--	0	--	++	0	0	0	0	0
Return Fan complete failure	8/23/2007	0	0	0	0	++	++	--	--	0	--	++	0	0	0	0	0
Cooling Coil Valve Control unstable (Reduce PID PB by	8/24/2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

half)																		
Cooling Coil Valve Reverse Action		9/3/2007	++	++	++	++	0	0	0	0	0	0	--	0	0	0	--	0
OA Damper Stuck (Fully Closed)		8/26/2007	0	0	0	0	++	++	++	++	0	+	-	0	0	0	0	0
Cooling Coil Valve Stuck	Fully closed	8/27/2007	0	0	--	--	++	++	++	++	++	++	++	++	+	+	0	0
	Fully open	8/31/2007	++	++	++	++	0	0	0	0	0	0	--	0	0	0	--	0
	Partially open (15%)	9/1/2007	0	0	--	--	++	++	++	++	++	++	++	++	+	+	0	0
	Partially open (65%)	9/2/2007	++	++	++	++	0	0	0	0	0	0	--	0	0	0	--	-
Heating Coil Valve Leaking	Stage 1 - 0.4GPM	8/28/2007	0	++	++	++	0	0	0	0	0	0	0	0	0	0	0	0
	Stage 2 – 1.0GPM	8/29/2007	0	++	++	++	0	0	0	0	0	0	0	0	0	0	0	0
	Stage 3 – 2.0GPM	8/30/2007	0	++	++	++	0	0	0	0	0	0	0	0	0	0	0	0
OA Damper Leak (45% Open)		9/5/2007	0	0	0	0	-	-	-	-	0	0	++	0	0	0	0	0
OA Damper Leak (55% Open)		9/6/2007	0	0	0	0	-	-	-	-	0	0	++	0	0	0	0	0
AHU Duct Leaking (after SF)		9/7/2007	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0
AHU Duct Leaking (before SF)		9/8/2007	0	0	0	0	--	--	--	--	0	--	--	0	0	0	0	0

For winter period, the table is reported.

Fault description	Date	HWC-VLV	P_E_hcoil	CHWC-VLV	P_E_ccoil	SF-SPD	P_E_SF	RF-SPD	P_E_RF	P_SA_CFM	P_RA_CFM	P_OA_CFM	SA-TEMP	MA-TEMP	RA-TEMP	DAMPERS	SA-SP
OA Damper Stuck (Fully Close)	2/12/2008	--	--	0	0	++	+	++	+	--	++	--	0	-	0	++	--
OA damper leaking (52% open)	2/13/2008	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0
OA damper leaking (62% open)	2/15/2008	0	0	0	0	0	0	0	0	0	0	+	0	0	0	0	0
EA Damper Stuck (Fully open)	2/2/2008	0	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0
EA Damper Stuck (Fully Close)	2/3/2008	-	--	0	0	0	0	0	--	0	--	--	0	0	0	0	0
Cooling Coil Valve Stuck (Fully Open)	2/10/2008	++	++	++	++	0	0	0	0	0	0	0	-	0	0	0	0
Cooling Coil Valve Stuck	2/11/2008	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0

(Partially Open - 20%)																	
Heating Coil Fouling Stage 1	2/5/2008	0	--	0	0	+	+	+	+	0	+	-	0	0	0	0	0
Heating Coil Fouling Stage 2	2/6/2008	0	--	0	0	+	+	+	+	0	+	-	0	0	0	0	-
Heating coil reduced capacity Stage 1	2/7/2008	+	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Heating coil reduced capacity Stage 2	2/8/2008	+	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Heating coil reduced capacity Stage 3	2/9/2008	+	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0

For spring period, the table is reported.

Fault description	Date	HWC-VLV	P_E_hcoil	CHWC-VLV	P_E_ccoil	SF-SPD	P_E_SF	RF-SPD	P_E_RF	P_SA_CFM	P_RA_CFM	P_OA_CFM	SA-TEMP	MA-TEMP	RA-TEMP	DAMPERS	SA-SP
OA temperature sensor bias (+3F)	5/29/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0
OA temperature sensor bias (-3F)	5/30/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0
OA Damper Stuck (Fully Close)	5/7/2008	0	0	0	0	+	+	+	+	-	+	--	0	0	0	0	--
OA Damper Stuck (40% open)	5/8/2008	0	0	0	0	+	+	+	+	-	+	--	0	0	0	0	-
EA Damper Stuck (Fully open)	5/27/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EA Damper Stuck (Fully Close)	5/10/2008	0	0	0	0	0	0	0	-	0	-	0	0	0	0	0	0
EA Damper Stuck (40% open)	5/11/2008	0	0	0	0	0	0	0	-	0	-	0	0	0	0	0	0
Cooling Coil Valve Stuck (Fully Closed)	5/6/2008	0	0	--	--	++	++	++	++	++	++	++	++	+	+	0	0
Cooling Coil Valve Stuck (Fully Open)	5/15/2008	0	0	--	--	++	++	++	++	++	++	++	++	+	+	++	0
Cooling Coil Valve Stuck (Partially Open - 50%)	5/16/2008	++	++	++	++	0	0	0	0	0	0	0	0	0	0	--	0
Return Fan complete failure	5/12/2008	++	++	++	++	0	0	0	0	0	0	0	0	0	0	--	0
Return Fan at fixed speed （20%spd）	5/18/2008	++	++	++	++	0	0	0	0	0	0	0	0	0	0	0	0
Return Fan at fixed speed （80%spd）	5/19/2008	0	0	0	0	0	0	--	--	0	--	0	0	0	0	0	0
Air filter block fault (10%)	5/22/2008	0	0	0	0	0	0	--	--	0	--	0	0	0	0	0	0
Air filter block fault (25%)	5/25/2008	0	0	0	0	0	0	++	++	0	++	0	0	0	0	0	0

Mixed air damper unstable	5/13/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mixed air damper unstable/ Cooling coil control unstable	5/14/2008	0	0	0	0	+	+	+	+	0	0	0	0	0	0	0	0
Sequence of heating and cooling unstable	5/17/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0
Supply fan control unstable	6/1/2008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0

Table 4 Symptoms matrix for summer, winter and spring (Wen & Li, 2011)

These table of symptoms have been very helpful in the choice of the variables for the first phase of development of the methodology, since the variables taken into account at the beginning were the ones reported in this table, matter of fact they were related to the related to the faults with deviation trends.

The exploitation of this data has been made possible thanks the permission given by the AHSRAE for the utilization of the validation data of ASHRAE- RP1312. (© ASHRAE www.ashrae.org ("ASHRAE 1312-RP Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units."), (2011).)

The selection of variables used for the test of the proposed methodology during its developing phases, incurred many modifications, so to find the most appropriate set for the description of the dataset with respect to the faults considered in the analysis. Therefore, an insight about the process of selection of the variables used for test and validation of the methodology has been reported below.

5.2. Selection of the variables

At the early phases of development of the methodology, the variables that have been chosen for the methodology development were taken from the paper of Ma et al. that uses the same dataset to construct a decision tree for AFDD. Then the variables involved which have a correlation with the faults considered described in the symptoms matrix have been chosen. Therefore, data are taken from the summer period and the days involving the same faults studied in the (Yan, et al., 2016)'s paper, that are reported in Table 5.

FAULTS ACRONYMS	DESCRIPTION
CCVSFC	Cooling coil valve stuck fully closed
CCVS65	Cooling coil valve stuck at 65%
DLBSF	Duct leakage before supply fan
EASFC	Exhaust air damper stuck fully closed
OASFC	Outdoor air damper stuck fully closed
HCVL	Heating coil valve leakage
RFCF	Return fan complete failure
RFF30	Return fan at fixed speed (30%)

Table 5 Faults considered in the analysis

The dataset used for the analysis from this point on, contains a selection of the days present in the whole ASHRAE RP-1312 dataset, since the original dataset includes 3 different seasons and several days of faults implementation.

For the sake of reduction of the dimensionality of the dataset to handle, a selection of the variables is conducted, as well as a selection of days. As a consequence, from the ASHRAE's dataset have been extracted only data related to AHU-A, i.e. the one which is subjected to the artificially introduced faults, in the days of faults in Table 5, in addition to the data of AHU-B, i.e. the one always running in fault free condition, in the same days of AHU-A. Moreover, data related to both the AHUs in normal operation conditions have been taken into account too, so have a larger dataset to handle, with a higher percentage of fault free data.

The distribution of the data among the faulty and faultless condition is resumed in Table 6.

TAG	NUMBER OF DAYS
CCVSFC	1
CCVS65	1
DLBSF	2
EASFC	1
OASFC	1
HCVL	3
RFCF	1
RFF30	1
NORMAL	19
	Total days = 30

Table 6 Distribution of data among faults and fault free condition

Even though the number of the data related to each fault is not the same, it does not affect the results in terms of validity of the single rule obtained or the diagnosis of the faults resulting from the rules assessment.

Successively, the variables that has been chosen as possibly correlated with the faults considered are selected on the basis of personal evaluation and expertise. The season and the days which data are taken from are the same of the first set of variables. Since two of the faults considered are related to the cooling coil valve, a focus on this component has been conducted, so to define the most appropriate set of variables able to describe the behavior of the component in the various situation which it can be run.

In order to understand the functioning of the cold deck and its related variables trend, the profiles of these parameters have been assessed deeply, so to identify the most relevant quantities to track. The control of the energy exchanged by the coil is associated to the water flow rate, rather than water temperature. In fact, the 3-way valve controls the flow rate at the input of the coil, by means of controlling the amount of water flowing through a bypass branch. Therefore, the input water temperature is fixed and directly controlled by the chiller, since it is the chiller outlet temperature. The outlet temperature of the water passing through the cold deck, it is related to the amount of energy exchanged, therefore it is minimum when the valve is fully open, so with the maximum water flow rate and vice versa.

The energy exchanged is proportional to the water flow rate and the water temperature difference between inlet and outlet. Since the inlet temperature is fixed, the parameters that may vary the amount of energy exchanged are only the water flow rate and the outlet temperature.

The outlet temperature cannot express alone a direct correlation with the energy since it depends on the flow rate in the coil, but it is influenced by the energy itself and the valve position as well.

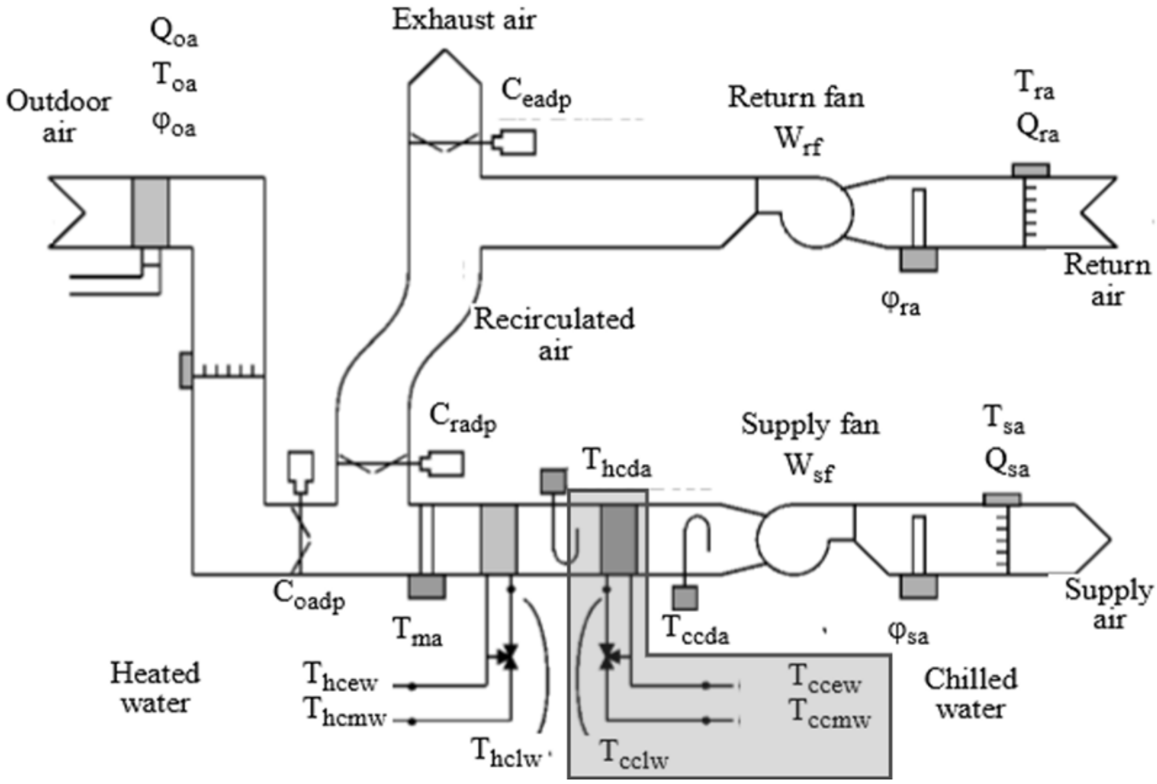


Figure 22 Cooling coil location schematic (Wen & Li, 2011)

In Figure 23 the points of temperature measuring are represented. Therefore, they are known the temperature of the water entering and leaving the deck, as well as after mixing downstream the valve. If the valve is completely open (100%), the leaving temperature is equal to the mixed temperature, matter of fact there is no mixing. On the contrary, if the valve is completely closed (0%) the mixed temperature is equal to the entering temperature, since the coil is totally by-passed. In this case the temperature measured in the leaving position has no meaning, since the water in that point it is not supposed to flow. As a consequence, the opening level of the valve determines the level of water flow rate passing through the coil. Therefore, the percentage of the opening can be read even as the percentage of the water flow rate in the deck. The functioning described is confirmed by the profile of the variables involved (Figure 24). In fact, as can be seen in the figure, the profiles of the temperature of water and the air flowing across the coil are very different in the two cases.

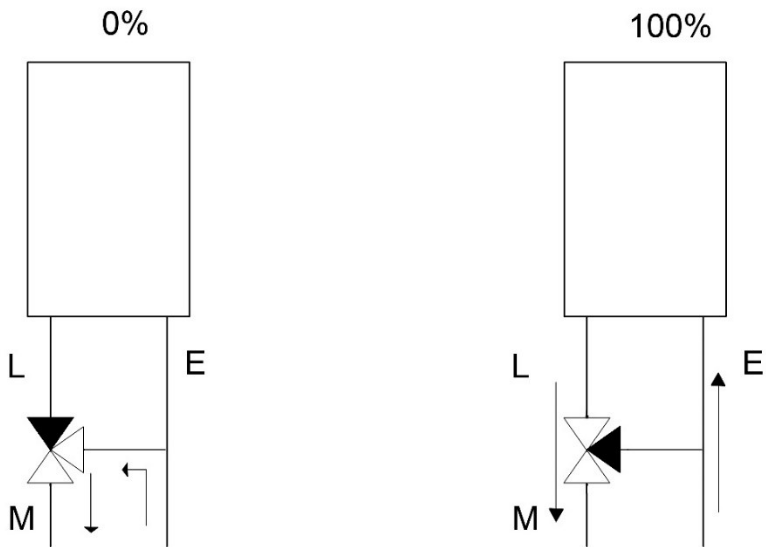


Figure 23 Three-way valve functioning

The upper side of the figure shows the behaviour with the valve completely open for the whole functioning time, while the bottom side refers to the completely closed condition. In the first case the water flow rate is not null, therefore the energy exchanged is not null. As a result, the temperature of the water leaving the coil deck is higher than the one in the entering point and coinciding with the mixed temperature. The temperature of the air crossing the deck has got a trend that follows the one of the energy exchanged and the leaving water temperature.

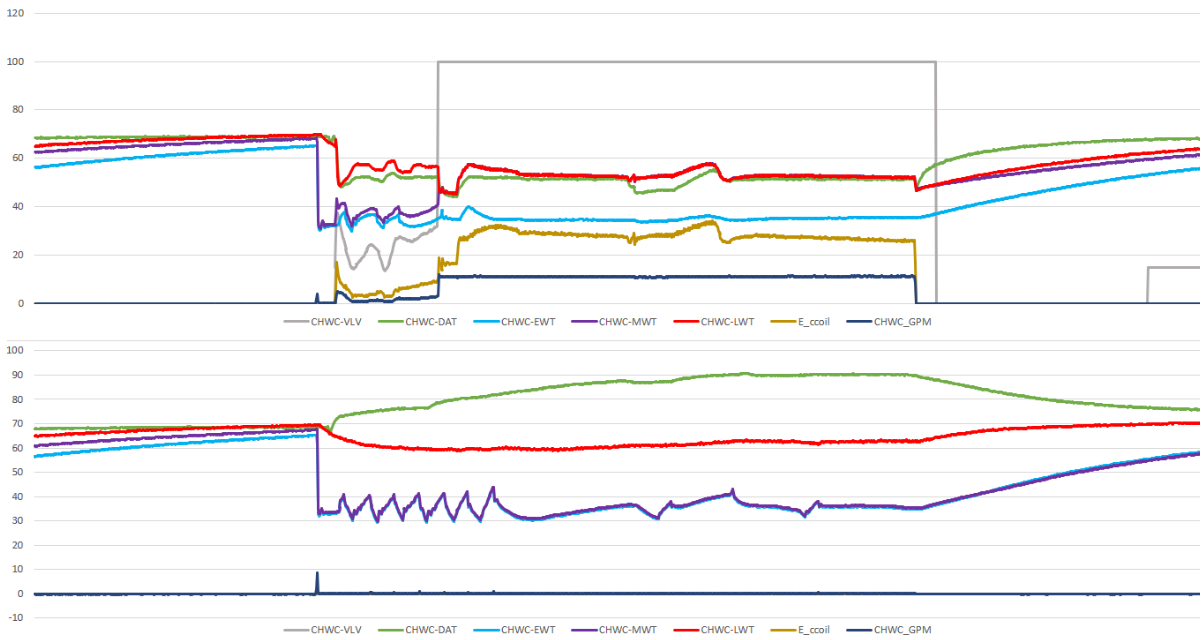


Figure 24 Profiles of cooling coil related variables in cooling coil valve stuck at fully open case (above) and fully closed case (bottom)

In the case of completely closed valve, the mixed temperature is attached to the entering temperature, due to the fact that the coil deck is completely by-passed, and the energy exchanged is null. The temperature of the water in the leaving position and the air temperature are not influenced by the entering water temperature, since they are totally separated by the by-pass branch. Therefore, the leaving temperature is approximately constant, since the water is stuck in that point, while the other increases influenced by other factors. The outlet temperature will be lower if the valve is completely open, that is without a bypass flow rate, so the outlet temperature is equal to the mixed temperature (downstream the 3-way valve).

As a consequence, it is useful to consider the energy exchanged in the coil, the water flow rate and the outlet temperature of the water. The latter it has been chosen considering the relation with the position of the valve, while the others have been chosen since can describe the effectiveness of the water-air exchange.

TAG	DESCRIPTION	CODE	UNIT (ORIGINAL)	UNIT (CONVERTED)
SA.TEMP	Supply air temperature	1	F	C
SA.HUMD	Supply air humidity	2	%	
SA.M3M	Supply air flow rate	3	CFM	m3/m
MA.TEMP	Mixed air temperature	4	F	C
RA.TEMP	Return air temperature	5	F	C
RA.HUMD	Return air humidity	6	%	
RA.M3M	Return air flow rate	7	CFM	m3/m
OA.TEMP	Outdoor air temperature	8	F	C
SF.WAT	Supply fan power	9	W	
SF.SPD	Supply fan speed	10	%	
RF.WAT	Return fan power	11	W	
RF.SPD	Return fan speed	12	%	
CHWC.VLV	Cooling coil valve opening	13	%	
CHWC.LM	Cooling coil water flow rate	14	GPM	l/m
CHWC.DAT	Cooling coil discharge air temperature	15	F	C
CHWC.LWT	Cooling coil water outlet temperature	16	F	C
E_CCOIL	Cooling coil power	17	KW	
HWC.VLV	Heating coil valve opening	18	%	
HWC.LM	Heating coil water flow rate	19	GPM	l/m
HWC.DAT	Heating coil discharge air temperature	20	F	C
E_HCOIL	Heating coil power	21	KW	

Table 7 Variables considered in the analysis

In addition to these chilled water parameters, an air related quantity has been added to the set, that is the temperature of the air downstream the coil. This quantity is strictly related to the efficiency of the heat exchange between air and water at coil level. Other than these specific components variables, non-specific component air parameters are taken in account selecting flow rate, temperature and humidity for supply, mixed and return air and outdoor air temperature. Fans too are considered in the analysis, so speed and power of return and supply fan are selected.

At the end of the selection process, 21 quantities have been selected for the application of the methodology, which are reported in the Table 7.

5.3. Application of the methodology

The methodology developed has been applied entirely to the dataset, so to test the methodology itself and evaluate the results obtained, both for the Fault Detection phase and the Fault Diagnosis phase.

First of all, the data has been prepared for the application of the methodology in the Data preparation stage, so to get variables' values suitable for the subsequent analysis.

The first operation carried out on the data has been a filtering of the time series, because some of the variables present several outliers, probably due to the way the monitoring devices providing these quantities have been collected the data. In fact, the actual profile of these variables was disturbed by some nearly 0 values located regularly along the entire time series. In order to do so, the Hampel filter has been applied, which replaces the outliers with the median calculated in the neighbourhood of the point analysed.

In fact, it computes a sliding window of user-set width, for the identification of a neighbourhood of the point analysed, then it finds the median and the standard deviation in this interval. After that, it substitutes each value deviating from the median more than a threshold with the median. The threshold considered for the filtering is nothing else the product of the standard deviation and a coefficient set by the user, which is often equal to 3.

The filtering by means of the Hampel filter has been performed on R, using the function "hampel" included in the "pracma v1.9.9" package. This function applies the Hampel filter setting the window length and the coefficient multiplying the standard deviation considered as threshold for sifting. The best results in data cleaning for the dataset considered have been obtained for the sliding window width set to 30 minutes and the coefficient for the threshold computation set to 1. Higher values of the coefficient determine a less tight threshold; therefore, the outliers are not filter out all. On the other hand, a smaller window width would give more importance to the outliers' values, influencing too much the calculation of the median and the standard deviation. By setting the parameters like so, the profiles of the variables after the data cleaning stage, do not present disturbances or noise observances. (Figure 25)

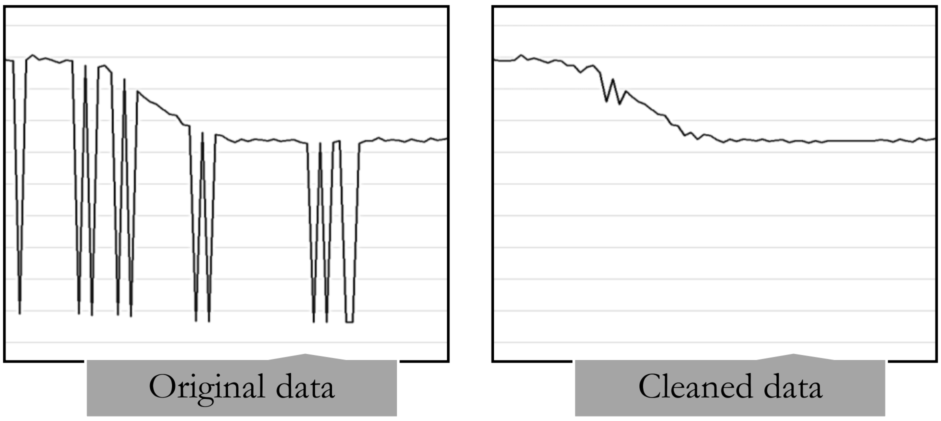


Figure 25 Example of data cleaning by means of Hampel filter

Once the data has been cleaned, the aggregation phase has been conducted, so to reduce the computational effort by reducing the size of the time series.

The aggregation procedure may involve a loss of information the reduction of the time series' size is pushed too far. In fact, the aggregation process substitutes the values of the profile with the mean value calculated with various interval. As a result, if the interval is too wide the approximation committed is too high, hence the choice of the aggregation interval represents a delicate point.

At this point, a new objective of the test and validation of the methodology comes out. In fact, since the sampling frequency of the dataset is very high, i.e. 1 minute, it can be possible to split the analysis on different aggregation levels, so to assess the influence of this operation on the results.

For this sake, the aggregation has been conducted 3 times, with 3 different aggregation intervals, that are 5 minutes, 15 minutes and 30 minutes. As a result, from this point on, the performing of the methodology is replicated for all the 3 aggregations, since from the original dataset has been derived 3 different datasets.

The aggregation has been implemented using the function "ddply" of the "plyr" package, collapsing time series of the variables around the mean values calculated in the aggregation interval. After the aggregation, the data has been transformed from continuous to discrete, since the ARM algorithm used successively, i.e. cSpade, can handle only categorical quantities, therefore this passage is mandatory for the application of this methodology.

The data discretization is process that can easily vitiate the results, making them useless, therefore to avoid mis-discretization caused by the user's choice, this phase has been thought has the most unsupervised possible.

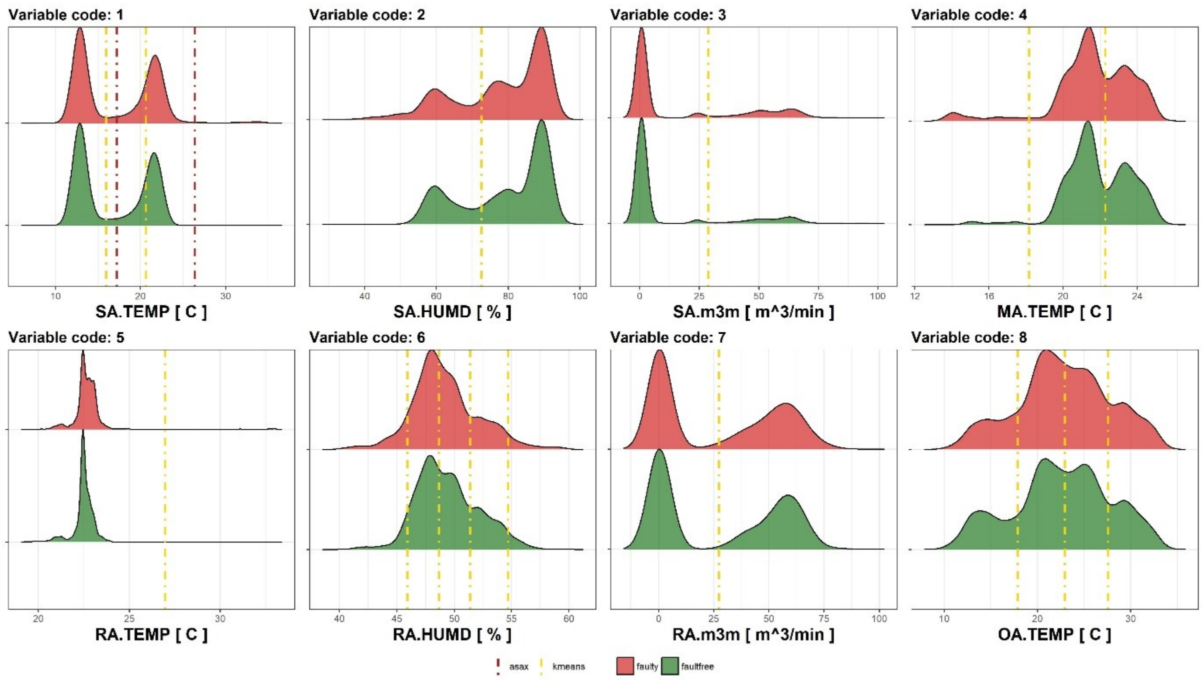
In this sense, the discretization of the variables is conducted using the aSAX optimization algorithm, with the one-dimensional k-means in charge of the initialization of the breakpoints.

This kind of initialization has been preferred to the Gaussian-like equal probability one of the SAX discretization and proposed by the authors of the aSAX algorithm, mainly because in this way the user is relieved of duty of the number of intervals definition for each variable.

In fact, the k-means clustering algorithm, using the Euclidean distance as similarity criterion, is employed for the identification of the initial breakpoints position, optimizing the number of cluster found. The optimization of the number of cluster, namely discretization intervals, has been conducted by means of the calculation of the Davies and Bouldin index to evaluate the quality of the clusters found. The optimization has been performed using the NbClust function of the "NbClust v3.0" package, while for the discretization has been used the "discretize" function from the "arules" package, which can compute the k-means discretization rather than the discretization with given values for the breakpoints.

The number of intervals of discretization has been limited from a value of 2 to a value of 5, in order to guarantee at least 2 levels, e.g. "off condition" and "on condition", and to avoid having too narrow intervals if the number would be too high, driving to poor representation of the variables and a high computational cost. The process of the discretization is performed on the entire dataset, i.e. including faulty and faultless data, even though the dataset has been split in faulty and faultless for the fault detection analysis, in the following phases of the procedure.

The distribution of the variables of the dataset with 15 minutes' aggregation, reporting the discretization's breakpoints, is reported in the Figure 26, while all the other cases are reported in the Appendix 0.



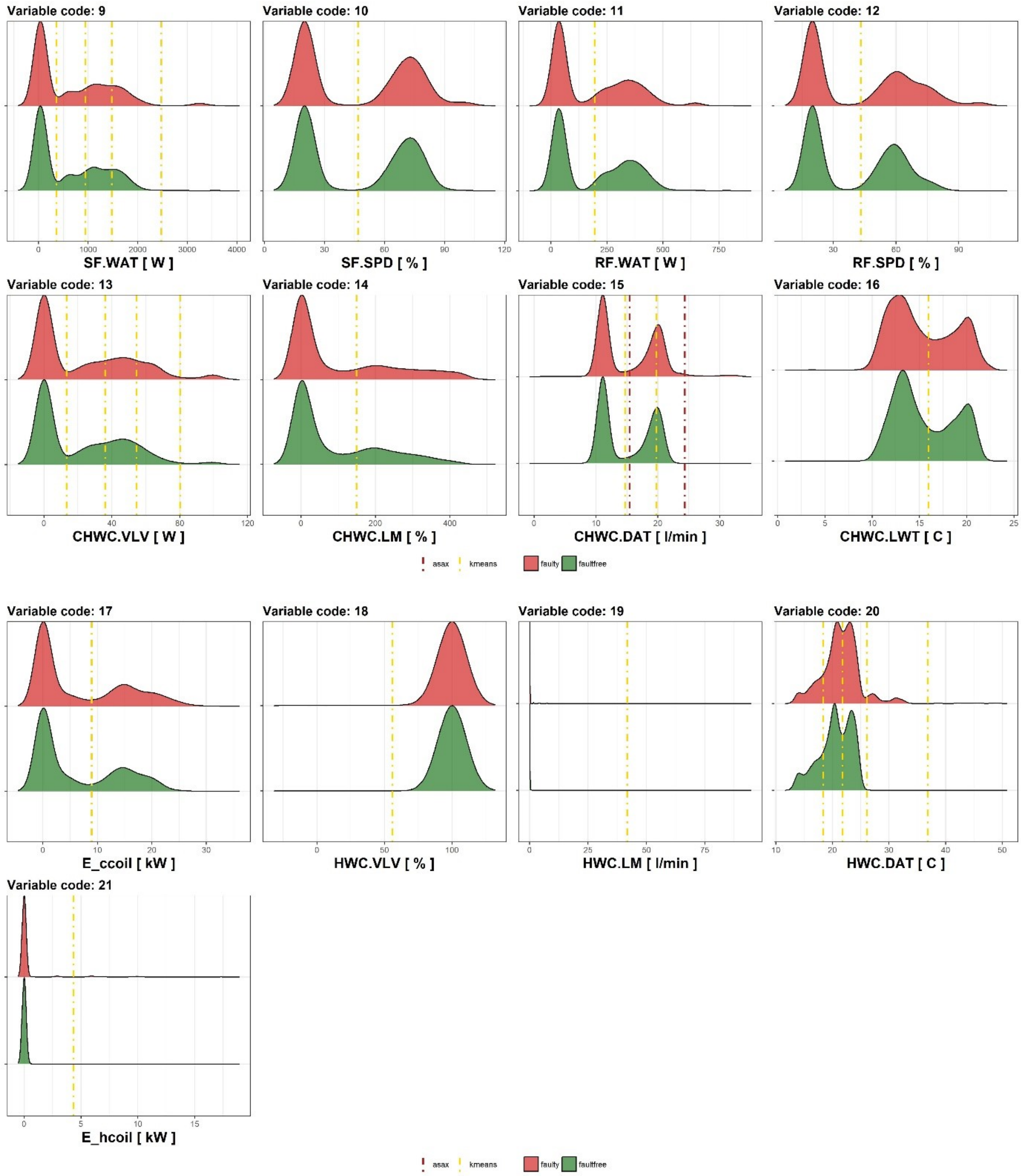


Figure 26 Variables distribution and discretization's breakpoints location

The representation of the distributions shows how the faulty ones are fuller and wider with respect to the fault free ones, since the amount of data is higher, and the greater range of values assumed by the variables in certain fault condition. The location of the breakpoints of discretization prefers zones at low density of points, thanks to the k-means similarity criterion.

The criterion of optimization used by aSAX, i.e. the SSE minimization, is the same of the k-means algorithm, therefore it is not surprising that the k-means initialized breakpoints location is often considered optimal by the aSAX algorithm, resulting in a superposition of the 2 kind of breakpoints representation.

At this point, data is ready to be treated to create a transitions' unique time series and then organized in an inter-transactional database, as described in the chapter 4.1.

The inter-transactional database has been created starting from the dataset purified from the faulty data, since the ARM process aims to retrieve rules related to the normal operation of the system. By using only the data tagged as "Normal", it is sure that the association rules extracted are describing a fault less condition of the system, therefore if the rules have high values of support and confidence can be considered as dominant, so a reference for the expected functioning condition.

In the ARM step, the limits and the effectiveness of the methodology are explored further, since in addition to the aggregation intervals assessment, the time window width analysis has been conducted as well.

By the construction of the inter-transactional database at different time window width, it has been possible to identify the minimum time span in which the association rules first appear, so characterising the association rules found with their related minimum time window width.

For this reason, the time window has been implemented for 5 different width, which are 30, 60, 120, 180 and 240 minutes. In the intermediate developing phases, longer time spans have been used, but results of longer time window can be related to narrower ones, since the association rules found in those cases are basically the same found for smaller time windows.

In the implementation of the sliding window, particular attention has been payed to the sliding across the days. In fact, given nature of the dataset used, it is prevented the case which the sliding window is located between 2 days, since they are consecutive in the dataset, but not necessary strictly consecutive in temporal sense.

Once the inter-transactional database has been constructed, the actual ARM process has been carried out, by means of the application of the cSpade algorithm for the search of the association rules in the database. The cSpade algorithm is implemented in the "cspade" function of the "arules" package, therefore its application is easy to perform.

The association rules mining process has not being constrained in any way by the cspade parameters, other than the minimum support and minimum confidence ones, which are mandatory for the ARM process. These 2 parameters are set to 0.1 both, even though the aim of the ARM in this case is to find dominant rules, hence with high values of support and confidence.

The threshold for the rules extraction parameters are set as low at this point of the methodology, because the algorithm can calculate the value of support and confidence on the basis of the number of transactions, that is the length of the inter-transactional database, i.e. the number of observances of the time series. This number is huge, so the corresponding values of support and confidence obtained for the rules extracted are very low. As a reference, it should be considered that values of support higher than 0.2 has not being extracted during all the developing phases. After the cspade rules extraction, support and confidence associated to the rules found have been recalculated and then the real thresholds for minimum support and confidence are applied, filtering out all the rules with support or confidence less than 0.5

At this point, the rules extracted should be filtered out to reduce to set to the only relevant and unique rules. The filtering has been conducted in accordance with the procedure described in the chapter 4.1.

In this way, the fault detection stage of the methodology has produced a set of association rules which is able to provide a reference for the normal operation condition of the system to the user.

The ARM has not produced a set of rules for each time window width, since the shorter the time span, the less the rules succeeding in exceeding the minimum support and confidence threshold.

In fact, were extracted rules only for the time span 120, 180 and 240 for the 5 and 15 minutes, while for the 30 minutes aggregation only 180 and 240 produced association rules. The number of rules extracted, and survived from the filtering process, for the whole range of cases, is reported in the Table 8.

By means of the representation of the patterns of the rules and the profile of the same variables when the rules are not respected, the user can interpret the information encapsulated in the rules, so to extract the knowledge necessary to identify whether the violation of a rule can bring to a fault or not, since rules in the set can result even not helpful for fault detection.

		Time window width		
		120	**180**	**240**
Aggregation	**5**	4	50	74
	15	2	18	47
	30	-	6	18

Table 8 Rules extracted for different aggregation and time span

After the identification of the occurrence of a fault, the methodology provides even a way to identify the cause of the fault by means of the Fault Diagnosis stage. Since the fault moment are tagged in the dataset used, it has been possible to perform this part of the procedure as well.

Although dataset contains the fault tagged moments, i.e. entire days reporting the same fault tag, the fault diagnosis phase cannot be performed in the sense which has been intended to be performed. In fact, cannot be constructed an actual error signals matrix as input for the fault diagnosis stage, but this kind of dataset can be used to implement a kind of sensitivity analysis, to evaluate the effectiveness of the methodology in the isolation of faults.

For this purpose, it has been constructed an error signals matrix for every fault present in the dataset, taking all the rules that present a violation for the fault considered. In this way, the signals considered are all the possible signals collectible if a fault is present. It is worth noting that if a fault is present in a real application, the error signals reporting the system should be even a fraction of the total rules considered in this analysis.

By taking a fault by time, it has been performed the score calculation analysis, paying particular attention on the difference between the score related to the fault considered and the scores of the other faults. In this sense, the diagnosis of the fault considered can have sense only if the difference among the scores is high, unless the identification of that fault cannot be performed.

In order to enhance the grade of detail of the analysis, the diagnosis stage has been split in accordance with the time of the day in which the rules act. In this sense, the association rules found have been classified in rules related to the “start-up”, the “operation” and the “shutdown” of the system.

The rationale between the classification is supervised, given that the operation rules are the one present in a time range higher than 3 hours from the 10:00 to the 16:00, while the other 2 classes embraces the cases which the time range is less than 3 hours and the occurrence of the rules has been reported before the 10:00 or after the 16:00 respectively.

This kind distinction is necessary to avoid the mixing of rules referring to different class, so reporting normal behaviour in different moments of the day, which can eventually opposite in terms of trend too. For this sake of clarity, it is worth noting that, by chance, among the rules of start-up can be present a rule reporting the same variables of one from the shutdown set, but with reversed trend.

After this last analysis, the methodology is completed, the interpretation of the results can bring to the reasoning of the effectiveness and limits of the whole methodology with its different facets coming from the exploration of different aggregation, time window width and time of the day.

6. Results

In this section, the results obtained for both the Fault Detection part and Fault Diagnosis are discussed. In this way, the interpretation of the output of the methodology is provided, giving an example of the knowledge that can be extracted from this kind of Fault Detection and Diagnosis methodology.

Firstly, it is reported a set of rules extracted with different aggregation and different minimum time window width, discussing their physical implication and interpretation.

Successively, the diagnosis procedure's application is reported, showing the results in terms of a grouping of the rules extracted and the results of the isolation of the faults considered in a kind of a sensitivity analysis.

6.1. Extracted rules interpretation

In this section it is reported an insight on the results obtained in the first stage of the methodology, that consists in the operation of Fault Detection. In particular, the assessment of the results has been focused on the interpretation of the information provided by the rules and contextualized by the system expert. In fact, the interpretation process cannot be carried out unless it is available a prior knowledge of the kind of system which the dataset refers to. Therefore, the awareness of the main functioning and the expected behaviour and correlations of the system's component is necessary.

As a consequence, the expert of the system plays a fundamental role at this point, since the grade of the knowledge extracted depends even on the grade of awareness of the expert interpreting the results, although the methodology provides easy to understand information to allow the utilization of results in any case.

Before going into the actual analysis of the rules extracted, it can get aware of a point that affects the whole methodology employment and the results interpretation as well.

At first glance, Table 8 let come out how large is the entire set of the rules extracted from all the aggregations, but it is worth noting that, since the datasets for the different aggregations derive from the same original dataset, there are many rules repeating themselves with different aggregation. Differently, as can be noticed by the lowering number of rules extracted with increasing aggregation interval length, some rules disappear in the approximation of wider aggregation. In fact, with higher aggregation interval, the resulting modified time series, i.e. the resulting PAA, is less affected by minor fluctuations of the original time series. The definition of minor fluctuation changes with the width of the aggregation interval. As a consequence, the 5 minutes aggregation fits more accurately to the original time series than the 30 minutes one.

Even though the number of the 5 minutes aggregation provides a better description of the system by reporting a higher number of rules, i.e. of situations considered, it is always desirable to obtain such large set of rules. In fact, in cases which the computational cost of the application of the methodology gives constraints in the choice of input parameters, the 5 minutes aggregation for this dataset would probably be avoided. In fact, the time elapsed for the whole process results much higher for this aggregation, although the accuracy in the system's behaviour description is higher too.

Nevertheless, the employment of aggregation this tight, may not represent an issue for some applications, so the analysis has been conducted even for this case.

Following has been reported a selection of relevant rules taken from the whole set of the extracted ones.

As a reference, the whole set of rules extracted for all the aggregations has been reported in Appendix D, even though in this section only a fraction of these rules has been investigated deeply.

The rules reported in this section have been selected in order to avoid repetitions in the exploration of the different aggregations and provide an insight of the most relevant ones sorted by the time of the day. Reporting the rules relating them to the moment of the day which they occur, allows to classify the rules in "start-up", "operation" and "shut-down" ones. This classification makes clear the context in which the rules operates, making them more useful in the cases of real-time implementation.

The rules are identified by an encoding reporting the aggregation, the minimum time window and the reference sequential number. The reference sequential number is specific for the aggregation and time window, since it starts over for every time window set of rules for the same aggregation.

The format of the following representations of the rules is set as constant for the sake of clarity, reporting, from the top to the bottom, an example of the profiles of the rule when it is verified, the time of the day distribution, the actual time window distribution and the examples of the kind of violations encountered in the dataset.

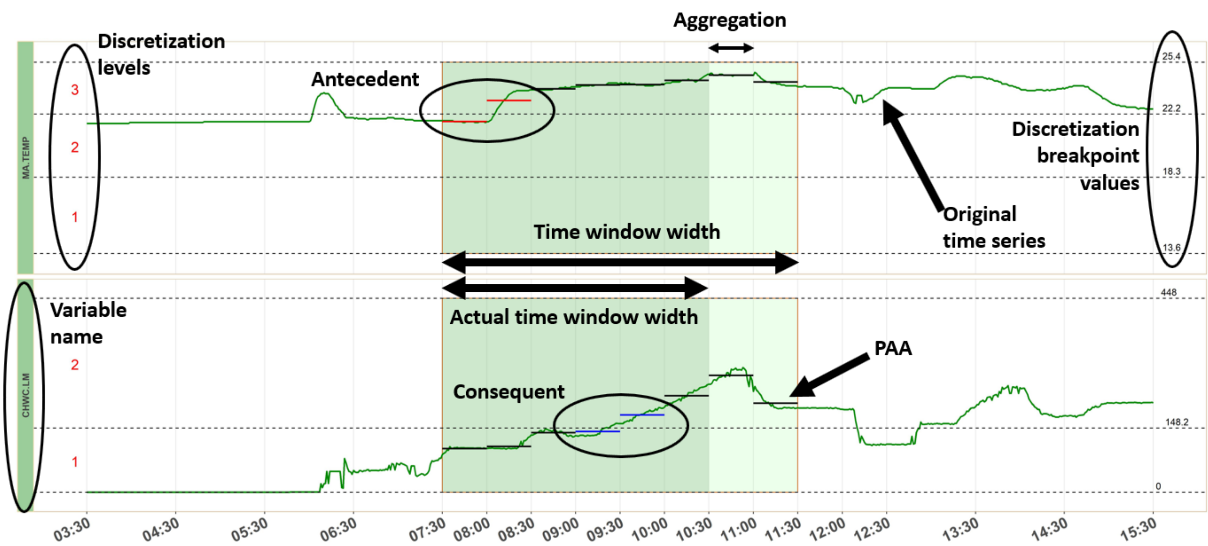
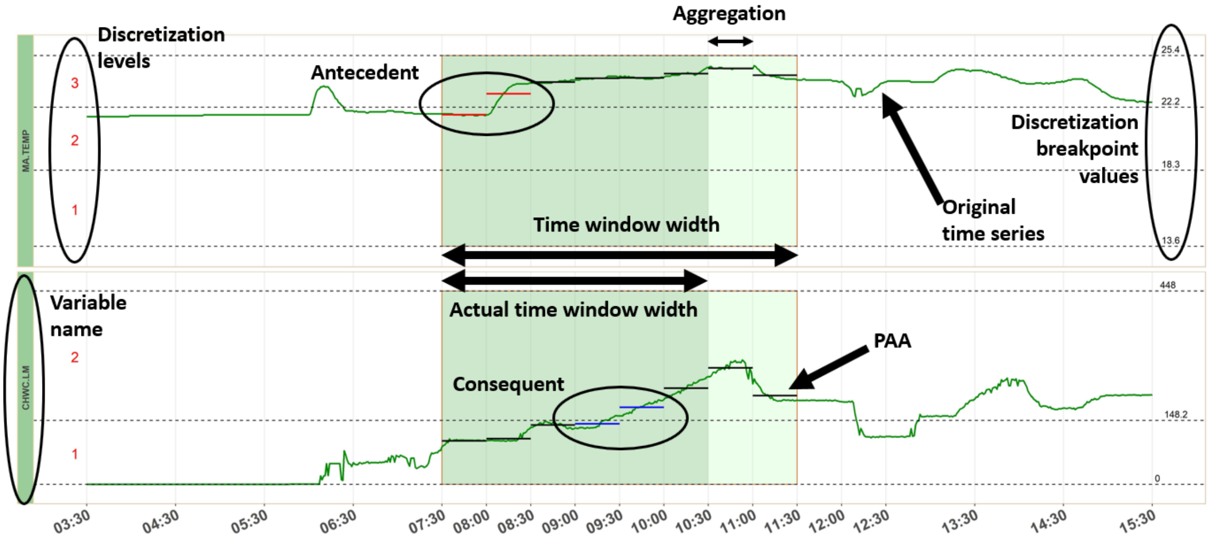


Figure 27 Rules' profile format

In Figure 27



it is shown the format of the profiles images, indicating the information reported in the representation. The profiles images show in green the original time series, i.e. the 1-minute frequency one, with a superposition of the PAA, constructed with the considered aggregation, only in the time window in which the rule's pattern is located. The pattern is highlighted on the PAA, showing in red the antecedent's transitions, while in blue the consequent's pattern. The profiles of the PAA and the original dataset are reported for the sake of reduction of the level of abstraction reached by the discretization encoding. In fact, to make clear the interpretation of transitions present in the pattern, the profiles representation reports also the breakpoints of the discretization of the variables implied in the pattern. The time window in which the pattern of the example is occurring is shown as green boxes surrounding the profiles, colouring differently the all-time window width and the time window width in which more of the 80% of the patterns extinguish, therefore showing the time window and the actual time window related to the rule.

The actual time window distribution is reported at the time of the day indication, specifying separately the actual distribution and the cumulative of the distribution. At the very bottom of the representation, the representation of the kind of violation related to the rule is reported, showing from left to right the case of the only presence of the antecedent, the only presence of the consequent and the absence of the whole pattern. The last case includes the other 2 since represents a general categorization of violation of the rule. If the violation of the rule is not related to one of the violation mode considered, space is left blank, maintaining the format of the representation.

The first rule represented is the rule 5-240-236 (Table 9,Figure 28), showing the correlation between the temperature of the water downstream the cooling coil and the power consumption of the supply fan, generally at the start-up, in fact in the profile reported the time window starts at 6:05 am. Even though according to Table 9, the rule may seem representative of the functioning on the morning, from the distribution of the occurrences along the day in Figure 28 it is clear that this rule is specifically describing a start-up situation.

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
5	**240**	**236**	162161	9293	0,10	0,40	0,86	0,97	160	0,93	1,97	05:45	12:20

Table 9 Rule 5-240-236

The correlation found in this rule highlights the functioning of the system at cooling coil level, since both the variables can be related to the heat exchanged between air and water in the cooling deck. In fact, it shows the common trend of these 2 variables when the water of the cooling coil is starting to be chilled and the supply fan increases its consumption, to enhance the heat exchange rate at the cooling coil level by means of increasing the air speed

This rule refers to a typical situation occurring at the start-up of the system, showing how strong is the correlation between these 2 events. In fact, looking at Table 9 it can be noticed that the support is high, as well as the percentage of days during which the rule occurs, i.e. support of day, and the confidence too. The values of these parameters induce in thinking that the behaviour represented is very common and the violation of this kind of rule represents a strong signal of fault since it means that the starting of the system encounters some problems.

As concerns the support and the confidence parameters, it should be considered that only the parameters calculated with the CONOTOOL (Martínez-de-Pisón Ascacíbar, et al., 2009) formulation should be considered valid as reference, since the other 2 parameters reported Table 9 are not a correct representative for the rule. In fact, the simple support and the simple confidence values are too low to be taken into account as they are, since they were calculated on the basis of the whole number of transactions which is equal to number of instances in the dataset. In Table 9 these 2 parameters have been reported too only for the sake of completeness, since these parameters have been calculated by the algorithm which extracts the rules.

From Table 9 it can be deduced even how the rule is frequent and how it is distributed along the day, since it reports the value of support of days and Mean Occurrences per Day. The first parameter gives an indication of how frequent the verification of the rule is, since this rule occurs a high percentage of days in the dataset. The second parameters states how the rule is generally distributed along the day, reporting how many times it occurs in a day. This parameter characterises the rule's showing up in statistical terms, providing helpful information for the complete interpretation of the rule.

The time window that large and the shape of the profile of the verified rule in the representation of Figure 28, let come out the reason why the MOD is nearly 2. In fact, most of the times, the first occurrence of the pattern, occurs almost contemporary, influenced by the presence of a peak at the moment of the turning on of the plant. The second occurrence of the pattern can be probably attributable to the less fast response to the change of state, after a time span of more or less 180 minutes. As a result, this rule is able to take into account both the fast and slow changes occurring typically at the start-up for these 2 variables.

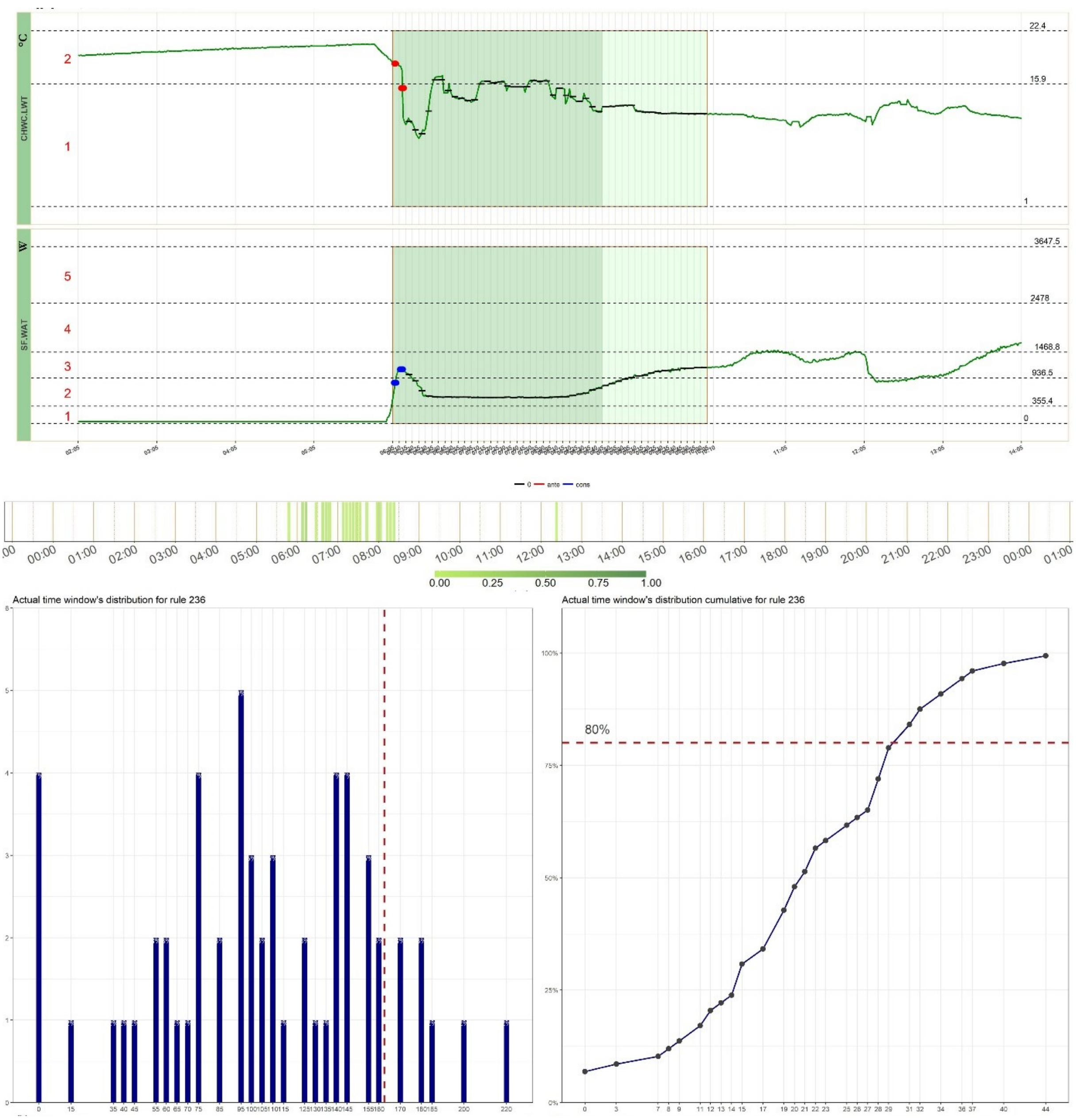


Figure 28 Rule 5-240-236 (a)

The strength of this rule lies even in the implication of the fault if violated, in fact, in the case which this rule is not respected, the only faults which can be correlated to is the complete blockage of the cooling coil valve. As a consequence, this rule allows diagnosing the CCVSFC fault (Appendix E) at a relatively short period of time from the start-up, by the only observance of the absence of the pattern (Figure 29)

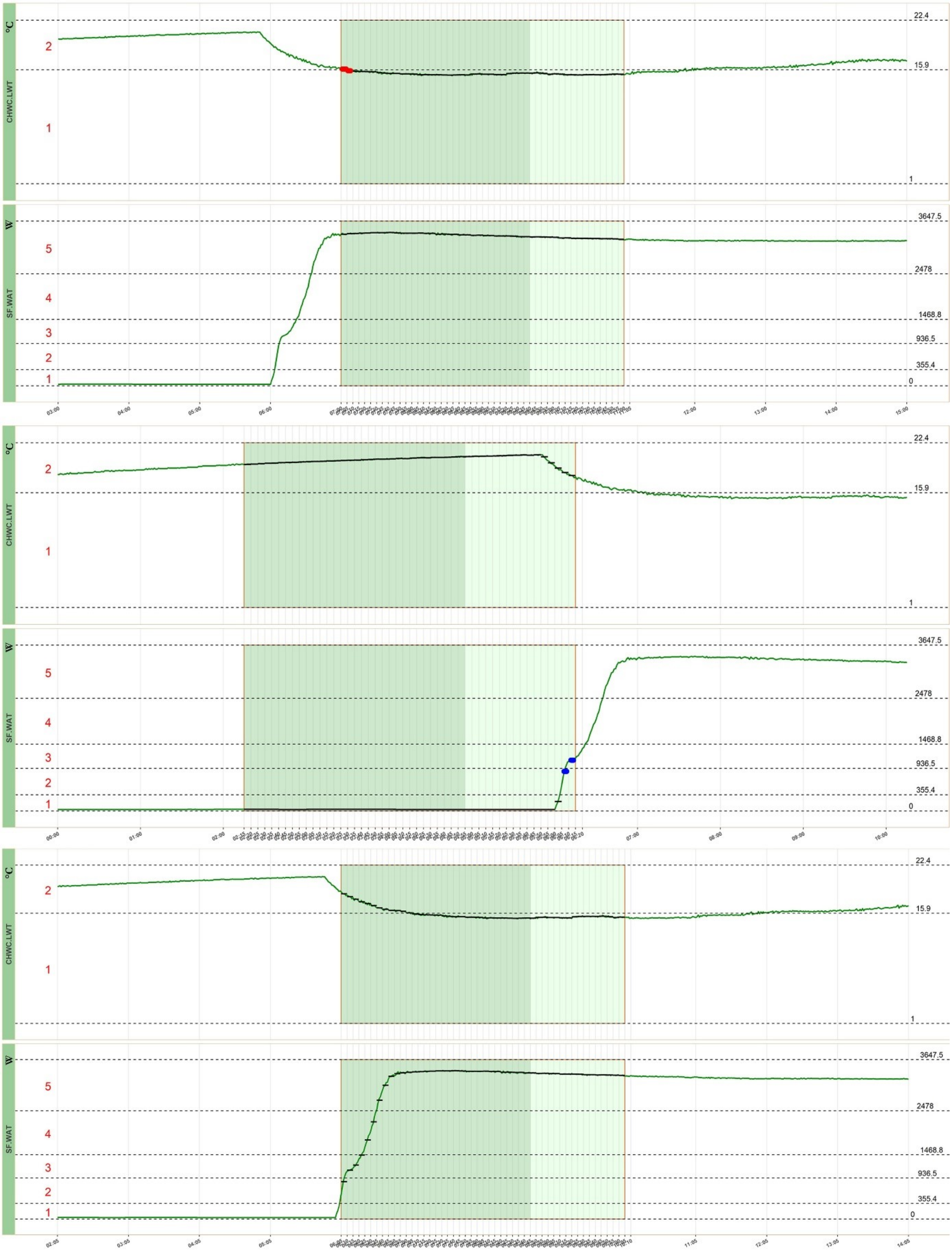


Figure 29 Rule 5-240-236 (b)

The set of rules extracted can provide information about the reference behaviour of the system in a less specific situation, reporting the common correlation among variables regardless the instant in which the pattern is observed. This is the case of the rule 5-240-62 (Table 10,Figure 30).

This rule gives an indication of normal functioning, as a trend of the profiles of CHWC.VLV, CHWC.LM and SF.WAT (see Appendix B for variables reference), mainly at any moment of the functioning of the morning.

The trend expressed by the pattern is in complete accordance with the physical expectance, matter of fact, all the variables influence the heat exchange at the cooling coil level in the same direction of their changes and the intervention of all the 3 variables is necessary to control the effectiveness of the heat exchange in the deck.

Even though support and confidence of the rule are both quite high but not extremely high, the implication of this rule is strong, since it is able to reduce the number of variables to check in the testing of some specific components. In fact, the implication due to the violation of this rule is specific to 2 different components, that although they are not strictly correlated physically, they are correlated in the general effect to the system caused by their malfunctioning. These 2 components are the cooling coil, which implication is clear from the presence of 2 variables strictly related to it in the antecedent, and the return fan, which is less expectable but equally important. (Appendix E)

The return fan failure implies the reduction of the recirculation rate, therefore the air supplied in the ambient is totally or almost totally coming from outside. Given that outside air is hotter than the indoor air, the temperature upstream the cooling coil is higher than expected, therefore, unless the cooling coil is set to exchange more than normal, the temperature of the supply air is higher too. This effect is easily attributable to a failure of the cooling coil as well, therefore the correlation of the violation of this rule with both the faults is justified.

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
5	**240**	**62**	132133, 141142	9394	0,13	0,48	0,81	0,72	135	0,73	1,60	06:05	13:10

Table 10 Rule 5-240-62

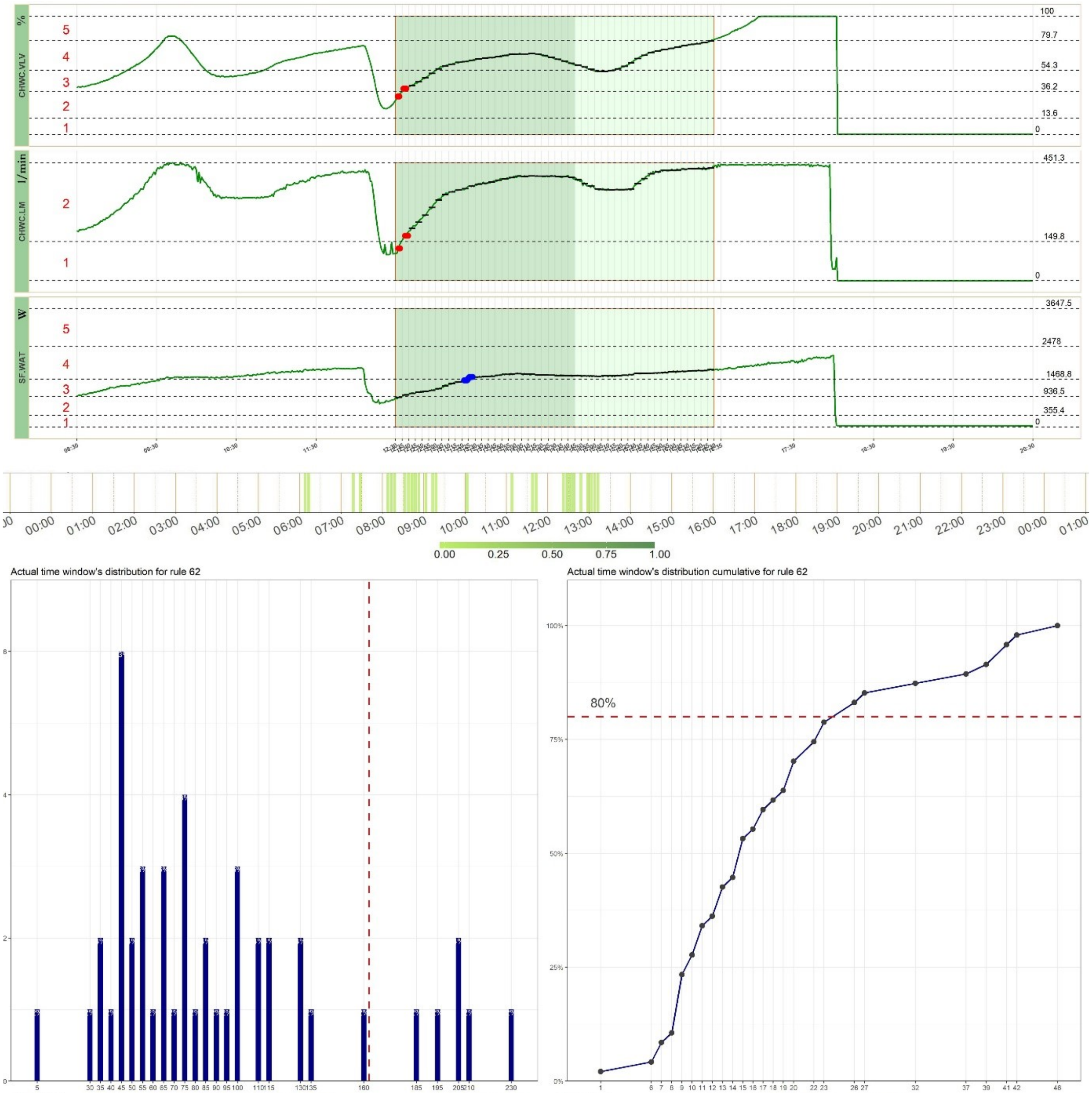


Figure 30 Rule 5-240-62 (a)

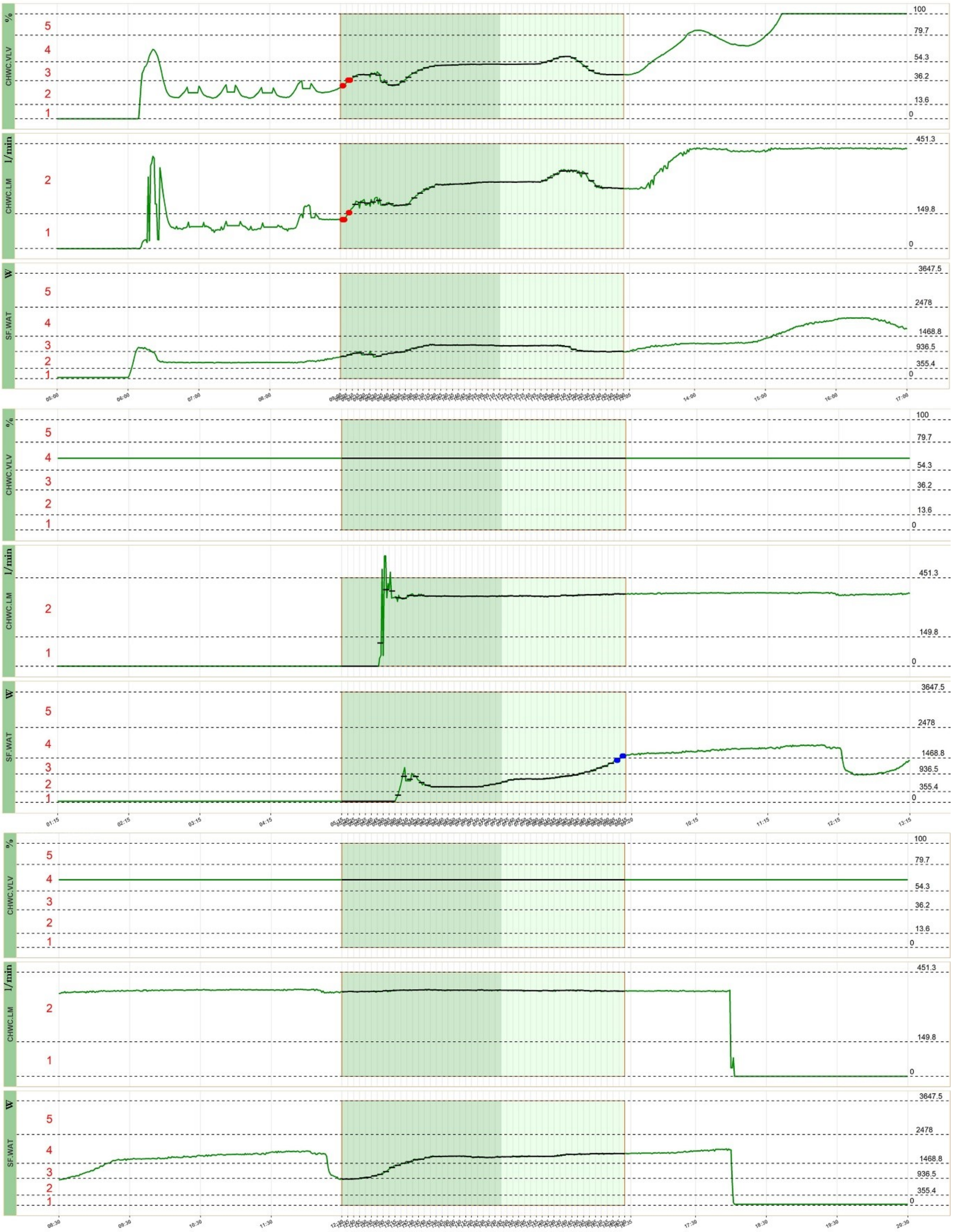


Figure 31 Rule 5-240-62 (b)

Following it is reported another non-time specific rule, that is rule 5-180-9 (Table 11,Figure 32), which can be found both in the early hours of the morning operation or in mid-day operation.

This rule is able to increase the awareness of the functioning of the system in terms of specification of the cause-effect logic between the variables monitored.

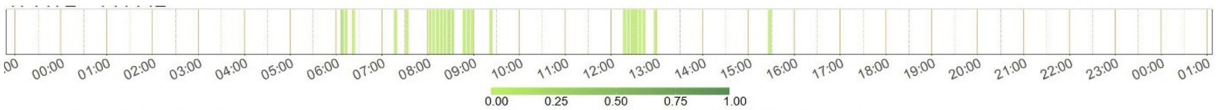
In fact, this rule states that there is a powerful correlation, sustained by high values of support, confidence, the percentage of verification days, between the E_ccoil and the CHWC.LM (see Appendix B for variables reference). This correlation is not surprising since the energy exchanged in the coil and the water flow rate are directly proportional. Furthermore, the actual time wind width that narrow demonstrates again how fast is the response of the AHU system, showing a 0-time lag in most of the cases.

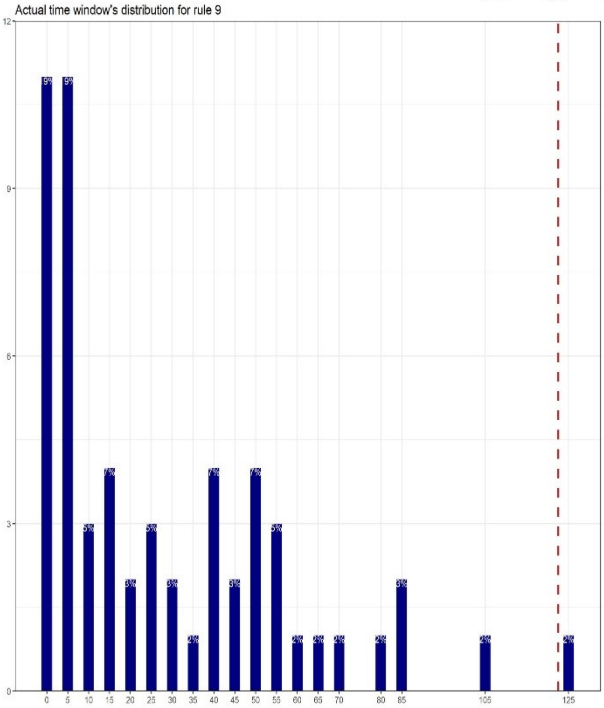
This rule is not specific for a time operation but provides an indication of the general functioning of a particular component that is the cooling coil, therefore the violation of this rule can imply only a malfunctioning at that component. (Figure 33**Errore. L'origine riferimento non è stata trovata.**) Furthermore, the almost twofold distribution along the day of the occurrence of this rule makes possible the detection of a fault in this component in more than one moment of the day. (Appendix E)

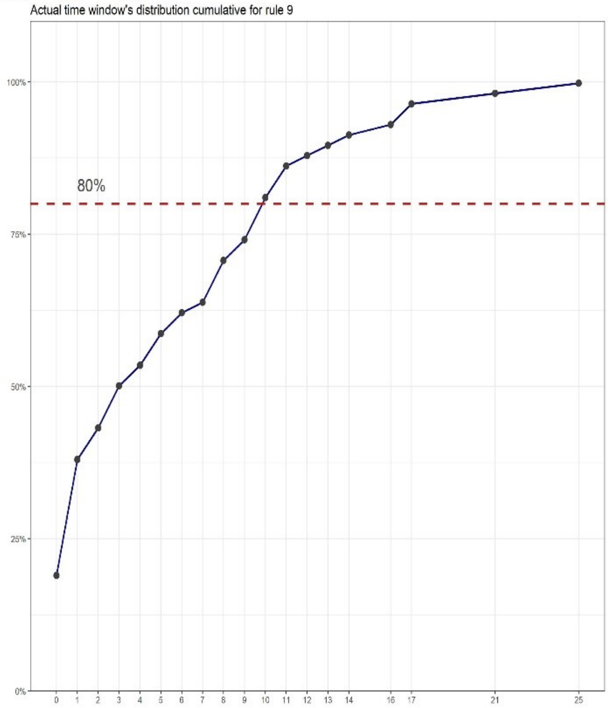
AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
5	**180**	**9**	171172	141142	0,15	0,55	0,81	0,91	55	0,90	1,97	06:05	15:25

Table 11 Rule 5-180-9


kW
E_ccoil
l/min
CHWC.LM
31.7
8.9
0
451.3
149.8
05:25
06:25
07:25
12:25
13:25
14:25


0.00
0.25
0.50
0.75
1.00


Actual time window's distribution for rule 9


Actual time window's distribution cumulative for rule 9
80%

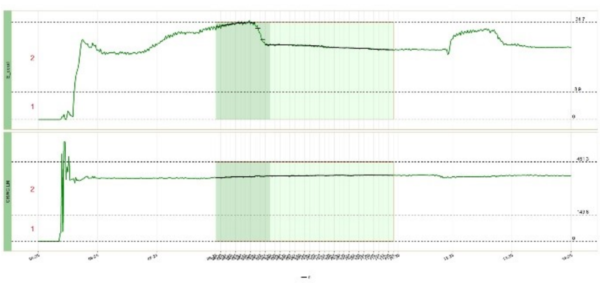


Figure 32 Rule 5-180-9 (a)

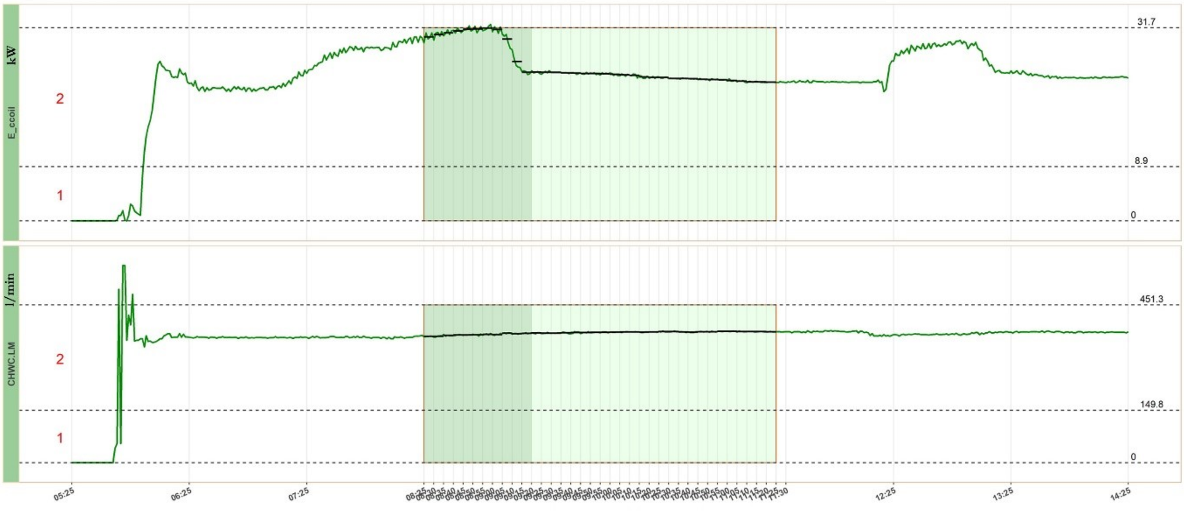


Figure 33 Rule 5-180-9 (b)

Another kind of non-time specific rules are the self-implying one, which rule 5-120-2 belongs to. (Table 12,Figure 34) The self-implying rules generally do not represent relevant rules, since they do not provide information of the correlation of various variables, so they are generally unable to describe the actual condition of the system. On the other hand, the self-implying rules can play a fundamental role in the detection and prompt diagnosis of faults strictly related to the quantity considered. This is the case of this rule, which takes into account the position of the valve of the cooling coil. This rule has a purely frequent and non-time specific occurrence since can be observed along the whole period of functioning of the system, more than 2 times a day when there are not any fault preventing the happening and with a strong implication (high values of support and confidence).

It is worth considering that, the stronger the implication, the higher the validity of the implication of its violation, because a strong implication makes justified the considering that implication as a general reference for normal condition. Given that, it is clear that this rule represents the golden rule for the identification of the stuck of the cooling coil valve, since the violation of that rule implies the lack of fluctuation of the position of the valve. (Appendix E) At support of this theory can be considered the evidence of the only absence of the pattern when the rule is violated. (Figure 35**Errore. L'origine riferimento non è stata trovata.**)

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
5	**120**	**2**	133134	134133	0,12	0,44	0,89	0,79	75	0,77	2,17	06:10	17:45

Table 12 Rule 5-120-2

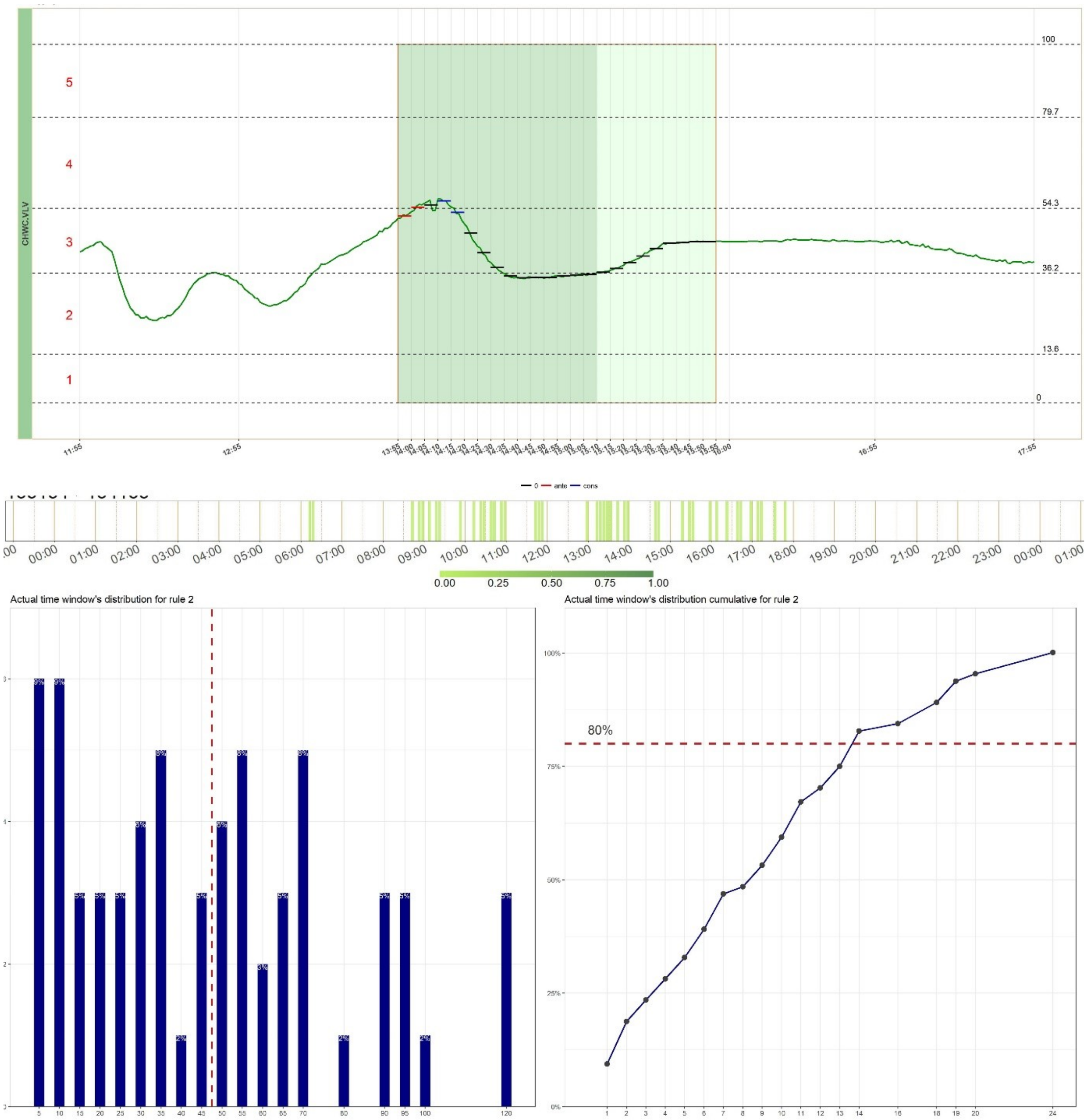


Figure 34 Rule 5-120-2 (a)

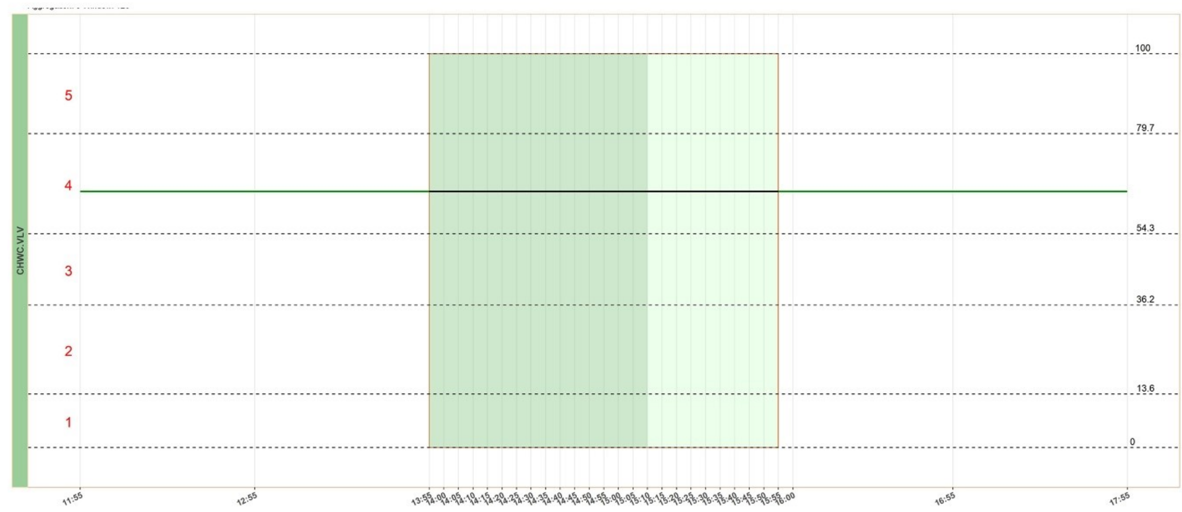


Figure 35 Rule 5-120-2 (b)

Another case of functioning at the start-up is expressed by the rule 15-240-102 (Table 13, Figure 36),which describes the behaviour of the SF.SPD, RA.m3m, RF.SPD, CHWC.DAT and SA.TEMP at the moment when the plant is turned on (see Appendix B for variables reference). The distribution of the time of the day and actual time windows give an indication of how specific this rule is since occurs all the time at the same time, i.e. the starting time, and almost instantaneously. Such low value for the actual time window confirms that the system has a very low inertia in the response since almost in the instant which the fans are turned on, the temperatures of air and water drops.

The extremely high values of support, confidence and percentage of days let deduce that this rule states a strong reference for the functioning of the system for normal conditions and for a specific moment of the day (this is supported even by the MOD assuming a value near to 1).

The faults implied by the violation of the rule is not specific to a fault, so the diagnosis is not straightforward as the previous rule but looking at the kind of faults implied it is clear how this rule is representative of the functioning of the components which its variables directly refer to. In fact, the violation of this rule can be attributable to a fault at the cooling coil valve or the return fan. (Appendix E, Figure 37**Errore. L'origine riferimento non è stata trovata.**)

This kind of rule results very helpful in the cases of online implementation since it embraces a high number of variables and reports a specific situation that is highly expected, therefore it represents a kind of checklist for the testing of the condition of the system at the moment which the plant is starting the operation. As a consequence, thanks to the narrow actual time window width too, the identification of the fault can be conducted almost in the exact moment of the starting of the system, guaranteeing a prompt intervention in case of malfunctioning.

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
15	**240**	**102**	101102, 7172, 121122	152151, 1211	0,1	0,88	0,9	0,9	30	0,9	0,9	05:45	05:45

Table 13 Rule 15-240-102

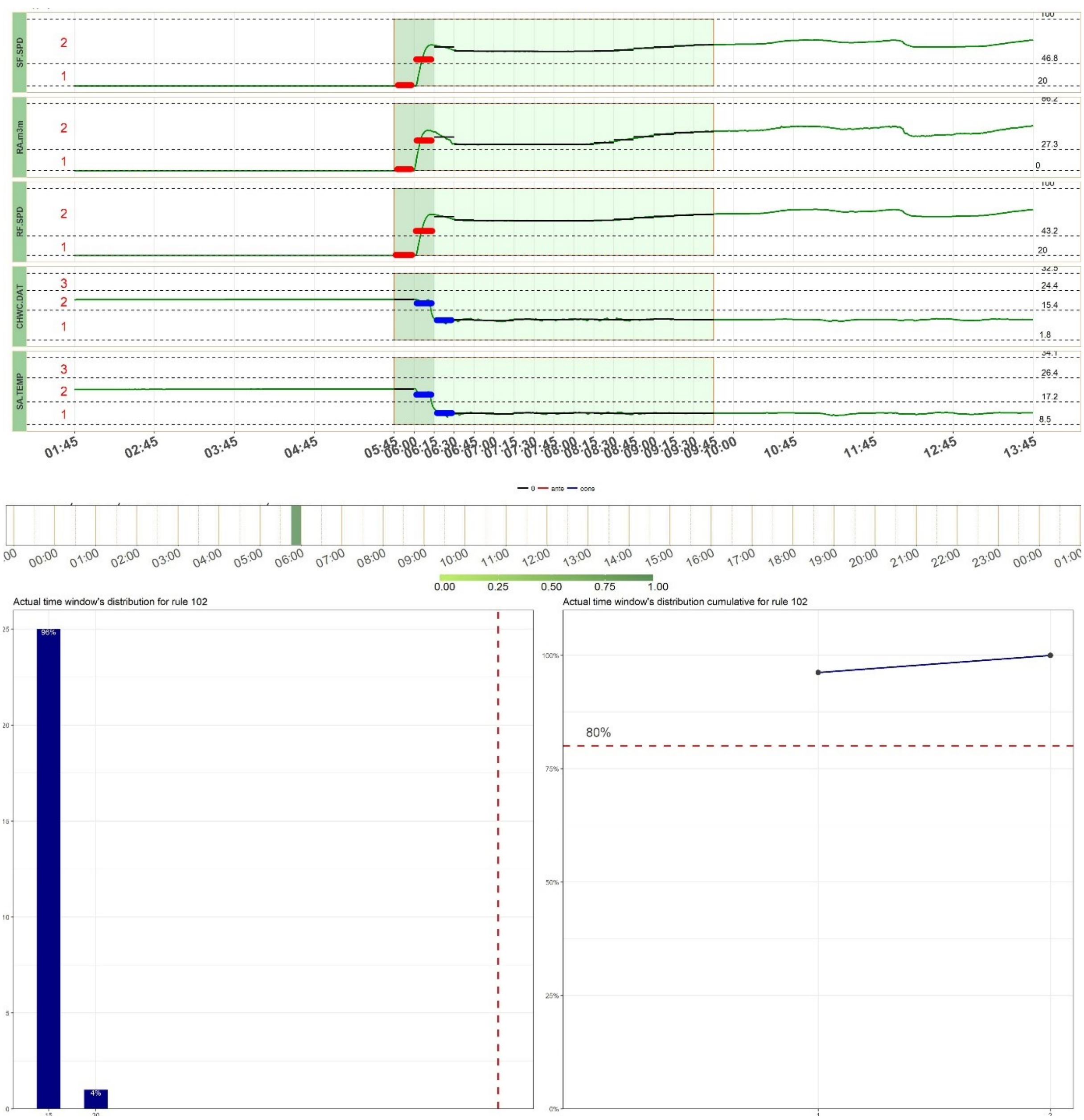


Figure 36 Rule 15-240-102 (a)

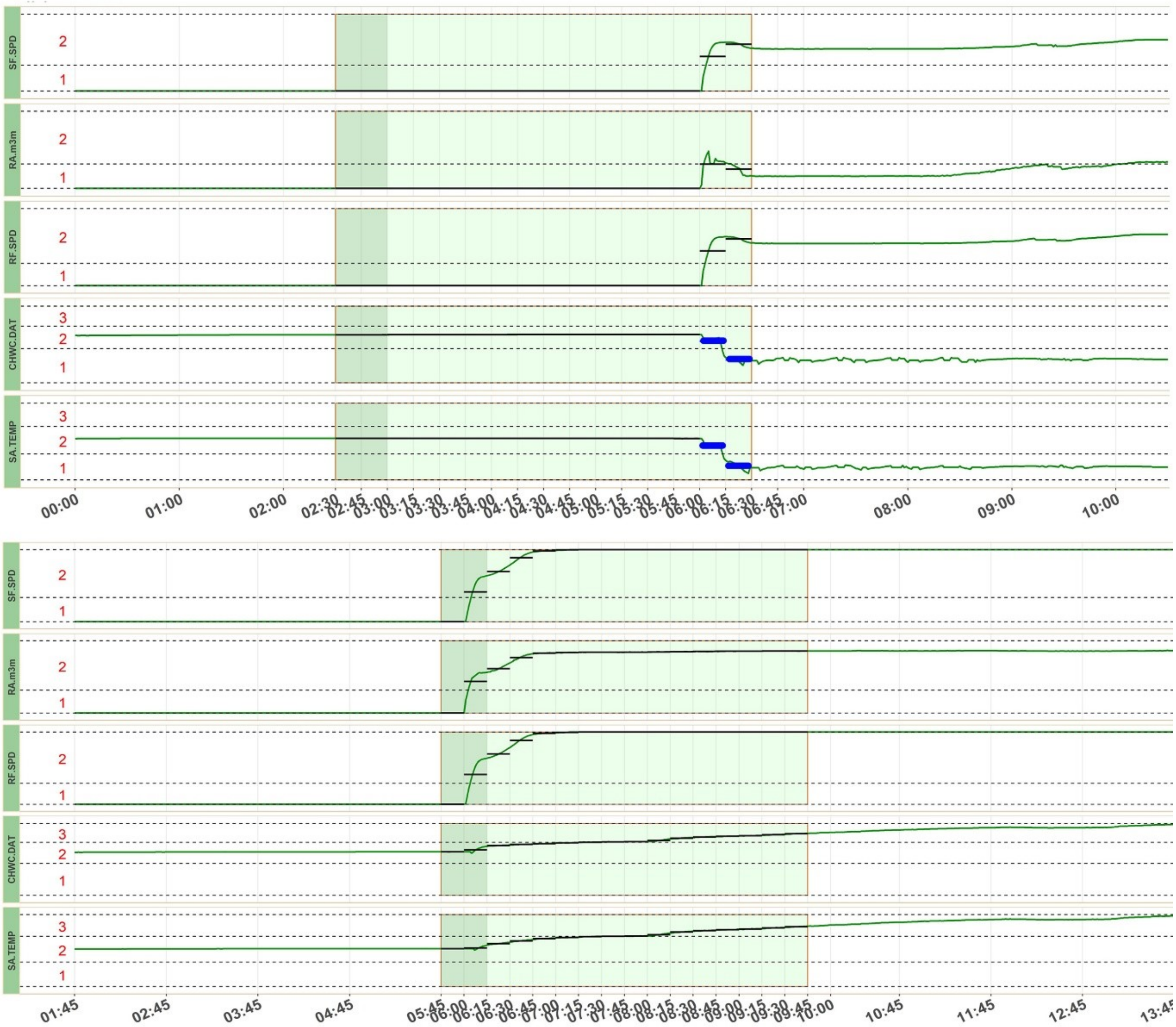


Figure 37 Rule 15-240-102 (b)

The rules that can be considered as reporting a general correlation between the system's variables during the operation along the day, may not refer necessarily to variables regarding quantities of the component. In fact, the unsupervised nature of the ARM drives to the identification of patterns implying variables unrelated at first glance.

This is the case of rule 15-240-19 (Table 14, Figure 38), which gives an indication of the correlation between E_ccoil and SF.WAT (see Appendix B for variables reference).

The correlation expressed by this rule results very strong and frequent, since the high values of confidence, support and MOD. The MOD of more than 1 states that it is spread along the day, supporting the frequency feature, but the time of the day in which it occurs tells that the rule is valid for the start-up and the mid-day operation.

This expression, even though it is not trivial to identify, it is not completely surprising, since these 2 quantities are both influenced by the heat exchange rate in the cooling coil. As a result, the violation of this rule can imply only a fault of the cooling coil. (Appendix E) In particular, the only fault implied is the stuck of the cooling coil valve at the closed position, as can be seen in the example of violations of the rule. (Figure 39**Errore. L'origine riferimento non è stata trovata.**)

Therefore, even though the 2 variables are not describing quantities of the same component, the rule is specific to a single component's behaviour.

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
15	**240**	**19**	171172	9293	0,14	0,48	0,81	0,96	195	0,83	1,6	05:45	12:45

Table 14 Rule 15-240-19

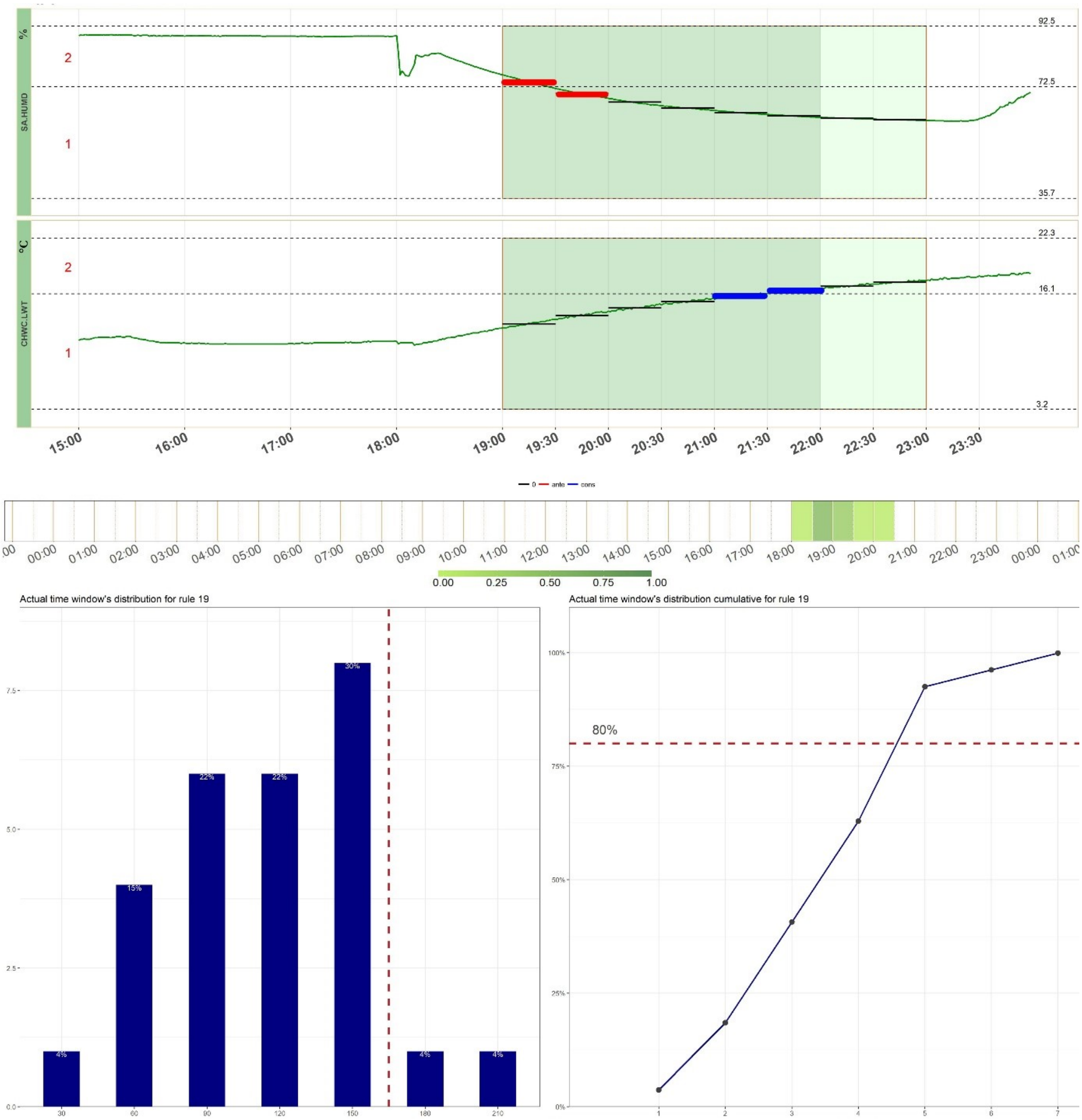


Figure 38 Rule 15-240-19 (a)

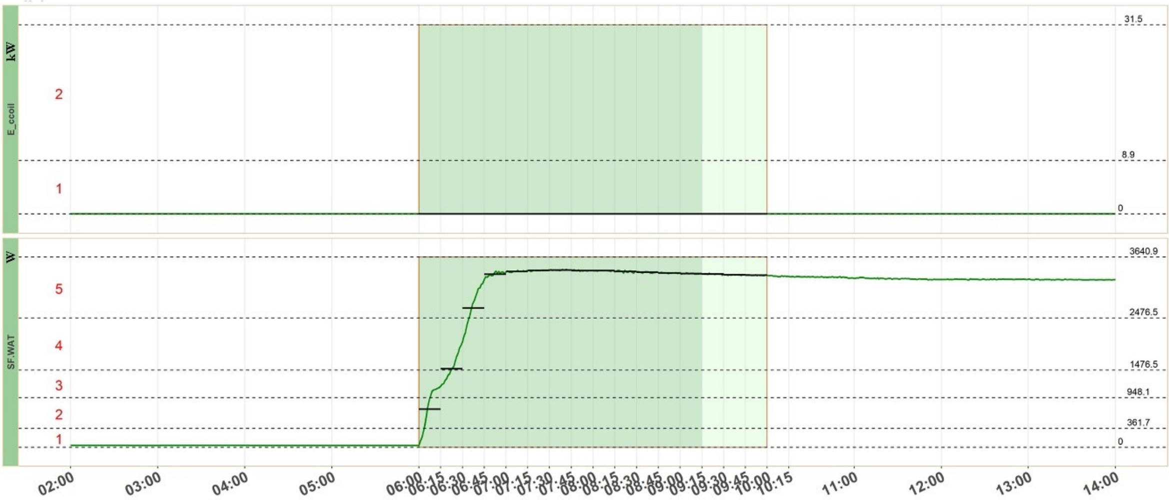


Figure 39 Rule 15-240-19 (b)

Another example of a specific rule for the start-up is the rule 30-240-34 (Table 15, Figure 40), which shows the correlation between CHWC.LWT and E_ccoil (see Appendix B for variables reference). This correlation is evident considering the way the heat is exchanged in cooling coil. In fact, the temperature of the water downstream the coil gives a first indication of the energy exchanged in the coil since the inlet temperature is constant.

Looking at the rule's parameters, it is clear that it is a frequent rule, reporting a strong correlation between the variables, but it is evident that it is able to detect the functioning of the system at the very moment of the start-up, detecting the typical peak of the profile for this specific system, and the less fast response of the system starting its operation.

Supporting this concept there is the value of the MOD stating that the occurrence of the rule is sure at least once a day.

The violation of this rules brings, not surprisingly, to the identification of a fault in the cooling coil, since the variables of the pattern are related only to that component. (Appendix E, Figure 41**Errore. L'origine riferimento non è stata trovata.**)

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
30	**240**	**34**	162161	171172	0,1	0,44	0,97	0,92	180	0,9	1,2	05:00	07:30

Table 15 Rule 30-240-34

Figure 40 Rule 30-240-34(a)

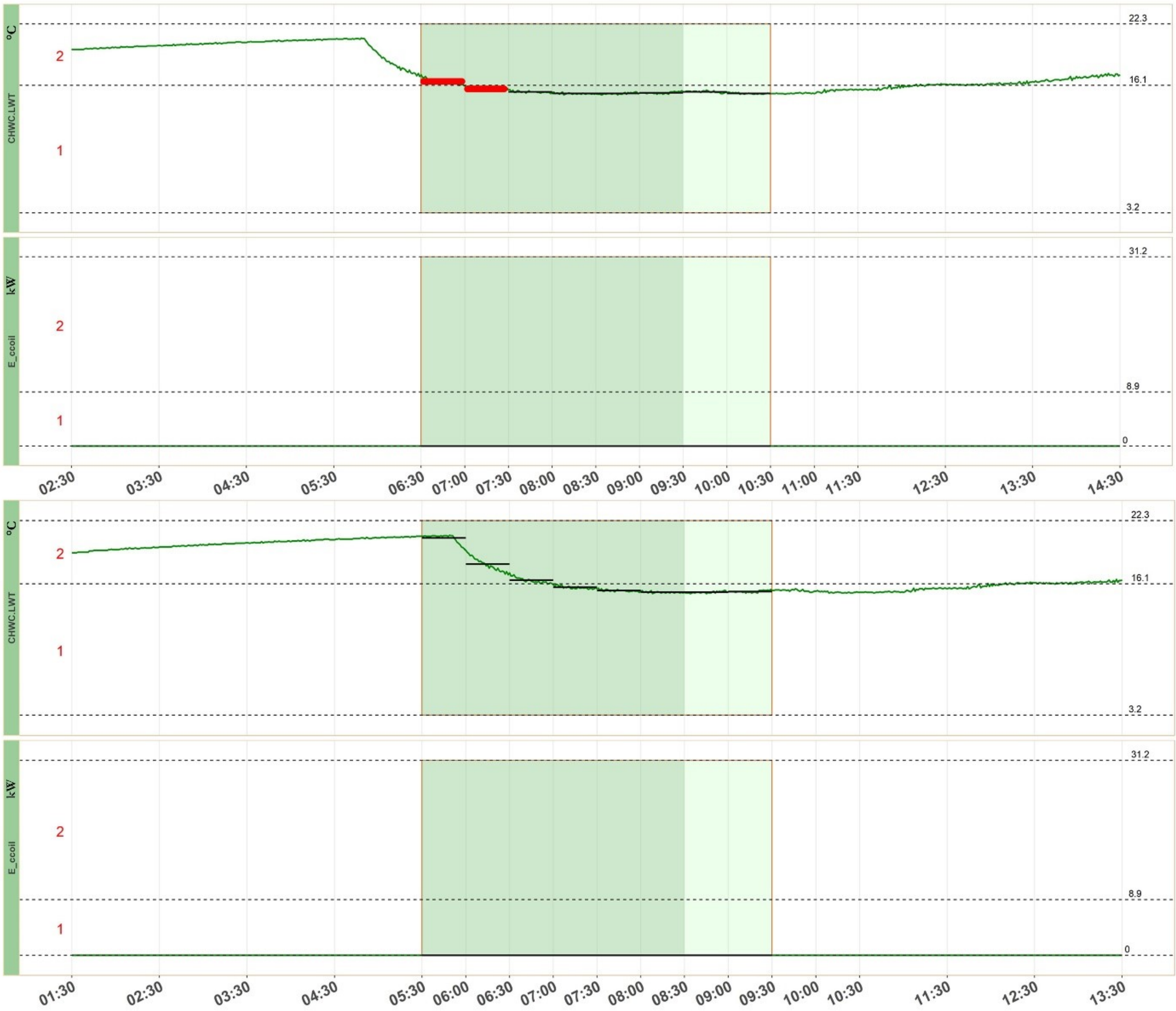


Figure 41 Rule 30-240-34(b)

It is worth noting that this rule has an aggregation of 30 minutes, hence, as can be seen in Figure 40, the PAA cannot following tightly the original profile, introducing an approximation to the analysis of the profiles. This problem is less evident for smaller aggregations since the original time series is approximated less.

On the other hand, the higher the aggregation, the less the rules are sensitive to minor fluctuations, therefore rules with higher aggregation stick more to the general trend of the variable, making them interesting as well.

This feature is particularly helpful in the cases of start-up or shut down, which the profiles are expected to register great variations.

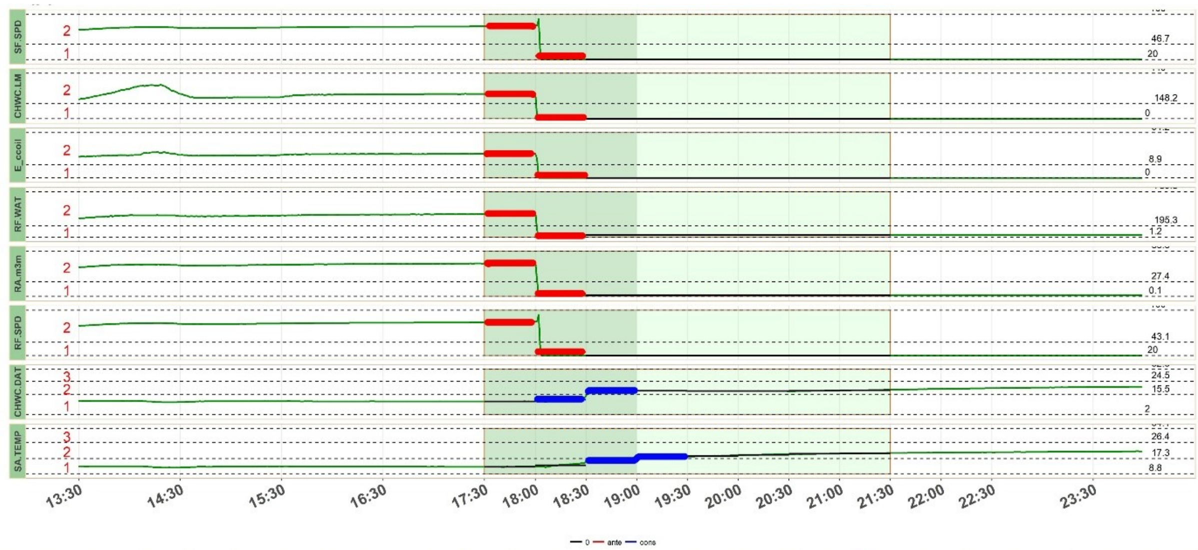
One of these cases is reported by the rule 30-240-20 (Table 16, Figure 42), which gives a reference for the behaviour of SF.SPD, CHWC.LM, E_ccoil, RF.WAT, RA.m3m, RF.SPD, CHWC.DAT and SA.TEMP at the shutdown of this system (see Appendix B for variables reference). This kind of rules is similar and specular to the specifical start-up type, since they can be used to test the system and make a roundup of the condition of the system at the end of the day.

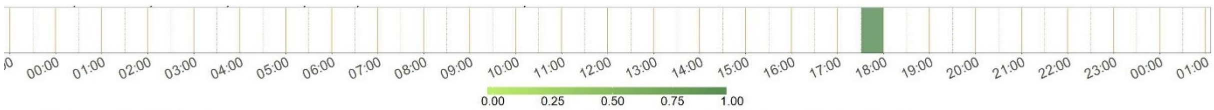
As the start-up ones, the shutdown rules are characterized by strong correlation, time specificity and narrow actual time window, as it is demonstrated by this rule.

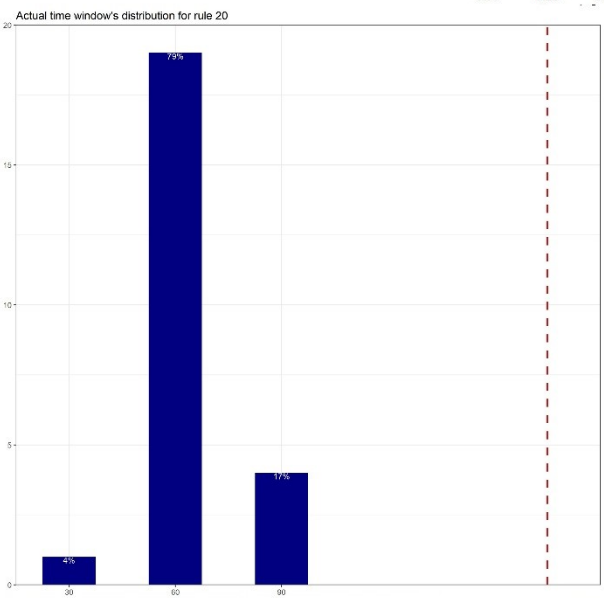
As the start-up rules, the higher the number of variables included in the pattern, the more detailed is a description of the system's condition, therefore rules with this large number of variables are good as a final checklist. This specific rule implies quantities related mainly to 3 components, which are the cooling coil, the supply fan and the return fan, therefore the violation of this rule may imply only a fault at these component level. (Appendix E)

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
30	**240**	**20**	102101, 142141, 172171, 112111, 7271, 122121	151152, 1112	0,11	0,63	0,86	0,89	90	0,83	0,83	17:30	17:30

Table 16 Rule 30-240-20







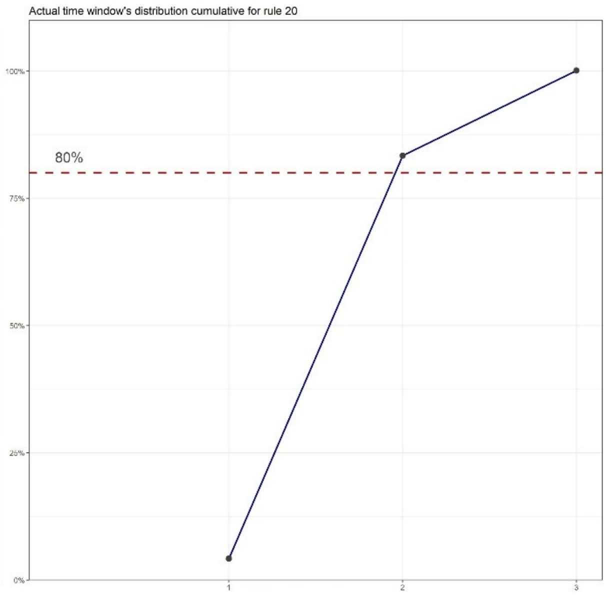


Figure 42 Rule 30-240-20 (a)

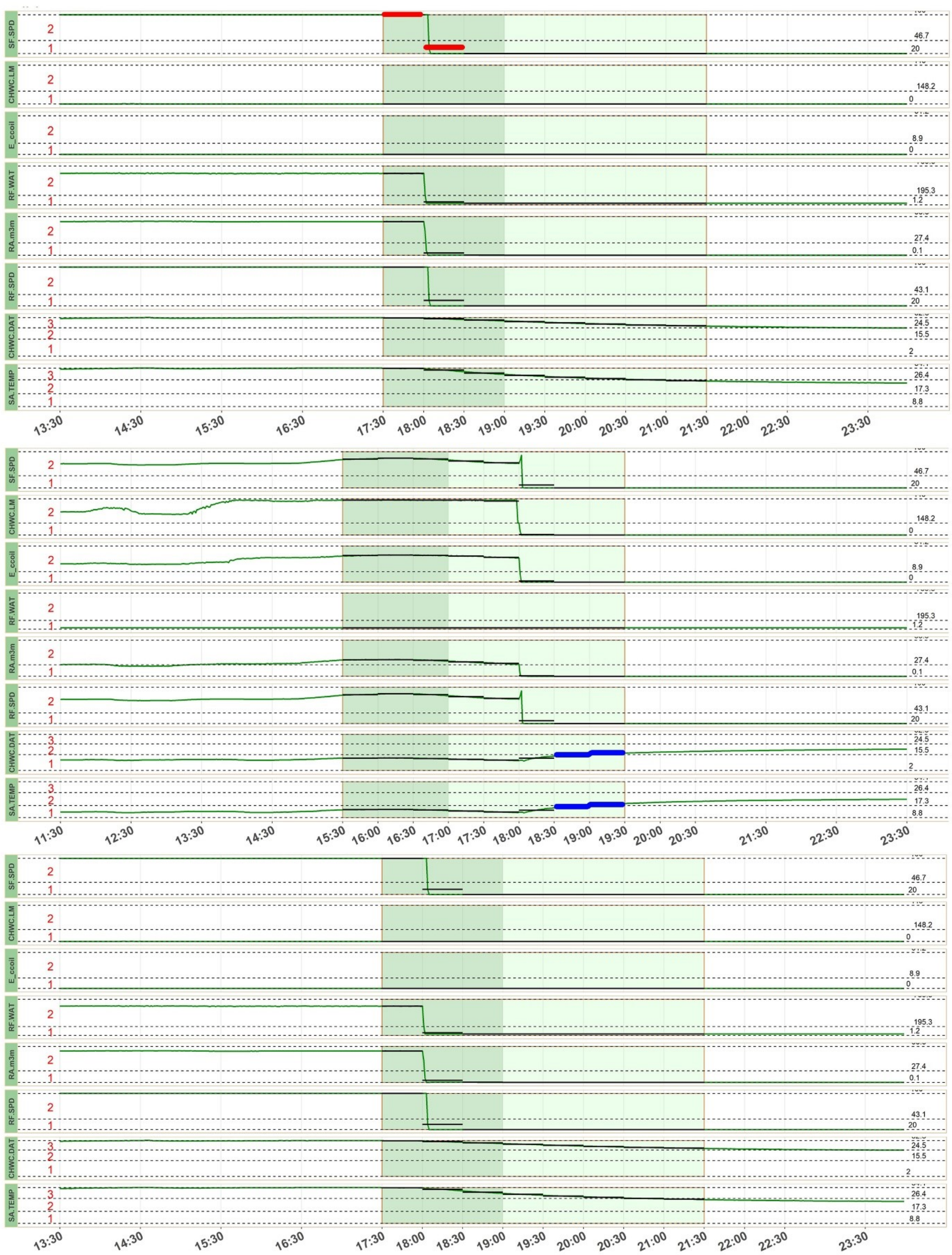


Figure 43 Rule 30-240-20 (b)

The last rule selected for this deeper insight is a off condition type, that is the rule 30-240-19 (Table 17, Figure 44). In fact, the methodology is applied on the whole dataset, without the segmentation of data on the basis of the running or not of the system, since common patterns can be found even when the system is turned off.

In particular, this rule expresses the strong correlation found between the SA.HUMD and CHWC.LWT (see Appendix B for variables reference). It is worth noting that the meaning of this quantities changes when the system is shut down, since the air is no more supplied and the water in the cooling coil is no more circulating, therefore these variables gives an indication of the temperature and air humidity in different points of the system free from the operational constraints.

In fact, it can be noticed from the profiles reported that the influence of the operation diminishes slowly and constantly, therefore the variables transitions' direction is not behaving unexpectedly unless some component of the system is not turned off as should be.

As a consequence, if all the system is turned off, the transitions identified by the patterns have to be intended as an indication of the level reached, not as actual transitions from a specific level to another, since the direction of the trend is known.

A demonstration of this concept, it can be considered the faults' implication of the shown rule.

In fact, the only implication of this rule regards the stuck of the cooling coil valve at fully closed position. (Appendix E, Figure 45**Errore. L'origine riferimento non è stata trovata.**)

If this fault occurs, the level of the CHWC.LWT is higher than expected at night time, while the SA.HUMD is lower, therefore the absence of the pattern, with the consequent violation of this rule, can diagnose the CCVSFC (see Appendix A for fault reference), fault even when the system is not running.

AGGR	Window	n	Antecedent	Consequent	Support	Confidence	Supp Conotool	Conf Conotool	Actual wind	SUPP DAY	MOD	H min	H max
30	**240**	**19**	2221	161162	0,11	0,57	0,7	0,82	180	0,93	0,93	18:00	20:00

Table 17 Rule 30-240-19

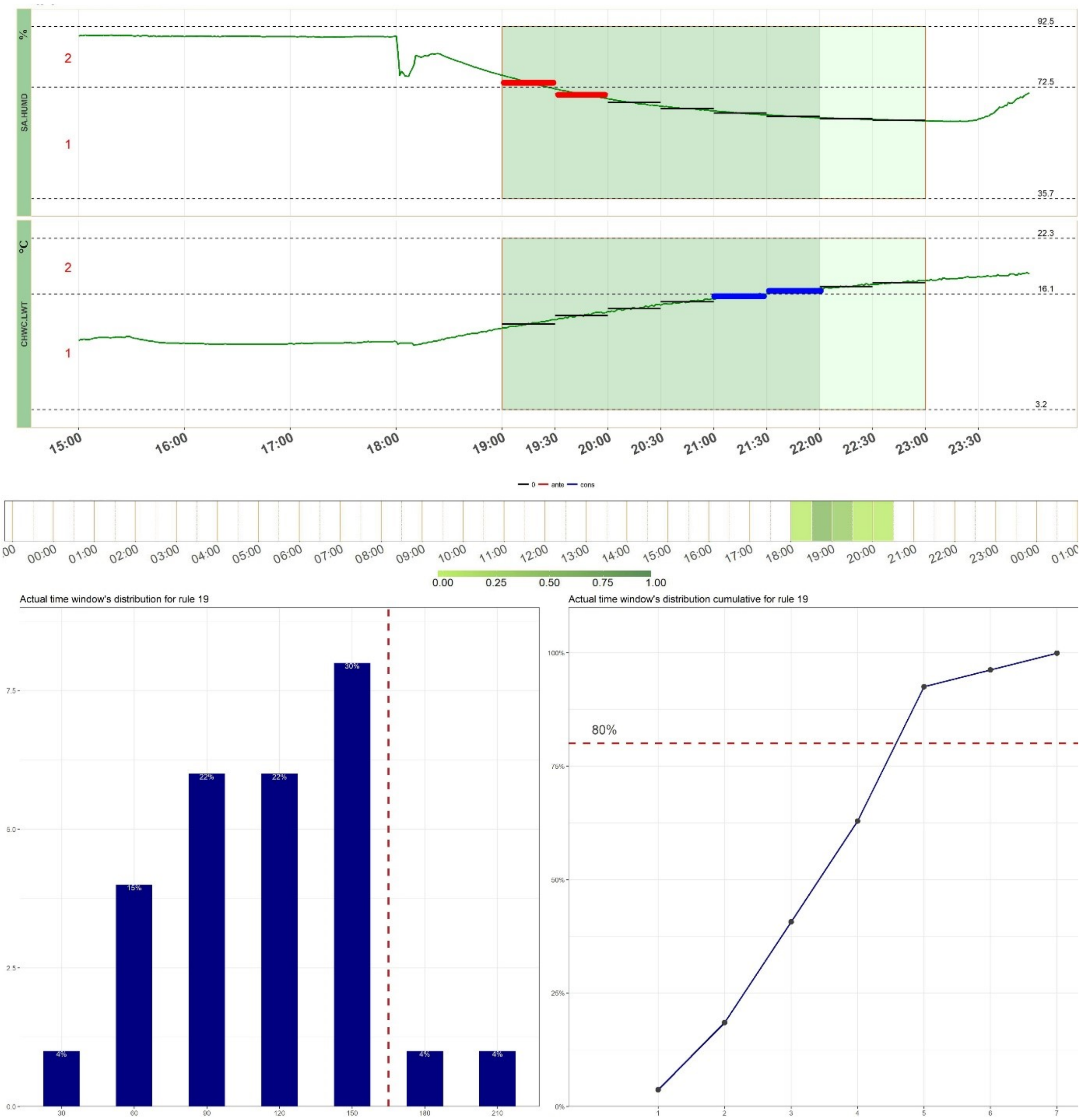


Figure 44 Rule 30-240-19 (a)

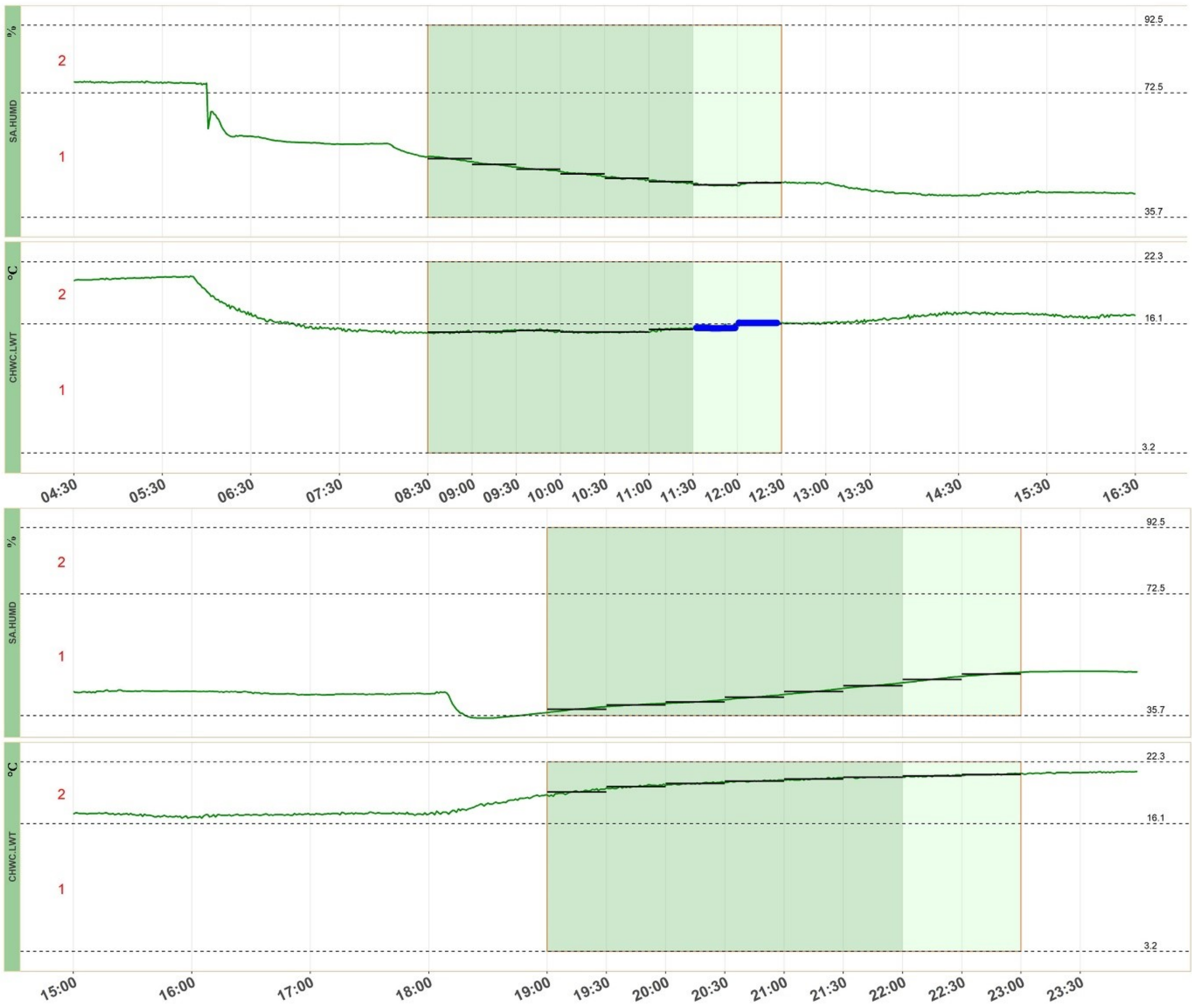


Figure 45 Rule 30-240-19 (b)

The rules described above are able to provide a set of rationales helpful for the identification of the malfunctioning of the system, by the interpretation of the rules taken singularly. This approach results effective since the rules found from the methodology are easily understandable and employable for the purpose.

6.2. Diagnosis results interpretation

In the previous section the results of the fault detection phase of the methodology have been presented and described, but they, in most cases, cannot provide a clear indication of the cause of the detected fault. Therefore, the diagnosis phase has been conducted starting from the fault detection phase output.

The possibility to perform the diagnosis is supported by the presence of tags of the fault running in the dataset used. In fact, each day is tagged as "Normal" or faulty with a code referring to the fault implemented in that day of operation. This particular feature of the dataset used to make it possible the performing of an inference procedure to search the occurrence of the rule along the whole dataset. In this way, the rules can be characterised in terms of presence of absence in the different faults day, so making it possible to intend the violation of the rule, i.e. the absence of the rule in a fault day, as a symptom for the presence of the fault. Furthermore, the inference analysis other than the presence or the absence of the rule in a day of fault, it is able to characterise the kind of violation is associated to the absence of the occurring of the rule. Therefore, in the days which the rule is not present, it has been searched if it is present at least one of the members of the pattern of the rule. Therefore, the violation of the rules has been tagged as the absence of the only consequent, the only antecedent or both. The representation of this characterisation for all the rules extracted is reported in the tables of Appendix E.

First of all, the rules extracted have been organized in a matrix reporting the correlation among faults and rules, sorting the faults with increasing diagnosability and the rules with increasing discrimination power grouping them according to the kind of implications.

This kind of matrix is represented colouring the cells referring to the violation of the rule in red, while its verification in green. In this way, the implications of the generic violation of the rule is graphically represented in a straightforward way. (Figure 46) On the top of the matrix is reported the dendrogram of the hierarchical clustering conducted to group the rules. (Figure 46) The grouping of the similar rules has been conducted by means of the hierarchical clustering with Single link method. After the clustering analysis, the matrix can be represented considered only one rule per cluster (cutting the dendrogram at height 0), in order to simplify the assessment of the matrix itself. In this way, every rule reported shows a unique kind of faults implication, being representative for its cluster.

Examples of these kind of matrix are reported in Figure 46 and Figure 47, showing the effect of the cut of the clustering dendrogram on the 15 minutes aggregation, while all the cases of the matrices, before and after the cut, are reported in Appendix F.

For the sake of performing a sort of sensitivity analysis of the diagnosis phase, only the kind of matrix reporting the whole range of rules has been taken into account. By doing that, a set of similar matrices, that can be considered as error signals matrices, has been constructed taking every time only the rules which have an implication for a specific fault. In this way, fault's specific matrices have been constructed by taking a fault by time, so these matrices have been used as input for the score evaluation step.

In order to test the capability of performing diagnosis of the methodology, the score calculation has been conducted exploring the rules even under the time specificity point of view. For this sake, the set of matrices has been divided so to construct, in the end, a set of fault-specific fault-rules matrices for "start-up", "operation" and "shut down".

By the assessment of the results of the score calculation, entirely reported in Appendix G, it is clear that the number of rules present in the set affects highly the capability to diagnose the fault considered. In Figure 48, they have reported the results of the score calculation for 3 moments of the functioning, with an aggregation of 5 minutes. Thanks to the high number of rules for this aggregation, these results are the best obtained, even though the faults are not always isolated.

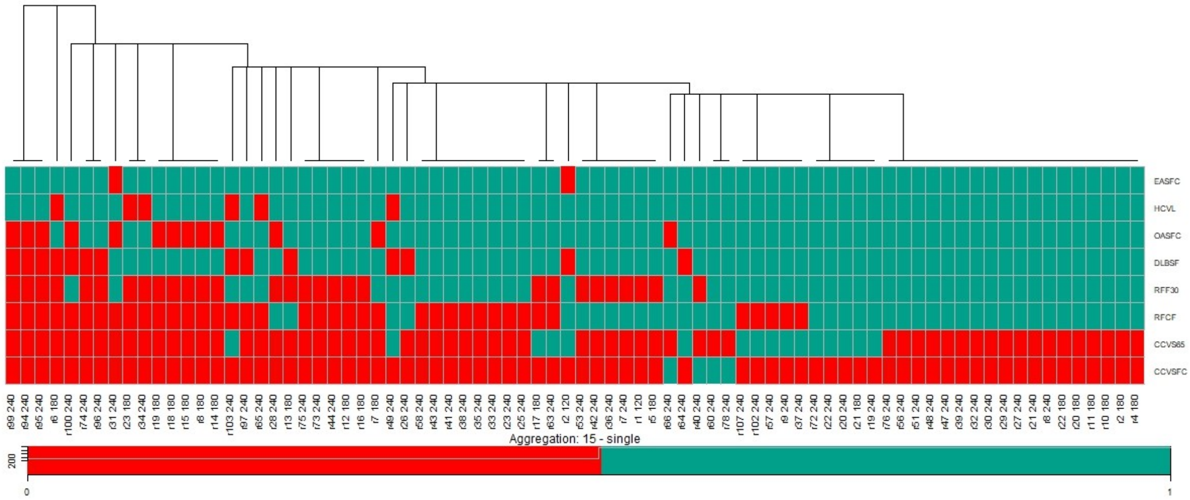


Figure 46 Faults-rules matrix for 15 minutes aggregation

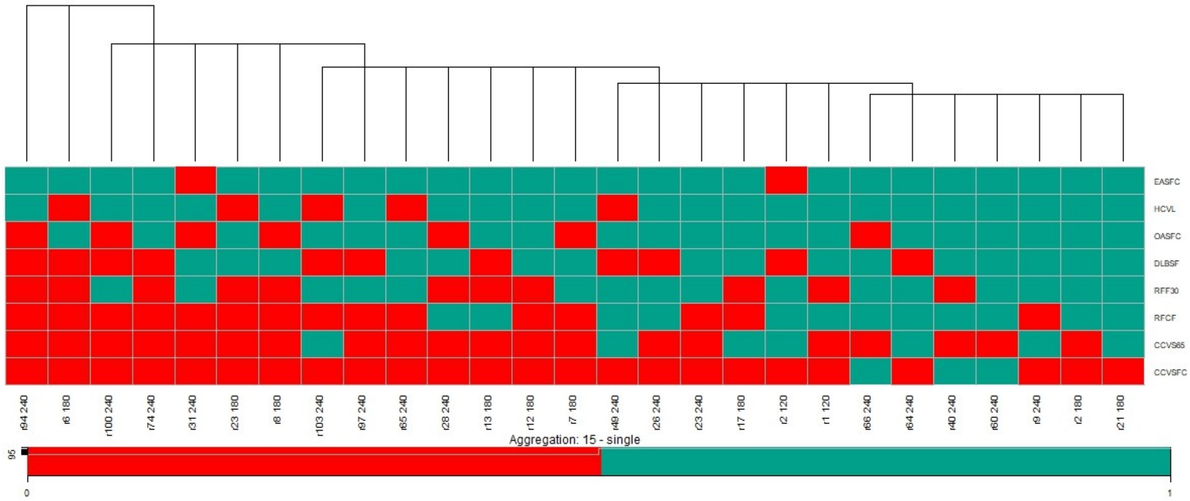
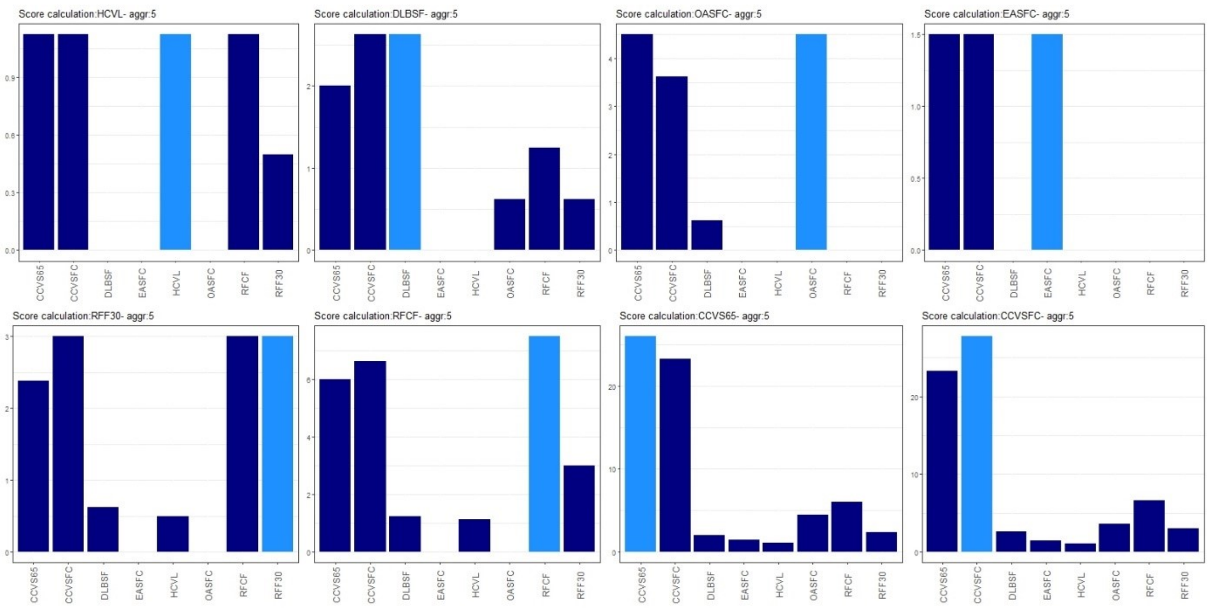


Figure 47 Faults-rules matrix for 15 minutes aggregation, cut at 0-height



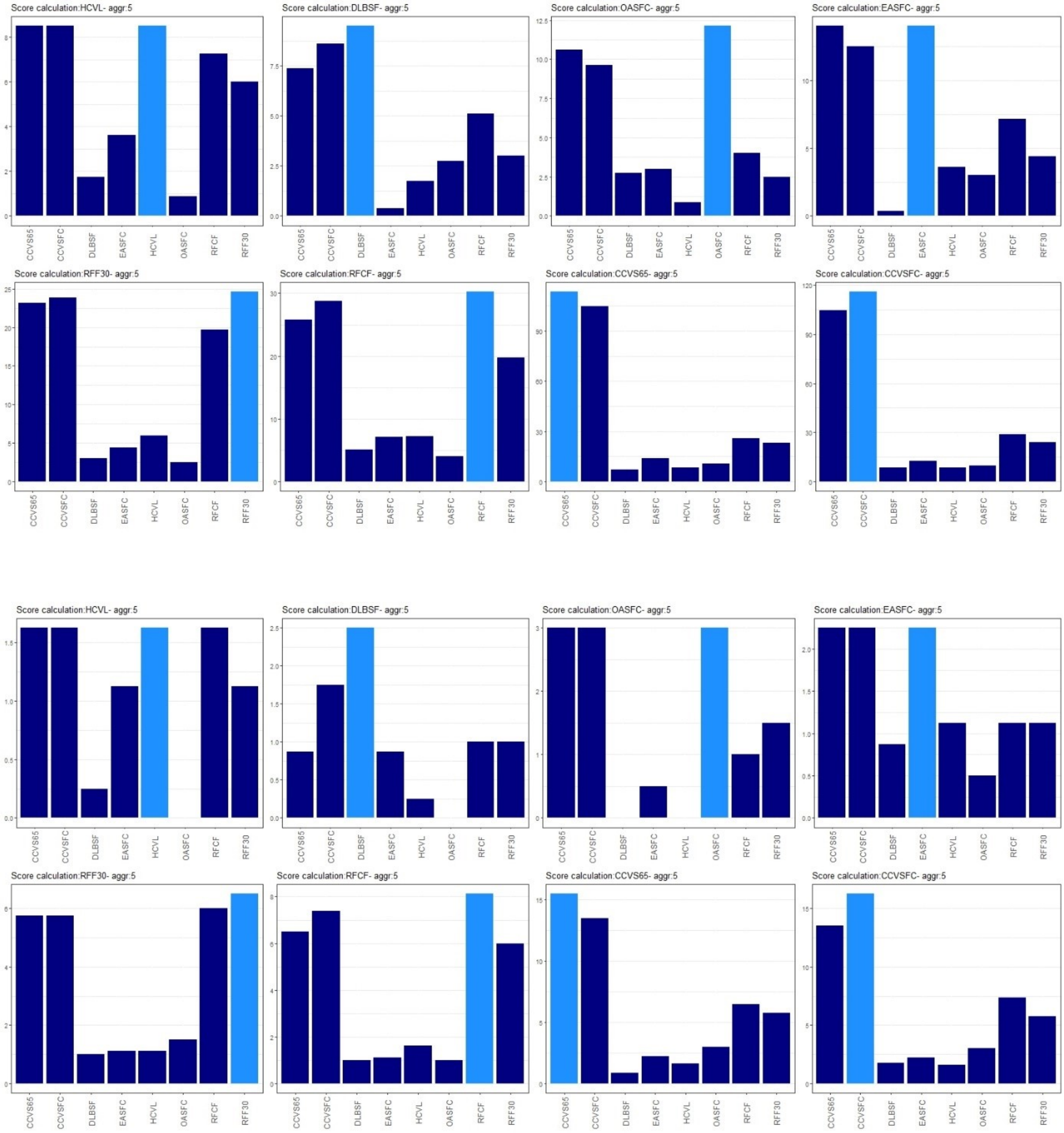


Figure 48 Score calculation results for start-up (top), operation (middle) and shut down (bottom) for 5 minutes aggregation

The results show that the isolation of cooling coil faults is generally possible, even though it is not always possible to discriminate easily which of the 2 faults considered is occurring.

The faults related to the return fan cannot be easily isolated excluding the other faults, since the score of cooling coil valve faults is quite high when a return fan fault is occurring, matter of fact the effect on the system of these faults can be considered similar. Dampers faults and duct leakage fault cannot be diagnosed unless the cooling coil valve fault is excluded priorly, matter of fact the scores associated to these faults are pretty high when dampers or duct leakage are tried to be diagnosed. The HCVL (see Appendix A for fault reference) fault has a poor statistical evidence in the dataset used, since the hot coil is not used in all the days considered, therefore the diagnosis of this fault is hard. The results with the other aggregation are similar in qualitative terms, but the diagnosis results get worse increasing the aggregation.

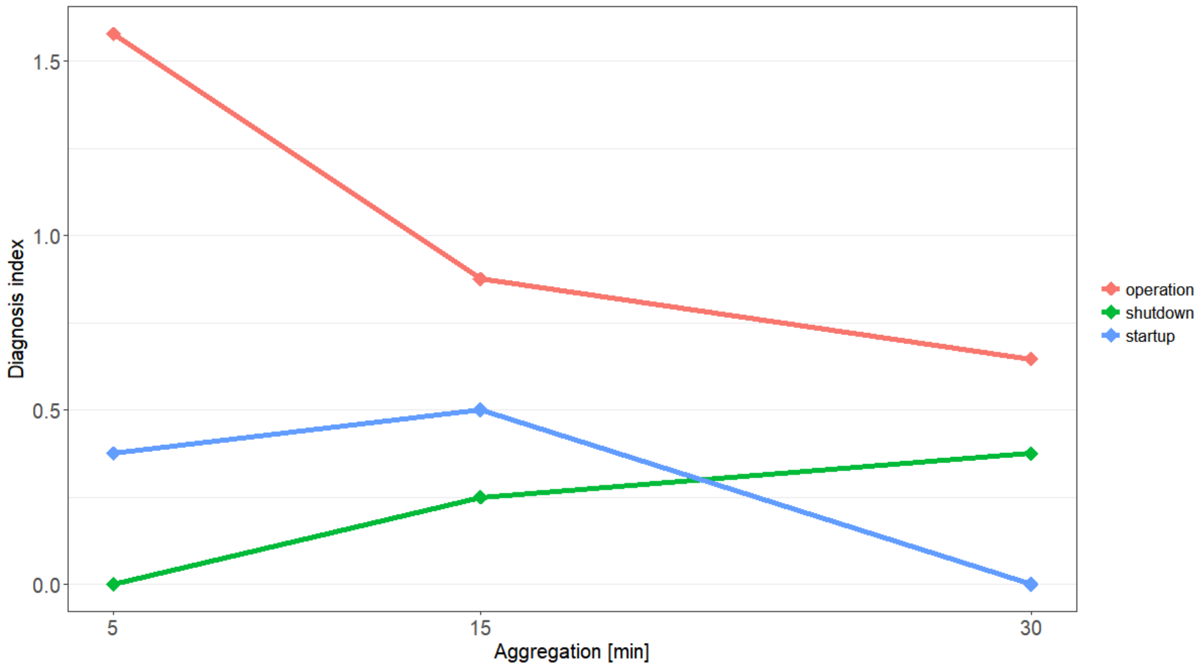


Figure 49 Isolation capability

In order to measure the capability to isolate the fault considered, it is computed the difference between the score associated to the fault analysed and the second highest score. The trend of the average of these differences is reported Figure 49, as diagnosis index for all the 3 moments of the functioning of the system. From the graph, it is clear the effectiveness of the diagnosis depends mainly on the number of rules present in the set. Even though the average value of the diagnosis index remains more or less constant, the values of the index for the different moments of functioning change in a different way, strongly affected by the number of the rules. In fact, the higher number of rules, the higher diagnosis index, probably because the effect of the noise generated by poorly discriminating rules diminishes when the number of rules used for analysis increases. As can be seen in Figure 49, increasing the aggregation, the values of the index for the 3 moments of functioning get closer, unless the number the calculation is completely biased by the low number of rules, as happened for the start-up functioning. In fact, the value of the Diagnosis index for the start-up moment with 30 minutes of aggregation is null, because the noise is too important with respect to the actual values of the faults to diagnose.

It should be pointed out that the difference in the trend of the Diagnosis index for the 3 moments of functioning can be attributable even to the relationship between the sensitivity to minor fluctuation of the different aggregations and the typical behaviour of the system came out from the profiles visualization. During the operation moment, the number of rules extracted for small aggregation is low because the minor fluctuations of the variable are caught almost as they are, therefore as the aggregation increases, the rules referring to small variations disappear. This kind of rule can be representative of the functioning of the system, even if the variation describing is not large and may not affect the whole set of system's variables. As a result, the loss of small variation rules determines a degradation of the description of the system's behaviour, thus justifying the decreasing trend in the Figure 49.

Regarding the shut down functioning, it may seem pointless, since it has a reversed trend with respect to the previously described moment, but this trend can be physically interpreted as well. At the shut down of the system, the changes are generally smooth and big, therefore this moment is represented well with large aggregations, because the minor fluctuations are not representative for this moment of functioning. In fact, approximating the minor fluctuation at this moment, makes that rules found more robust, therefore the number of rules identified for this moment of functioning increases with the aggregation.

Even though a similar observation can be done for the start-up situation, the typical functioning at this moment influenced the trend in such a way that it starts as the shutdown but drops to 0 for the largest aggregation. In fact, in exact moment when the system is turned on, generally, it is registered a peak in the variables' profile, which may disappear for the 30 minutes aggregation. This peak is the basis of the start-up rules extracted for small aggregation, but while it is no more recognized, the number of rules extracted for this moment of functioning drops dramatically, since otherwise it is not possible to characterise the actual functioning of the system.

7. Discussion

The results just described highlight how powerful are the implications expressed, making possible the exploitation of these rules as a reference for the normal condition of the system's functioning. In fact, the great potentiality of this approach is that in a completely unsupervised way it makes it possible to characterise the normal operation of the system, with the identification of patterns expressing even physical correlations, without any intervention of the user by the injection of prior expertise of the system in the implementation's parameters. In addition to that, it is worth noting that the characterisation of the rules according to their time window width results very helpful for the identification of different kinds of relationships, such as the fast response of rapid transients or the slow influence among the variables.

It is worth stressing the point that even though the rules have been extracted by means of an unsupervised method, the relationships found are completely meaningful in physical terms, demonstrating how effective is the adaptation of the method to the data which it is applied to. In fact, since the methodology does not accept any input parameters, making the procedure specific to the kind of system analysed, it is clear that the methodology can be potentially employed in any system obtaining meaningful results for any of them. This feature goes in favour of the generalization of the methodology since it is able to adapt itself to many systems and conditions.

Moreover, the exploration of different aggregations has driven to the idea that, generally, low aggregation implies a better description of the behaviour of the system, matter of fact it provides a higher number of rules which can embrace a higher range of situations. On the other hand, it should be considered that the level of fidelity in the following the original profile, in some circumstances is preferred to be less accurate, in order to reduce the influence of minor fluctuations. Furthermore, in the evaluation of the different aggregations, it should be considered the computational cost required since the 5 minutes aggregation provides better results but with a high time-consuming implementation. This feature may represent an obstacle for real-time implementation since the delay in the obtaining the results would be too high to exploit them.

The methodology has been demonstrated capable to detect the cause of the anomaly observed, providing an indication of single or multiple faults probably occurring in the system. Moreover, it should be taken into account that a diagnosis process constructed in this way is able to perform simultaneous faults identification as well. This feature is very important since the association of the symptoms observed to different faults it is not always clearly possible to perform. In fact, it is not always possible to separate the effects of multiple faults on a real system, so to analyse them independently. For this reason, many of the Fault Diagnosis procedures present in literature are incapable to identify more than a fault occurring contemporary in the system.

On the other hand, considering the results obtained for the diagnosis phase of the methodology it can be deduced that the isolation of the faults is not always possible, due to the lack of a significant distance between the fault intended to diagnose and the other faults. This issue is probably attributable to the lack of specificity of discrimination of some rules, which cannot distinguish clearly the effect of different faults on the basis of the violation of the patterns. In addition to that, another factor that increases the noise in the score calculation is the small number of rules related to the same moment of functioning. This point is evident by looking at high aggregation results since for some faults the calculation of the score has been impossible to perform, because the number of the rules was null, even though the signals error matrices were constructed without cutting the dendrogram of the rules' clustering.

Considering that, it can be possible to elect the 15 minutes aggregation as the best choice, since it provides a quite large number of rules for the characterization of the normal operation conditions, useful for the Fault Detection phase and then it provides Fault Diagnosis results not far from the 5 minutes aggregation, but with a much lower computational cost.

Despite the Fault Diagnosis results may be considered poor in absolute terms, they can be used as an indication for the construction of a checklist for the manual fault diagnosis. Basically, the checklist to provide to the system's operator should report the faults taken into account in the analysis sorted by the scores calculated. In this way, the faults are not isolated but the procedure provides support to the manual diagnosis, establishing the priority of components checking.

8. Conclusions

The methodology proposed fulfils completely the Automated Fault Detection and Diagnosis goal, since it is able to detect anomalies and diagnose the faults in a completely unsupervised way, thanks to its pattern recognition nature. The great advantages of the employment of this methodology lie on the automated approach which provides results relevant to the system's functioning and physical behaviour.

The encoding approach used by this methodology gives it great generalization power since the encoding of the transitions allows to use the same patterns for different scales of the same kind of system. In fact, the relationships found by the ARM can be considered more or less the same for a different size, by only changing the values of the boundaries of the variables' discretization.

The generalization of the results is attributable even to the abstraction of the representation of the results themselves, which may induce to think that they are a poor description of the system's operation. On the contrary, by their simplified description, the trends of the system's quantities are made clear and easily understandable even for non-expert users. Although the description of the trends provided by rules' patterns is simplified, the complete awareness of the system's functioning is not jeopardized, since the grade of detail of the patterns can be high in any case.

Thanks to the high quality of the results and the straightforward retrieving of merely applicable results, this methodology can be taken into account even for real-time implementation of fault detection and diagnosis. In this sense, it is worth considering the implementation cost on the basis of the kind of data used. In fact, too tight aggregation cannot be employed for on-line implementation, matter of fact that the delay in the obtaining the results would be too high to exploit them. On the other hand, larger aggregations can be employed for this purpose with great results.

This methodology can provide even high-level knowledge about the fault occurring, by providing a guess on the cause of the fault. In order to perform this phase, it is necessary to exploit a map of faults, since the correlation between faults and rules can be generated only if it is known the condition of the system in previously encountered faults' situations. This requirement is quite constraining since often it is not available such kind of information related to the data used, due to the lack information or tools automatically providing this kind of information.

Although the methodology can perform both fault detection and fault diagnosis, the development of the methodology can be pushed further, since there is still a margin of improving, especially considering the diagnosis results. In fact, a great improvement in this sense may be obtained by possibly implementing a more effective classification for the isolation of the faults. In order to improve the efficiency of the identification of the causes of the faults, future developments of the methodology may imply, for example, the use of Bayesian Belief Network. In this way, it would be possible to operate on a solid structure that may help in the characterisation of the relationship between the faults and the rules. A better characterisation may drive to better separation of the faults since the noise in the implications would be reduced.

Furthermore, the grade of detail of the violation of the rules can be enhanced more, by the exploitation of the violation mode of the rules. In fact, the characterisation of the way the rules are violated in the presence of the faults can be used to split the rules in different instances of the faults-rules matrix, so to take advantages from the possible specificity characterising some rule's violation. Therefore, by this operation, more importance would be given to the cases which a particular kind of violation is specific to a single fault.

Another way to enhance the isolation's effectiveness and increase the difference between the scores calculated could be the employment of the fuzzy logic. In fact, by the introduction of a less stringent criterion for the faults' implication of the rules' violation, the score's calculation may produce more separate results.

Finally, considering that the methodology proposed is thought with the aim of performing Fault Detection and Diagnosis in an automatic and unsupervised way, so to provide a straightforward support in the system's management and operation, it should be given a further effort in the refinement of the methodology in terms of computational efficiency. This operation is fundamental for the sake of its application to the real monitoring data, especially for a real-time implementation perspective.

9. Acknowledgement

Particular thanks are directed to ASHRAE for the permission given for the use of the data and documents from ASHRAE RP-1312.

10. References

Agrawal, R. & Srikant, R., 1994. Fast algorithms for mining association rules in large databases. *VLDB '94 Proceedings of the 20th International Conference on Very Large Databases.*

Arseniev, D., . Lyubimov, B. & Shkodyrev, V., 2009. Intelligent fault detection and diagnostics system on Rule-based neural network approach. *IEEE International Symposium on Intelligent Control.*

Bhuvaneswari, V. & Umajothy, P., 2013.
Mining Temporal Association Rules from Time Series Microarray Using Apriori Algorithm . *Review of Bioinformatics and Biometrics.*

Bynum, J., Claridge, D. & Curtin, J., 2012. Development and testing of an automated building commissioning analysis tool (ABCAT).. *Energy and Building.*

Capozzoli, A., Cerquitelli, T. & Piscitelli, M., 2016. Enhancing energy efficiency in buildings through innovative data analytics technologies. In: *Pervasive Computing.* s.l.:Ciprian Dobre and Fatos Xhafa .

Day, A., Jones, P. & Turton, J., 2013. Development of a UK centre for efficient and renewable energy in buildings (CEREB). *Renewable Energy.*

Dey, D. & Dong, B., 2016. A probabilistic approach to diagnose faults of air handling units in buildings.. *Energy and Buildings.*

DOE, 2016. *Building Energy Data Book.* [Online]
Available at: http://buildingsdatabook.eren.doe.gov/

Fan, C. & Xiao, F., 2016. *Assessment of building operational performance using data mining techniques: a case study.* s.l., s.n.

Fan, C. & Xiao, F., 2017. Mining big building operational data for improving building energy efficiency: A case study. *Building services engineering research and technology.*

Fan, C., Xiao, F., Madsen, H. & Wang, D., 2015. Temporal knowledge discovery in big BAS data for buildingenergy management. *Energy and Buildings.*

Fan, C., Xiao, F. & Yan, C., 2015. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction.*

Fontugne, R. et al., 2013. *Strip,bind,andsearch:A method for identifying abnormal energy consumption in buildings..* s.l., s.n.

Fournier-Viger, P. et al., 2015. Mining partially-ordered sequential rules common to multiple sequences. *IEEE Transactions on Knowledge and Data Engineering.*

Fournier-Viger, P., Wu, C., Tseng, V. & Nkambou3, R., 2012. *Mining Sequential Rules Common to Several Sequences with the Window Size Constraint.* s.l., s.n.

Han, J., Kamber, M. & Pei, J., 2011. In: *Data mining:concepts and techniques.* s.l.:Morgan Kaufmann.

Han, J., Pei, J., Yin, Y. & Mao, R., 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery.*

HVACSWGSpin12007, 2011. *HVACSWGSpin12007.* [Online]
Available at: http://www.seai.ie/Your Business/Large Energy Users/Special Initiatives/Special Working Groups/HVAC SWG 07/

Kang, C. & Golay, M., 1999. A Bayesian belief network- based advisory system for operational availability focused diagnosis of complex nuclear power systems. *Expert Systems and Applications.*

Katipamula, S. & Brambley, M., 2005. Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems— A Review, Part I-II. *HVAC&R RESEARCH.*

Katipamula, S. & Kim, W., 2017. A review of fault detection and diagnostics methods for building systems.. *Science and Technology for the Built Environment..*

Kaur, G., 2014. (IJCSIT) International Journal of Computer Science and Information Technologies.

Kocyigit, N., 2015. Fault and sensor error diagnostic strategies for a vapor compression refrigeration system by using fuzzy inference systems and artificial neural network.. *International journal of refrigeration.*

Liang, J. & Du, R., 2007. Model-based fault detection and diagnosis of HVAC systems using Support Vector Machine method.. *International journal of refrigeration.*

Li, G. et al., 2017. Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. *Applied Energy.*

Li, H. & Braun, J., 2007. A methodology for diagnosing multiple simultaneous faults in vapor-compression air conditioners.. *HVAC&R Research.*

Li, H. & Lee, S., 2009. Mining frequent itemsets over data steams using efficient window slideing techniques. *Expert Systems with Applications.*

Lin, G. & Claridge, D., 2015. A temperature-based approach to detect abnormal building energy consumption.. *Energy and Buildings.*

Lin, J., Keogh, J. E., Lonardi, S. & Chiu, B., 2003. *A Symbolic Representation of Time Series, with Implications for Streaming Algorithms.* s.l., s.n.

Martínez-de-Pisón Ascacíbar, F. et al., 2009. *System for uncovering hidden knowledge in real time for the analysis of environmental and agricultural processes.* Badajoz, s.n.

Martínez-de-Pisón Ascacíbar, F. et al., 2012. Mining association rules from time series to explain failures in a hot-dip galvanizing steel line. *Computers & Industrial Engineering.*

Miller, C., Nagyb, Z. & Schlueter, A., 2018. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews.*

Pérez-Lombard, 2008. A review buildings energy consumption information. *Energy Building.*

Pham, N., Le, Q. & Dang, T., 2010. *HOT aSAX: A Novel Adaptive Symbolic Representation for Time Series Discords Discovery.* s.l., s.n.

Qin, J. & Wang, S., 2005. A fault detection and diagnosis strategy of VAV air-conditioning systems for improved energy and control performances.. *Energy and Buildings.*

Tan, P., Steinbach, M. & Kumar, V., 2005. In: *Introduction to Data Mining.* s.l.:Pearson.

U.S. EIA, E. I. A., 2013. *Annual Energy Outlook 2013,* s.l.: s.n.

Wang, L. et al., 2013. Monitoring-based HVAC commissioning of an existing office building for energy efficiency.. *Applied Energy.*

Wang, S. & Cui, J., 2006. A robust fault detection and diagnosis strategy for centrifugal chillers.. *HVAC&R Research* .

Wen, J. & Li, S., 2011. *ASHRAE 1312-RP Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units.,* s.l.: s.n.

Wu, S. & Sun, J., 2011. Cross-level fault detection and diagnosis of building HVAC systems.. *Building and environment.*

Xue, P. et al., 2017. Fault detection and operation optimization in district heating substations based on data mining techniques.. *Applied Energy.*

Yang, X., Jin, X., Du, Z. & Zhu, Y., 2011. A novel model-based fault detection method for temperature sensor using fractal correlation dimension.. *Building and Environment.*

Yan, R., Ma, Z., Zhao, Y. & Kokogiannakis, G., 2016. A decision tree-based data-driven diagnostic strategy for Air Handling Unit. *Energy and Buildings.*

Yu, Y., Woradechjumroen, D. & Yu, D., 2014. A review of fault detection and diagnosis methodologies on air-handling units.. *Energy and Building.*

Yu, Z., Haghighat, F., Fung, B. & Zhou, L., 2012. A novel methodology for knowledge discovery through mining associations between building operational data.. *Energy and Buildings.*

Zaki, M., 2001. SPADE:an efficient algorithm for mining frequent sequences. *Machine Learning.*

Zhai, L., Tang, X., Li, L. & W., J., 2018. Temporal association rule mining based on t-apriori algorithm and its typical application.

Zhao, X., Yang, M. & Li., H., 2014. Field implementation and evaluation of a decoupling-based fault detection and diagnostic method for chillers.. *Energy and Buildings.*

Zhao, Y., Wen, J. & Wang, S., 2015. Diagnostic Bayesian networks for diagnosing air handling units faults – Part II: Faults in coils and sensors. *Applied Thermal Engineering.*

Zhao, Y. et al., 2017. Diagnostic Bayesian networks for diagnosing air handling units faults – part I: Faults in dampers, fans, filters and sensors. *Applied Thermal Engineering.*

11. Glossary

ABCAT	Automated Building Commissioning Analysis Tool
AFDD	Automated Fault detection and Diagnosis
AHU	Air Handling Unit
APAR	Air handling unit Performance Assessment Rules
ARM	Association Rules Mining
aSAX	adaptive Symbolic Aggregation Approximation
BBN	Bayesian Belief Network
BMS	Building Management System
CART	Classification and regression tree
CCVS65	Cooling coil valve stuck at 65%
CCVSFC	Cooling coil valve stuck fully closed
CHWC.DAT	Cooling coil discharge air temperature
CHWC.LM	Cooling coil water flow rate
CHWC.LWT	Cooling coil water outlet temperature
CHWC.VLV	Cooling coil valve opening
DLBSF	Duct leakage before supply fan
DM	Data Mining
E_ccoil	Cooling coil power
E_hcoil	Heating coil power
EASFC	Exhaust air damper stuck fully closed
EMD	Empirical Mode Decomposition
ERS	Energy Resource Station
FDD	Fault Detection and Diagnosis
FP-Growth	Frequent Pattern- Growth
FP-tree	Frequent Pattern tree
HCVL	Heating coil valve leakage
HWC.DAT	Heating coil discharge air temperature
HWC.LM	Heating coil water flow rate
HWC.VLV	Heating coil valve opening
ICT	Information and Communication Technology
IMF	Intrinsic Mode Function
KDD	Knowledge Discovery in Databases
LHS	Left-hand Side
MA.TEMP	Mixed air temperature
MFI-TransW	Mining Frequent Item sets within a Transaction-sensitive Sliding Window
NIST	National Institute of Standards and Technology
OA.TEMP	Outdoor air temperature
OASFC	Outdoor air damper stuck fully closed
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis

RA.HUMD	Return air humidity
RA.m3m	Return air flow rate
RA.TEMP	Return air temperature
RF.SPD	Return fan speed
RF.WAT	Return fan power
RFCF	Return fan complete failure
RFF30	Return fan at fixed speed (30%)
RHS	Right-hand Side
SA.HUMD	Supply air humidity
SA.m3m	Supply air flow rate
SA.TEMP	Supply air temperature
SAX	Symbolic Aggregation Approximation
SF.SPD	Supply fan speed
SF.WAT	Supply fan power
SVM	Support Vector Machine
TARM	Temporal Association Rules Mining
VAV	Variable Air Volume

12. List of figures

Figure 1 General Data Mining framework 3
Figure 2 Classification of energy applications-related BMS data (Capozzoli, et al., 2016) 5
Figure 3 Classification of AFDD methods (Katipamula & Brambley, 2005) 9
Figure 4 Sliding window schematic 25
Figure 5 Apriori pruning from infrequent itemset (Tan, et al., 2005) 26
Figure 6 Construction of FP-tree (Tan, et al., 2005) 27
Figure 7 Application of FP-growth algorithm to find frequent item sets ending in “e” (Tan, et al., 2005) 28
Figure 8 Example of SAX application with Word=8 and Alphabet size=3 (Lin, et al., 2003) 34
Figure 9 FDD method main characteristics 37
Figure 10 AFDD methods for AHU classification (Yu, et al., 2014) 39
Figure 11 Framework of the Fault Detection and Diagnosis methodology 55
Figure 12 Dataset of discrete encoded variables 57
Figure 13 Transitions’ time series 58
Figure 14 Inter-transactional item sets database creation 59
Figure 15 Rule-fault matrix with violation characterisation 62
Figure 16 Fault-rules matrix of violation 63
Figure 17 Score visualization for fault diagnosis 65
Figure 18 Decision-tree obtained with Ma et al. approach 66
Figure 19 ERS facility (Wen & Li, 2011) 70
Figure 20 AHU- A and B schematic (Wen & Li, 2011) 70
Figure 21 Match test of AHU-A and B for summer, winter and spring cases (Wen & Li, 2011) .. 74
Figure 22 Cooling coil location schematic (Wen & Li, 2011) 79
Figure 23 Three-way valve functioning 80
Figure 24 Profiles of cooling coil related variables in cooling coil valve stuck at fully open case (above) and fully closed case (bottom) 80
Figure 25 Example of data cleaning by means of Hampel filter 83
Figure 26 Variables distribution and discretization’s breakpoints location 85
Figure 27 Rules’ profile format 90
Figure 28 Rule 5-240-236 (a) 93
Figure 29 Rule 5-240-236 (b) 94
Figure 30 Rule 5-240-62 (a) 96
Figure 31 Rule 5-240-62 (b) 97
Figure 32 Rule 5-180-9 (a) 100

Figure 33 Rule 5-180-9 (b) 100
Figure 34 Rule 5-120-2 (a)....................101
Figure 35 Rule 5-120-2 (b) 102
Figure 36 Rule 15-240-102 (a) 103
Figure 37 Rule 15-240-102 (b)....................104
Figure 38 Rule 15-240-19 (a) 105
Figure 39 Rule 15-240-19 (b)....................106
Figure 40 Rule 30-240-34(a)....................107
Figure 41 Rule 30-240-34(b) 108
Figure 42 Rule 30-240-20 (a) 110
Figure 43 Rule 30-240-20 (b)....................110
Figure 44 Rule 30-240-19 (a) 112
Figure 45 Rule 30-240-19 (b)....................113
Figure 46 Faults-rules matrix for 15 minutes aggregation....................115
Figure 47 Faults-rules matrix for 15 minutes aggregation, cut at 0-height 115
Figure 48 Score calculation results for start-up (top), operation (middle) and shut down (bottom) for 5 minutes aggregation....................116
Figure 49 Isolation capability....................117

Appendix

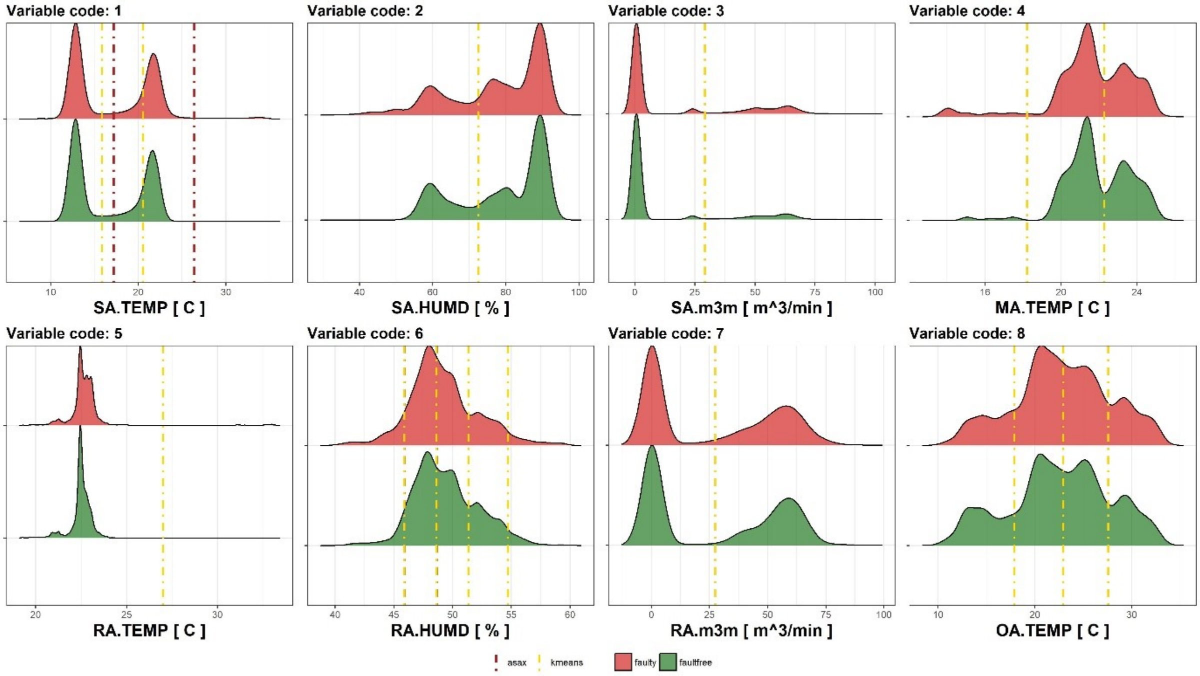
A. Faults coding

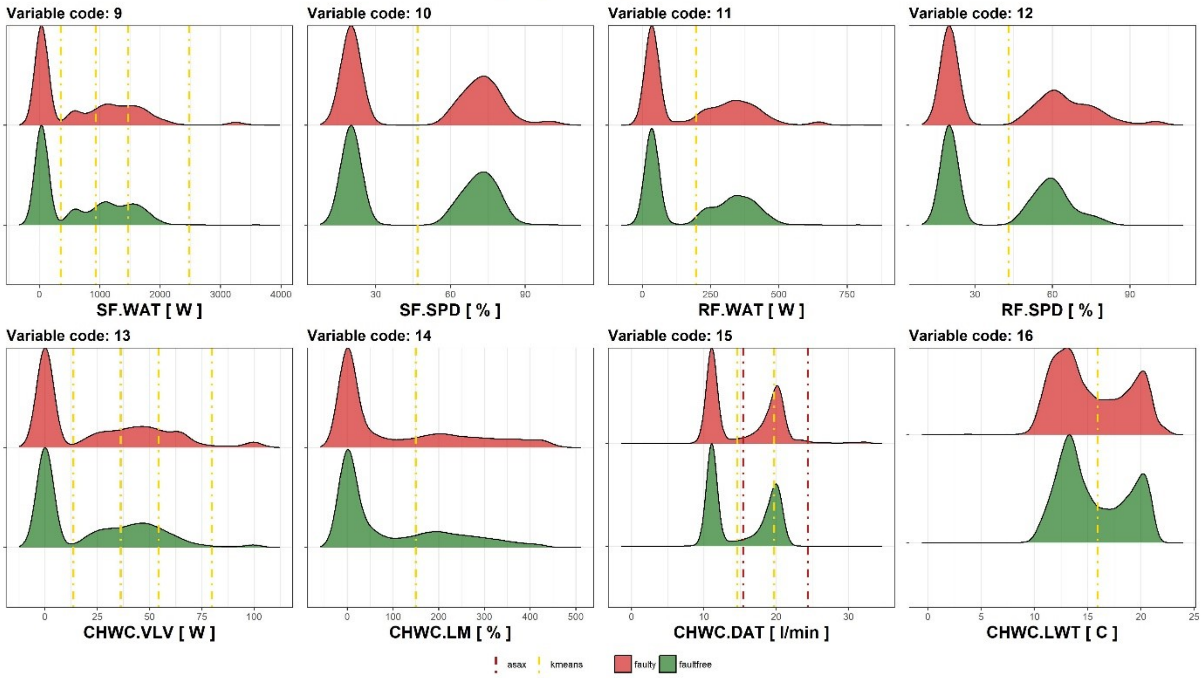
FAULTS ACRONYMS	DESCRIPTION
CCVSFC	Cooling coil valve stuck fully closed
CCVS65	Cooling coil valve stuck at 65%
DLBSF	Duct leakage before supply fan
EASFC	Exhaust air damper stuck fully closed
OASFC	Outdoor air damper stuck fully closed
HCVL	Heating coil valve leakage
RFCF	Return fan complete failure
RFF30	Return fan at fixed speed (30%)

B. Variables coding

TAG	DESCRIPTION	CODE	UNIT (ORIGINAL)	UNIT (CONVERTED)
SA.TEMP	Supply air temperature	1	F	C
SA.HUMD	Supply air humidity	2	%	
SA.M3M	Supply air flow rate	3	CFM	m3/m
MA.TEMP	Mixed air temperature	4	F	C
RA.TEMP	Return air temperature	5	F	C
RA.HUMD	Return air humidity	6	%	
RA.M3M	Return air flow rate	7	CFM	m3/m
OA.TEMP	Outdoor air temperature	8	F	C
SF.WAT	Supply fan power	9	W	
SF.SPD	Supply fan speed	10	%	
RF.WAT	Return fan power	11	W	
RF.SPD	Return fan speed	12	%	
CHWC.VLV	Cooling coil valve opening	13	%	
CHWC.LM	Cooling coil water flow rate	14	GPM	l/m
CHWC.DAT	Cooling coil discharge air temperature	15	F	C
CHWC.LWT	Cooling coil water outlet temperature	16	F	C
E_CCOIL	Cooling coil power	17	KW	
HWC.VLV	Heating coil valve opening	18	%	
HWC.LM	Heating coil water flow rate	19	GPM	l/m
HWC.DAT	Heating coil discharge air temperature	20	F	C
E_HCOIL	Heating coil power	21	KW	

C. Distribution and discretization


Variable code: 1
SA.TEMP [C]
Variable code: 2
SA.HUMD [%]
Variable code: 3
SA.m3m [m^3/min]
Variable code: 4
MA.TEMP [C]
Variable code: 5
RA.TEMP [C]
Variable code: 6
RA.HUMD [%]
Variable code: 7
RA.m3m [m^3/min]
Variable code: 8
OA.TEMP [C]
asax
kmeans
faulty
faultfree


Variable code: 9
SF.WAT [W]
Variable code: 10
SF.SPD [%]
Variable code: 11
RF.WAT [W]
Variable code: 12
RF.SPD [%]
Variable code: 13
CHWC.VLV [W]
Variable code: 14
CHWC.LM [%]
Variable code: 15
CHWC.DAT [l/min]
Variable code: 16
CHWC.LWT [C]
asax
kmeans
faulty
faultfree

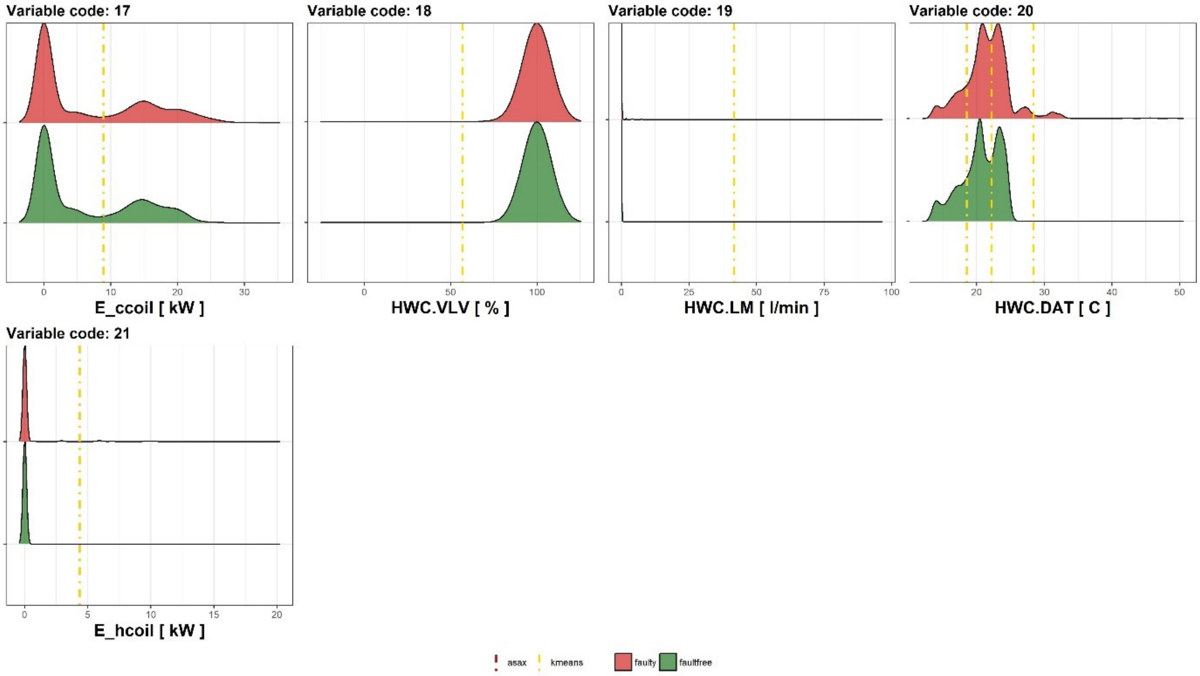
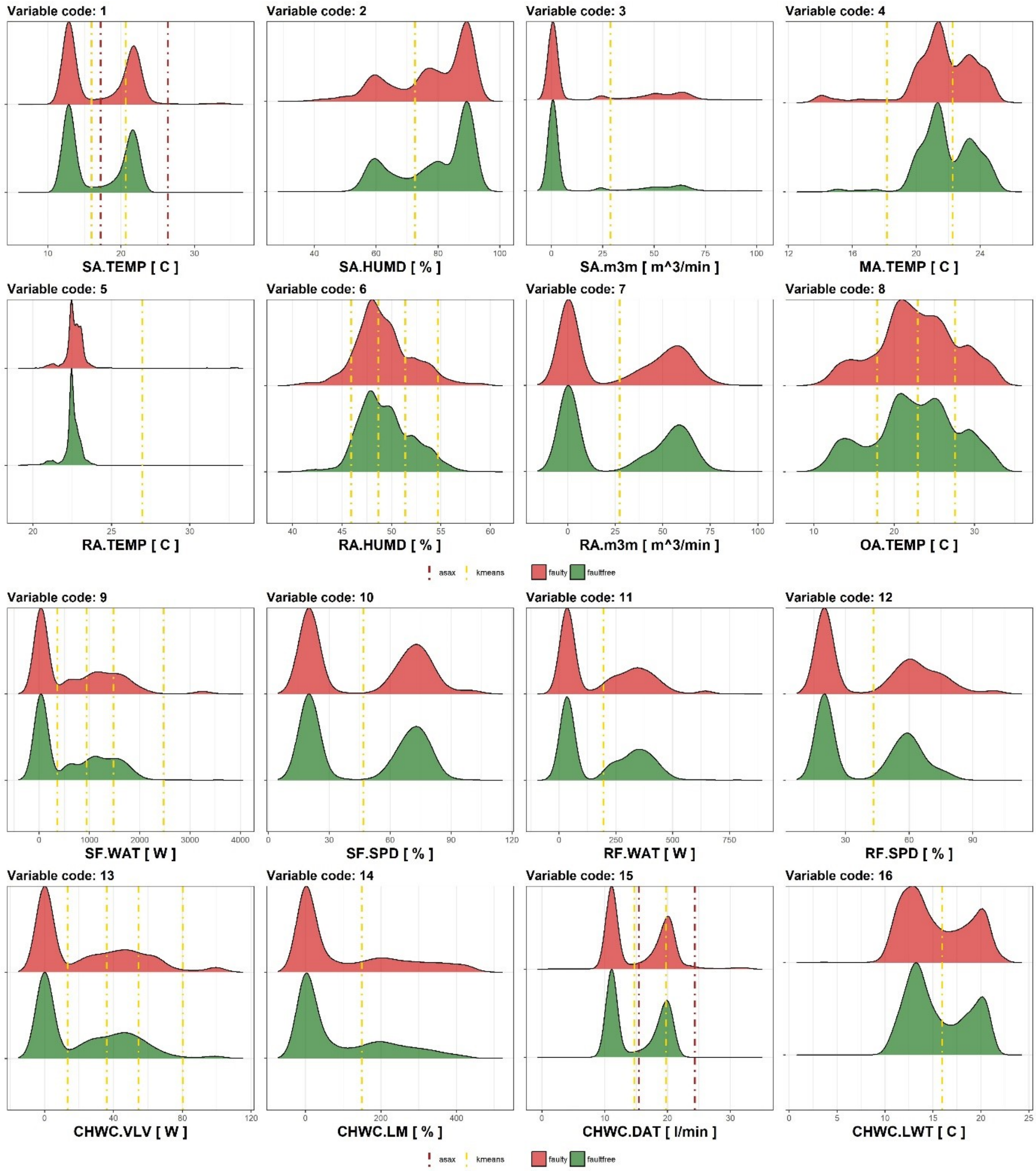


FIG. 11 Distribution and discretization with aggregation at 5 minutes



Variable code: 1
SA.TEMP [C]
Variable code: 2
SA.HUMD [%]
Variable code: 3
SA.m3m [m^3/min]
Variable code: 4
MA.TEMP [C]
Variable code: 5
RA.TEMP [C]
Variable code: 6
RA.HUMD [%]
Variable code: 7
RA.m3m [m^3/min]
Variable code: 8
OA.TEMP [C]
asax
kmeans
faulty
faultfree
Variable code: 9
SF.WAT [W]
Variable code: 10
SF.SPD [%]
Variable code: 11
RF.WAT [W]
Variable code: 12
RF.SPD [%]
Variable code: 13
CHWC.VLV [W]
Variable code: 14
CHWC.LM [%]
Variable code: 15
CHWC.DAT [l/min]
Variable code: 16
CHWC.LWT [C]
asax
kmeans
faulty
faultfree

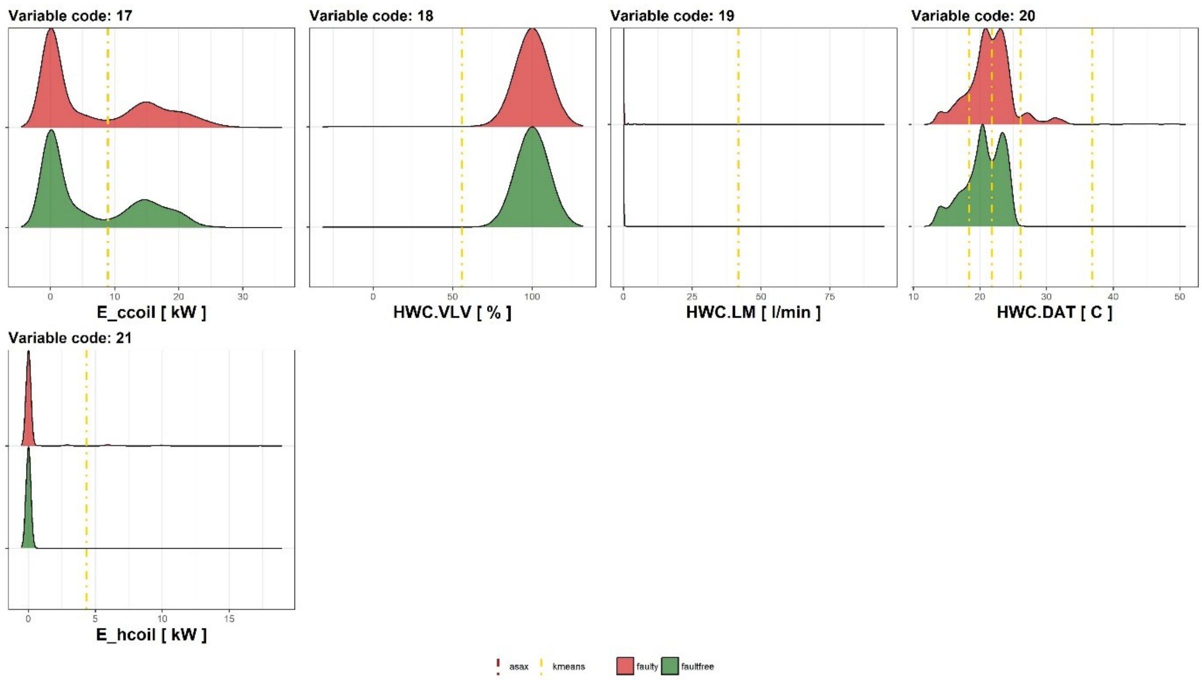
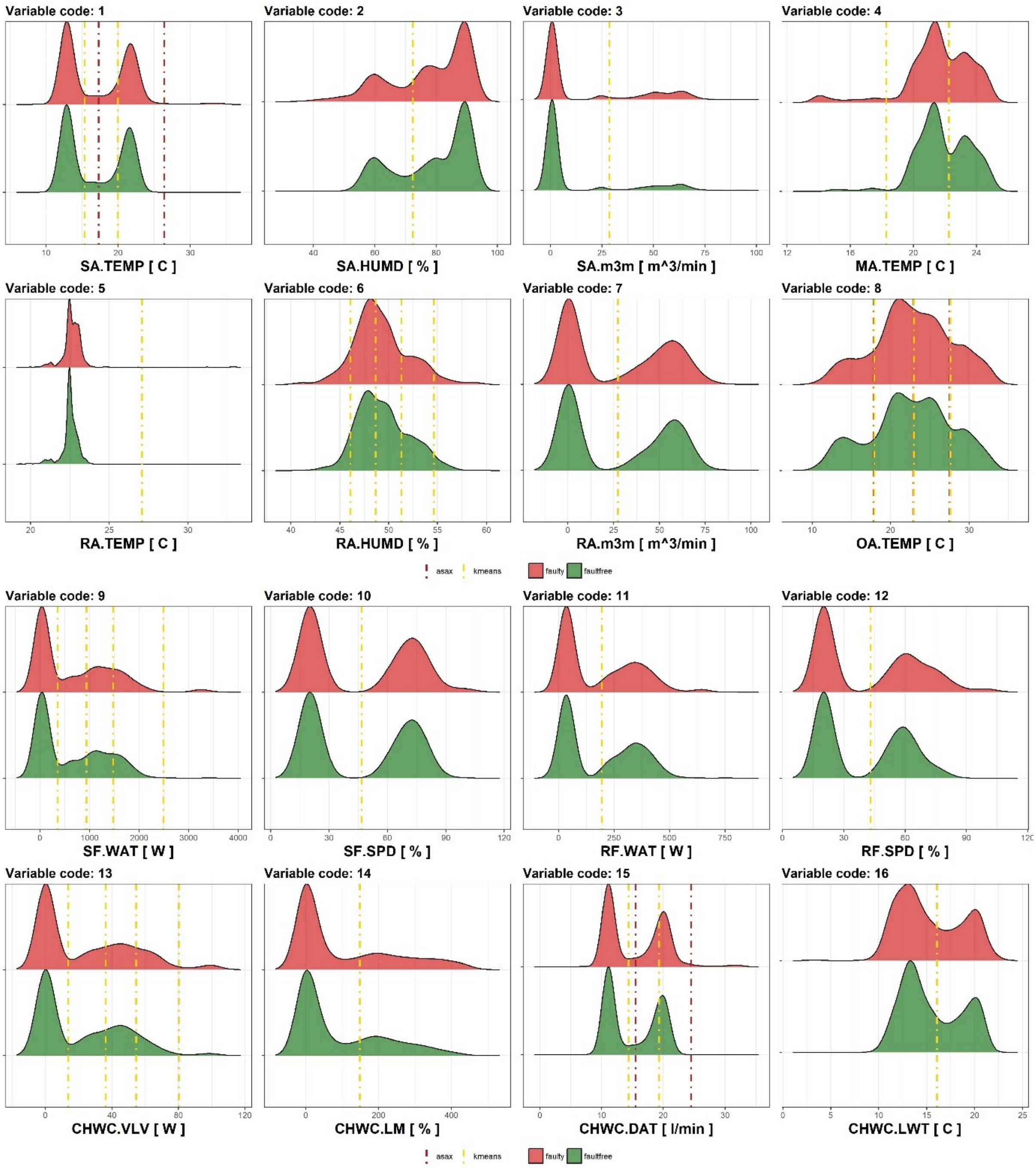


FIG. 2 Distribution and discretization with aggregation at 15 minutes


Variable code: 1
SA.TEMP [C]
Variable code: 2
SA.HUMD [%]
Variable code: 3
SA.m3m [m^3/min]
Variable code: 4
MA.TEMP [C]
Variable code: 5
RA.TEMP [C]
Variable code: 6
RA.HUMD [%]
Variable code: 7
RA.m3m [m^3/min]
Variable code: 8
OA.TEMP [C]
asax
kmeans
faulty
faultfree
Variable code: 9
SF.WAT [W]
Variable code: 10
SF.SPD [%]
Variable code: 11
RF.WAT [W]
Variable code: 12
RF.SPD [%]
Variable code: 13
CHWC.VLV [W]
Variable code: 14
CHWC.LM [%]
Variable code: 15
CHWC.DAT [l/min]
Variable code: 16
CHWC.LWT [C]
asax
kmeans
faulty
faultfree

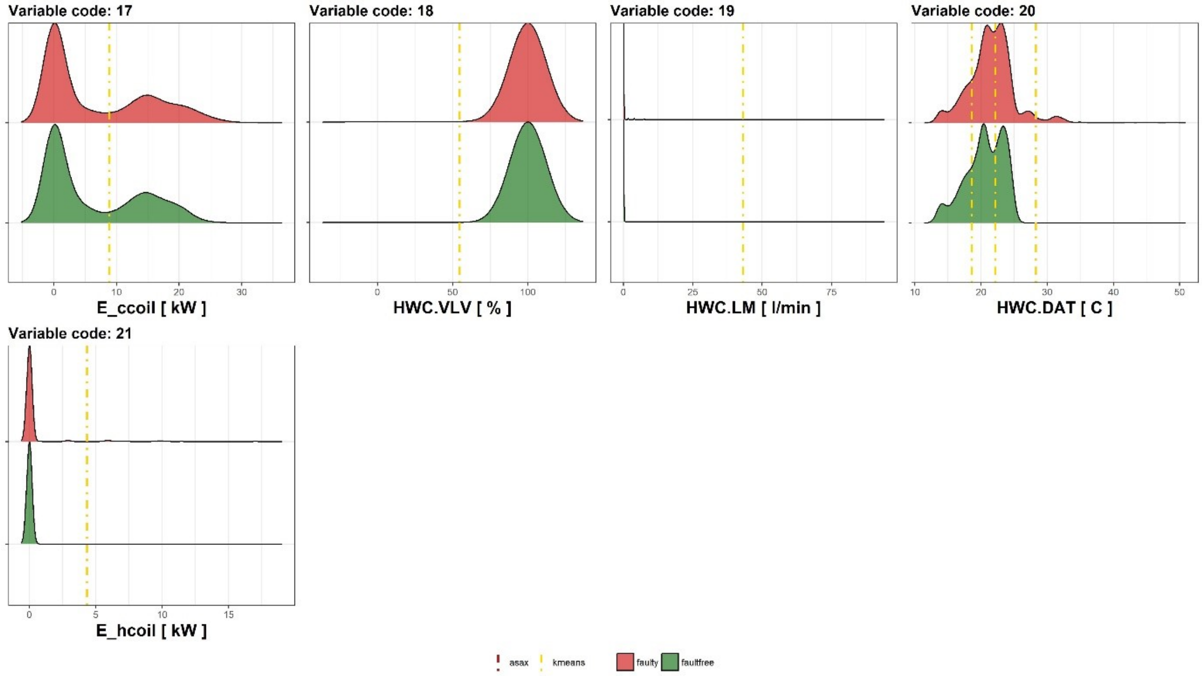


FIG. 3 Distribution and discretization with aggregation at 30 minutes

D. Extracted rules resume

The set of extracted rules for 5 minutes aggregation is reported.

Antecedent	Consequent	Supp	Conf	Actual wind	SUPP DAY	MOD	Supp Conotool	Conf Conotool	n	Window	H min	H max
133132	132133	0,13	0,41	60	0,87	2,23	0,76	0,74	1	120	06:15	15:35
133134	134133	0,12	0,44	75	0,77	2,17	0,89	0,79	2	120	06:10	17:45
132133	141142	0,12	0,44	55	0,83	1,97	0,81	0,68	3	120	06:05	15:45
132133	9293	0,10	0,39	50	0,80	1,70	0,74	0,59	6	120	06:10	15:45
132133	133132	0,16	0,48	115	0,67	1,70	0,56	0,59	5	180	06:05	16:15
6261	6162	0,16	0,66	60	0,93	1,67	0,83	0,89	6	180	06:25	18:25
133132	141142	0,16	0,37	120	0,83	1,90	0,78	0,63	7	180	06:15	15:35
171172	141142	0,15	0,55	55	0,90	1,97	0,81	0,91	9	180	06:05	15:25
133132	9293	0,15	0,35	130	0,90	1,87	0,81	0,62	10	180	06:10	15:35
171172	9293	0,14	0,53	115	0,90	1,90	0,83	0,88	11	180	05:55	15:25
141142	133134	0,14	0,44	85	0,80	1,73	0,63	0,72	12	180	06:05	15:50
9394	133134	0,14	0,55	110	0,70	1,40	0,51	0,72	13	180	06:05	17:40
142141	132133	0,14	0,39	110	0,83	1,47	0,50	0,61	14	180	06:20	15:35
132133	133134	0,13	0,38	125	0,80	2	0,72	0,69	16	180	07:15	15:45
9293	141142	0,13	0,42	55	0,80	1,47	0,60	0,65	17	180	06:05	15:50
202203	171172	0,13	0,63	30	0,70	1,23	0,56	0,70	18	180	06:00	12:15
134133	133134	0,13	0,39	110	0,73	1,63	0,59	0,68	19	180	06:20	17:15
133132, 132133	141142	0,13	0,59	130	0,73	1,67	0,68	0,67	20	180	06:15	15:35
171172	132133	0,13	0,46	130	0,90	2,03	0,69	0,94	21	180	06:05	15:25
141142	9394	0,12	0,36	95	0,73	1,53	0,78	0,64	29	180	06:05	13:20
6465	6564	0,12	0,74	45	0,70	1,23	0,86	0,84	31	180	06:00	23:55
132133	9394	0,11	0,33	105	0,80	1,97	1	0,68	34	180	07:15	16:15
132133	132133	0,11	0,33	5	0,90	2,90	0,99	1	36	180	06:05	16:15
101102, 111112, 7172, 121122	152151, 1211	0,11	0,84	15	0,73	0,73	0,73	0,76	39	180	06:00	6:00
142141	141142	0,11	0,32	115	0,83	1,47	0,60	0,61	40	180	06:20	15:35
133134	9394	0,11	0,33	155	0,77	1,57	0,80	0,57	42	180	06:10	16:35
151152	2221	0,11	0,75	30	0,90	0,90	0,69	0,93	43	180	18:25	19:00
142141, 7271	102101, 203202, 122121	0,11	0,89	10	0,67	0,70	0,72	0,51	47	180	15:35	17:55
142141, 112111	102101, 203202, 122121	0,11	0,89	10	0,63	0,67	0,69	0,54	48	180	15:35	17:55
112111, 7271	132131, 122121	0,11	0,71	10	0,60	0,63	0,90	0,73	49	180	08:25	17:55
112111, 7271	132131, 102101, 122121	0,11	0,71	10	0,60	0,63	0,86	0,73	50	180	08:25	17:55
112111, 7271	132131, 102101	0,11	0,71	10	0,60	0,63	0,90	0,73	51	180	08:25	17:55
7271	132131, 102101, 122121	0,11	0,67	10	0,63	0,67	0,91	0,56	52	180	08:25	17:55
112111	132131, 102101, 122121	0,11	0,71	10	0,60	0,63	0,86	0,58	53	180	08:25	17:55
6263	6362	0,11	0,31	125	0,73	2,30	0,68	0,66	54	180	00:40	23:20
133134	142141	0,11	0,32	125	0,70	1,40	0,58	0,51	55	180	06:10	17:45
102101, 9291, 122121	202201	0,11	0,93	10	0,63	0,63	0,54	0,66	57	180	08:25	18:00
6362	6263	0,11	0,32	95	0,67	2,17	0,61	0,65	61	180	00:20	23:00
133132	133134	0,11	0,26	150	0,80	1,60	0,58	0,53	63	180	06:15	15:35
9392	9293	0,11	0,46	130	0,90	1,43	0,62	0,90	64	180	06:05	15:30
101102, 7172, 121122	152151, 162161, 1211	0,11	0,71	15	0,77	0,77	0,77	0,77	66	180	06:00	6:00
101102, 111112, 7172, 121122	162161, 1211	0,11	0,79	15	0,70	0,70	0,62	0,72	67	180	06:00	6:00
111112, 7172, 121122	162161, 1211	0,11	0,79	15	0,70	0,70	0,62	0,66	68	180	06:00	6:00
101102, 111112, 7172	162161, 1211	0,11	0,79	15	0,70	0,70	0,62	0,70	69	180	06:00	6:00

101102, 111112, 121122	162161, 1211	0,11	0,74	15	0,70	0,70	0,62	0,70	70	180	06:00	6:00
111112, 7172	162161, 1211	0,11	0,79	15	0,70	0,70	0,62	0,84	71	180	06:00	6:00
112111, 7271	102101, 203202, 9291, 122121	0,10	0,66	10	0,57	0,57	0,59	0,65	83	180	17:55	17:55
9293	133134	0,10	0,33	80	0,80	1,40	0,51	0,62	87	180	06:05	15:50
202203	141142	0,10	0,50	70	0,77	1,40	0,58	0,79	89	180	06:00	12:25
9394	134133	0,10	0,41	130	0,73	1,37	0,56	0,71	96	180	08:55	17:40
142141	9293	0,10	0,29	125	0,83	1,30	0,57	0,54	99	180	06:20	15:35
161162	162161	0,10	0,39	30	0,50	1,03	0,50	0,51	100	180	06:25	12:15
202203	132133	0,10	0,48	140	0,80	1,47	0,50	0,83	105	180	06:00	12:25
172171	171172	0,10	0,31	125	0,77	1,23	0,56	0,57	106	180	06:10	15:15
9293	132133	0,19	0,46	140	0,87	1,50	0,51	0,66	1	240	06:05	15:35
133132	6362	0,16	0,31	180	0,83	1,87	0,55	0,62	13	240	06:10	17:55
132133, 133132	132133	0,15	0,66	180	0,70	1,73	0,59	0,72	18	240	06:05	14:35
9394	9493	0,15	0,46	175	0,87	1,17	1	0,60	22	240	06:05	16:20
172171	161162	0,15	0,34	205	0,87	1,33	0,65	0,62	23	240	06:20	18:00
9293	9394	0,14	0,36	145	0,93	1,57	0,80	0,69	26	240	06:05	13:05
171172	133134	0,14	0,42	165	0,73	1,50	0,54	0,69	27	240	06:05	12:55
133132, 132133	9293	0,14	0,47	165	0,73	1,60	0,70	0,64	31	240	06:15	15:35
101102, 111112, 7172, 121122	152151, 162161, 1211	0,14	0,76	15	0,60	0,60	0,60	0,62	33	240	06:00	6:00
133132, 132133	9394	0,14	0,46	155	0,80	1,70	0,86	0,68	35	240	06:15	15:25
4243	9293	0,14	0,53	170	0,87	1,17	0,51	0,88	37	240	05:55	12:20
133134	9493	0,14	0,34	180	0,77	1,47	1,26	0,54	40	240	06:10	16:35
133132, 142141	132133	0,13	0,47	155	0,83	1,60	0,55	0,73	48	240	06:15	15:35
202203	9293	0,13	0,51	155	0,90	1,63	0,71	0,92	50	240	06:00	12:25
203202	161162	0,13	0,41	200	0,87	1,30	0,63	0,71	55	240	06:05	18:50
132133, 141142	9394	0,13	0,48	135	0,73	1,60	0,81	0,72	62	240	06:05	13:10
171172, 132133	141142	0,13	0,67	160	0,77	1,63	0,67	0,79	63	240	06:05	15:25
133134, 134133	133134	0,13	0,43	170	0,77	1,63	0,59	0,65	65	240	06:10	16:05
6364	6463	0,13	0,44	100	0,80	1,60	0,79	0,75	66	240	00:55	23:55
134133	6261	0,13	0,32	165	0,67	1,33	0,70	0,56	67	240	06:20	17:55
6564	6463	0,13	0,63	110	0,63	1,10	0,54	0,79	70	240	00:05	8:00
4243	202203	0,12	0,48	15	0,67	0,93	0,52	0,70	79	240	05:55	12:20
162161	6463	0,12	0,48	115	0,80	1,40	0,69	0,69	80	240	06:05	7:55
6463	171172	0,12	0,47	140	0,70	1,17	0,53	0,58	83	240	02:25	14:50
141142	9293	0,12	0,29	175	0,87	1,83	0,80	0,76	84	240	05:45	15:50
133132	134133	0,12	0,23	200	0,70	1,50	0,62	0,50	85	240	06:15	15:35
6465	171172	0,12	0,59	90	0,70	1,20	0,55	0,82	91	240	06:00	7:50
133132	142141	0,12	0,23	155	0,87	2,37	0,97	0,79	92	240	06:15	17:55
132133, 9293	141142	0,12	0,49	170	0,63	1,50	0,62	0,74	93	240	06:10	15:45
133132, 9293	141142	0,12	0,52	180	0,63	1,47	0,60	0,71	94	240	06:10	15:35
142141	172171	0,12	0,25	10	0,93	1,87	0,85	0,78	97	240	06:20	17:55
133132, 141142	9394	0,12	0,49	175	0,70	1,37	0,69	0,64	99	240	06:15	12:50
134133	134133	0,12	0,30	5	0,80	2,40	0,99	1	101	240	06:15	17:55
9394	142141	0,12	0,37	190	0,83	1,30	0,53	0,67	108	240	06:05	17:40
171172	142141	0,11	0,33	230	0,87	1,40	0,58	0,65	118	240	06:05	15:25
4342, 151152	2221	0,11	0,81	75	0,83	0,83	0,64	0,63	119	240	18:00	18:20
141142	9392	0,11	0,28	215	0,83	1,30	0,80	0,54	121	240	06:05	12:40
142141, 6261	6162	0,11	0,64	150	0,70	1	0,50	0,55	124	240	06:20	17:55
171172, 9293	141142	0,11	0,55	160	0,70	1,43	0,59	0,67	125	240	06:05	15:25
133132	172171	0,11	0,22	145	0,80	1,87	0,85	0,62	126	240	06:10	17:55
142141, 112111, 7271, 9492	102101, 203202, 9291, 122121	0,11	0,96	10	0,50	0,50	0,52	0,56	127	240	17:50	17:55
172171, 171172	9293	0,11	0,72	160	0,77	1,20	0,52	0,71	131	240	06:10	15:15
142141	161162	0,11	0,24	205	0,87	1,20	0,58	0,50	133	240	06:20	17:55
133132, 142141	9293	0,11	0,39	170	0,83	1,33	0,58	0,61	136	240	06:15	15:35

132133	142141	0,11	0,26	235	0,77	1,57	0,64	0,54	142	240	06:05	16:15
171172	202203	0,11	0,32	140	0,87	1,30	0,72	0,60	144	240	05:55	9:20
6463	202203	0,11	0,42	160	0,67	1,07	0,59	0,53	156	240	02:00	8:20
141142	133132, 142141	0,11	0,26	195	0,87	1,30	0,58	0,54	158	240	06:05	15:50
6463	141142	0,11	0,41	175	0,77	1,30	0,53	0,65	175	240	02:25	14:50
102101, 122121, 151152	2221	0,11	0,83	85	0,90	0,90	0,69	0,93	176	240	18:00	18:00
162161	132133	0,11	0,42	135	0,90	1,93	0,66	0,95	179	240	06:05	12:20
171172, 132133	9293	0,11	0,55	175	0,77	1,73	0,75	0,84	180	240	06:05	15:25
132133, 141142, 133134	134133	0,11	0,70	160	0,70	1,43	0,59	0,60	193	240	07:15	15:45
141142, 133132	132133	0,11	0,53	195	0,87	1,60	0,55	0,79	200	240	06:05	13:20
162161	6362	0,11	0,41	155	0,93	1,90	0,56	0,93	201	240	06:05	12:20
171172	9392	0,11	0,30	230	0,83	1,27	0,78	0,58	205	240	06:05	12:30
162161	171172	0,11	0,41	130	0,93	1,97	0,89	0,97	207	240	05:45	12:20
171172, 141142	9394	0,10	0,47	200	0,70	1,33	0,68	0,66	208	240	06:05	12:55
171172	134133	0,10	0,30	215	0,70	1,30	0,53	0,60	213	240	06:05	12:55
112111, 7271, 151152	2221	0,10	0,83	90	0,83	0,83	0,64	0,81	226	240	17:55	17:55
132133, 9394	133134	0,10	0,55	165	0,63	1,47	0,53	0,56	229	240	07:15	15:00
134133	203202	0,10	0,26	230	0,70	1,23	0,66	0,51	232	240	06:15	17:55
162161	9293	0,10	0,40	160	0,93	1,97	0,86	0,97	236	240	05:45	12:20
112111, 7271, 9492	132131, 102101, 9291, 122121	0,10	0,80	10	0,43	0,43	0,65	0,50	265	240	17:55	17:55
7271, 9492	132131, 102101, 203202, 9291, 122121	0,10	0,80	10	0,40	0,40	0,60	0,57	274	240	17:55	17:55
112111, 9492	132131, 102101, 203202, 9291, 122121	0,10	0,80	10	0,40	0,40	0,60	0,55	275	240	17:55	17:55
202203	9392	0,10	0,39	200	0,73	0,93	0,57	0,53	283	240	06:00	11:45
6564	171172	0,10	0,51	135	0,67	1,10	0,50	0,79	287	240	06:05	8:00
2221	161162	0,10	0,48	150	0,97	1,03	0,50	0,82	288	240	05:55	20:25
112111, 7271, 102101, 203202, 122121, 4342	151152	0,10	0,68	65	0,63	0,63	0,63	0,68	291	240	17:55	17:55
112111, 7271, 102101, 203202, 122121, 202201	151152	0,10	0,65	65	0,60	0,60	0,60	0,72	292	240	17:55	17:55
133134	142141, 172171	0,10	0,25	190	0,73	1,37	0,80	0,50	293	240	06:10	17:45
132131, 102101, 122121	151152	0,10	0,59	60	0,63	0,63	0,63	0,90	316	240	18:00	18:00
202203, 171172	141142	0,10	0,57	165	0,73	1,23	0,51	0,90	318	240	06:00	12:15

The set of extracted rules for 15 minutes aggregation is reported.

Antecedent	Consequent	Supp	Conf	Actual wind	SUPP DAY	MOD	Supp Conotool	Conf Conotool	n	Window	H min	H max
133132	132133	0,11	0,68	60	0,73	1,4	0,53	0,89	1	120	06:15	15:15
6261	6162	0,1	0,65	75	0,7	1,23	0,76	0,82	2	120	06:15	18:00
132133	141142	0,15	0,47	60	0,8	2,03	0,91	0,78	2	180	06:00	15:30
132133	9293	0,12	0,39	75	0,8	1,77	0,9	0,68	4	180	06:00	15:30
133134	134133	0,12	0,46	90	0,7	1,5	0,98	0,69	5	180	08:30	17:00
102101, 142141, 172171, 203202, 112111, 7271, 122121	202201	0,11	0,83	30	0,63	0,63	0,56	0,68	6	180	17:45	17:45
101102, 7172, 9192	152151, 162161, 1211	0,11	0,79	30	0,73	0,73	0,76	0,79	7	180	05:45	05:45

101102, 111112, 7172	152151, 162161, 1211	0,11	0,83	30	0,7	0,7	0,72	0,7	8	180	05:45	05:45
202203	171172	0,11	0,58	45	0,67	1,07	0,63	0,73	11	180	05:45	09:45
133132	141142	0,11	0,49	135	0,6	1,2	0,54	0,77	12	180	06:15	12:45
134133	133134	0,11	0,51	90	0,57	1,1	0,5	0,73	13	180	09:30	17:00
101102, 111112, 7172, 9192	152151, 162161	0,1	0,83	30	0,63	0,63	0,68	0,7	14	180	05:45	05:45
101102, 111112, 7172, 9192	152151, 162161, 1211	0,1	0,83	30	0,63	0,63	0,66	0,7	15	180	05:45	05:45
101102, 111112, 7172, 9192	162161, 1211	0,1	0,83	30	0,67	0,67	0,65	0,74	16	180	05:45	05:45
101102, 111112, 7172, 9192	152151, 1211	0,1	0,83	30	0,7	0,7	0,7	0,78	17	180	05:45	05:45
111112, 7172, 9192	152151, 162161, 1211	0,1	0,83	30	0,63	0,63	0,66	0,7	18	180	05:45	05:45
101102, 111112, 9192	152151, 162161, 1211	0,1	0,83	30	0,63	0,63	0,66	0,66	19	180	05:45	05:45
141142	133134	0,1	0,34	135	0,73	1,37	0,62	0,62	20	180	06:00	15:45
171172	141142	0,1	0,44	60	0,87	1,57	0,7	0,94	22	180	06:00	12:45
102101, 142141, 172171, 203202, 112111, 7271, 122121	4342	0,1	0,76	30	0,7	0,7	0,53	0,75	23	180	17:45	17:45
132133	133132	0,16	0,41	195	0,73	1,5	0,94	0,58	7	240	06:00	14:30
142141	141142	0,15	0,36	135	0,83	1,3	0,58	0,59	8	240	06:15	15:30
141142	9394	0,15	0,4	180	0,8	1,6	0,94	0,73	9	240	05:30	13:15
171172	9293	0,14	0,48	195	0,83	1,6	0,81	0,96	19	240	05:45	12:45
151152	2221	0,14	0,7	45	0,97	0,97	0,76	1	20	240	18:15	19:15
132133	133134	0,14	0,37	150	0,73	1,83	0,83	0,71	21	240	06:00	13:15
9392	9293	0,14	0,72	90	0,87	1	0,51	1	22	240	06:15	15:15
132133	9394	0,14	0,36	135	0,8	1,87	1,1	0,72	23	240	06:00	16:00
9293	141142	0,14	0,37	105	0,77	1,43	0,64	0,74	25	240	06:00	15:45
133134	9394	0,13	0,43	180	0,63	1,4	0,82	0,65	26	240	08:30	17:30
142141	9293	0,13	0,31	135	0,8	1,23	0,63	0,56	27	240	06:15	15:30
133132	9293	0,13	0,49	180	0,63	1,27	0,64	0,81	28	240	06:15	12:45
202203	141142	0,13	0,57	90	0,8	1,23	0,55	0,84	29	240	05:45	09:45
132133	132133	0,13	0,35	15	0,87	2,6	0,99	1	30	240	06:00	16:00
133134	133134	0,13	0,42	15	0,77	2,17	0,98	1	32	240	08:30	17:30
133132, 132133	141142	0,13	0,64	150	0,53	1,13	0,51	0,72	34	240	06:15	12:45
9394	133134	0,13	0,45	105	0,67	1,2	0,55	0,72	35	240	08:45	17:30
171172	133132	0,13	0,42	195	0,7	1,1	0,69	0,66	36	240	06:00	12:15
4342	151152	0,12	0,44	60	0,9	0,9	0,9	0,69	37	240	17:45	18:15
171172	132133	0,12	0,42	75	0,87	1,63	0,62	0,98	39	240	06:00	12:45
6263	6362	0,12	0,3	135	0,6	1,43	0,61	0,55	40	240	05:45	23:00
132133	6362	0,12	0,32	150	0,77	1,5	0,64	0,58	42	240	06:00	16:00
134133	142141	0,12	0,47	225	0,67	1,3	0,58	0,87	44	240	09:30	17:30
9293	132133	0,12	0,33	150	0,87	1,37	0,52	0,71	47	240	06:00	15:30
9293	133134	0,12	0,33	195	0,73	1,33	0,61	0,69	48	240	06:00	15:45
202201	151152	0,12	0,44	60	0,73	0,73	0,73	0,67	49	240	18:00	18:15
141142	142141	0,12	0,31	210	0,83	1,33	0,6	0,61	51	240	06:00	15:45
141142	133132	0,12	0,31	195	0,73	1,2	0,75	0,55	53	240	06:00	12:45
133134	142141	0,11	0,36	210	0,7	1,47	0,66	0,68	58	240	08:30	17:30
102101, 112111, 7271, 122121	151152	0,11	0,51	75	0,9	0,9	0,9	0,96	63	240	17:45	17:45
142141	6162	0,11	0,25	165	0,77	1,3	0,8	0,59	64	240	06:15	17:45
6362	6263	0,11	0,3	180	0,53	1,47	0,56	0,64	66	240	04:45	22:45
9293	9394	0,11	0,29	150	0,87	1,33	0,78	0,69	72	240	06:00	13:00
9394	134133	0,11	0,37	135	0,6	0,97	0,63	0,58	73	240	08:45	16:15
121122	162161	0,11	0,5	30	0,97	0,97	0,6	0,94	78	240	05:45	06:00

141142	132133	0,1	0,27	150	0,87	1,77	0,67	0,8	81	240	06:00	15:30
9394	9493	0,1	0,36	165	0,77	0,9	0,96	0,54	90	240	08:30	16:15
101102, 111112, 7172, 131132	152151, 162161	0,1	0,88	30	0,63	0,63	0,68	0,73	94	240	05:45	05:45
101102, 111112, 7172, 131132	152151, 162161, 1211	0,1	0,88	30	0,63	0,63	0,66	0,73	95	240	05:45	05:45
101102, 111112, 7172, 131132	162161, 1211	0,1	0,88	30	0,67	0,67	0,65	0,77	96	240	05:45	05:45
101102, 7172, 9192, 131132	152151, 1211	0,1	0,88	30	0,73	0,73	0,73	0,81	97	240	05:45	05:45
101102, 111112, 7172, 131132	152151, 1211	0,1	0,88	30	0,67	0,67	0,67	0,77	98	240	05:45	05:45
111112, 7172, 131132	152151, 162161, 1211	0,1	0,88	30	0,63	0,63	0,66	0,73	99	240	05:45	05:45
101102, 7172, 131132	152151, 162161, 1211	0,1	0,8	30	0,73	0,73	0,76	0,85	100	240	05:45	05:45
101102, 111112, 131132	152151, 162161, 1211	0,1	0,8	30	0,7	0,7	0,72	0,84	101	240	05:45	05:45
101102, 7172, 121122	152151, 1211	0,1	0,88	30	0,9	0,9	0,9	0,9	102	240	05:45	05:45
7172	171172	0,1	0,48	165	0,87	0,93	0,55	0,82	107	240	05:45	08:00

The set of extracted rules for 30 minutes aggregation is reported.

Antecedent	Consequent	Supp	Conf	Actual wind	SUPP DAY	MOD	Supp Conotool	Conf Conotool	n	Window	H min	H max
133134	134133	0,12	0,55	120	0,67	1,3	0,93	0,76	1	180	08:00	16:30
6261	6162	0,12	0,66	90	0,63	1	0,71	0,75	2	180	06:00	22:00
141142	9394	0,11	0,48	120	0,67	1,03	0,65	0,63	3	180	05:30	13:00
4342	151152	0,11	0,6	90	0,9	0,9	0,9	0,77	4	180	17:30	18:00
132133	141142	0,1	0,52	60	0,77	1,07	0,64	0,76	5	180	07:30	12:30
4342	2221	0,1	0,57	90	0,87	0,87	0,74	0,74	7	180	17:30	18:00
171172	141142	0,14	0,68	90	0,87	1,17	0,7	0,97	6	240	05:30	08:30
4342	151152, 1112	0,13	0,54	90	0,9	0,9	0,93	0,77	7	240	17:30	18:00
134133	133134	0,13	0,57	120	0,6	1	0,58	0,73	8	240	09:30	16:00
132133	9394	0,12	0,49	150	0,67	1	0,63	0,71	9	240	07:30	16:00
9394	133134	0,12	0,44	120	0,53	0,87	0,5	0,55	10	240	08:30	15:30
102101, 142141, 172171, 112111, 7271, 122121	2221	0,12	0,68	120	0,8	0,8	0,69	0,86	11	240	17:30	17:30
9394	134133	0,12	0,43	180	0,63	0,9	0,64	0,57	13	240	08:30	16:00
9394	9493	0,12	0,42	180	0,77	0,9	0,96	0,57	14	240	08:30	16:00
141142	133134	0,12	0,4	240	0,6	1,03	0,6	0,63	17	240	05:30	13:00
2221	161162	0,11	0,57	180	0,93	0,93	0,7	0,82	19	240	18:00	20:00
102101, 142141, 172171, 112111, 7271, 122121	151152, 1112	0,11	0,63	90	0,83	0,83	0,86	0,89	20	240	17:30	17:30
6263	6362	0,11	0,31	180	0,57	1,1	0,59	0,57	22	240	05:30	21:30
1112	6263	0,11	0,54	150	0,8	1	0,51	0,5	24	240	05:30	19:00
4243	141142	0,11	0,48	180	0,77	0,97	0,58	0,83	27	240	05:30	10:00
1112	161162	0,1	0,52	180	0,97	1,27	0,95	0,63	28	240	05:30	19:00
9293	9394	0,1	0,37	180	0,8	0,93	0,58	0,65	30	240	06:00	12:30
102101, 142141, 172171, 112111, 7271, 122121	4342	0,1	0,57	60	0,8	0,8	0,67	0,86	32	240	17:30	17:30
162161	171172	0,1	0,44	180	0,9	1,2	0,97	0,92	34	240	05:00	07:30

E. Faults implication of extracted rules violation

An indication of the fault mode for the violation of the rules is reported, providing a correlation between faults and rules. The blank cells refer to the lack of the implication in case of fault, the encoding gives an indication of the presence of the only antecedent," A", the only consequent "C" or the absence of the pattern, "X".

The table of the violation characterisation of the rules with 5 minutes aggregation is reported.

aggr	wind	n	CCVS65	CCVSFC	DLBSF	EASFC	HCVL	OASFC	RFCF	RFF30
5	120	1	X	X						
5	120	2	X	X						
5	120	3	X	X						
5	120	6	CX	CX						
5	180	5	X	X		ACX				
5	180	6		AX						
5	180	7	X	X						
5	180	9	X	X						
5	180	10	CX	CX						
5	180	11		CX						
5	180	12	X	X						
5	180	13	AX	AX					ACX	ACX
5	180	14	X	X						
5	180	16	X	X						
5	180	17	AX	AX						
5	180	18	AX	AX						
5	180	19	X	X						
5	180	20	X	X		ACX				
5	180	21	X	X						
5	180	29	CX	CX					ACX	
5	180	31	X					X		
5	180	34	CX	CX					ACX	
5	180	36	X	X						
5	180	39	AX	AX					ACX	ACX
5	180	40	X	X						
5	180	42	CX	CX						
5	180	43	X	X						
5	180	47	X	X			AX			AX
5	180	48	X	X			AX		X	X
5	180	49	X	X		AX	ACX		X	X
5	180	50	X	X		AX	ACX		X	X
5	180	51	X	X		AX	ACX		X	X
5	180	52	X	X		X	CX		AX	
5	180	53	AX	X		AX	ACX		X	X
5	180	54	X							
5	180	55	X	X						
5	180	57		X	CX				X	X
5	180	61	X							
5	180	63	X	X						
5	180	64		CX						
5	180	66	AX	AX		ACX				ACX
5	180	67	AX	ACX					ACX	ACX
5	180	68	AX	ACX					CX	CX
5	180	69	AX	ACX					ACX	ACX
5	180	70	AX	ACX					ACX	ACX
5	180	71	X	ACX					CX	CX
5	180	83	X	X			AX		X	X
5	180	87	AX	AX						
5	180	89	AX	AX						
5	180	96	AX	AX					ACX	
5	180	99	CX	CX						
5	180	100	X	ACX				CX	CX	CX
5	180	105	AX	AX						
5	180	106	X	X						
5	240	1	AX	AX						
5	240	13	X	CX						
5	240	18	X	X		CX				
5	240	22		AX						
5	240	23	X	CX						
5	240	26	X	X	X	X	X	X	X	X
5	240	27	X	X						
5	240	31	CX	CX						
5	240	33	AX	AX	ACX	ACX			ACX	ACX

5	240	35	CX	CX					ACX	
5	240	37		ACX				ACX		
5	240	40	CX	X						
5	240	48	X	X						
5	240	50	X	X	X	X	X	X	X	X
5	240	55	X	CX						
5	240	62	CX	CX					ACX	ACX
5	240	63	X	X	ACX					
5	240	65	X	X						
5	240	66	X							
5	240	67	CX	CX						ACX
5	240	70	X					CX		
5	240	79	ACX	ACX		ACX		ACX		
5	240	80	X							
5	240	83	X	AX						
5	240	84		CX						
5	240	85	X	X						ACX
5	240	91	X	AX				CX		
5	240	92	X	X						
5	240	93	X	X			ACX	ACX	ACX	
5	240	94	X	X		ACX	ACX	ACX	ACX	
5	240	97		X						
5	240	99	CX	CX					ACX	
5	240	101	X	X						
5	240	108	AX	AX						
5	240	118	X	X						
5	240	119	X	X					AX	
5	240	121	CX	X						
5	240	124	CX	X	AX					
5	240	125	X	X				ACX		
5	240	126	X	X						
5	240	127	X	X			AX		X	X
5	240	131	CX	CX						
5	240	133	X	CX						
5	240	136	CX	CX						
5	240	142	X	X			ACX		ACX	
5	240	144		CX						
5	240	156	CX	ACX						
5	240	158	X	X						
5	240	175	X	AX						
5	240	176	X	X						
5	240	179	X	AX						
5	240	180	CX	CX	ACX					
5	240	193	X	X						ACX
5	240	200	X	X						
5	240	201	X							
5	240	205	CX	X						
5	240	207		AX						
5	240	208	CX	CX					ACX	
5	240	213	X	X						
5	240	226	AX	X					X	X
5	240	229	X	X					ACX	ACX
5	240	232	X	X						
5	240	236		ACX						
5	240	265	AX	X		AX	ACX		X	X
5	240	274	X	X		X	CX		X	X
5	240	275	X	X		AX	ACX		X	X
5	240	283	ACX	AX						
5	240	287	X	AX				CX		
5	240	288		CX						
5	240	291	X	X			AX		X	X
5	240	292	X	X	X		X		X	X
5	240	293	X	X				ACX		
5	240	316	X	X		X	AX		X	
5	240	318	X	AX						

The table of the violation characterisation of the rules with 15 minutes aggregation is reported.

aggr	wind	n	CCVS65	CCVSFC	DLBSF	EASFC	HCVL	OASFC	RFCF	RFF30
15	120	1	X	X						CX
15	120	2		AX	X	X				
15	180	2	X	X						
15	180	4	CX	X						
15	180	5	X	X						AX
15	180	6	X	X	CX		CX		X	X
15	180	7	X	X				CX	CX	

15	180	8	X	X				CX	CX	CX
15	180	11	X	X						
15	180	12	X	X					ACX	CX
15	180	13	X	X	ACX					CX
15	180	14	X	X				CX	CX	CX
15	180	15	X	X				CX	CX	CX
15	180	16	X	CX					CX	CX
15	180	17		X					CX	CX
15	180	18	AX	X				CX	CX	CX
15	180	19	X	X				CX	CX	CX
15	180	20	X	X						
15	180	22	X	X						
15	180	23	X	X			CX		X	CX
15	240	7	X	X						AX
15	240	8	X	X						
15	240	9		X					ACX	
15	240	19		X						
15	240	20		X						
15	240	21	X	X						
15	240	22		X						
15	240	23	CX	X					ACX	
15	240	25	AX	X					ACX	
15	240	26	CX	X	ACX					
15	240	27	CX	X						
15	240	28	CX	X				ACX		CX
15	240	29	X	X						
15	240	30	X	X						
15	240	32	X	X						
15	240	34	X	X			ACX		ACX	CX
15	240	35	AX	X					ACX	
15	240	36	X	X						AX
15	240	37		X					X	
15	240	39	X	X						
15	240	40	X							ACX
15	240	42	X	CX						ACX
15	240	44	X	X					ACX	CX
15	240	47	AX	X						
15	240	48	AX	X						
15	240	49		X	AX		AX			
15	240	51	X	X						
15	240	53	X	X						AX
15	240	58	X	X					ACX	
15	240	63		X					X	X
15	240	64		X	AX					
15	240	66	X					ACX		
15	240	72		X						
15	240	73	AX	X					ACX	AX
15	240	78	X							
15	240	81	X	X						
15	240	90		X						
15	240	94	X	X	ACX			CX	CX	CX
15	240	95	X	X	ACX			CX	CX	CX
15	240	96	X	CX	ACX				CX	CX
15	240	97	X	X	ACX				CX	
15	240	98	X	X	ACX				CX	CX
15	240	99	X	X	ACX			CX	CX	CX
15	240	100	X	X	ACX			CX	CX	
15	240	101	X	X	ACX			CX	CX	CX
15	240	102		X					CX	
15	240	107		X					ACX	

The table of the violation characterisation of the rules with 30 minutes aggregation is reported.

aggr	wind	n	CCVS65	CCVSFC	DLBSF	EASFC	HCVL	OASFC	RFCF	RFF30
30	180	1	X	X	X					CX
30	180	2		AX		X				CX
30	180	3	CX	X					ACX	
30	180	4		X					X	
30	180	5	X	X						
30	180	7		X					X	
30	240	6	X	X						
30	240	7		AX					CX	
30	240	8	X	X	X					AX
30	240	9	CX	X					ACX	
30	240	10	AX	X	AX				ACX	AX
30	240	11		AX					ACX	ACX

30	240	13	AX	X	AX				ACX	
30	240	14		X	AX					
30	240	17	X	X	AX					AX
30	240	19		CX						
30	240	20		AX					ACX	ACX
30	240	22	X	ACX						
30	240	24	AX	CX						ACX
30	240	27	AX	AX		ACX			CX	
30	240	28		CX						
30	240	30		X						
30	240	32		ACX					AX	ACX
30	240	34		AX						

F. Faults-rules matrices

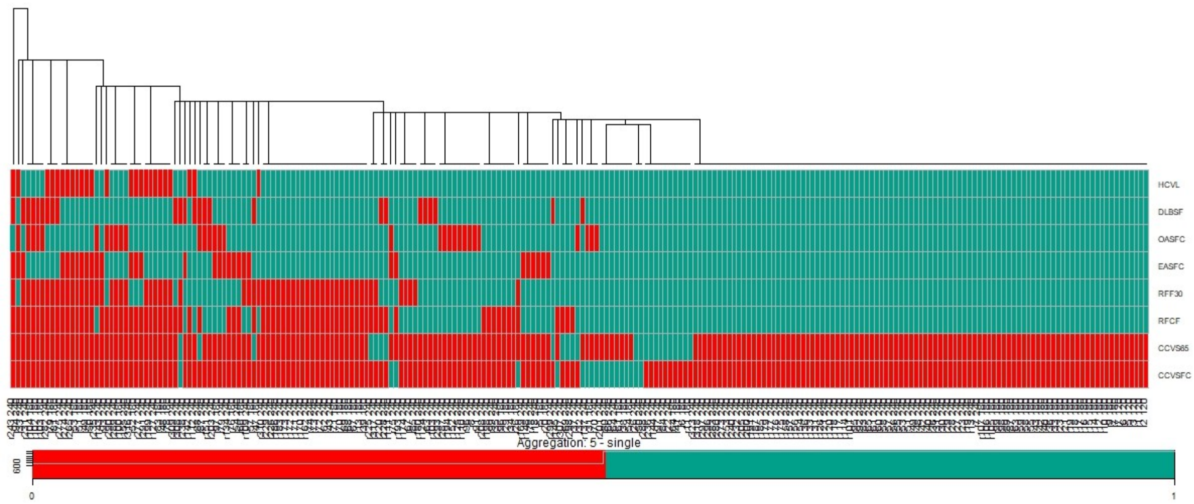


FIG. 4 Faults-rules matrix for 5 minutes aggregation

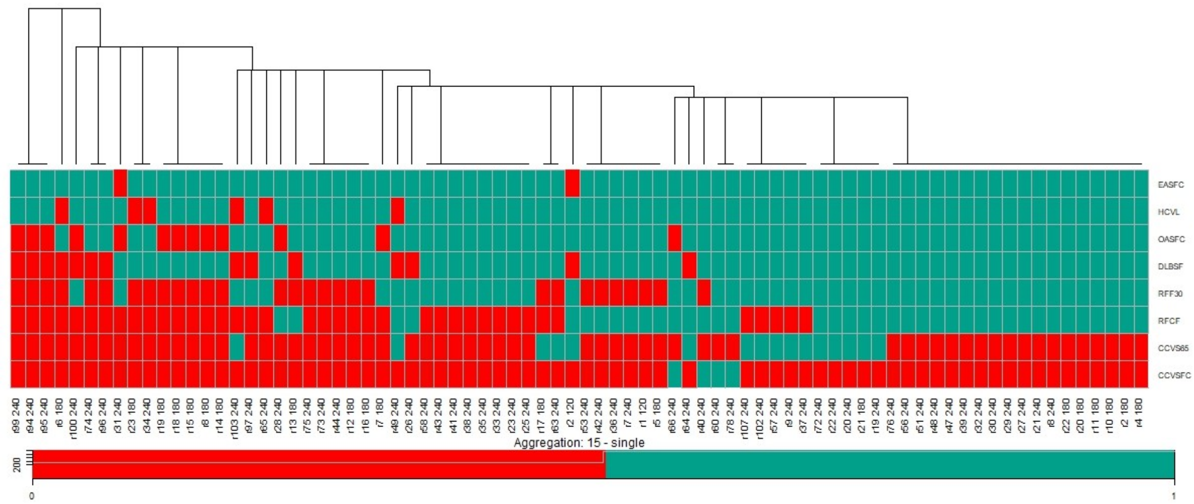


FIG. 5 Faults-rules matrix for 15 minutes aggregation

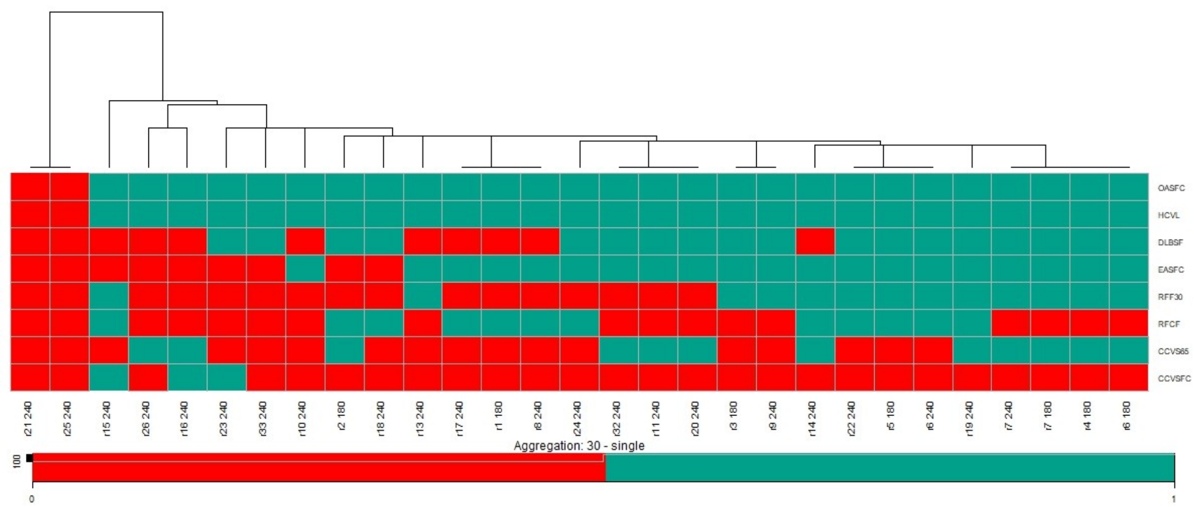


FIG. 6 Faults-rules matrix for 30 minutes aggregation

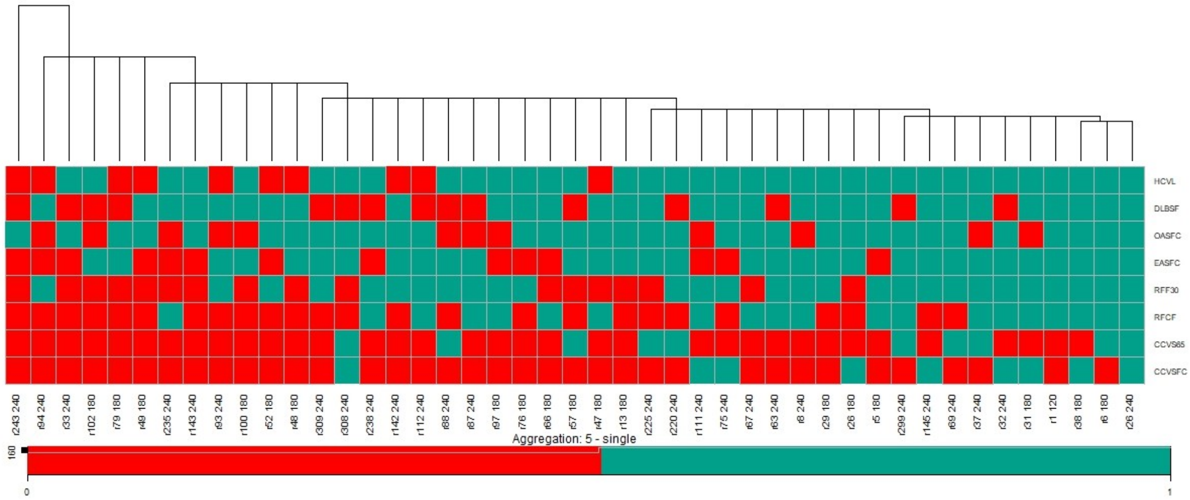


FIG. 7 Faults-rules matrix for 5 minutes aggregation, after the cut at 0-height rules clustering

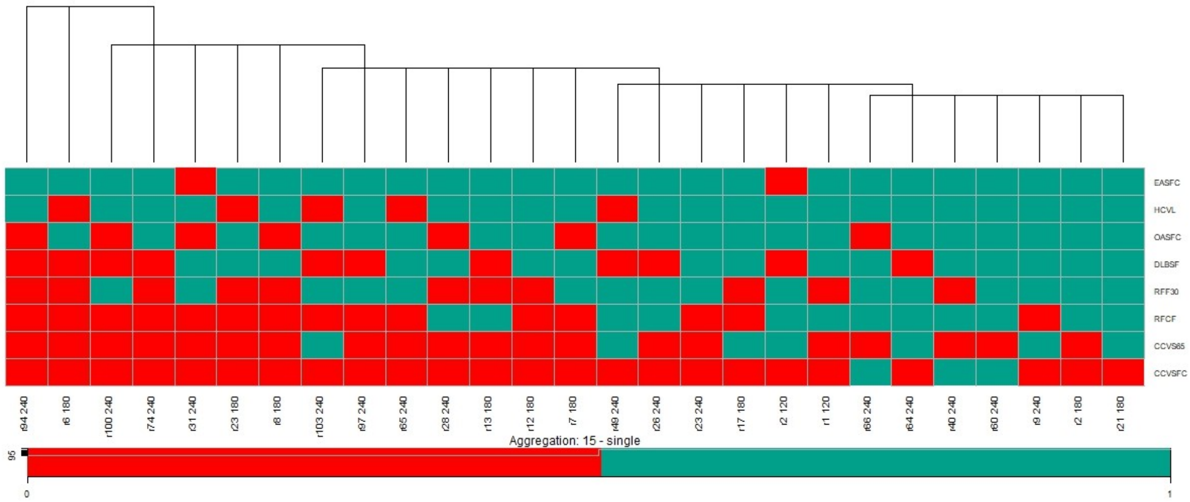


FIG. 8 Faults-rules matrix for 15 minutes aggregation, after the cut at 0-height rules clustering.

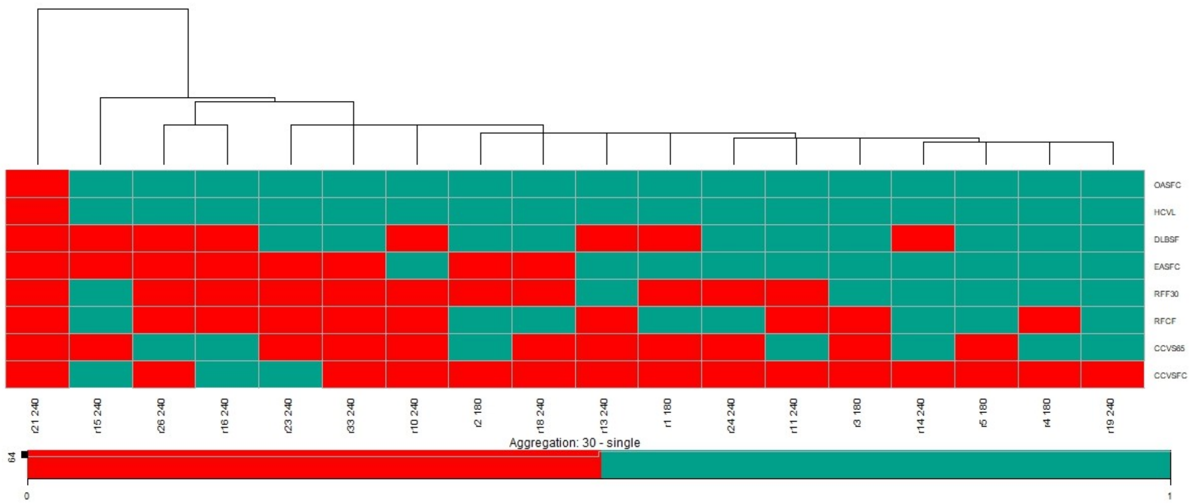


FIG. 9 Faults-rules matrix for 30 minutes aggregation, after the cut at 0-height rules clustering.

G. Score calculation

The score calculation's results for all the cases are reported.

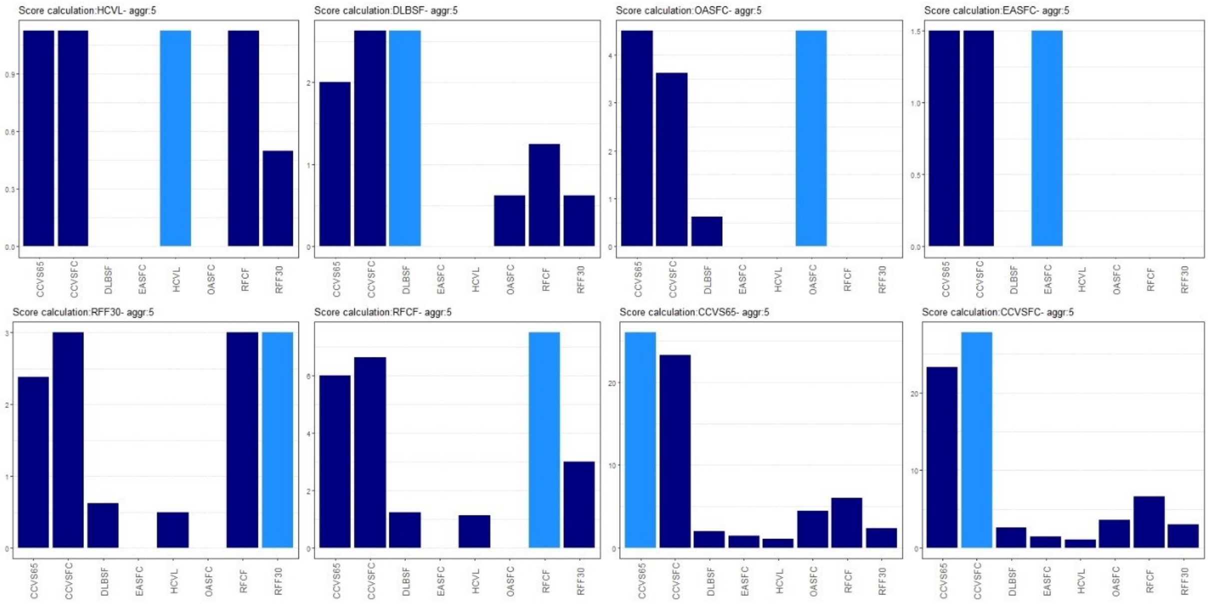


FIG. 10 Start-up of 5 minutes aggregation.

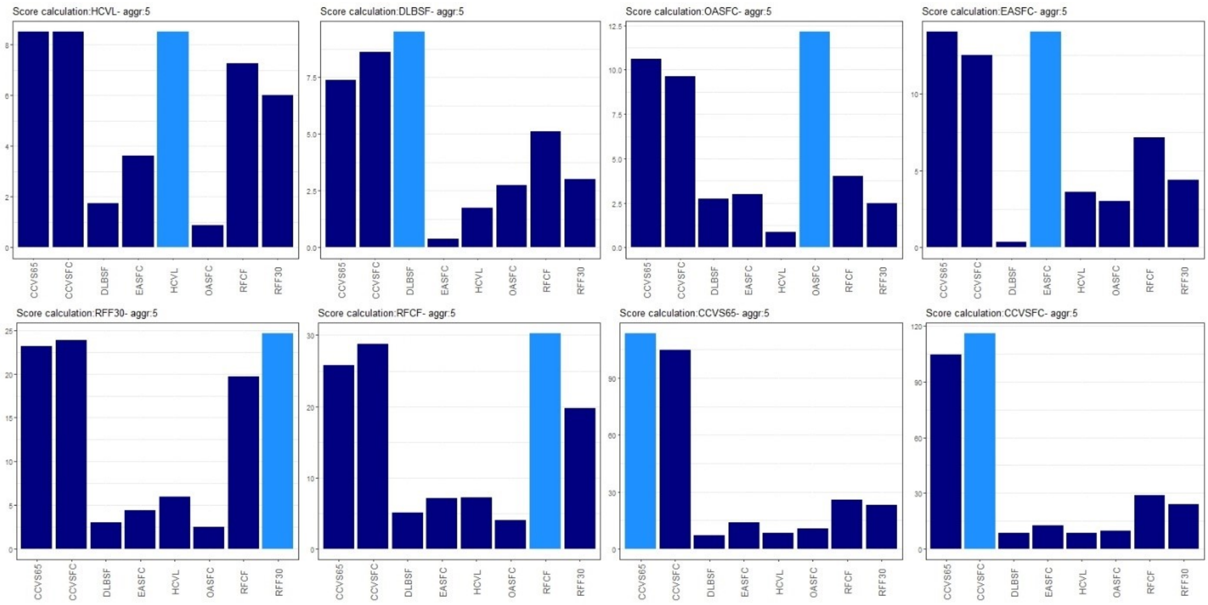


FIG. 11 Operation of 5 minutes aggregation.

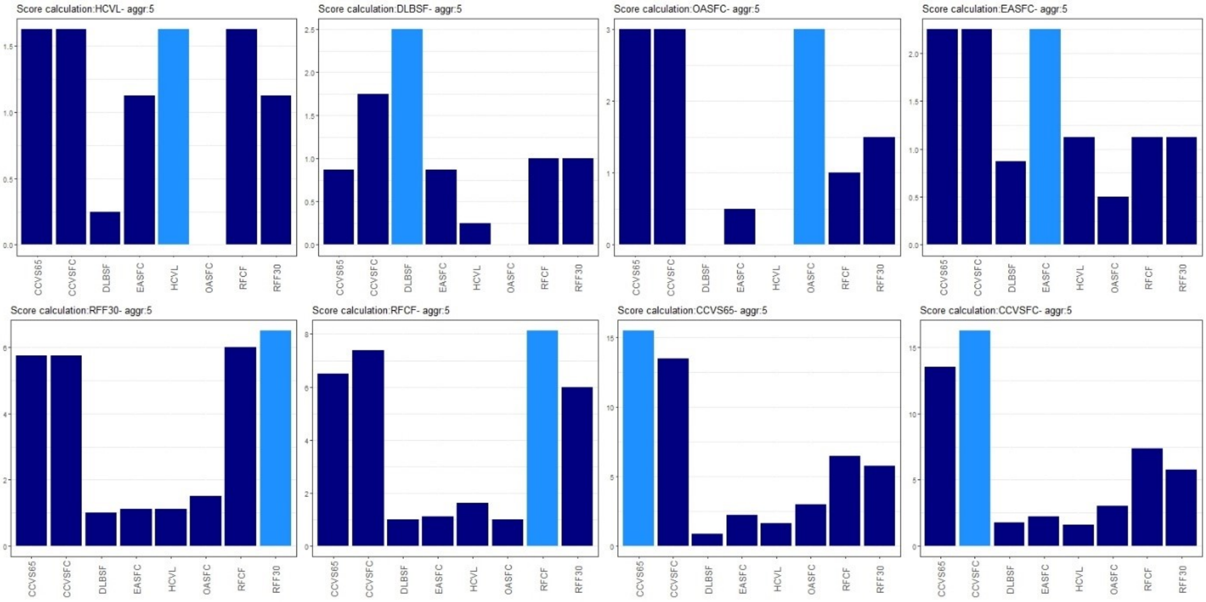


FIG. 12 Shut down of 5 minutes aggregation.

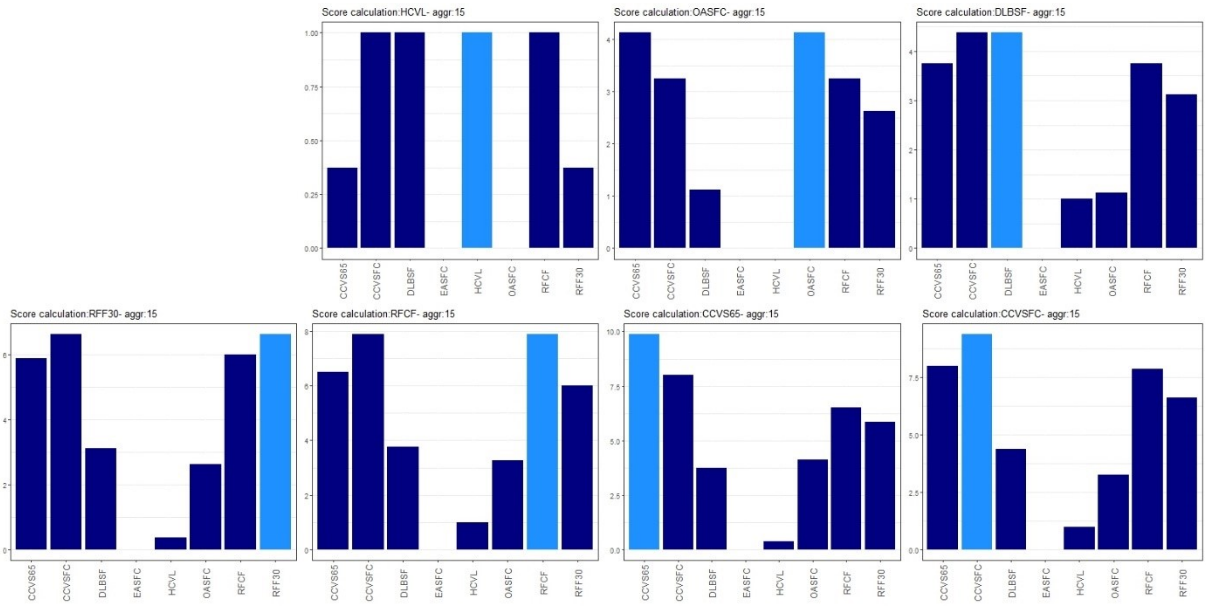


FIG. 13 Start-up of 15 minutes aggregation.

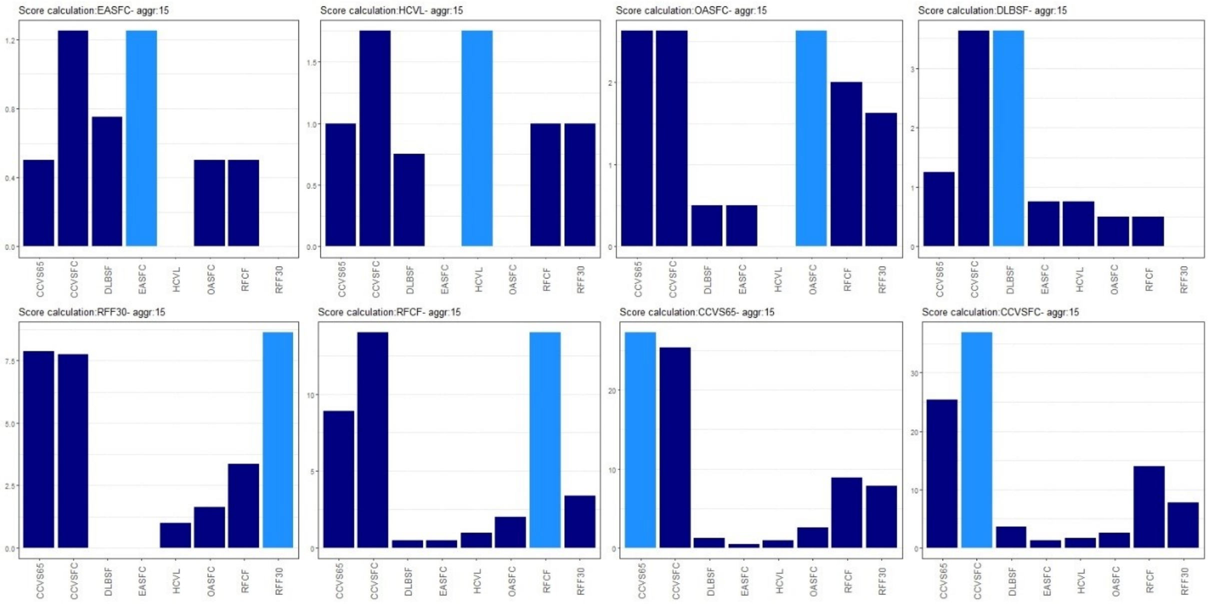


FIG. 14 Operation of 15 minutes aggregation.

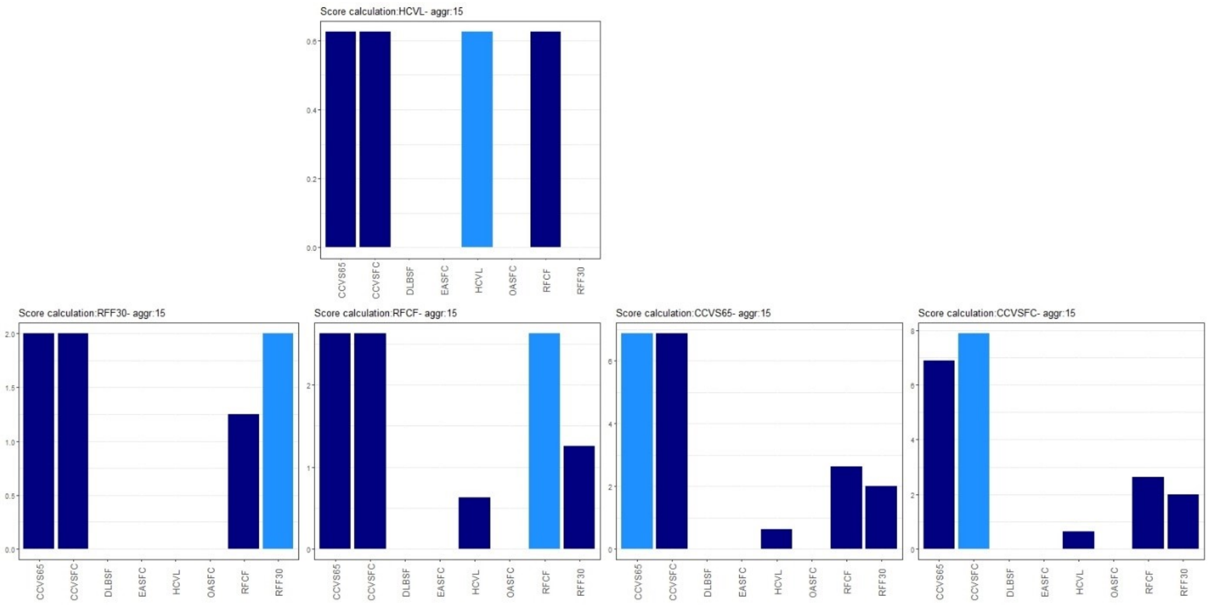


FIG. 15 Shut down of 15 minutes aggregation.

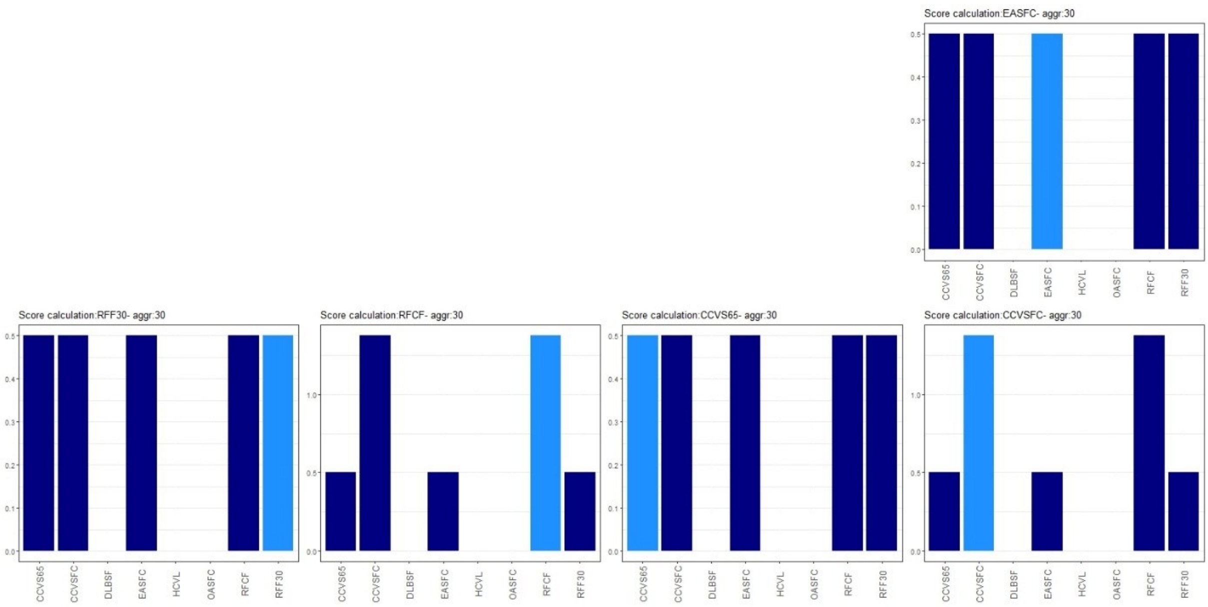


FIG. 16 Start-up of 30 minutes aggregation.

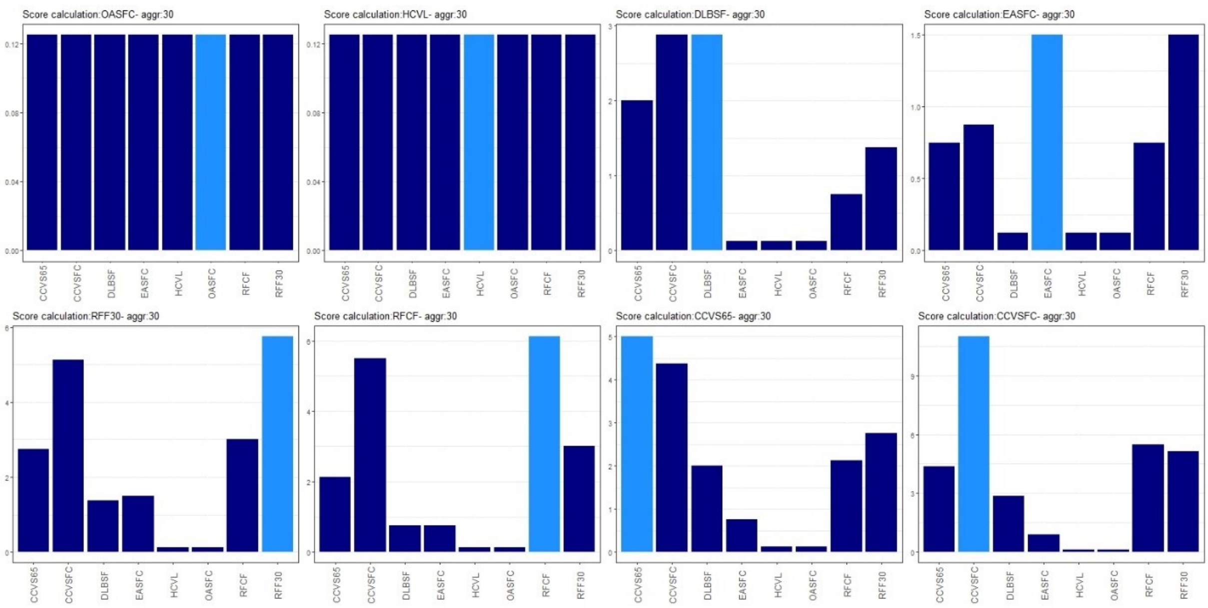


FIG. 17 Operation of 30 minutes aggregation.

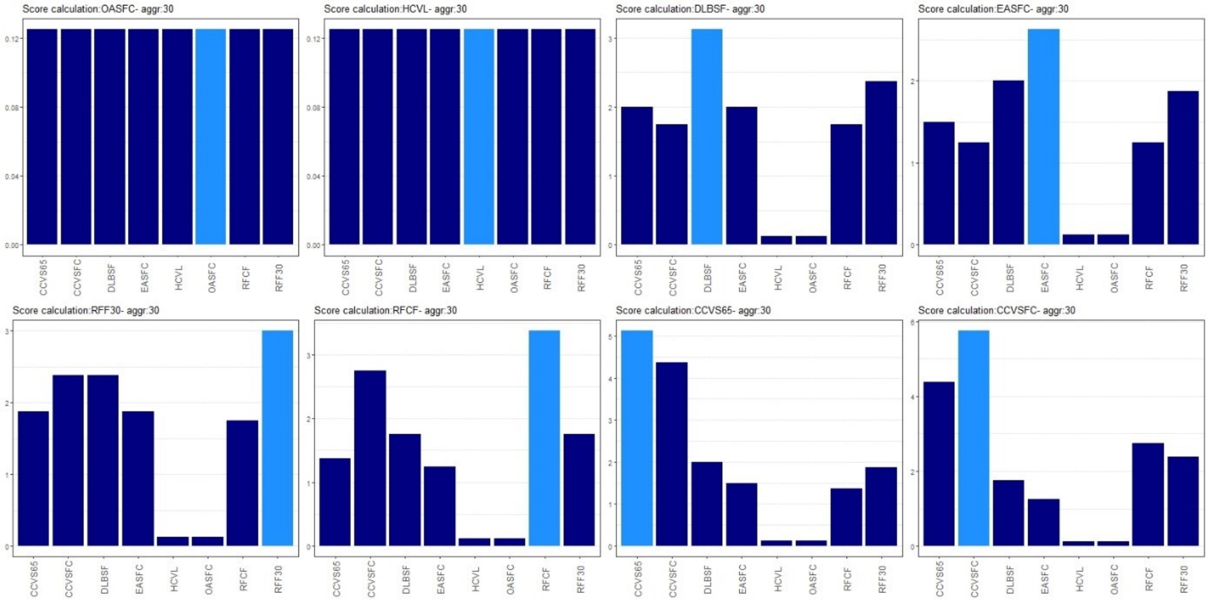


FIG. 18 Shut down of 30 minutes aggregation.