Characterizing thermal energy consumption through exploratory data mining algorithms

Tania Cerquitelli[†], Evelina Di Corso[†]
[†] Dipartimento di Automatica e Informatica, Politecnico di Torino - Torino, Italy
[†] {name.surname}@polito.it

ABSTRACT

Nowadays large volumes of energy data are continuously collected through a variety of meters from different smart-city environments. Such data have a great potential to influence the overall energy balance of our communities by optimizing building energy consumption and by enhancing people's awareness of energy wasting. This paper presents FARTEC, a data mining engine based on exploratory and unsupervised data mining algorithms to characterize building energy consumption together with meteorological conditions. FARTEC exploits a joint approach coupling cluster analysis and association rules. First, a partitional clustering algorithm is applied to weather conditions to discover groups of thermal energy consumption that occurred in similar weather conditions. Each computed cluster is then locally characterized through a set of association rules to ease the manual inspection of the most interesting correlations between thermal consumption and weather conditions. FARTEC also includes a categorization of the rules into a few groups according to their meaning. Each group is determined by the data features appearing in the rule. The experimental evaluation performed on real datasets demonstrates the effectiveness of the proposed approach in discovering interesting knowledge items to raise people's awareness of their energy consumption.

1. INTRODUCTION

Nowadays the demand for energy in the main urban sectors is driven by human activities and by people's awareness of wasting energy. It is challenging to increase people's awareness and persuade them to pursue energy-saving behaviours but it is fundamental to have a positive impact on the global energy balance. Many research activities have been carried out to use database technologies and statistical tools to store and analyze energy data to evaluate the efficiency of buildings. Research contributions on energy-related data have been carried out for: (i) supporting data visualization and warning notification [12]; (ii) efficient storing and retrieval operations based on NoSQL databases [11]; (iii) characterizing building consumption [2] and consumption profiles among different users [6]. 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), grouping data with similar properties (e.g., clustering), and extracting information for prediction (e.g., classification, regression). The first two classes of algorithms are the most interesting ones for their exploratory nature, as 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.

This paper presents an exploratory data mining engine, named FARTEC (From Association Rules To Energy Consumption), targeted at energy-related data. FARTEC analyzes energy data collections enriched with meteorological data through a two-level methodology based on cluster analysis and association rules. The clustering analysis allows the discovery of groups of thermal energy consumption that occurred with similar weather conditions. Each cluster is then locally characterized by a set of interesting patterns to summarize cluster content and to highlight correlations among thermal energy consumption and meteorological conditions. Specifically, FARTEC includes the K-means algorithm [8] to cluster weather data while using the association rule miner [4] to model correlations among energy data and meteorological conditions. A categorization of rules into a few reference classes according to their meaning has been also proposed. As a case study, FARTEC has been validated on real energy consumption collected in a major Italian city. These data have been integrated with meteorological data. Preliminary experimental results show that the proposed approach is effective in discovering interesting correlations to raise people's awareness of their energy consumption.

In this paper, Section 2 introduces an overview of the FARTEC system, while a thorough description of its main components is presented in Section 3. Section 4 discusses the preliminary experimental results obtained on real data, and Section 5 draws conclusions and presents the future development of this work.

2. THE FARTEC SYSTEM

Figure 1 shows the overall architecture of the FARTEC

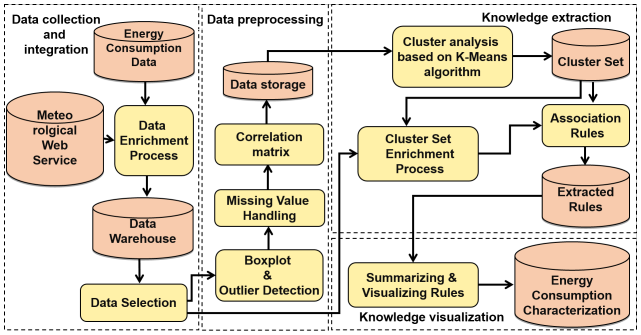

Figure 1: The FARTEC **system architecture.**

system to collect, integrate, characterize, and analyze energy-related data by making people aware of their energy and thermal consumption, as well as encouraging them to pursue energy saving strategies. FARTEC 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 3. In FARTEC the *Data collection and integration* component stores measurements on energy consumption every 5 minutes and aggregates them in hourly thermal energy consumption. These data are enriched with spatial and temporal information at different abstraction levels as well as with various hourly meteorological conditions. The enriched dataset is stored in a datawarehouse as proposed in [3]. Different phases of *Data preprocessing* are then performed to prepare data for the subsequent analysis. The *Knowledge extraction* component discovers groups of energy consumption levels associated with similar meteorological conditions as well as correlations among thermal energy consumption and meteorological conditions. Discovered correlations, in the form of association rules [4], are categorized into a few reference classes according to their meaning. Lastly, the *Knowledge visualization* component shows user-friendly plots to summarize building performance over time.

3. THE FARTEC COMPONENTS

The analysis process in FARTEC is applied on data as modeled in [3]. Thus, the *Data collection and integration* component collects thermal energy consumption, roughly every 5 minutes, from a large number of smart meters deployed in a major Italian city, and aggregates them every hour. As proposed in [3], these data are enriched with temporal information at different granularity levels as well as with various meteorological conditions available as open data sources. Weather data include temperature, relative humidity, precipitation, wind direction, UV index, solar radiation and atmospheric pressure. In this paper we mainly focus on the exploitation of exploratory and unsupervised data mining algorithms to characterize energy consumption at different coarse granularities. Different criteria can be exploited to select only a portion of data (e.g., daily energy consumption in a winter season) stored in the datawarehouse to address a targeted analysis. The FARTEC components, addressing the main phases of the analysis process, are described in the next sections.

3.1 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 for the thermal energy consumption). To address this issue, FARTEC exploits the boxplot (also known as whiskers plot) to graphically show groups of numerical data through their quartiles. The boxplot sums up data distribution through a few numbers (i.e. median, quartiles, min and max values) modeling the frequency distribution. The median summarizes the central tendency of the distribution and compared to quartiles provides information about the asymmetry of the distribution. The quartiles give an indication of the variability through the difference interquantile. Extremes not only provide information on the maximum and minimum value but also on the possible presence of data with abnormal characteristics, plotting them as individual points.

Missing value handling is an important step that significantly affects the mining process. Since we focus on the characterization of thermal energy consumption, we only consider data records where the corresponding consumption value is available. However, FARTEC exploits two strategies to handle missing values on other considered features (e.g., meteorological data): (i) replace them with the daily average value or (ii) replace them with the hourly average value computed in the last week. The choice is mainly driven by the physical meaning of each attribute. For example, case (i) is exploited for the precipitation and wind direction attributes, while case (ii) is for the solar radiation and UV index attributes.

Correlation analysis. Correlated attributes have similar impact in the analysis process. Thus, they are usually removed to reduce the space and time complexity of data mining algorithms. FARTEC leverages the correlation matrix to analyze the dependence between multiple variables at the same time. Each correlation coefficient between each variable and the others is computed through the Pearson correlation defined as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

where $\text{cov}(X,Y)$ is the covariance between X and Y, σ_X is the standard deviation of X and analogously σ_Y for Y. Correlation coefficients are not influenced by the measurement unit of the attributes. The higher the coefficient values the stronger the correlation.

3.2 Knowledge extraction

To extract meaningful and interesting knowledge items from data while maintaining the number of extracted results within manageable limits, the analysis should be performed on the most interesting subsets of input data and the results manually evaluated by a 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, FARTEC 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 rules to be manually validated. FARTEC uses a clustering algorithm to partition data in 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 data. FARTEC also includes a categorization of extracted rules in a few groups to ease manual inspection by the domain expert.

3.2.1 Clustering

Clustering algorithms divide data into groups/subsets (clusters) so that objects within the same group are more similar to each other than objects assigned to different groups [10]. In FARTEC, groups are identified by analyzing records of meteorological conditions and the distance between two objects is computed with the Euclidean distance. The aim is to discover records of energy consumption that occurred with similar weather data. FARTEC integrates a partitional algorithm, the *K-means* algorithm [8], to subdivide the input dataset into K groups, where K is defined by the user and each object is assigned to a single cluster. Each group is represented by its centroid computed as the average of all the objects in the cluster. First, the algorithm sets K initial centroids, chosen randomly. Then each point is assigned iteratively to the closest centroid. Next, the centroids are recalculated. The algorithm repeats the previous steps until the centroids no longer change. K-means is probably the most popular clustering algorithm [5, 13], although it has a bias towards clusters with a spherical shape. However, it identifies the cluster set in a limited computational time by producing a quite good cluster set. K-means requires the number of clusters to be specified in advance, which is one of the biggest drawbacks. To address this issue, FARTEC analyzes the trend of the SSE quality index and the optimal value of K must be selected at the coordinates where the marginal decrease in the SSE curve is maximized. The SSE index [10] measures the cluster quality in terms of cluster cohesion. It is is computed as the total sum of squared errors for all objects in the collection, where for each object the error is computed as the squared distance from the closest centroid.

3.2.2 Association rules extraction

FARTEC discovers correlations from the cluster set identified by the K-Means algorithm. Discovered correlations, in terms of association rules, model interesting relationships among the data under analysis. A transactional dataset $\mathcal{D}$ is a set of transactions in which each one is a set of items (also called itemset). An item is represented in the form *attribute* = *value*. Since we are interested in analyzing energy-related data, each attribute may describe energy consumption, meteorological data (e.g., wind direction, UV index), temporal data (e.g., daily time slot). Since the associaton rule mining requires a transactional dataset of categorical attributes, FARTEC applies the discretization step to convert continuously valued measurements into categorical bins. An association rule is expressed in the form $X \rightarrow Y$, where X and Y are disjoint itemsets, i.e. $X \cap Y = \emptyset$. X is also called rule antecedent and Y rule consequent. The rule quality is measured through two basic indices, named *support* (s) and $confidence$ (c). The *rule support* is the percentage of records containing both X and Y. It represents the prior probability of $X \cup Y$ (i.e. its observed frequency) in the dataset. The *rule confidence*, instead, is the conditional probability of finding Y given X.

Given a set of transactions $\mathcal{D}$, FARTEC 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. To rank the most interesting rules, FARTEC uses the lift index [10], which measures the (symmetric) correlation between antecedent and consequent of the extracted rules. When a rule has lift equal to one, the occurrence probability of the antecedent and the consequent are independent, so X and Y are not correlated. Lift values above 1 show a positive correlation between itemsets X and Y, while values below 1 indicate a negative correlation. FARTEC ranks rules according to their lift value to focus on the subset of most positively correlated rules.

3.3 Association rule categorization

FARTEC includes a categorization of the rules into a few groups according to their meaning to ease the manual inspection of the domain expert. The meaning of a rule is determined by its template which includes the attributes characterizing data. We defined three basic classes of rules that progressively provide more detailed information. Templates are summarized in Table 1, where a basic example rule is reported for each of them.

Specifically, the first template models the *Correlations among cluster and weather conditions included in it*, as shown in Table 1 at row $T1$. This template mainly focuses on the weather conditions that characterize each cluster, without considering the other aspects. We only consider 2-length rules to extract the peculiar characteristics of the climatic conditions of each cluster. This rule set is extracted from the complete cluster set. At row $T2$ the template models the *Correlations among weather conditions included in the cluster*. This template models the cluster content based on the most frequent weather conditions. This kind of rule is locally extracted from each cluster content. The third template at row $T3$ in Table 1 models the *Correlations among energy consumption level, time, and weather conditions*. This template models the correlation between weather conditions and energy consumption level at a different time granularity. This kind of rule is locally extracted from each cluster content enriched with the energy consumption information.

4. EXPERIMENTAL RESULTS

We performed a preliminary analysis of energy consumption on a real dataset, including energy consumption of 15 residential buildings, using the FARTEC engine. We considered energy data related to a complete winter period from October 15^{th}, 2014 to April 15^{th}, 2015. Data collected through the smart meters are integrated with meteorological information collected from the Weather Underground web service[7], which gathers data from Personal Weather Stations (PWS) registered by users. These data are ana-

TId	Question	Rule template	Rule example	Rule meaning
T1	What is the main weather phenomenon that characterize each cluster?	$\{cluster\}$ ⇒ {weather condition}	$\{cluster = Cluster\_4\}$ ⇒ $\{Temperature = warm\}$	It means that the Cluster_4 is characterized by warm temperatures
T2	What are the association rules that are the most representative for each cluster?	$\{weather\ conditions\}$ ⇒ {weather conditions}	$\{Temperature = cold, Wind\ Direction = North\}$ ⇒ $\{Precipitation = no\ rain, Pressure = High\}$	It means that the analized cluster is characterized by cold winds that blow from the North that lead to clear sky and dry weather. In fact, the north winds are strong winds that bring good weather sweeping away the clouds, in agreement with the lack of rain and high humidity.
T3	Given a fortnight and a daily time slot, what kind of consumption level characterizes them under varying atmospheric conditions of each cluster?	$\{fortnight, daily\ time\ slot$, weather conditions} ⇒ {consumption level}	$\{Fortnight = 16-31\ December, Daily\ time\ slot = Day, UV\ index = minimum, Precipitation = no\ rain, Humidity = very\ high, Temperature = very\ cold, Wind\ Direction = North\}$ ⇒ $\{Consumption\ level = very\ high\}$	It means that in the fortnight = 16-31 December and in the daily time slot = Day, a high consumption occurred. Very cold temperatures and high humidity make the body feel a greater sense of cold and then physical discomfort, and the winds blow from the North which are strong and cold winds.

Table 1: Rule template

lyzed for each building separately. We addressed three issues: (i) *outlier detection and correlation analysis* (Section 4.1); (ii) *cluster characterization* in terms of data distribution in each cluster (Section 4.2) and representative association rules (Section 4.3); (iii) knowledge visualization (Section 4.4); (iv) FARTEC sensitivity and robustness to parameter setting (Section 4.5). Here we discuss a given building which is representative of the group of buildings in the considered dataset.

Based on the experimental evaluation discussed in Section 3.2, parameter setting (K=4, $minconf$=1%, $minsup$=1%, $minlift$=1.1) has been used as reference default configuration for FARTEC . To address the problem of centroids initialization for the K-means algorithm we performed multiple runs, with randomly chosen initial centroids and the number of iterations set to 20. The open source RapidMiner toolkit [1] has been used for the correlation analysis, cluster analysis and association rule extraction. The toolkit MATLAB has been used to perform the analysis of data distribution. Experiments were performed on a 2.66-GHz Intel(R) Core(TM)2 Quad PC with 8 GBytes of main memory.

4.1 Oulier detection and correlation analysis

Here we discuss the preliminary results performed to address the outlier detection and removal phase as well as the correlation analysis step performed by FARTEC. Since data collected from sensors are expected to be dirty, collected measurements are analyzed one phenomenon at a time through boxplot. Humidity measurements are discussed as a representative example. Figure 2 shows the humidity distribution of measurements related to a winter period before and after outlier removal. In the left part of the figure is shown the boxplot with the presence of outliers. The plot highlights the presence of incompliant (with humidity percentage values) measurements. To ease the manual inspection of values outside the allowable range, the boxplot shows outliers as individual points in the graph. Figure 2 (right) shows the humidity distribution in the absence of values classified as outliers. The boxplot has the median value close to 70% and 50% of data falls in the interquartile range $[55\% - 85\%]$.

FARTEC exploits the correlation matrix to analyse the

Figure 2: Humidity data distribution

dependence between multiple variables at the same time. The correlation matrix shown in Table 5 contains the correlation coefficients between each couple of attributes computed as discussed in Section 3.1. 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 data. Elements on the diagonal of the matrix are always 1, since they represent the correlation of an attribute with itself. The matrix shown in Table 5 has been computed on data, available for a given building, of a complete winter period. These results highlight two strong correlations: (1) a positive and strong correlation (0.967) between *External Temperature*, i.e. the mean external temperature monitored through PWS, and *Mean Temperature* monitored through a sensor deployed on the roof of the considered building. (2) A high correlation, greater than 0.90, exists between *UV index* and *Solar Radiation*. Since highly correlated attributes are similar in behaviour, for each couple of attributes highlighted in the matrix one is removed from the analysis to reduce both the computational cost and the cardinality of

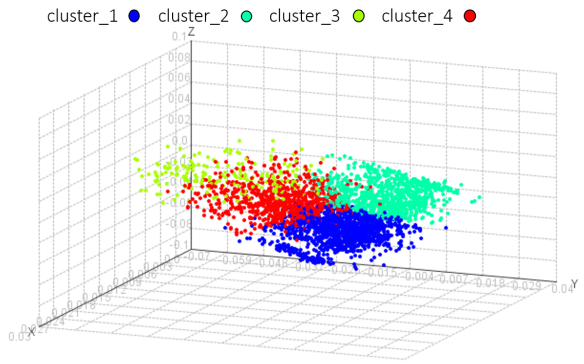

Figure 3: Cluster set representation through SVD

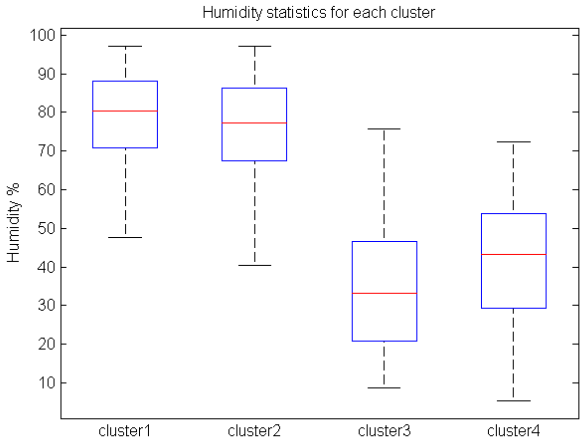

Figure 4: Humidity data distribution for each cluster

the extracted knowledge. Based on the above results, we do not consider *Mean Temperature* and *Solar Radiation* in the subsequent analysis process.

4.2 Cluster characterization

FARTEC exploits the cluster analysis to identify groups of energy consumption that occurred in similar meteorological conditions. The K-Means algorithm has been applied on meteorological data related to a winter period. FARTEC represents the cluster set through (i) the singular value decomposition (SVD) [10] to show the results in a graphical and friendly way; (ii) the comparison of boxplots (one for each cluster) for each attribute separately.

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 3 shows the SVD decomposition of the cluster set discovered by K-means with K=4. Since all clusters in Figure 3 are well-separated, K-means is able to identify a good cluster set.

Figure 4 shows the *Humidity* distribution in the four discovered clusters. The set of clusters is characterized by both positive and negative skewness and groups of observations are quite different, i.e. Cluster_1 and Cluster_2 have quite high median values while Cluster_3 and Cluster_4 have lower median values. In case of positive skewness, observations increase in correspondence with the lowest values, while in the case of negative skewness, the observations increase in correspondence with the highest ones. Cluster_1 and Cluster_2 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. Data are more concentrated between the median and the third quartile, as the same percentage of observations falls in a smaller range. These clusters have higher relative humidity than Cluster_3 which instead has a positive skewness due to the presence of lower relative humidity values.

4.3 Analysis of extracted patterns

Here we discuss the most interesting association rules classified according to the rule template presented in Section 3.3. Since association rule mining requires a transactional dataset of categorical values, FARTEC performs the discretization step to convert continuously valued measurements into categorical bins. In our case study, we used fixed-size discretized bins determined by a domain expert based on

RId	Rule	Supp %	Conf %	Lift
R_1	{cluster = Cluster_1} ⇒ {Precipitations = drizzling}	8.1	20.5	1.8
R_2	{cluster = Cluster_2} ⇒ {Humidity = high}	13.2	45.3	1.3
R_3	{cluster = Cluster_3} ⇒ {Temperature = warm}	2.8	41.7	5.8
R_4	{cluster = Cluster_4} ⇒ {Humidity = low}	8.0	33.1	3.1

Table 2: Rule subset according to the first template.

the significance in the energy and meteorological context. The used fixed-size bins have been determined below. (1) *Energy consumption per unit of volume* (denoted as consumption level): two bins until 15.5 KW/m^3 (off until 0.05 KW/m^3, low until 15.5 KW/m^3), a bin each 10 KW/m^3 for values until 35.5 (medium consumption until 25.5, high consumption until 35.5) and an additional bin for values exceeding 35.5 KW/m^3. (2) *Humidity*: a bin each 20% from 0 to 100%. (3) *Temperature*: values are discretized in five bins (very cold up to 5° Celsius, cold up to 10° Celsius, mild up to 18° Celsius, hot up to 25° Celsius, very hot up to 45° Celsius). (4) *Temporal data*: timestamp is aggregated into the corresponding *daily time slot* (e.g., morning, day, afternoon, evening). Each day is classified as holiday or working, and aggregated in week, fortnight, month, 2-month, 3-month, 6-month time periods. (5) The last *meteorological data* have been discretized based on meteorology criteria available in [9]: precipitation level values and wind direction in have been categorized in eight bins each; likewise UV index in six bins; and atmospheric pressure in two bins.

Table 2 shows the top interesting rule (with the highest lift value) characterizing each cluster according to the first template. These rules are extracted from the complete set of energy consumption related to a given building enriched with cluster labels. Rules $R_1 - R_4$ identify the most representative meteorological item in each cluster. Through the second template, these weather items are subsequently combined with other meteorological items to characterize each cluster in more detail. $R_1 - R_4$ include different meteorolog-

CId	RId	Rule	Supp %	Conf %	Lift
C_1	R_5	{Pressure = low, Precipitations = drizzling} ⇒ {UV index = minimum, Humidity = very high}	10.3	86.4	1.7
C_2	R_6	{Temperature = cold, Wind direction = North} ⇒ {Precipitations = no rain, Pressure = high}	10.5	87.2	1.4
C_3	R_7	{Precipitations = no rain, UV index = medium, Humidity = low} ⇒ {Temperature = warm}	11.8	67.6	1.6
C_4	R_8	{Precipitations = no rain, Temperature = mild, Wind direction = South} ⇒ {Pressure = high}	12.5	72.9	2.0

Table 3: Rule subset according to the second template.

ical items to characterize each cluster and this result further highlights that the discovered groups are well-separated.

Table 3 shows the most positively correlated rules ($R_5 - R_8$) summarizing each cluster content. These rules, examples of the second template, show a strong correlation among various meteorological features, and compactly model each discovered cluster. For example, Cluster_1 includes meteorological data related to cold days, while Cluster_4 regards mild days. Specifically, Cluster_1 is characterized by very high humidity, low pressure and rain, with the presence of clouds and low UV index, while Cluster_4 is characterized by mild temperatures, high pressure and light winds.

Table 4 reports a subset of extracted rules according to the third template. The rules, one for each energy consumption level, are sorted by decreasing lift values. Rules R_9 and R_{12} highlight a high level of thermal energy consumption together with various weather conditions. Specifically, the former means that, during rainy days, the relative humidity of the air tends to increase as very high humidity and low pressure imply the presence of clouds. Also the south wind is a very weak and moist wind and cold temperature accentuates the body's discomfort. Thus, the energy consumption level is very high. Rules R_{10} and R_{13} instead characterize lower thermal energy consumption. Specifically, rule R_{13} means that the wind from the Southeast is a warm and moist wind, the humidity is high and the temperature is mild. So the thermal energy consumption level is negligible. It is October and the low consumption is also motivated by the fact that temperatures are not low, despite being in the evenings.

According to the discussed set of patterns, the selected building chosen as representative has a good thermal energy consumption level which is in line with the meteorological factors that influenced it.

4.4 Summarizing and comparing energy consumption

To enhance the user energy awareness of its energy consumption, FARTEC summarizes the building energy consumption levels over time grouped to similar meteorological conditions. Different symbols and colors (see Figure 5, right) are used for different energy consumption levels. Figure 5 shows the proposed graphical representation to simplify and synthesize the energy consumption patterns (according to the third template) over time in a compact, human-readable, detailed and exhaustive model. This representation also sim-

CId	RId	Rule	Supp %	Conf %	Lift
C_1	R_9	{Fortnight = 16-31 January, Daily time slot = Evening, UV index = minimum, Humidity = very high, Temperature = cold, Pressure = low, Wind direction = South, Precipitations = drizzling} ⇒ {Consumption level = very high}	0.2	100.0	153.5
C_3	R_{10}	{Fortnight = 1-15 April, Daily time slot = Evening, Precipitations = no rain, UV index = low, Pressure = high, Humidity = low, Temperature = warm, Wind direction = South} ⇒ {Consumption level = off}	0.5	100.0	52.8
C_2	R_{11}	{Fortnight = 16-31 December, Daily time slot = Day, UV index = minimum, Precipitations = no rain, Pressure = high, Temperature = cold} ⇒ {Consumption level = medium}	0.6	62.5	11.3
C_1	R_{12}	{Fortnight = 1-15 December, Daily time slot = Morning, UV index = minimum, Pressure = low, Humidity = very high, Temperature = cold, Wind direction = South} ⇒ {Consumption level = high}	0.2	66.7	7.5
C_4	R_{13}	{Fortnight = 16-31 October, Daily time slot = Evening, Temperature = mild, Wind direction = Southeast, Humidity = high} ⇒ {Consumption level = low}	0.1	100.0	2.7

Table 4: Rule subset according to the third template.

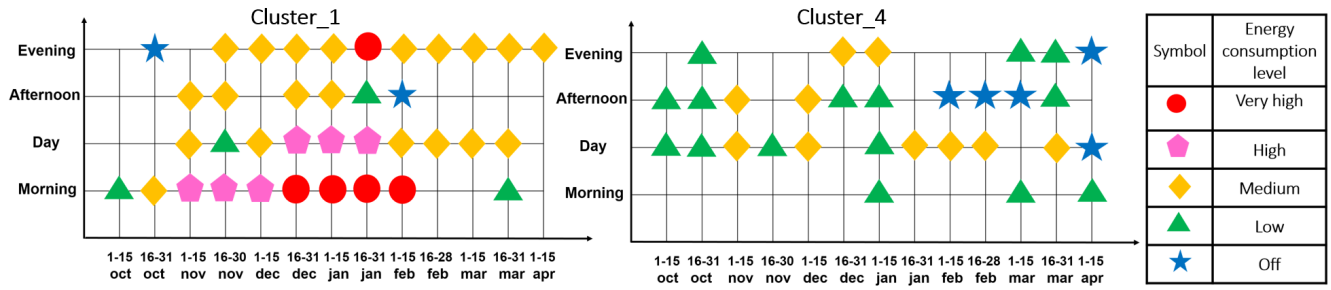

Figure 5: Energy consumption levels over time grouped according to similar meteorological conditions.

plifies the comparison of thermal energy consumption levels between two buildings. Figure 5 shows two graphs of the four discovered clusters for the selected building. Each graph reports the thermal energy consumption level for each couple (daily time slot, fortnight). Specifically, for each cluster, rules in the form of the third template are partitioned for each 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_4 are discussed as representative because they represent orthogonal weather conditions (cold days versus mild days). The Cluster_1 graph (Figure 5 left) includes a large number of symbols modeling high average consumption levels. In fact 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 in every month the evenings are characterized by a medium consumption level. Instead the Cluster_4 graph (Figure 5 center) 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 happen in correspondence with some mild days.

The graphical model that FARTEC uses to display the extracted knowledge can simultaneously compare the energy consumption levels among different buildings. In the presence of different behaviours, users can expand the corresponding rules compactly represented in the graph. Table 6 shows a subset of rules comparing the energy efficiency between the previously discussed building (b_i) and a new one (building b_j). For example, R_{14} shows as rule antecedent bad weather conditions that correspond to a different energy efficiency of the two buildings. The former has a *very high* energy consumption level while the latter *high*. This is due to the different building size (6,297 m^3 and 3,120 m^3) and different populations behaviour. Rule R_{17} instead shows an example in which the consumption of building b_i is far lower than that of building b_j. Since the fortnight corresponds to the Christmas holidays, perhaps the people living in b_i take a holiday period away and turn off the heating system.

4.5 Analysis of parameter setting

We analyzed the robustness of the FARTEC engine to parameter settings for both phases of analysis (i.e. cluster analysis and association rules). The K-means algorithm requires as input parameter the number of clusters (K), which

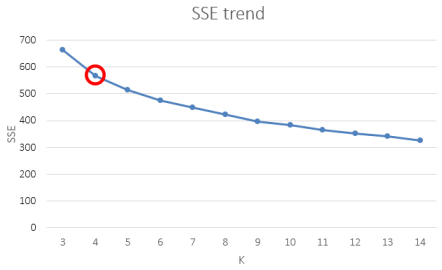

Figure 6: SSE trend against K

is in general very difficult to define, given the wide range in which it may vary. To address this issue we performed many runs of the algorithm with varying values of K, and for each run, the cluster set is evaluated by computing the SSE. Figure 6 shows the SSE value against the K parameter. The smaller the SSE, the better the quality of discovered clusters. However, as the number of cluster increases, the SSE decreases because smaller and more cohesive clusters are identified. To identify a good trade-off between the number of clusters and their significance, we selected $K = 4$ corresponding to the maximization of the marginal decrease in the SSE curve.

To analyze the impact of traditional rule quality measures (i.e. support, confidence and lift) on the cardinality of the mined rule set, we performed many experiments by varying $minsup$, $minconf$, and $minlift$. We recommend users to set low support and confidence threshold values (e.g., 1% and 1% respectively) to avoid pruning some interesting rules with low confidence but a high lift value. We also recommend a minimum *lift* threshold equal to 1.1 to prune both negatively correlated and uncorrelated item combinations.

5. CONCLUSIONS AND FUTURE WORKS

In this paper we presented FARTEC, a data mining engine to analyze energy-related data through exploratory data mining algorithms. Preliminary results on a real dataset demonstrate the potential of the proposed methodology. We are currently extending the FARTEC system with a social platform where users are proactively engaged to pursue energy-saving behaviours as well as in the act of generating data. Users could be engaged with rewards, promoting virtuous behaviours shared with social peers, and introducing gaming approaches (e.g., a shared ranking of energy ratings among neighbours). Engaged users could also provide contextual information useful to optimize building energy

Attributes	External Temperature	Mean Temperature	Precipitation	Wind Direction	Solar Radiation	UV Index	Humidity	Pressure
External Temperature	1	0.967	-0.061	-0.026	0.482	0.477	-0.488	-0.004
Mean Temperature	0.967	1	-0.031	-0.011	0.414	0.403	-0.463	-0.031
Precipitation	-0.061	-0.031	1	0.083	-0.069	-0.064	0.145	-0.057
Wind Direction	-0.026	-0.011	0.083	1	0.018	0.008	-0.084	-0.115
Solar Radiation	0.482	0.414	-0.069	0.018	1	0.913	-0.485	0.056
UV Index	0.477	0.403	-0.064	0.008	0.913	1	-0.423	0.040
Humidity	-0.488	-0.463	0.145	-0.084	-0.485	-0.423	1	-0.068
Pressure	-0.004	-0.031	-0.057	-0.115	0.056	0.040	-0.068	1

Table 5: Correlation Matrix

RId	Rule body	Rule head	
		Building b_i **(6,297** m^3**)**	**Building** b_j **(3,120** m^3**)**
R_{14}	Fortnight = 16-31 December, Daily time slot = Morning, Humidity = Very high, Pressure = High, Wind direction = North, Uv index = Minimum, Temperature = Very cold, precipitation = No rain	Consumption level = very high	Consumption level = high
R_{15}	Fortnight = 1-15 November, Daily time slot = Morning, Uv index = Minimum, Precipitations = No rain, Humidity = Very high, Temperature = Cold, Wind direction = North	Consumption level = high	Consumption level = off
R_{16}	Fortnight = 16-30 November, Daily time slot = Afternoon, Pressure = Low, Uv index = Minimum, Temperature = Cold, Precipitation = light rain	Consumption level = medium	Consumption level = high
R_{17}	Fortnight = 16-31 December, Daily time slot = Afternoon, Humidity = Medium, Pressure = High, Wind direction = South, precipitations = No rain, Temperature = Mild	Consumption level = low	Consumption level = high
R_{18}	Fortnight = 16-31 October, Daily time slot = Afternoon, Humidity = Low, Pressure = High, Wind direction = North, Uv index = Low, Temperature = Warm, precipitation = No rain	Consumption level = off	Consumption level = low

Table 6: Rule comparison between two different buildings.

consumption.

6. REFERENCES

[1] R. M. P. . The Rapid Miner Project for Machine Learning. Available: http://rapid-i.com/ Last access on December 2015.

[2] A. Acquaviva, D. Apiletti, A. Attanasio, E. Baralis, L. Bottaccioli, F. B. Castagnetti, T. Cerquitelli, S. Chiusano, E. Macii, D. Martellacci, and E. Patti. Energy signature analysis: Knowledge at your fingertips. In *IEEE International Congress on Big Data,*, pages 543–550, 2015.

[3] A. Acquaviva, D. Apiletti, A. Attanasio, E. Baralis, F. B. Castagnetti, T. Cerquitelli, S. Chiusano, E. Macii, D. Martellacci, and E. Patti. Enhancing energy awareness through the analysis of thermal energy consumption. In *Workshops of the EDBT/ICDT*, pages 64–71, 2015.

[4] R. Agrawal, T. Imielinski, and Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 1993*, pages 207–216, 1993.

[5] D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, and L. Venturini. SaFe-NeC: a Scalable and Flexible system for Network data Characterization. In *NOMS*, 2016.

[6] O. Ardakanian, N. Koochakzadeh, R. P. Singh, L. Golab, and S. Keshav. Computing electricity consumption profiles from household smart meter data. In *EDBT/ICDT Workshops'14*, pages 140–147, 2014.

[7] W. Data. Weather Underground: Weather Forecast & Reports. Available: http://www.wunderground.com/ Last access on December 2015.

[8] B.-H. Juang and L. Rabiner. The segmental k-means algorithm for estimating parameters of hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(9):1639–1641, Sep 1990.

[9] Meteo. Information about metereological data. Available: https://en.wikipedia.org/wiki/Rain, https://en.wikipedia.org/wiki/Wind, https://en.wikipedia.org/wiki/Ultraviolet_index, https://en.wikipedia.org/wiki/Atmospheric_pressure Last access on December 2015.

[10] Pang-Ning T. and Steinbach M. and Kumar V. *Introduction to Data Mining.* Addison-Wesley, 2006.

[11] J. van der Veen, B. van der Waaij, and R. Meijer. Sensor data storage performance: SQL or NoSQL, physical or virtual. In *IEEE Cloud Computing conference*, pages 431–438, June 2012.

[12] D. Wijayasekara, O. Linda, M. Manic, and C. Rieger. Mining building energy management system data using fuzzy anomaly detection and linguistic descriptions. *Industrial Informatics, IEEE Transactions on*, 10(3):1829–1840, Aug 2014.

[13] B. Zheng, S. W. Yoon, and S. S. Lam. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482, 2014.