energies

MDPI

Article

METATECH: METeorological Data Analysis for Thermal Energy CHaracterization by Means of Self-Learning Transparent Models

Evelina Di Corso *, Tania Cerquitelli * and Daniele Apiletti *

Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi, 24–10129 Torino, Italy

* Correspondence: evelina.dicorso@polito.it (E.D.C.); tania.cerquitelli@polito.it (T.C.); daniele.apiletti@polito.it (D.A.)

Received: 29 March 2018; Accepted: 17 May 2018; Published: 24 May 2018

Abstract: In the last few years, a large number of smart meters have been deployed in buildings to continuously monitor fine-grained energy consumption. Meteorological data deeply impact energy consumption, and an in-depth analysis of collected and correlated data can uncover interesting and actionable insights to improve the overall energy balance of our communities and to enhance people's awareness of energy wasting. To effectively extract meaningful and interpretable insights from large collections of energy measurements and multi-dimensional meteorological data, innovative data science methodologies should be devised. Research frontiers are addressing self-learning approaches, which allow non-experts to exploit machine learning techniques more easily, and algorithmic transparency of models, hence providing actionable, explicit, declarative knowledge representation. This paper presents METeorological Data Analysis for Thermal Energy CHaracterization (METATECH), a data mining engine based on both exploratory and unsupervised data analytics algorithms, devised to build transparent models correlating weather conditions and energy consumption in buildings. METATECH exploits a joint approach coupling cluster analysis and generalized association rules to allow a deeper yet human-readable understanding of how meteorological data impact heating consumption. First, a partitional clustering algorithm is applied to weather conditions. Then, resulting clusters are characterized by means of generalized association rules, which provide a self-learning explainable model of the most interesting correlations between energy consumption and weather conditions at different granularity levels. The experimental evaluation performed on real datasets demonstrates the effectiveness of the proposed approach in automatically extracting interesting knowledge from data, and provide it transparently to domain experts.

Keywords: data exploration; clustering algorithms; correlation analysis; pattern extraction; energy data; meteorological data; sensor data

1. Introduction

Nowadays large volumes of energy data are continuously collected through a variety of smart meters from different smart-city environments. The analysis of energy-related data collections has received increasing attention from different and cross-research communities, including energy, data mining, databases and statistics communities. These data collections have great potential because an interesting subset of actionable knowledge (e.g., detailed patterns and models to characterize energy consumption at different granularity levels) can be discovered to support the decision-making process of different stakeholders (e.g., energy managers, energy analysts, consumers, building occupants).

Data mining emerged during the late 1980s and focused on studying algorithms to find implicit, previously unknown, and potentially useful information from large volumes of data. Data mining activities include studying correlations among data (e.g., association rules at different levels of abstraction [1]), grouping data with similar properties (e.g., clustering [2]), and extracting information for prediction (e.g., classification, regression [3,4]). The first two classes of algorithms are also known as exploratory methods because they do not require a-priori knowledge (such as the target class to be predicted), thus supporting different and interesting targeted analyses. The exploitation of these approaches on energy-related data is of paramount importance to bring interesting, actionable, and hidden knowledge to the surface. Extracted knowledge items have a great potential to influence the overall energy balance of our communities, in particular by optimizing the building thermal energy consumption, which mainly consists of (i) a static contribution, that is determined by the building structure (e.g., walls, windows, materials, captured by the building energy signature) and appliance energy ratings, and (ii) a dynamic component, that is provided by the usage behaviors and the lifestyle of the people living inside the buildings. With the aim of reducing energy demand, people should be more aware about their building consumption to pursue energy-saving actions. Innovative analytics methodology should be devised to provide interesting and actionable knowledge items about energy consumption in buildings. The knowledge items should be easily interpretable by people to be effectively exploitable.

Furthermore, the influence of multi-dimensional weather data on energy consumption has been condensed into few attributes (e.g., the temperature and humidity) in most existing approaches, due to the complex nature of the full set of meteorological conditions, and the difficulty of automatically identifying the most relevant correlations with many variables. Hence the need to address such correlations with self-learning transparent approaches, which harness the power of complex algorithms to the benefit of energy-domain experts and citizens.

In this paper, Section 2 discusses related works on heating consumption in buildings. Section 3 introduces an overview of the METATECH approach, while a thorough description of its main components is presented in Section 4. An experimental evaluation performed on real data collected in a major Italian city is presented in Section 5. Finally, Section 6 draws conclusions and presents the future development of this work.

2. Related Work

The wide diffusion of smart meters in recent years allows monitoring indoor and outdoor environmental parameters in buildings and collecting huge archives of measures with temporal and spatial references. The analysis of such data collections brings to the facility managers interesting and useful knowledge items to support them in the decision making process. A lot of research activities have been carried out to exploit database management systems, data mining and machine learning techniques, and statistical tools in the field of storage and analysis of energy-related data with different research challenges: (i) identifying the main factors that increase energy consumption (e.g., floors and room orientation [5], location [6]); (ii) supporting data visualization and warning notification [7]; (iii) efficient storing and retrieval operations based on NoSQL databases [8]. Differently from the above research works, this paper proposes a data mining engine to understand thermal energy consumption in buildings by exploiting both supervised and unsupervised algorithms. In [9], the authors analyze the major cause of high energy consumption for air conditioning in indoor space, analyzing the physiological signals (temperature and humidity) within concrete structures. Authors in [10] focus on the cost-recovery of WSNs (Wireless sensor networks) and on the reduction of air conditioning energy consumption in convenience stores. Our main goal is the analysis of residential building thermal energy consumption data enriched with weather condition information. Moreover, data-driven models are also promising in other domains, such as gas utilization ratio (GUR) prediction. To measure the operating status and energy consumption of blast furnaces, the authors in [11] present a soft-sensor approach, i.e., a novel on-line sequential extreme learning machine based on the GUR indicator. Unsupervised techniques are also used for estimating consumption in other environments. Authors in [12] detail two models for estimating small power consumption in office buildings, alongside typical power demand profiles. Both models were tested through a blind validation demonstrating a good correlation between metered data and monthly predictions of energy consumption.

A great deal of research attention has been devoted to characterizing energy consumption profiles among different users [6,13] or buildings [14,15]. The works in [14,15] presented two Big data oriented systems exploiting scalable technologies to compute a variety of key performance indicators (KPIs): basic KPIs in [15] (e.g., energy consumption per unit of volume during specific outdoor conditions) and advanced KPIs in [14] (e.g., inter/intra-building KPIs based on the energy signature that estimates the total heat loss coefficient of a building) have been proposed. Such previous works [14,15] proposed by authors have completely different target and analysis approach, and a substantially different architecture (the only similarity lays in the datawarehouse design). The current work aims at understanding energy consumption in buildings through unsupervised algorithms.

The first implementation of METATECH tailored to energy-related data was first introduced in [16]. The current approach is more focused on capturing multi-dimensional correlations in meteorological data. To this aim, the engine proposed in this work significantly enhances the data analytics techniques proposed in [16], by providing different exploratory algorithms. In particular, (i) METATECH exploits the DBSCAN algorithm to automatically identify the subset of outliers, thus reducing the manual interaction with an expert or the user during the outlier detection phase; (ii) generalized association rules are exploited instead of the traditional-only rules, hence bringing to the surface energy consumption patterns at different abstraction levels; (iii) a different methodology (i.e., the Silhouette-based cohesiveness gain) has been exploited to automatically identify the input parameter of the K-means clustering algorithm (i.e., the desired number of clusters) with respect to the approach presented in [16,17].

Several studies have been made to analyze the electricity system. New technologies such as sensor networks have been incorporated into the management of buildings for organizations and cities. Consumption patterns should be extracted for the purpose of energy and monetary savings. Electricity smart-meter consumption data is enabling utilities to analyze consumption information at unprecedented granularity [18]. The authors in [19] present an interesting analysis related to the reliability of the electricity system. A major cause of the increasing of energy consumption in residential buildings is the growing home comfort. The purpose of the author in [19] is to assess how network reliability and distribution efficiency can be improved through the reduction of building energy consumption. On the other hand, authors in [18] enhance the K-Means clustering performance including time series analysis and wavelets by harvesting inherent structure from the smart meter data.

Rural buildings have been analyzed by authors in [20]. Their studies aim to establish an appropriate strategic plan for promoting rural building energy efficiency by conducting a strength-weakness-opportunity-threat analysis. The authors propose several strategies obtained by the analysis of multiple sources which can contribute to the customization and prioritization of policy recommendations for governments. Moreover, the methodology proposed in [21] extracts electric energy consumption patterns in big-data time series, to draw valuable conclusions for managers and governments. Authors in [22] propose a methodology to determine the appropriate time interval and time length for the analysis, based on the weather characteristics, clustering analysis methods and statistical principles.

Authors in [23] propose a methodology for the study of the envelope airtightness of residential buildings. The paper presents a statistical sampling method to determine the most useful dwellings to be tested, including several variables concerning airtightness (e.g., climate zone, year of construction, and typology). These variables are also being studied by authors in [24] to extract interesting knowledge information on the standard energy performance, thermo-physical and geometrical-related properties of existing buildings at different coarse granularities. Additionally, authors in [25] focus their attention in the analysis of the daily wind patterns and their relational associations with other metocean variables (i.e., oceanographic and meteorological) to capture the seasonal pattern from the hourly observed meteorological covariates. Finally, authors in [26] define a model predictive control (MPC) formulation framework able to critically discuss the outcomes of different existing MPC algorithms for heating ventilation and air conditioning systems.

The huge amounts of data generated by heterogeneous transactions are too many and too complex to be processed and analyzed by traditional methods in several different domains. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. In healthcare, data mining and machine learning are becoming increasingly popular generating information that is very useful to all parties involved in the healthcare industry [27]. Authors in [28] propose learning models to characterize the specificity-determining residue-nucleotide interactions of different known DNA-binding domain families.

Moreover, large volumes of textual data are being collected at an ever increasing rate in various modern applications (e.g., social networks like Twitter, Facebook, e-learning platforms, digital libraries) [17]. Authors in [29] propose a text classification model based on convolutional neural networks for cyber-bullying and hate speeches and observe significant improvements thanks to the proposed 2D TF-IDF features. Authors in [30] proposed a distributed self-tuning engine to analyze and characterize a real crisis tweet collection. Experimental results show the effectiveness of the engine in discovering interesting groups of correlated tweets without selecting neither the algorithms nor their parameters. Moreover, emergency management is a dynamic process conducted under stressful conditions, requiring flexible and rigorous planning, cooperation, and vigilance [31]. All human endeavors involve uncertainty and risk. Risk management has become a vital topic both in academia and practice during the past several decades. Data mining is demonstrated on a financial risk set of data for the basic classification algorithms as presented by the authors in [31]. The authors have demonstrated small-scale application of the basic algorithms. The intent is to make data mining less of a black-box exercise, thus hopefully enabling users to be more intelligent in their application of data mining.

Increasing amounts of data are being collected in all kinds of sports, and automated data analysis has become an important and rapidly developing field [32]. The contribution of the authors in [32] is to build a variety of learning approaches (e.g., deep learning, Bayesian networks, archetypal analysis) focusing on spatio-temporal player trajectories, regularly conducted physiological measurements, or player career data from independently drawn instances.

3. The METATECH Approach

This paper presents a data mining engine, named METATECH (METeorological data Analysis for Thermal Energy CHaracterization), covering the whole analytics work-flow of energy-related data. METATECH analyzes energy data collections enriched with meteorological data through a two-fold methodology based on cluster analysis and generalized association rules to automatically extract and transparently describe energy consumption patterns correlated with meteorological data. The joint approach based on both cluster analysis and generalized association rules allows an efficient characterization of the energy consumption. Specifically, the clustering analysis targets the unsupervised discovery of groups of different thermal energy consumption that occurred with similar weather conditions. Each cluster is then locally characterized by a set of interesting patterns at different granularity levels to summarize the cluster content and to highlight interesting correlations among thermal energy consumption and meteorological conditions. METATECH exploits the K-means algorithm [33] to cluster weather data, jointly with a self-tuning strategy to automatically discover the desired number of groups, while the generalized association rule miner [34] extracts correlations among energy data and meteorological conditions. A categorization of rules into few reference classes according to their meaning is proposed to ease the manual inspection of the results and their understanding. The model of the data is transparent as it consists of rules, in the form of correlations among different attribute values, which are directly readable by humans. The full process is designed to self-learn from the data how to proceed at each step, by tuning parameters, partitioning the data, and identifying the most relevant rules among the full set of correlations that exist in the data.

Extracted knowledge items can support energy managers in the decision-making process, for example through the definition of proper strategies to efficiently satisfy the energy demand for different buildings. Furthermore, extracted knowledge items can enhance people's (consumers and building occupants) awareness of energy consumption and plan ad-hoc strategies to reduce the building consumption during some critical time slots (e.g., energy peak demand) or when rooms are empty.

The main novelties of METATECH are twofold. (1) It is a self-learning joint approach, based on both cluster analysis and generalized association rules, able to automatically extract interesting knowledge patterns and make them easily interpretable to characterize thermal energy consumption. In particular, the model self-learns, i.e., automatically infers from data the patterns and their correlations, without prior knowledge and with limited user interaction, thanks also to the automatic tuning strategies of the algorithm parameters. (2) It analyzes real-world data collected in a heating system available in a major Italian city and presents experimental results of interest for domain experts.

Figure 1 shows the overall architecture of the METATECH system, and presents the whole data analytics work-flow, from input data sources to result presentation. METATECH includes four main components, named *Data collection and integration*, *Data preprocessing*, *Knowledge extraction* and *Knowledge visualization*.

These components are briefly described below and a more detailed description is given in Section 4. In METATECH the *Data collection and integration* component stores measurements from sensors as they asynchronously arrive, and it is in charge of their temporal synchronizations and aggregation. For the purpose of the analysis, these data are enriched with spatial and temporal information at different abstraction levels. The enriched dataset is stored in a datawarehouse as proposed in [15]. Different phases of *Data preprocessing* are then performed to prepare data for the subsequent analysis. The *Knowledge extraction* component exploits a joint approach based on both clustering and generalized association rule mining to automatically identify and describe the patterns in the data.

Lastly, the *Knowledge visualization* component presents the results with special attention to highlighting the rationale of the extracted patterns and to make them easily interpretable by people and effectively exploitable.

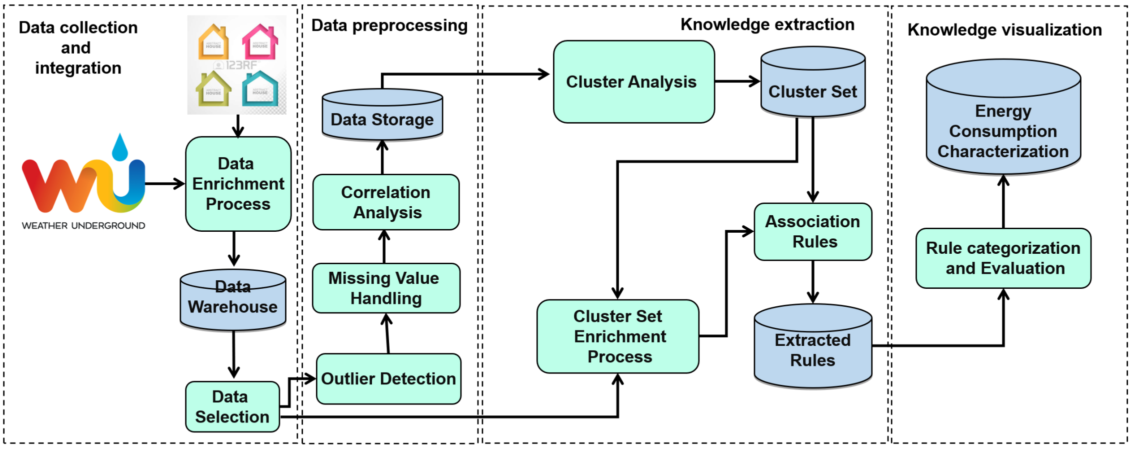


Figure 1. The METATECH system architecture, with its main building blocks.

4. The METATECH Components

METATECH is a data analytics engine aimed at characterizing correlations between meteorological data and energy consumption. The analysis process is applied on data as modeled in [15].

The METATECH components addressing the different phases of the analysis process are described in the following subsections.

4.1. Data Collection and Integration

The *Data collection and integration* component collects energy consumption measurements as they asynchronously arrive from sensors, then aggregates data at hourly intervals. These data are enriched with temporal information at different granularity levels as well as with various meteorological conditions available from open-data sources.

In our case study, energy measurements are sampled every 5 min from a large number of smart meters deployed in a major Italian city, and meteorological data was collected from the Weather Underground service [35], which gathers data from Personal Weather Stations (PWS) registered by users. In Turin, a major Italian city where buildings providing energy data are located, tens of PWSs are present. Weather data associated with a specific building are computed as a distance-based weighted mean of the values provided by the three nearest PWSs. The weight is inversely proportional to the distance from the PWS to the building location, hence three equally distant PWSs would have the same weight in determining the outdoor values of a given building. When a high concentration of PWSs is available, they reasonably reflect the real conditions registered in their precise neighborhood, as opposed to other services providing more precise values, but related to a much wider area.

4.2. Data Preprocessing

Extracting actionable knowledge from data is a multi-step process. The knowledge extraction phase is preceded by a preprocessing phase, which aims to smooth the effect of possibly unreliable measurements. Preprocessing entails the following steps: (i) *outlier detection and removal*, (ii) *missing value handling*, and (iii) *correlation analysis*.

Outlier detection and removal. An outlier is an observation that lies outside the expected range of values. It may occur either when a measurement does not fit the model under study or when an error in measurement happens (e.g., faulty sensors may provide unacceptable measurements of thermal energy consumption). For identifying and removing outliers, METATECH exploits a clustering algorithm based on the density concept, named DBSCAN (Density-Based Spatial Clustering of Application with Noise) [36]. DBSCAN allows the detection of clusters of arbitrary shape, and the automatic identification of noise points and outliers. Specifically, DBSCAN detects clusters on the basis of a density reachability concept, where clusters are defined as higher-density regions separated by lower-density regions. DBSCAN needs two parameters to be provided, the minimum number of nearby points ($MinPts$) and a distance: the epsilon radius (Eps). Each data point is either marked as (i) *core point*, or (ii) *border point*, or (iii) *noise point*. A core point has more than $MinPts$ within Eps. A border point has less than $MinPts$ within Eps, but is in the neighborhood of a core point (the epsilon radius determines the neighborhood distance). All other points are noise points.

Missing value handling is an important and crucial step that significantly impacts the mining process. Since our aim is the characterization of energy consumption, we disregarded data records with missing consumption values. Instead, for meteorological data, METATECH exploits two strategies to handle missing values: (i) replacement with the daily average value or (ii) replacement with the hourly average value computed at the same time in the previous week. The choice is mainly determined by the physical meaning of each considered attribute: case (i) is applied to the precipitation and wind direction attributes, while case (ii) is applied to the solar radiation and UV index attributes.

Correlation analysis. Couples of strongly correlated attributes provide no additional contribution to the analysis process. Hence, to reduce the space and time complexity of data mining algorithms, we remove one of each pair of correlated meteorological attributes before executing the analysis. METATECH leverages the correlation matrix [37] to analyze the dependence between multiple variables at the same time. For each pair of attributes (X,Y), METATECH computes the correlation coefficient through the Pearson correlation defined as $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$, where $\mathrm{cov}(X,Y)$ is the covariance between X and Y, σ_X is the standard deviation of X and analogously σ_Y for Y. The higher the absolute coefficient value is, the stronger the correlation becomes: for each couple of correlated attributes, whose value is higher than a threshold (0.9 in our use case), METATECH we remove one of them. Since we aim at characterizing energy consumption, to select the meteorological attribute to be removed between the two, its correlation with the energy consumption data is exploited and the most correlated is kept.

4.3. Knowledge Extraction

To extract meaningful and interesting knowledge from data while maintaining the number of extracted results within manageable limits, METATECH must be able to automatically identify the most interesting subsets of the input data, so that the specific results can be manually evaluated by a human domain expert. Selecting specific subsets from which interesting knowledge can be independently derived is of paramount importance to bring hidden knowledge to the surface. For this purpose, METATECH exploits a clustering algorithm to identify specific data subsets from which interesting data correlations can be discovered. Specifically, since energy consumption is strongly influenced by weather conditions, the identification of energy consumption records that occurred with similar weather conditions reduces both the complexity of the correlation analysis and the cardinality of the extracted knowledge to be manually validated. METATECH uses a clustering algorithm to partition weather data into relevant subsets. Before the clustering phase, the dataset is normalized with the range transformation (0, 1). Each cluster is then locally characterized by a set of association rules to model the most interesting correlations among weather data and energy consumption. METATECH also includes a categorization of the extracted rules according to specific templates, to ease manual interpretation by domain experts, and to drive the knowledge extraction. The template-driven extraction of association rules builds a so-called transparent model of the data, which is directly readable by humans, easily interpretable, and actionable, differently from black-box models such as Artificial Neural Networks.

4.3.1. Cluster Analysis

Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that objects within the same group are similar to each other and different from the objects in other groups. The greater the similarity within a group, the better the clustering result [4]. METATECH computes the similarity between two objects by using the Euclidean distance, and integrates a partitional algorithm, the *K-means* algorithm [33], which divides the input dataset into K non-overlapping subsets (i.e., clusters) such that each data object is in exactly one subset. The procedure is to randomly define K centroids, one for each cluster. Then each point is iteratively associated to the nearest centroid. Next, the centroids are updated. These steps are repeated by the algorithm until the centroids do not move any further.

Even if K-means identifies the clusters in a limited computational time by producing a quite good cluster set, it requires the number of clusters to be specified in advance, which is one of its main drawbacks. To address this issue, METATECH automatically tests several configurations by varying the input parameter K of the algorithm (i.e., number of desired clusters). These solutions are then compared through the analysis of Silhouette-based indices to measure the cohesion and separation of each cluster set. METATECH includes a variation of the standard Silhouette index [38] to evaluate the quality of the discovered cluster set, which is presented in [17]. This variation is the weighted distribution of the silhouette index (WS). The Silhouette index measures both intra-cluster cohesion and inter-cluster separation by evaluating the appropriateness of the assignment of a meteorological measurement to one cluster rather than to another. It assumes values in $[-1; 1]$. Negative and positive Silhouette values represent wrong and good record placements, respectively: the higher the index, the better the clustering. However, smaller values of K reduce the probability of error. Instead, the WS index (assuming values in [0; 1]) represents the percentage of meteorological records in each positive bin properly weighted with an integer value w ∈ [1; 10] (the highest weight is associated with the bin [1 −0.9] and so on) and normalized within the sum of weights. The higher the weighted silhouette index, the better the identified partition. In this study, METATECH includes a new index to measure the *cohesiveness gain* obtained using the weighted Silhouette with respect to the Silhouette index. The cohesiveness gain is represented by means of two values: (i) the *ratio* and (ii) the *delta* between the weighted and the standard Silhouettes. METATECH chooses the clustering characterized by the highest value of cohesiveness gain, since it represents a good trade-off between the number of selected clusters and the values of both Silhouettes.

4.3.2. Association Rules

Association rule extraction [39] is one of the most powerful exploratory techniques in data mining, it aims at finding interesting relationships among data. Since clusters are anonymous groups of records of energy consumption that occurred with similar weather data, METATECH characterizes each cluster with a set of relevant patterns, i.e., association rules, able to summarize interesting correlations. Such approach improves the understanding of the analysis results.

An association rule is expressed in the form $X \rightarrow Y$, where X and Y are disjoint and non-empty itemsets, i.e., $X \cap Y = \emptyset$. X is also called rule antecedent or rule body and Y rule consequent or rule head. Typical correlations involve energy consumption, meteorological data (e.g., wind direction, UV index) and temporal data (e.g., daily time slot).

Association rule mining requires a transactional dataset of categorical attributes. A transactional dataset $\mathcal{D}$ is a set of transactions in which each one is a set of items (also called itemset). To this aim, a discretization step is applied to convert the original continuously-valued measurements into categorical bins of a transactional dataset.

METATECH includes a two-fold characterization: (i) fine-grained correlations based on traditional association rules, and (ii) high-level correlations based on generalized association rules.

Traditional association rule mining extracts fine-grained correlations because its results are recurring patterns (i.e., rules) among specific categorical values. Such level of details provides clear advantages when the phenomena under exam exhibit patterns with specific values, but presents the drawbacks of depending on the discretization bins and to limit the abstraction capabilities. For this reason, a second characterization block exploits generalized association rule mining to capture higher-level correlations.

Generalized association rule mining [34] is an exploratory data mining technique that has been largely used to extract hidden correlations at different granularity levels. To introduce the concept of generalized association rules, we first recall the notion of generalized itemset. A generalized itemset is a set of generalized items (attribute = *generalized value*) where *generalized value* is defined through a taxonomy. A taxonomy is a forest of generalization trees, each one representing a hierarchy of aggregations defined on an attribute domain. Traditional (non-generalized) itemsets are a special case of generalized itemsets in which all items assume values in the lowest levels of the corresponding taxonomy (i.e., leaves of the generalization trees).

Figure 2 shows an example of the generalization tree for the temperature attribute. Leaf nodes are labeled with values in the temperature attribute domain (after a discretization step), while non-leaf nodes are aggregations of lower nodes, up to the root which represents all values in the domain.

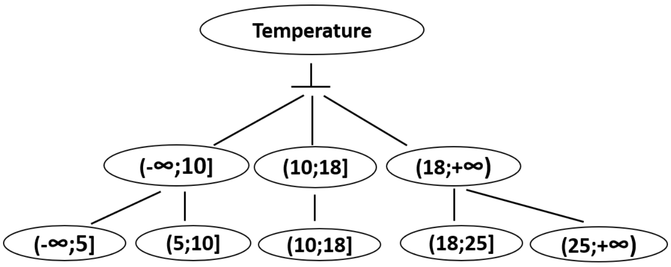


Figure 2. Example of a generalization tree for the temperature attribute.

4.4. Association Rule Categorization

Even if rules are a human readable result, experience shows that experts from application domains, such as energy, benefit from semantic frameworks that drive them into interpreting the results. To this aim, the traditional approach is to sort the rules according to an interestingness score. In METATECH, resulting rules are sorted by score and additionally grouped together by their meaning. The meaning of a rule is determined by its template, which includes specific attributes characterizing the data. Two templates currently provided by METATECH and presented in this paper stem from our experience and experimental results. They are reported in Table 1. More templates can be easily added to address specific needs and questions by domain experts.

The first template, at row $T1$, highlights the most peculiar weather conditions characterizing each cluster. Hence, the rule body must contain the cluster id, whereas the attributes considered in the rule head are all those describing the weather. No other attributes must be present, and only rules of length 2 are considered, so only a single specific meteorological attribute is highlighted in each rule head, to keep it focused. An example of $T1$ is $\{cluster = Cluster\_4\} \Rightarrow \{temperature = warm\}$. It means that the Cluster_4 is characterized by warm temperatures.

Template $T2$ in Table 1 models temporal periods, weather conditions, and energy consumption. In particular, we noticed that rules identified as interesting by domain experts often correlate energy consumption with weather conditions in specific time periods, which can be as short as intra-day periods, or as long as weeks. To capture such richness, this template targets rules having in the body both a daily time slot and a fortnight (14-day period), together with any number of attributes describing weather conditions. The rule head, instead, must contain only the energy consumption level, which is the end goal of the analysis. An example of $T3$ is $\{fortnight$ = 16–31 $December,\ daily\ time\ slot = Midday,\ UV\ index = minimum,\ precipitation = no\ rain,\ humidity = very\ high,\ temperature = very\ cold,\ wind\ direction = North\} \Rightarrow \{energy\ consumption\ level = very\ high\}$. It means that in the period from 16 to 31 December, in the given context of weather conditions during the day, a high energy consumption occurred. Very cold temperatures and high humidity make the body feel a greater sense of cold and then physical discomfort, and the winds from North are strong and cold.

Table 1. Association rule templates included in METATECH and their interpretation.

TId	Question	Rule Template
T1	What are the most specific weather conditions characterizing each cluster?	$\{cluster\} \Rightarrow$ {weather condition}
T2	Given a fortnight and a daily time slot, what kind of consumption level characterizes them under different weather conditions?	$\{fortnight,\ daily\ time\ slot$, weather conditions} $\Rightarrow$ {consumption level}

Association Rule Evaluation

To rank the most interesting rules, METATECH uses three quality scores named *support, confidence* and *lift*.

The *rule support* is the percentage of records containing both X and Y. It represents the prior probability and the support condition of transaction $X \cup Y$ is defined as $s(X \cup Y)/N$ of $X \cup Y$, where $s(X \cup Y)$ is the observed frequency in the full dataset and N is the total number of transactions. The rule support is an indication of how frequently the itemset appears in the dataset.

The *rule confidence* is the conditional probability that the consequent Y is true under the condition of the antecedent X. It is computed as $c(X \rightarrow Y) = s(X \cup Y)/s(X)$. Given a set of transactions $\mathcal{D}$, METATECH finds all the rules having support $\geq minsup$ and confidence $\geq minconf$, where $minsup$ and $minconf$ are the corresponding support and confidence thresholds that are user-specified parameters. The rule confidence is an indication of how often the rule has been found to be true.

High-confidence rules can sometimes be misleading because the confidence measure ignores the support of the itemset appearing in the rule consequent. A way to address this problem is the analysis of the lift value. The *lift index* [4] is defined as $lift(X \rightarrow Y) = s(X \cup Y)/(s(X) \cdot s(Y))$, which computes the ratio between the rule's confidence and the support of the itemset in the rule consequent. Lift measures how many times more often X and Y occur together than expected if they were statistically independent. A lift ratio larger than 1.0 implies that the relationship between the antecedent and the consequent is more significant than would be expected if the two sets were independent. The larger the lift ratio, the stronger the association.

5. Experimental Results

We performed an experimental meteorological data analysis for thermal energy characterization on a real dataset, including energy consumption of 15 residential buildings, using the METATECHẽngine. We considered energy data related to a complete winter period from 15 October 2014 to 15 April 2015 because, in Italy, central heating systems are operated only in such period, hence dates outside this range were not considered as they would collect only zero consumption values. Data collected through the energy smart meters are integrated with meteorological information collected from the Weather Underground web service [35], which gathers data from Personal Weather Stations (PWS) registered by users. Experiments addressed the following issues:

1. Feature-correlation analysis (Section 5.2)
2. Thermal energy characterization in terms of data distribution (Section 5.3)
3. Cluster characterization in terms of data distribution within each cluster (Section 5.4)
4. Cluster characterization in terms of association rules (Section 5.5)
5. Knowledge visualization (Section 5.6)

Since data collected from sensors are expected to be dirty, collected measurements are analyzed through the DBSCAN algorithm, which is able to automatically identify outliers. Outliers are often considered noise points when proper density parameters are set. To select the algorithm parameters (i.e., Eps and $MinPoints$) the k-distance plot has been analyzed. We performed many runs with varying values of k (i.e., $MinPoints$ parameter) between 2 and 20. We noticed that the resulting curve was very similar with values k = 12 and k = 13. For both plots we have looked for the knee. The Y-axis value in which the knee is formed corresponds to a good Eps value for that particular $MinPoints$ value. If Eps is chosen too small, a large part of the data will not be clustered; whereas for a too high value of Eps, clusters will merge and the majority of objects will be in the same cluster. METATECH sets as good possible configuration for the DBSCAN algorithm $MinPoints$ = 12 and Eps = 0.2.

To address the problem of centroids initialization for the K-means algorithm, we randomly chose the initial centroids.The K-means algorithm requires the number of clusters (K) as input parameter. Generally, this is a difficult parameter value to choose, given the wide range in which it may vary. To address this issue, METATECH automatically performs many runs of the algorithm with varying values of K. For each run, the cluster set is evaluated by computing the standard and weighted Silhouettes. Table 2 shows the cohesiveness gain obtained for each run of the algorithm. Finally, METATECH selects the best value for the parameter, which is $K = 4$ in our experiments, providing the maximization of the cohesiveness gain.

Also association rule extraction requires parameters to be defined. However, these are less important, since they act as a pruning factor of the result set, by automatically removing the less important rules based on their value of support, confidence and lift. Hence, the best values of these parameters depend on how much time and resources domain experts can devote to the manual inspection of the rules. Based on the experimental results, the following parameter settings have been used as the reference default configuration for METATECH:

- Number of clusters $K = 4$,
- Minimum confidence value $minconf = 10\%$,
- Minimum support value $minsup = 0.1\%$,
- Minimum lift value $minlift = 1.1$.

The *minsup* parameter has been intentionally set to a low value to avoid missing relevant correlations. The resulting number of rules is then very large (e.g., more than 20 thousands), and the most interesting ones have been selected according to decreasing lift. Support, confidence, and lift have been computed on the overall dataset when characterizing the clusters (in Table 5), and on the subset of records of each cluster to identify interesting correlations within each group (in Table 6).

The Java-based RapidMiner toolkit [40] has been used for correlation analysis, cluster analysis, and association rule extraction, while the data distribution analysis has been performed using MATLAB [41]. Experiments performed on a 2.66-GHz Intel(R) Core(TM)2 Quad PC with 8 GBytes of main memory and Linux Ubuntu 14.04 yielded to an average execution time of 48 s, considering the complete work flow, from clustering to rule extraction and ranking, over the whole season of data.

Table 2. Cohesiveness gain trend of the clustering results for different values of the K parameter (number of clusters).

			Cohesiveness Gain	
K	**Standard Silhouette**	**Weighted Silhouette**	**Ratio**	**Delta (%)**
3	0.40	0.45	1.13	5.39
4	0.37	0.44	**1.17**	**6.34**
5	0.34	0.39	1.15	5.14
6	0.33	0.38	1.14	4.83
7	0.32	0.36	1.15	4.80
8	0.32	0.37	1.15	4.88
9	0.32	0.37	1.15	4.78
10	0.32	0.36	1.13	4.35

5.1. Data Description

To address the temporal analysis of the thermal energy consumption, each record, whose raw version has only a timestamp, is enriched with the following information:

- Date, holiday (yes or no), week of the year (1–52), month (1–12), 2-month, 3-month, 4-month, 6-month periods;
- Daily time slots, i.e., morning [4–8], midday [9–13], afternoon [14–17], evening [18–22]; during the night, from 22 to 4, the heating system is switched off in the buildings under study.

Weather attributes and their corresponding units of measure, including a briefly description are listed in Table 3; their mean values are collected once every hour. Summarizing, each building is characterized by 10 attributes.

Table 3. Weather data features included in the experimental dataset.

Attributes	Unit of Measurement	Description
Air Temperature	°C	mean hourly temperature, provided by PWS (Personal Weather Stations)
Outdoor Temperature	°C	mean hourly temperature, provided by a sensor on the roof of buildings
Precipitation	mm	mean hourly value of precipitation
Wind Direction	azimuth	mean hourly value of Wind Direction
Solar Radiation	W/m^2	mean hourly value of Solar Radiation
UV index	-	mean hourly value of UV index
Humidity	percentage	mean hourly value of Humidity
Pressure	hPa	mean hourly value of pressure

5.2. *Feature-Correlation Analysis*

METATECH exploits the correlation matrix to analyze the dependence between multiple variables at the same time. The correlation matrix shown in Table 4 contains the correlation coefficients between each couple of attributes computed as discussed in Section 4.2. This matrix is symmetric (i.e., the correlation of column *i* with column *j* is the same as the correlation of column *j* with column *i*), and its generic element (i, j) models the correlation between the attribute in row *i* and the one in column *j*. Correlation coefficients always lie in the range $[-1, 1]$. A positive value ($[0, 1]$) implies a positive correlation between attributes *i* and *j*. Thus, large (small) values of attribute *i* tend to be associated with large (small) values of attribute *j*. A negative value ($[-1, 0]$) means a negative or inverse association. In this case, large values of *i* tend to be associated with small values of *j* and vice versa. A value near 0 indicates weakly correlated or uncorrelated data.

Table 4. Correlation matrix among weather data features.

Attributes	Air Temperature	Outdoor Temperature	Precipitation	Wind Direction	Solar Radiation	UV Index	Humidity	Pressure
air temperature	1	**0.97**	−0.06	−0.03	0.48	0.48	−0.49	−0.01
outdoor temperature	**0.97**	1	−0.03	−0.01	0.41	0.4	−0.46	−0.03
precipitation	−0.06	−0.03	1	0.08	−0.07	−0.06	0.15	−0.06
wind direction	−0.03	−0.01	0.08	1	0.02	0.01	−0.08	−0.12
solar radiation	0.48	0.41	−0.07	0.02	1	**0.91**	−0.4	0.06
UV index	0.48	0.4	−0.06	0.01	**0.91**	1	−0.42	0.04
humidity	−0.49	−0.46	0.15	−0.08	−0.49	−0.42	1	−0.07
pressure	−0.01	−0.03	−0.06	−0.12	0.06	0.04	−0.07	1

The matrix shown in Table 4 highlights two strong correlations, i.e., whose value is above the experimental threshold set at 0.9: (1) a positive and strong correlation (0.97) between *air temperature*, i.e., the mean external temperature monitored through PWS, and *outdoor temperature* monitored through a sensor deployed on the roof of the building; (2) a very high correlation (0.91) between *UV index* and *Solar Radiation*.

Since highly correlated attributes are similar in behavior, for each couple of attributes highlighted in the matrix, the attribute which is less correlated with the thermal energy consumption is removed from the analysis to reduce both the computational cost and the cardinality of the extracted knowledge. Based on the above results, we do not consider *Outdoor Temperature* and *Solar Radiation* in the subsequent analysis process.

5.3. Thermal Energy Consumption Distribution

To describe the thermal energy consumption behavior during the full winter period, we exploit the histogram of the response variable. To construct the histogram, the first step is to bin the range of values (i.e., to divide the entire range of values into a series of intervals) and then count how many values fall into each interval. The bins are usually specified as consecutive and non-overlapping intervals of a variable. We divided the entire interval into 10 bins to analyze the deciles.

In Figure 3 we reported the histogram and the cumulative distribution. The histogram shows a skewed distribution to the right, i.e., it is positively skewed. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side). Most of the values (90%) are below 12.23 KW/m^3, i.e., within the 4th decile, and almost 70% of the values are below 8.73 KW/m^3, i.e., within the 3rd decile.

METATECH provides such basic quantitative information to domain experts to allow a better understanding of the next analysis steps. In particular, as discussed in Sections 4.4 and 5.5, values of thermal energy consumption will be discretized into bins to allow association rule extraction.

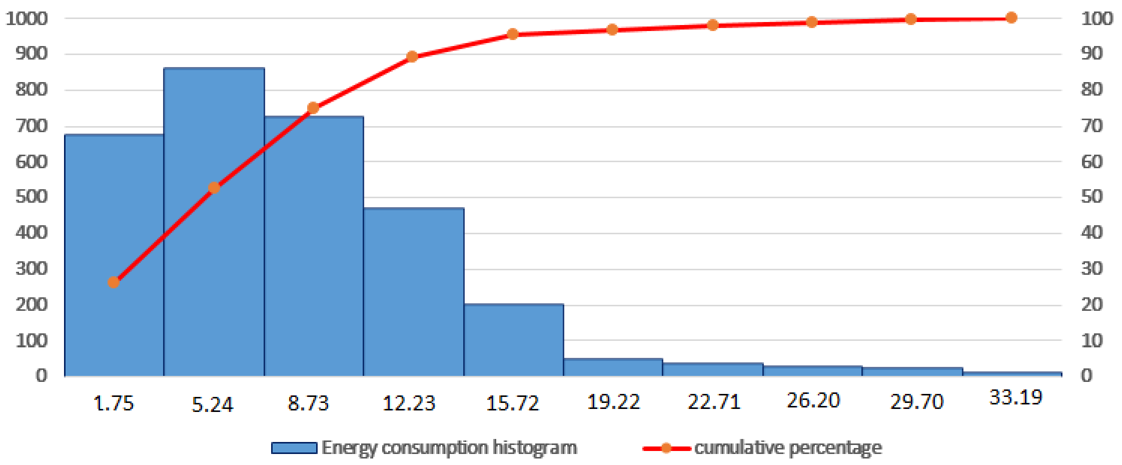


Figure 3. Histogram and cumulative distribution of thermal energy consumption.

5.4. Cluster Characterization

The cluster analysis is exploited by METATECH to identify energy consumption patterns occurred in similar meteorological conditions. The K-Means clustering algorithm has been applied to meteorological data related to a complete winter period. METATECH supports domain experts in capturing the rationale of the clustering results by exploiting two representations: (i) the singular value decomposition (SVD) [4] to show the clustered points in a graphical and friendly two-dimensional space; (ii) an attribute-based box-plot comparison, to better understand the distribution of the attribute values characterizing each cluster.

SVD is a matrix factorization method that factorizes the input data matrix into three matrices. It can be easily exploited to reduce the data dimensions by only considering the most representative attributes. Figure 4 shows the SVD decomposition of the cluster set discovered by K-means with K = 4.

All clusters are well-separated, and indeed K-means was able to identify a good partition of records that occurred with similar meteorological conditions.

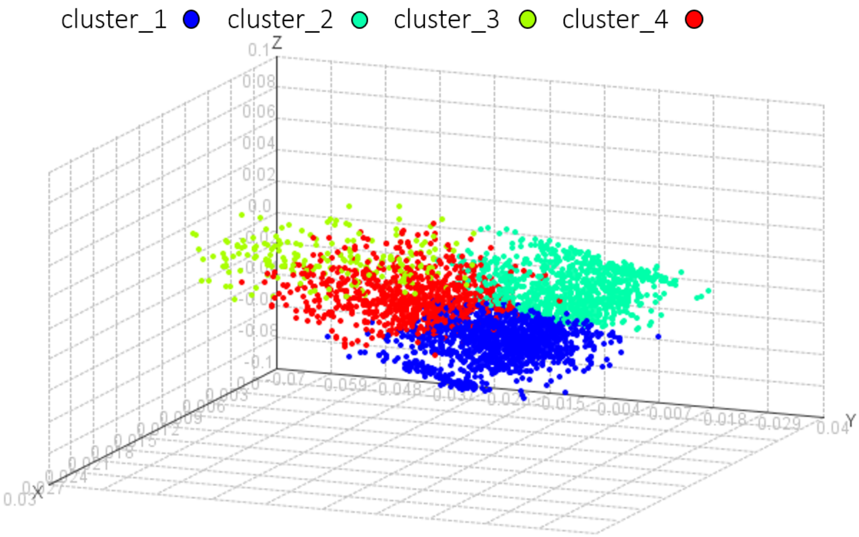


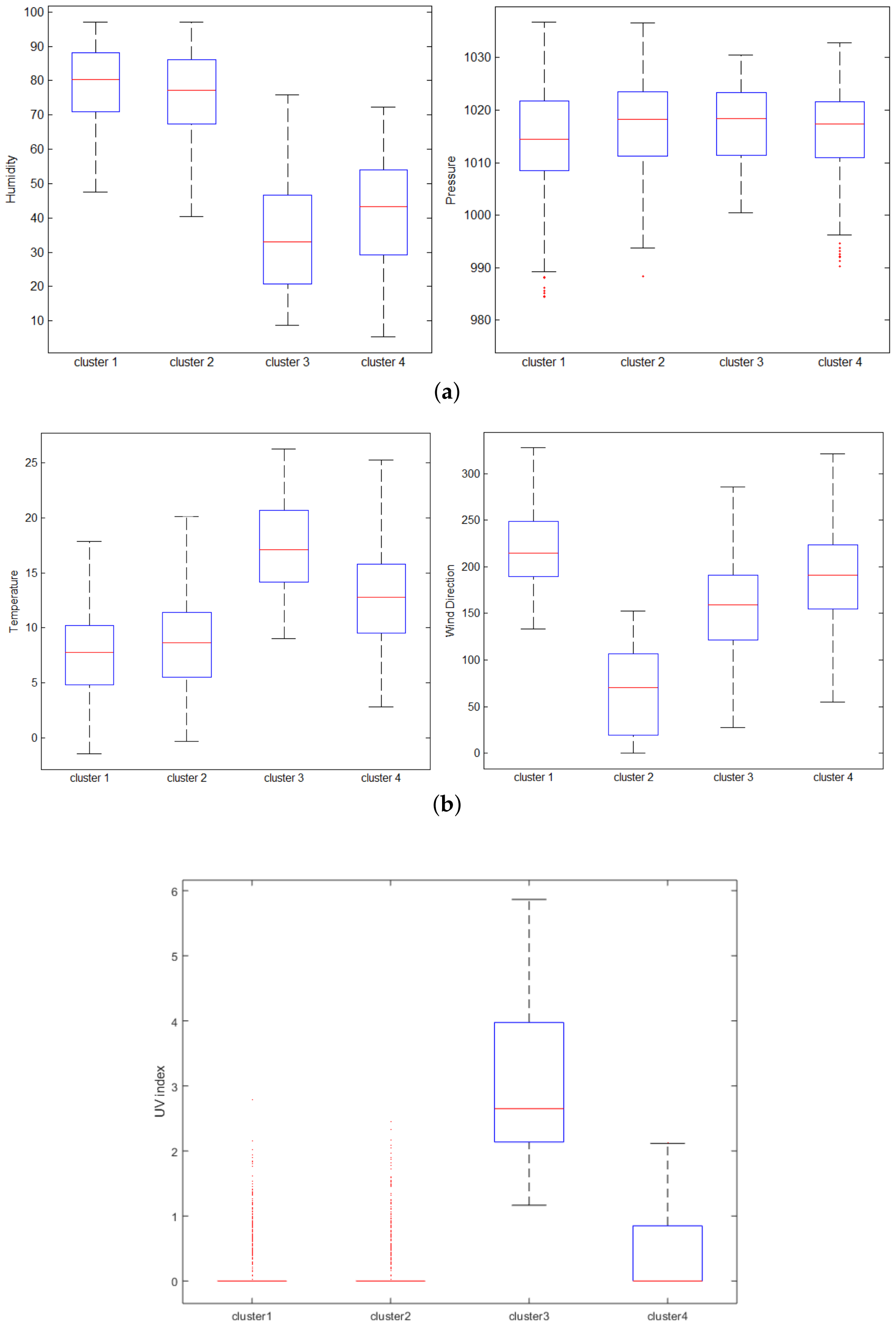
Figure 4. Cluster set representation through SVD, where each point is represented with the color of the cluster it belongs to.

Figure 5 compares the value distributions of the meteorological attributes to characterize the clustering results through the boxplot analysis [42]. In more detail, Figure 5a (left) shows the *humidity* distribution in the four clusters. Cluster$_1$ and Cluster$_2$ have high median values and are characterized by positive skewness. Cluster$_3$ and Cluster$_4$ have low median values, with the former exhibiting negative skewness. In case of positive skewness, more observations with lower values are present, while in the case of negative skewness, more observations fall in correspondence of the highest values. For instance, considering Cluster$_1$ and Cluster$_2$ that have a negative skewness, $(Q_3 - Me) < (Me - Q_1)$, where Me is the median, Q_1 the first quartile and Q_3 the third quartile.

Figure 5a (right) shows the *pressure* distribution. All clusters exhibit a similar behavior in terms of both skewness and median values, hence the pressure is not a characterizing attribute for the clustering result. Figure 5b (right) shows the *wind direction* distribution separately for each cluster. With respect to the humidity distribution, Cluster$_1$ and Cluster$_3$ are characterized by positive skewness. Instead Cluster$_2$ is characterized by negative skewness, while Cluster$_4$ is almost symmetric. In more details, Cluster$_1$ has 212.5 as median value, a value that is related to winds that blow from the South-West. Half of records of Cluster$_1$ fall within the range [187.5, 250.0], corresponding to winds that blow from South-East, East and South-West.

Overall, on a per cluster basis, we can see that each cluster is characterized by specific ranges of values for different variables. For instance, Cluster$_3$ shows low humidity and high temperature values, whereas pressure is not a characterizing feature of the cluster. On the contrary, Cluster$_1$ exhibits high humidity, low temperature, and "high" wind direction values.

The current cluster characterization, as provided by the boxplots, is coarse. However, it is provided as a support for the following association-rule extraction experiment (Section 5.5). Association rules and their corresponding quality metrics allow to describe not only the information provided by boxplots, but also deeper insights on the data, in a more human-readable fashion, also for non-expert end-users.



(**a**)

(**b**)

(**c**)

Figure 5. (**a**) (**left**) humidity distribution and (**right**) pressure distribution; (**b**) (**left**) air temperature distribution and (**right**) wind direction distribution; (**c**) uv index distribution. Characterization of the cluster set through the box-plot distribution of the attribute values.

5.5. Analysis of Extracted Patterns at Different Abstraction Levels

In this subsection we discuss the most interesting correlation patterns found by METATECH, in the form of association rules. Since association rule mining requires a transactional dataset of categorical values, METATECH performs the discretization of continuously-valued measurements to obtain categorical bins.

In our case study, the knowledge discovery process is driven by a taxonomy. The taxonomy in the context of the association rule mining is called generalization tree, since it allows rules to be generalized. Discretization bin values are provided by a domain expert, so that they are based on their meaning in the energy and meteorological contexts, as described in the following.

(1) *Energy consumption per unit of volume* (denoted as consumption level): two bins until 5.5 KW/m^3 (off until 0.05 KW/m^3, low until 5.5 KW/m^3), a bin each 10 KW/m^3 for values until 25.5 (medium consumption until 15.5, high consumption until 25.5) and an additional bin for values exceeding 25.5 KW/m^3 (very high). Thus, the corresponding generalization tree includes 5 leaf values ([0.0, 0.05] (0.05, 5.5] (5.5, 15.5] (15.5, 25.5] (25.5, +∞)), each one associated to a range of non-overlapping values. The tree also includes an intermediate level with three aggregate values (i.e., [0, 5.5] (5.5, 15.5] (15.5, +∞)) and the root including all values in the corresponding domain.

(2) *Humidity*: a bin each 20% from 0 to 100% (i.e., very low until 20%, low until 40%, medium until 60%, high until 80% and very high until 100%). The corresponding generalization tree includes 5 leaf values ([0.0, 0.20] (0.20, 0.40] (0.40, 0.60] (0.60, 0.80] (0.80, 1.0]) and the root (representing all values).

(3) *Temperature*: values are discretized in five bins (very cold up to 5 °C, cold up to 10 °C, mild up to 18 °C, warm up to 25 °C, hot up for higher values). The corresponding generalization tree includes 5 leaf values ((−∞, 5] (5, 10] (10, 18] (18, 25] (25, +∞)), an intermediate level with values (−∞, 10], (10, 18], and (18, +∞)), and the root including all values in the corresponding domain.

(4) *Temporal data*: the timestamp is aggregated into the corresponding *daily time slot* (e.g., morning, midday, afternoon, evening). Each day is classified as holiday or working, and aggregated in week, fortnight, month, 2-month, 3-month, 4-month and 6-month periods.

(5) *Meteorological measurements* have been discretized based on the criteria available in [43–46]: precipitation level values and wind direction have been categorized in eight leaf values each, UV index in six leaf values, and atmospheric pressure in two leaf values.

From experimental experience, to avoid pruning interesting correlations with low confidence but high lift, recommended values of support and confidence thresholds for association rule mining in the current context are 0.1% and 1% respectively. Moreover, we also recommend a minimum lift threshold equal to 1.1 to prune both negatively correlated and uncorrelated item combinations.

5.5.1. Fine-Grained Association Rule Extraction

This section presents the most interesting correlations in the form of traditional (fine-grained) association rules. To this aim, the rule templates presented in Section 4.4 are exploited.

Table 5 shows the top-three rules, sorted by descending lift, characterizing each cluster according to the first template. Support, confidence, and lift are computed on the overall dataset, as the cluster is a feature of the dataset itself. Rules $R_1 - R_{12}$ identify the most representative meteorological items in each cluster.

Rules describe Cluster$_1$ as the group modeling "bad" weather data (drizzling, cloudy, low UV index), Cluster$_2$ has cold humid measurements, Cluster$_3$ warm sunny days, and Cluster$_4$ mild dry ones. The characterization of the clusters by means of the rules provides insights that from a boxplot would be hard to spot. For instance, Cluster$_1$ from the boxplot seems to have zero UV index as main value. However, the proportion of zero UV index records in Cluster$_1$ is lower than the overall presence in the dataset. Hence, Cluster$_1$ is actually characterized by the minimum UV index instead of the zero value, because minimum UV index values are more present in Cluster$_1$ than in other clusters. Such information is provided by the lift quality index, which is above 1.0, specifically in rule R_3.

These weather items are subsequently combined with other meteorological items to characterize each cluster in more detail through the second template.

Table 5. Fine-grained traditional rules according to the first template.

RId	Rule	Supp %	Conf %	Lift
R_1	{cluster = Cluster$_1$} ⇒ {Precipitations = drizzling}	8.1	20.5	1.8
R_2	{cluster = Cluster$_1$} ⇒ {Pressure = low}	17.4	43.9	1.2
R_3	{cluster = Cluster$_1$ } ⇒ {UV index = minimum}	37.5	94.6	1.1
R_4	{cluster = Cluster$_2$ } ⇒ {Humidity = high}	13.2	45.3	1.3
R_5	{cluster = Cluster$_2$ } ⇒ {Precipitations = no rain}	26.1	89.2	1.1
R_6	{cluster = Cluster$_2$ } ⇒ {Temperature = cold}	12.3	42.1	1.1
R_7	{cluster = Cluster$_3$ } ⇒ {Temperature = warm}	2.8	41.7	5.8
R_8	{cluster = Cluster$_3$ } ⇒ {UV index = medium}	2.9	43.1	1.5
R_9	{cluster = Cluster$_3$ } ⇒ { Pressure = high}	4.6	68.2	1.1
R_{10}	{cluster = Cluster$_4$ } ⇒ {Humidity = low}	8.0	33.1	3.1
R_{11}	{cluster = Cluster$_4$ } ⇒ {Temperature = mild}	13.9	57.0	1.5
R_{12}	{cluster = Cluster$_4$ } ⇒ {Wind Direction = South}	7.9	32.7	1.5

Table 6 reports a subset of extracted rules according to the third template. Support, confidence, and lift are computed separately on the subset of the dataset of each cluster.

Rules R_1, R_2 and R_5 describe the weather conditions correlated with a very high level of thermal energy consumption. For instance, the first rule of Cluster$_1$ (R_1) applies to drizzling evenings in January, with very high humidity, and cold temperature, besides the presence of South wind, which is a very weak and moist wind, accentuating the body's discomfort. All three rules correlate very high energy consumption with minimum UV index, very high humidity, and cold or very cold temperatures. Daily time slot changes from evening (for two rules) to morning (for the third one), as well as the fortnights, from December to January. Two rules have very high confidence values, from 92% to 100%, while the other rule has a relatively high confidence at 73.4%.

Rules R_7, R_8 and R_{11} instead characterize periods with no thermal energy consumption (*off* value). Common conditions are absence of rain, high pressure, warm or mild temperatures, winds from the South or Southeast. The period is in March or April. All rules have very high confidence values, from 95% to 100%, meaning that when the meteorological conditions are met, then the thermal energy consumption is almost always *off*.

Identified correlations are confirmed by domain experts and for some aspects are obvious, e.g., the energy consumption is higher in December and January when it is colder. However, the interestingness of such results is twofold. First, the correlations are automatically inferred from data, showing that they correctly model a more or less known phenomenon, i.e., they actually make sense. Second, the results are human readable, and add meaningful details to trivial correlations, e.g., they specify the most correlated daily time slots and wind directions.

Table 6. Fine-grained traditional rules according to the second template.

		Rule Body								Rule Head			
CId	**RId**	**Forthnight: 15 Days**	**Daily Time Slot**	**Temperature**	**UV Index**	**Humidity**	**Pressure**	**Wind Direction**	**Precipitation**	**Consumption Level**	**Supp %**	**Conf %**	**Lift**
C1	R_1	16–31 January	Evening	Cold	Minimum	Very high	Low	South	Drizzling	Very high	0.2	100.0	153.5
	R_2	1–15 January	Morning	Cold	Minimum	High	Low	South		High	0.2	66.7	7.5
	R_3	16–31 December	Morning	Very cold	Minimum	Very high		SouthWest		Very high	0.2	73.4	5.6
C2	R_4	16–31 December	Midday	Cold	Minimum		High		No rain	Medium	0.6	62.5	11.3
	R_5	1–15 December	Evening	Cold	Minimum	Very high	High	North	No rain	Very high	0.2	92.0	3.8
	R_6	1–15 January	Morning	Very cold	Minimum	Very high		North	No rain	High	0.1	100.0	3.8
C3	R_7	1–15 April	Evening	Warm	Low	Low	High	South	No rain	Off	0.5	100.0	52.8
	R_8	16–31 March	Afternoon	Warm	Medium	Very low	High	South	No rain	Off	0.4	95.0	52.8
	R_9	1–15 March	Midday	Warm	Low	Very low	High	South	No rain	Medium	0.9	95.0	4.2
C4	R_{10}	16–31 October	Evening	Mild		High		SouthEast		Low	0.1	90.2	2.7
	R_{11}	1–15 March	Afternoon	Mild	Low	Medium	High	SouthEast	No rain	Off	0.1	100.0	2.3
	R_{12}	1–15 February	Midday	Mild	Low		High	South		Medium	0.4	90.1	1.9

5.5.2. High-Level Generalized Association Rules

This Section discusses the most relevant generalized association rules, extracted by METATECH and classified according to the rule template presented in Section 4.4. These kind of rules allow us to extract interesting relationships at a higher level among data under analysis, capturing correlations that in the fine-grained extraction would be missed.

Table 7 shows the top-three interesting generalized association rules (with the highest lift value) characterizing each cluster. We concentrate directly on the second template, which yields the most interesting rules. Resulting rules can contain both original leaf values (e.g., morning, afternoon, cold, hot, etc.), and generalized values, such as "root" to indicate the full domain of the attribute, e.g., any value of temperature, or different levels of aggregation, i.e., 4-week period or 8-week period aggregating two or four adjacent fortnights.

We remind that rules described Cluster_1 as the group modeling "bad" weather data (drizzling, cloudy, low UV index), Cluster_2 has cold humid measurements, Cluster_3 warm sunny days, and Cluster_4 mild dry ones.

Rules with the highest confidence typically correlate low consumption levels. For instance, all Cluster_3 and Cluster_4 top correlations (R_8 to R_{12}) present low consumption levels (only R_7 has a medium level). R_{11} stems out from this group of rules because it targets a very large 8-week period from October to December (late Autumn), and states that independently of the temperature ("root" level of generalization), during the midday time slot (from 9 a.m. to 1 p.m.), the consumption level is low, with confidence 74%. A similar behavior is presented by R_8, which states that in afternoons from mid February to mid March (4 week aggregation of two fortnights), independently of the temperature ("root" value), the consumption level is low, with a very high confidence (90%).

Other rules stem out due to their high support. For instance R_9 presents a correlation verified for 16% of the Cluster_3 observations. Typically, generalized association rules, since collect more observations, present higher support than fine-grained rules, which are more specific and intuitively describe quantitatively-limited conditions.

Table 7. Generalized association rules according to the second template.

		RuleBody			**Rule head**			
CId	**RId**	**Fortnight (4 or 8 Weeks)**	**Daily Time Slot**	**Temperature**	**Consumption Level**	**Supp %**	**Conf %**	**Lift**
C1	R_1	December-January	Midday	Cold	High	0.4	27.8	2.5
	R_2	January-February	Midday	Cold	High	1.1	21.0	1.9
	R_3	February-March	Midday	root	Medium	2.1	96.3	1.5
C2	R_4	October-November	Midday	Mild	Low	3.1	80.0	2.31
	R_5	October-November	Afternoon	Mild	Low	1.5	66.7	1.9
	R_6	November-December	Midday	root	Medium	3.3	65.2	1.2
C3	R_7	February-March	Midday	Mild	Medium	10.9	79.3	3.3
	R_8	February-March	Afternoon	root	Low	4.7	90.1	1.2
	R_9	March-April	Afternoon	Hot	Low	16.1	87.2	1.14
C4	R_{10}	March-April	Evening	Mild	Low	5.0	79.2	1.8
	R_{11}	October-December (8 w)	Midday	root	Low	3.1	74.2	1.8
	R_{12}	February-April (8 w)	Afternoon	Mild	Low	2.1	29.6	1.7

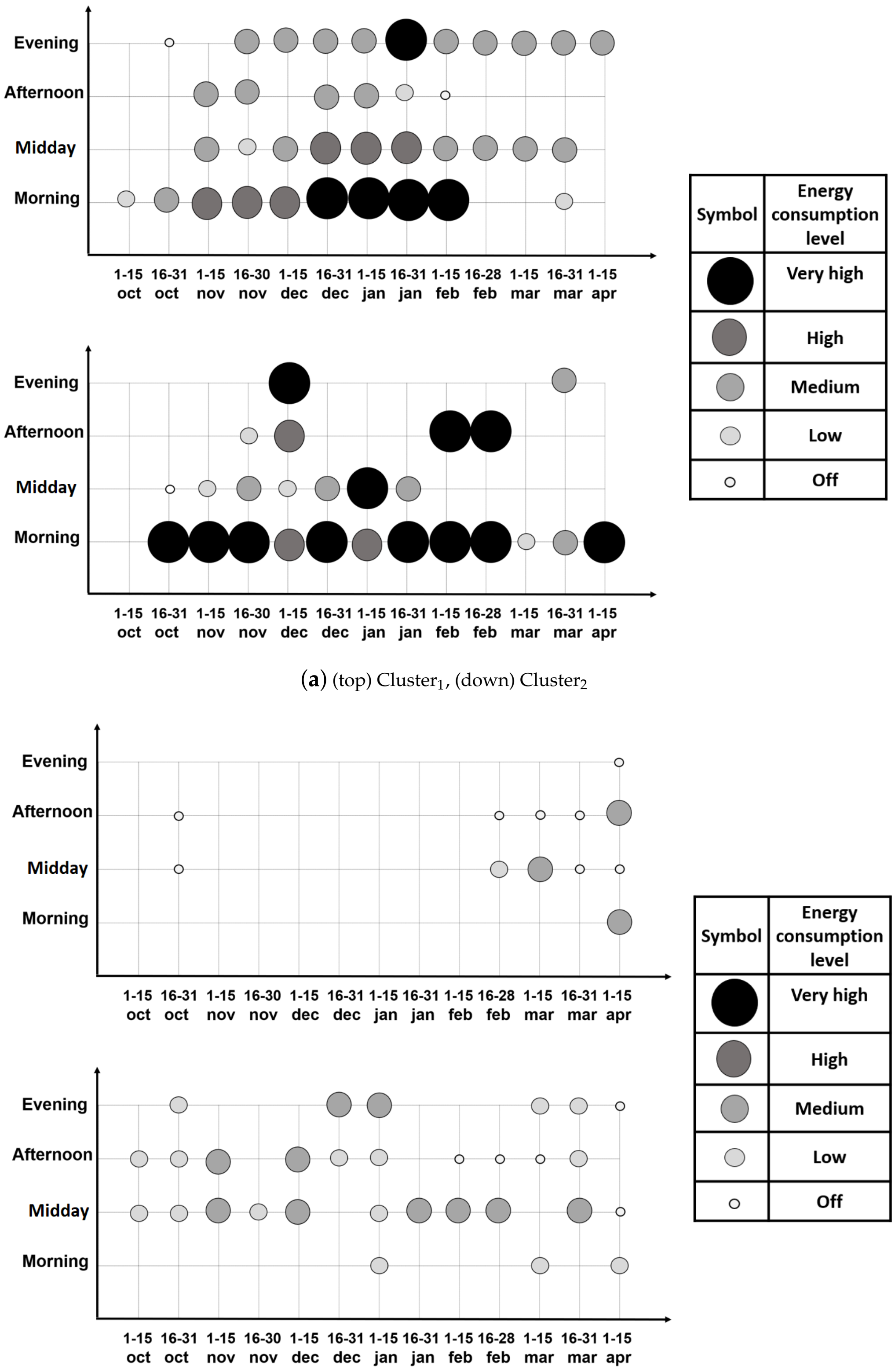
5.6. Summarizing and Comparing Energy Consumption

To present the rule results at a glance, METATECH summarizes energy consumption levels over time in similar meteorological conditions by exploiting a graphical representation, where self-explaining bubble symbols are used for different energy consumption levels.

Figure 6 shows the proposed graphical representation to simplify and synthesize the energy consumption patterns over time in a compact, human-readable, detailed and exhaustive representation.

The four graphs refer to the four clusters identified by the experimental session. Specifically, for each cluster, rules in the form of the second template are partitioned for each daily time slot and fortnight. The rule with the highest lift value is selected and the symbol associated with the corresponding energy consumption level is reported in the graph.

Cluster_1 and Cluster_2 graphs (Figure 6 top) include a large number of symbols modeling very high and high consumption levels. In particular, in the mornings of the winter months consumption is high due to the bad weather conditions. In spring and autumn there was a reduction of the consumption level, while evenings are typically characterized by a medium consumption level (in Cluster_1). Instead the Cluster_4 graph is characterized by lower consumption levels because this cluster represents mild weather conditions. Especially in spring and autumn, consumption levels are low or negligible during the day and afternoon time slots, while during the winter, low or medium consumption levels are frequent.



(**a**) (top) Cluster$_1$, (down) Cluster$_2$

(**b**) (top) Cluster$_3$, (down) Cluster$_4$

Figure 6. Energy consumption levels over time grouped according to similar meteorological conditions.

Hierarchical Graphical Representation

The hierarchical graphical model that METATECH uses to display the extracted knowledge can simultaneously compare the energy consumption levels at different granularity levels as shown in Figure 7. From top to bottom, the three graphs are characterized by coarse to detailed time periods: the upper graph has an 8-week granularity, the middle one presents 4-week periods, and the lower one details fortnights. The first two graphs differentiate three energy consumption levels: "high" aggregates the values "very high" and "high" in a five-level scale; and "low" aggregates "low" and "off". The third graph, besides the more detailed time granularity, presents energy consumption levels on a five-value scale (very high, high, medium, low, and off).

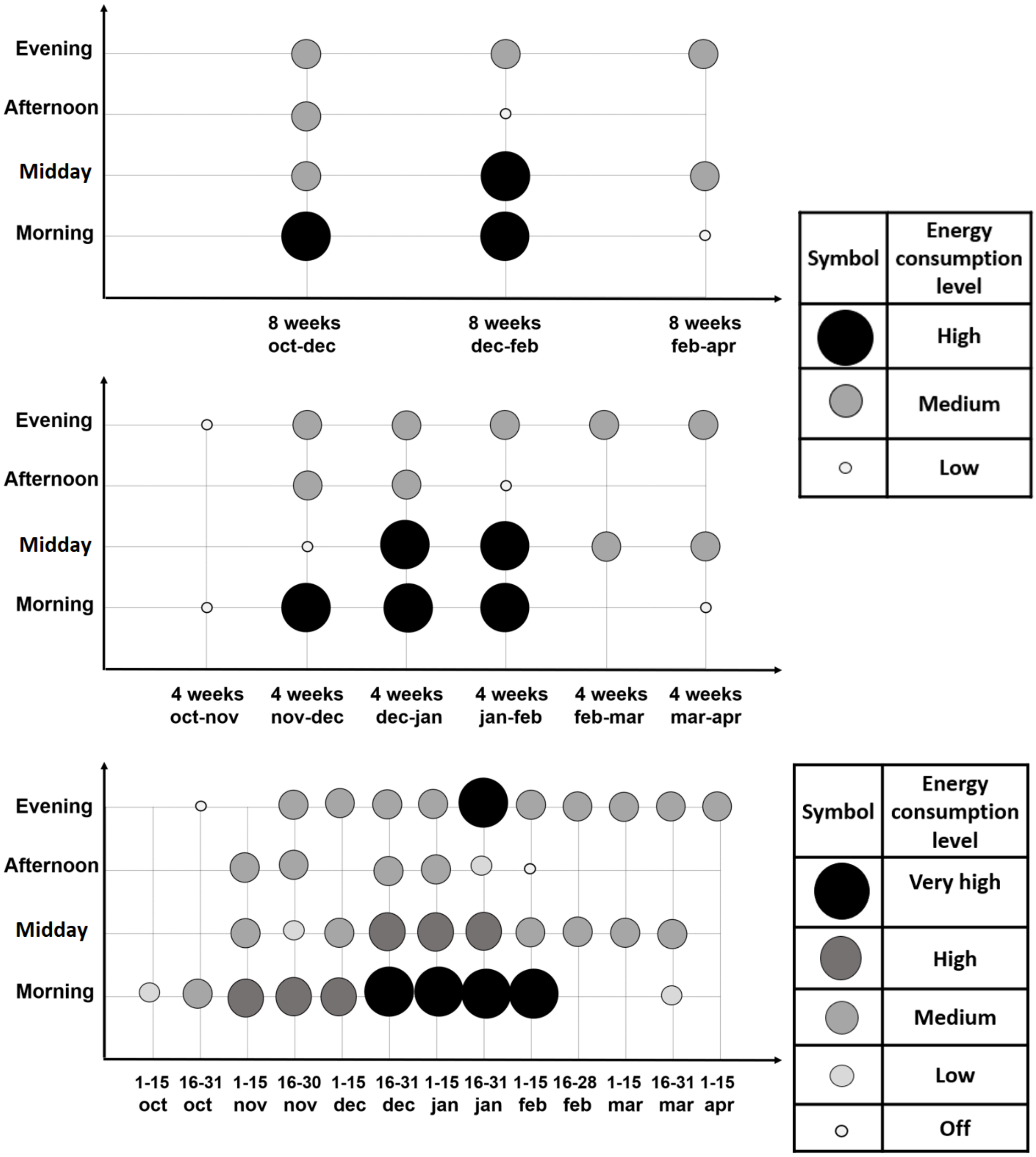


Figure 7. Cluster$_1$ energy consumption over different time-period aggregation levels.

As an example, Cluster$_1$ results are reported. The graphs are built from rules extracted from the dataset and belonging to the second template. When no rules are present, no bubble is indicated.

For instance, in the period 1–15 October, rules characterize the morning, while no correlations have been identified for the other periods of the day.

The graph can be analyzed in two ways: a (i) bottom-up approach and a (ii) top-down approach. The lower graph is obtained using the third template (i.e., correlations between weather conditions and energy consumption level at a different time granularity) of the traditional association rules, while the two upper graphs are obtained using the generalized association rules with the fortnight aggregation to 4 weeks and 8 weeks respectively. The graphical model is able to summarize in a friendly and simple way the consumption of each cluster. Specifically, for each cluster, rules in the form of the second template with the highest lift value are selected and the symbol associated with the corresponding energy consumption level is reported in the graph.

6. Conclusions and Future Works

In this paper we presented METATECH, a data mining engine devised to build transparent models correlating weather conditions and energy consumption. METATECH exploits a joint approach coupling cluster analysis and generalized association rules to allow a deeper yet human-readable understanding of how meteorological data impact heating consumption. Experimental results on a real dataset demonstrate the effectiveness of the proposed methodology in automatically extracting interesting transparent knowledge for domain experts.

We are currently extending the METATECH system to actively engage users to pursue energy-saving behaviors within a social platform, and to measure their changes in energy consumption over time. Users could be engaged with rewards, promoting virtuous behaviors, and introducing gaming approaches. We are also adding new predictive features to METATECH, such as exploiting data mining algorithms (e.g., Artificial Neural Networks and Support Vector Machines) to forecast fine-grained energy consumption. Such extension would allow a better comparison with the real consumption patterns, enriching also the planned activity of collection and measurement of inhabitants reactions and changes in energy-related behaviors when their awareness is risen as a result of the proposed approach.

Author Contributions: All authors contributed equally to this work. All authors wrote, reviewed and commended on the manuscript. All authors have read and approved the final manuscript.

Funding: This research was funded by Piedmont Region under the POR FESR 2007/2013 n. 281-79 (EDEN Project).

Acknowledgments: The authors wish to thank colleagues and other partners involved in the EDEN project (i.e., CSP Innovazione nel ICT, COMMITWORLD S.r.l., Capetti Elettronica s.r.l., Consorzio TOP-IX, Experientia, Sisvel Technology s.r.l.) for their advices and fruitful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kotsiantis, S.; Kanellopoulos, D. Association rules mining: A recent overview. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 71–82.
2. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1988.
3. Venturini, L.; Baralis, E.; Garza, P. Scaling associative classification for very large datasets. *J. Big Data* **2017**, *4*, 44. [CrossRef]
4. Pang-Ning, T.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Addison-Wesley: Boston, MA, USA, 2006.
5. Filippín, C.; Larsen, S.F. Analysis of energy consumption patterns in multi-family housing in a moderate cold climate. *Energy Policy* **2009**, *37*, 3489–3501. [CrossRef]
6. Depuru, S.; Wang, L.; Devabhaktuni, V.; Nelapati, P. A hybrid neural network model and encoding technique for enhanced classification of energy consumption data. In Proceedings of the Power and Energy Society General Meeting, San Diego, CA, USA, 24–29 July 2011; pp. 1–8.
7. Wijayasekara, D.; Linda, O.; Manic, M.; Rieger, C. Mining Building Energy Management System Data Using Fuzzy Anomaly Detection and Linguistic Descriptions. *Ind. Inf. IEEE Trans.* **2014**, *10*, 1829–1840. [CrossRef]

8. Van der Veen, J.; van der Waaij, B.; Meijer, R. Sensor Data Storage Performance: SQL or NoSQL, Physical or Virtual. In Proceedings of the IEEE 5th International Conference on Cloud Computing (CLOUD), Honolulu, HI, USA, 24–29 June 2012; pp. 431–438.
9. Hung, S.S.; Chang, C.Y.; Hsu, C.J.; Chen, S.W. Analysis of Building Envelope Insulation Performance Utilizing Integrated Temperature and Humidity Sensors. *Sensors* **2012**, *12*, 8987–9005. [CrossRef] [PubMed]
10. Chen, C.S.; Lee, D.S. Energy Saving Effects of Wireless Sensor Networks: A Case Study of Convenience Stores in Taiwan. *Sensors* **2011**, *11*, 2013–2034. [CrossRef] [PubMed]
11. Li, Y.; Zhang, S.; Yin, Y.; Xiao, W.; Zhang, J. A Novel Online Sequential Extreme Learning Machine for Gas Utilization Ratio Prediction in Blast Furnaces. *Sensors* **2017**, *17*, 1847. [CrossRef] [PubMed]
12. Menezes, A.; Cripps, A.; Buswell, R.; Wright, J.; Bouchlaghem, D. Estimating the energy consumption and power demand of small power equipment in office buildings. *Energy Build.* **2014**, *75*, 199–209. [CrossRef]
13. Ardakanian, O.; Koochakzadeh, N.; Singh, R.P.; Golab, L.; Keshav, S. Computing Electricity Consumption Profiles from Household Smart Meter Data. In Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference, EDBT/ICDT Workshops, Athens, Greece, 28 March 2014; pp. 140–147.
14. Acquaviva, A.; Apiletti, D.; Attanasio, A.; Baralis, E.; Bottaccioli, L.; Castagnetti, F.B.; Cerquitelli, T.; Chiusano, S.; Macii, E.; Martellacci, D.; et al. Energy Signature Analysis: Knowledge at Your Fingertips. In Proceedings of the IEEE International Congress on Big Data (BigData Congress), New York, NY, USA, 27 June–2 July 2015; pp. 543–550.
15. Acquaviva, A.; Apiletti, D.; Attanasio, A.; Baralis, E.; Castagnetti, F.B.; Cerquitelli, T.; Chiusano, S.; Macii, E.; Martellacci, D.; Patti, E. Enhancing Energy Awareness Through the Analysis of Thermal Energy Consumption. In Proceedings of the Workshops of the EDBT/ICDT 2015 Joint Conference, EDBT/ICDT Workshops, Brussels, Belgium, 27 March 2015; pp. 64–71.
16. Cerquitelli, T.; Di Corso, E. Characterizing Thermal Energy Consumption through Exploratory Data Mining Algorithms. In Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops, Bordeaux, France, 15 March 2016.
17. Di Corso, E.; Cerquitelli, T.; Ventura, F. Self-Tuning Techniques for Large Scale Cluster Analysis on Textual Data Collections. In Proceedings of the 32nd Annual ACM Symposium on Applied Computing, Marrakesh, Morocco, 3–7 April 2017; pp. 771–776.
18. Tureczek, A.; Nielsen, P.S.; Madsen, H. Electricity Consumption Clustering Using Smart Meter Data. *Energies* **2018**, *11*, 859. [CrossRef]
19. Favuzza, S.; Ippolito, M.G.; Massaro, F.; Musca, R.; Riva Sanseverino, E.; Schillaci, G.; Zizzo, G. Building Automation and Control Systems and Electrical Distribution Grids: A Study on the Effects of Loads Control Logics on Power Losses and Peaks. *Energies* **2018**, *11*, 667. [CrossRef]
20. Zhang, L.; Guo, S.; Wu, Z.; Alsaedi, A.; Hayat, T. SWOT Analysis for the Promotion of Energy Efficiency in Rural Buildings: A Case Study of China. *Energies* **2018**, *11*, 851. [CrossRef]
21. Pérez-Chacón, R.; Luna-Romera, J.M.; Troncoso, A.; Martínez-Álvarez, F.; Riquelme, J.C. Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities. *Energies* **2018**, *11*, 683. [CrossRef]
22. Wang, M.; Zheng, X. Sensitivity Analysis of Time Length of Photovoltaic Output Power to Capacity Configuration of Energy Storage Systems. *Energies* **2017**, *10*, 1616. [CrossRef]
23. Jesús, F.M.; Irene, P.C.; Roberto Alonso, G.L.; Cristina, P.; Víctor, E.; Rafael, A.D.L.; Jesica, F.A.; María Jesús, D.V.; Víctor José, D.C.D.; Manuel, M.C.; et al. Methodology for the Study of the Envelope Airtightness of Residential Buildings in Spain: A Case Study. *Energies* **2018**, *11*, 704. [CrossRef]
24. Di Corso, E.; Cerquitelli, T.; Piscitelli, M.S.; Capozzoli, A. Exploring Energy Certificates of Buildings through Unsupervised Data Mining Techniques. In Proceedings of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, UK, 21–23 June 2017; pp. 991–998.
25. Wu, H.T.; Fushing, H.; Chuang, L.Z. Computing and Learning Year-Round Daily Patterns of Hourly Wind Speed and Direction and Their Global Associations with Meteorological Factors. *Entropy* **2015**, *17*, 5784–5798. [CrossRef]
26. Serale, G.; Fiorentini, M.; Capozzoli, A.; Bernardini, D.; Bemporad, A. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies* **2018**, *11*, 631. [CrossRef]

27. Koh, H.C.; Tan, G. Data mining applications in healthcare. *J. Healthc. Inf. Manag.* **2011**, *19*, 65.
28. Wong, K.C.; Li, Y.; Peng, C.; Moses, A.M.; Zhang, Z. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res.* **2015**, *43*, 10180–10189. [CrossRef] [PubMed]
29. Chen, J.; Yan, S.; Wong, K.C. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Comput. Appl.* **2018**, 1–10. [CrossRef]
30. Di Corso, E.; Ventura, F.; Cerquitelli, T. All in a Twitter: Self-Tuning Strategies for a Deeper Understanding of a Crisis Tweet Collection. In Proceedings of the IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 3722–3726.
31. Olson, D.L.; Wu, D.D. Data Mining Models and Enterprise Risk Management. In *Enterprise Risk Management Models*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 119–132.
32. Brefeld, U.; Zimmermann, A. Guest editorial: Special issue on sports analytics. *Data Min. Knowl. Discov.* **2017**, *31*, 1577–1579. [CrossRef]
33. Juang, B.H.; Rabiner, L. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Signal Proc.* **1990**, *38*, 1639–1641. [CrossRef]
34. Srikant, R.; Agrawal, R. Mining Generalized Association Rules. *Future Gener. Comput. Syst.* **1997**, *13*, 161–180. [CrossRef]
35. Data, W. Weather Underground: Weather Forecast & Reports. Available online: http://www.wunderground.com/Last (accessed on 1 March 2018).
36. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Oregon, Portland, 2–4 August 1996; pp. 226–231.
37. Casella, G.; Berger, R.L. *Statistical Inference*; Duxbury: Pacific Grove, CA, USA, 2002.
38. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
39. Agrawal, R.; Imielinski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 25–28 May 1993; pp. 207–216.
40. Rapid Miner. The Rapid Miner Project for Machine Learning. Available online: http://rapid-i.com/ (accessed on 1 March 2018).
41. MathWorks. MATLAB and Simulink for Technical Computing. Available online: www.mathworks.com (accessed on 1 March 2018).
42. Ross, S.M. *Introduction to Probability Models*; Academic Press: Cambridge, MA, USA, 2014.
43. Meteo. Information About Metereological Data. Rain. Available online: https://en.wikipedia.org/wiki/Rain (accessed on 1 March 2018).
44. Meteo. Information About Metereological Data. Wind. Available online: https://en.wikipedia.org/wiki/Wind (accessed on 1 March 2018).
45. Meteo. Information About Metereological Data. Ultraviolet_Index. Available online: https://en.wikipedia.org/wiki/Ultraviolet_index (accessed on 1 March 2018).
46. Meteo. Information About Metereological Data. Atmospheric_Pressured. Available online: https://en.wikipedia.org/wiki/Atmospheric_pressure (accessed on 1 March 2018).