Statistical indicator for the detection of anomalies in gas, electricity and water consumption: Application of smart monitoring for educational buildings

Mostafa Akil, Pierre Tittelein, Didier Defer, Frédéric Suard

Statistical indicator for the detection of anomalies in gas, electricity and water consumption: Application of smart monitoring for educational buildings

Mostafa Akil*[1], Pierre Tittelein[1], Didier Defer[1], Frédéric Suard[2]

[1] Laboratory of Civil Engineering and Geo-Environment LGCgE, University of Artois, Béthune Pole, Faculty of Applied Sciences, Technoparc Futura, 62400 Béthune, France.

[2] WiseBIM, F-73375 Bourget-du-Lac

* Corresponding author: mostafa.akil@univ-artois.fr

HIGHLIGHTS:

- This paper presents a methodology to assist building managers detect and identify operating anomalies.
- The data generated consist of three consumption sectors (water, gas and electricity).
- The methodology employed is based on combining data mining with machine learning, in the aim of creating a model that allows defining every day's status (atypical vs. frequent).
- The expert flow manager receives an alert and analyzes the situation, with the possibility of implementing actions to correct the problem should an anomaly be detected.

ABSTRACT

Building facility managers are increasingly equipping their buildings with extensive sets of sensors. This article aims to develop an analysis decision-making methodology based on the production of statistical indicators. The tracking of such indicators allows detecting any systems performance problems. The automatic pinpointing of malfunctions can serve to activate alerts. Our approach focuses on the processing of data stemming from secondary schools managed by departmental services in the Pas-de-Calais, where 117 secondary school buildings have been instrumented with various sensors and supplying data since 2015. This article starts with a close-up on data mining for water, gas and electricity consumption. Data

mining and machine learning methods, including the Clustering approach (K-Means), have been used to extract information from the measurements conducted in 2015 and 2016. This information is used to classify the 2017 measurements according to supervised approaches (SVM). The specificity of this work is to delve deeper into the analysis by combining into the same algorithm a set of various sensors related to both energy use and building occupancy. The data classification results have allowed highlighting "atypical" operations during the daytime, through interpreting data classification results in an effort to define the status of every day in year 2017.

KEYWORDS: Data mining, K-Means, Decision tree, Data energy, Technical building management.

1. Introduction

The Pas-de-Calais Department (France) manages 117 secondary school buildings. To improve knowledge of their energy impact, a several-year instrumentation plan was implemented to track both consumption (water, electricity, gas) and temperature (indoor, outdoor) in practically real time. A preliminary work on this dataset demonstrated this measurement potential by means of detecting water leaks using a single sensor. It is now being proposed to extend this effort to the application of other measurements, which could quickly become cumbersome and complicated, particularly when considering the dependence on both uses and the environment. To analyze such measurements, we have applied several data mining techniques compatible with such multiple-sensor configurations.

These data-oriented techniques have already been employed on a frequent basis in order to support and improve upon the fundamental aspects of energy efficiency management. Let's cite for example Yu *et al.* (2010), who proposed using decision trees to develop predictive models of the energy demands of structures since such an approach provides a more

straightforward interpretation than other classification techniques. Xiao and Fan (2014) made use of Clustering to identify daily energy consumption patterns. Morbitzer, Strachan and Simpson (2004) also applied Clustering algorithms to process the building monitoring data and wound up discovering some more obscure factors of energy overconsumption in building infrastructure. (Capozzoli, Lauro and Khan 2015) described a simplified approach to automatically detect defects in a building's utility equipment. (Chicco et al. 2004) grouped customers into classes according to their electricity consumption behavior. (Li *et al.* 2017) made use of clustering methods in order to identify the VRF (Variable Refrigerant Flow) energy consumption models by means of partitioning the full set of characteristic data. Moreover, the ARM (Association Rule Mining) algorithm has been employed find effective rules associated with VRF energy consumption.

Let's note that the objectives behind these works consisted of either predicting or processing the data and then classifying it by means of clustering methods. The underlying notion therefore is to develop an analytical methodology based on a mix of data processing and classification through combining data mining and machine learning methods. Application of this methodology can serve to detect systems behavior or performance problems. The approaches referenced generally focus on just one information element, like electricity consumption, and fail to take advantage of heterogeneous measurements. In this effort however, the approach has consisted of classifying various types of consumption and analyzing them simultaneously.

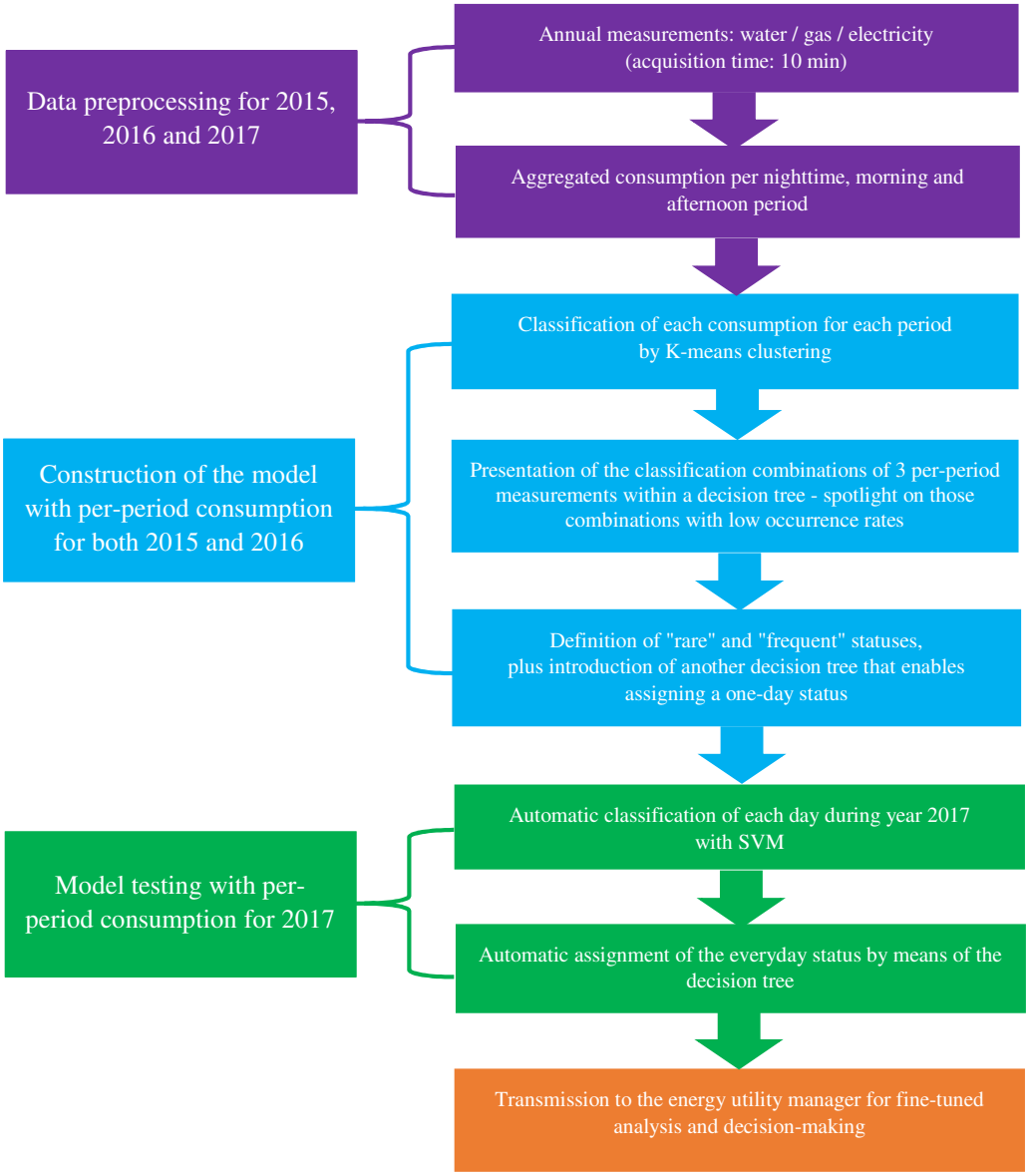


Figure 1: Flowchart of our work program

Figure 1 summarizes the overall approach presented in this document: meter readings of daily water, gas and electricity consumption at 10-minute intervals split into 3 periods (nighttime, morning and afternoon) were aggregated into total consumption levels for each period, yielding 9 data streams. First, for each measurement, the periods were combined into families sorted by similar consumption level through applying the K-Means algorithm. By visualizing the classification combinations obtained for the 3 measurements in a decision tree, some of those with a low occurrence rate could be highlighted. In defining a specific criterion, the rare and frequent operating days were identified by period. On this basis, a new decision tree could

be built, offering the possibility of assigning a status (rare or frequent) to the days contained in the new period. In this aim and by relying on the classes defined during the two prior years (learning), the days of 2017 were classified by Support Vector Machine (SVM) based on their per-period daily consumption level (nighttime, morning, afternoon). A status is then automatically assigned by the tree. Should the rare status be obtained, then the expert flow manager would be able to analyze the situation in greater detail and ultimately implement a set of actions to correct the problem and ensure that the problem never recurs.

This paper will describe the first stage of the above approach, in focusing on each secondary school independently of the others. The second stage consists of conducting an analysis on secondary schools taken all together (not presented herein). This procedure targets measurements of water, gas and electricity consumption and, more specifically, the steps for merging all individual analyses into a global analysis. The purpose here would be to make use of this entire dataset in order to develop an easily automated decision-making aid while highlighting "atypical" behavior.

2. Data Processing

This work is illustrated for one of the secondary schools over the period 2015 to 2017. Some periods are unusable due to measurement problems (on March 17th 2015, the interval from Oct. 26-31, 2015, and December 2017). These periods however only concern less than 2% of all the data, so the data quality does not require a thorough corrective processing.

Consumption recordings are typically studied daily. Water, gas and electricity consumption readings are available by 10-minute interval. Nevertheless, in considering the use patterns of these schools, full days of activity can be distinguished (on Mondays, Tuesdays, Thursdays and Fridays) from half-day operations on Wednesdays and completely idle weekend days. Moreover, the heating system schedule is closely correlated with the adopted usage profile.

The restart is triggered at 5 am and the setting is lowered at 7 pm. This usage breakdown led to separating every day into 3 periods, as follows:

- Nighttime: from 7 pm the prior evening until 5 am in the morning (not including reactivation of the heating system);
- Morning: from 5 am to 12 pm;
- Afternoon: from 12 pm to 7 pm.

This breakdown was established after consultation with the facility manager and by taking into account the heating schedule and the weekly usage profile of the secondary school for instruction-based activities. The data measured with a 10-minute acquisition step were aggregated into consumption levels for each period, yielding 9 datasets per day, i.e. 3 for water consumption, 3 for electricity and 3 for gas. This set-up led to processing 9 datasets over an entire year, such that each value contained in these datasets represents the cumulative consumption of the measurement conducted during each period of the day.

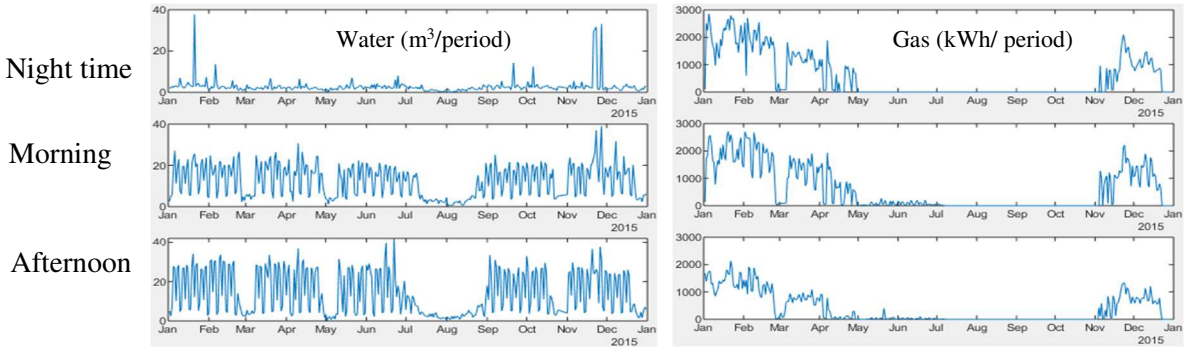


Figure 2: Overall consumption of water and gas in each division of the day in 2015

Figure 2 shows the total daily consumption levels for water and gas respectively, broken down by the three divisions of the day, for year 2015. As an initial observation, the total daily water consumption profiles underscore the effect of the presence of students at the school. The water consumption level always remains in positive territory. During holiday periods, this level drops to reach its school year minimum. Water consumption peaks can nonetheless be

noticed at night, along with the fact that afternoon water consumption runs higher than the morning level.

The total daily gas consumption profiles point to the seasonal effect associated with heating the secondary school. The decrease in heating temperatures during the 2-week holiday period straddling the end of February and beginning of March is clearly displayed by the low gas consumption level. The homogeneity of heating system operations is also evident across the three divisions of the day at this school, although during the spring break and Christmas school holiday periods, the heating system is turned off completely.

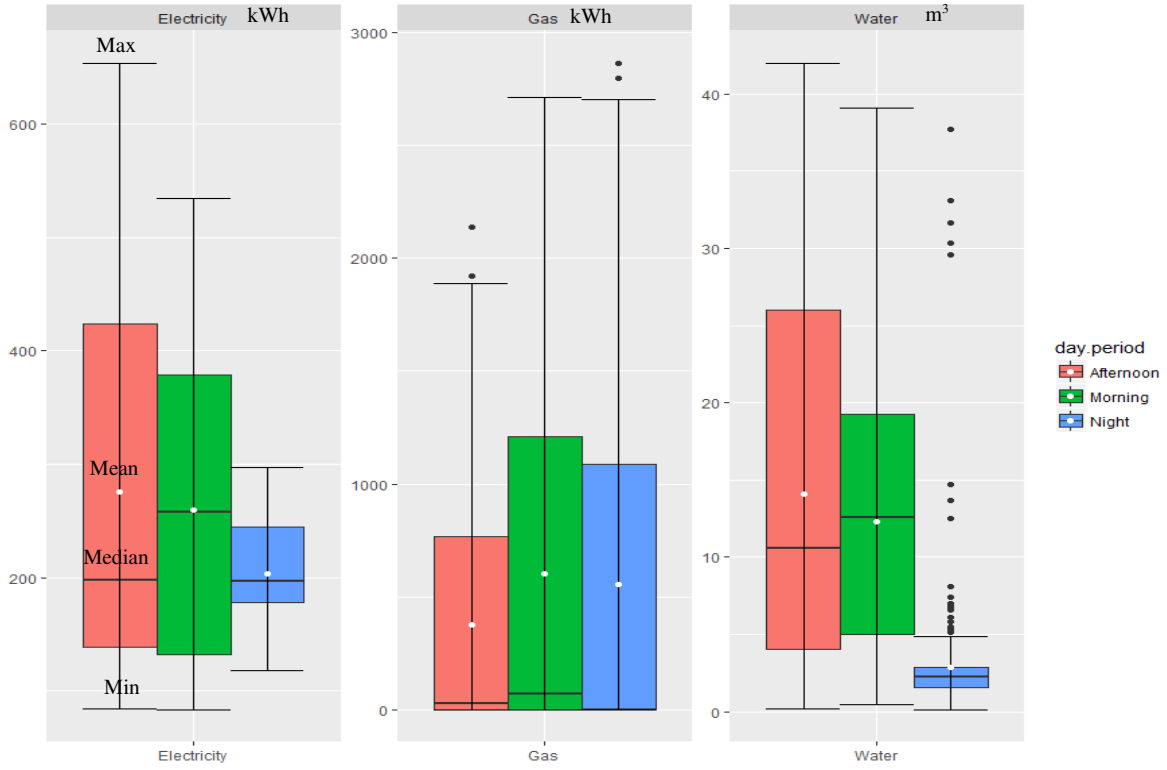


Figure 3: Three-panel boxplot of electricity, gas and water consumption vs. time of day

According to Figure 3, these boxplots serve to visualize several distribution parameters associated with the consumption measurements used in our study, namely the median, the interquartile interval and the maximum and minimum distribution values. A tremendous variability in both water and electricity consumption initially stands out during the mornings and afternoons, contrasting with extremely low nighttime variability. This finding is due to

147 the intense use rate during the morning and afternoon hours at the school compared with an
148 empty building space at night. Yet the water graph still exhibits several extreme nighttime
149 values, some of which may be due to a water leak (as identified on the peaks displayed in
150 Figure 2). On the other hand, the gas consumption graph reveals a median located in the lower
151 part of the box for all 3 times of day. This result can be explained by the seasonal effect of gas
152 consumption at a school, where the heating system is turned off nearly half the year.
153

154 3. Model construction with 2015/2016 per-period consumption

155 3.1. Classification of each type of utility consumption for each period

156 The objective of this work is to propose an automated model for application to all 117
157 secondary schools in the study. Our proposed initial stage entailed developing a method
158 capable of grouping every day of years 2015 and 2016 into families according to consumption
159 level. The problem encountered at the outset however was the lack of information available in
160 the data on consumption broken down by the 3 sectors during each period of the day.
161 Consequently, an unsupervised approach to treat the data without information, in the absence
162 of prior knowledge, was preferred. Several methods were tested, including the hierarchical
163 ascendant classification (HAC) (Farrelly et al. 2017) and the K-Means clustering method
164 (Usman, Ahmad, et Ahmad 2013). The K-Means method seeks to optimize a global objective
165 (variance of clusters) and reach a local optimum, while HAC is aimed at finding the best step
166 in every cluster fusion, which is achieved exactly yet winds up with a potentially suboptimal
167 solution. Meanwhile, K-Means offers a greater advantage in grouping large datasets, and its
168 performance improves as the number of clusters increases. Hence, the K-Means algorithm
169 outperforms the HAC algorithm (Joshi et Kaur 2013). Ultimately, we opted to utilize K-
170 Means, which is intended to provide a grouping into *k* families containing "like" individuals
171 (in our case, a series of measurements over a given period), e.g. the number "k" of classes

must be set ahead of time. This set-up will serve to define specific consumption classes. The procedure implemented in the K-Means method conforms to the following sequence:

1. First step: Applied to the full set of elements to be classified, a sorting routine serves to randomly choose k elements constituting the initial centers of the k classes.
2. Class assignment: Each individual or element in the set is assigned to the class whose center lies closest. Various measurements enable quantifying this proximity. The Euclidian distance selected for this exercise expresses the distance between element $x_i^{(j)}$ and the center of class c_j by: $\left\| x_i^{(j)} - c_j \right\|^2$.

The objective function J to be minimized seeks to minimize the dispersion of each class and is written as the sum of the distances of each element to the center of the associated class:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i^{(j)} - c_j \right\|^2 \quad \textbf{(1)}$$

3. In each class that has been assembled in this manner, the new barycenter is determined and designated as the new class center. Two cases become possible:
    - The k new class centers remain unchanged, hence the classification step is over.
    - The new group of class centers has been modified, hence phase 2 of the assignment step is repeated with the new group.

Upon completion of this iterative process, the k classes are all determined.

Within the framework of an automated processing procedure, which does not necessarily require configuration by an expert user, our method is intended to obtain the optimal number of classes. Silhouette (Subbalakshmi *et al.*, 2015) and Davies-Bouldin (DB) (Davies and Bouldin, 1979) were both employed and yielded the same results. We opted for DB since the Davies-Bouldin Index calculation is much more straightforward than that of the Silhouette

Index. This decision provided a major advantage in terms of real-time operations using clustering.(Petrovic, 2018).

The DB criterion is calculated as follows:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left\{\frac{I(C_i)+I(C_j)}{I(C_i,C_j)}\right\} \quad \textbf{(2)}$$

For each class i of the partition, the focus consists of identifying the class j that maximizes the "similarity index", as described below:

$$R_{ij} = \frac{I(C_i)+I(C_j)}{I(C_i,C_j)} \quad \textbf{(3)}$$

$I(C_i)$ denotes the average distance between the individuals belonging to class C_i and the class center, whereas $I(C_i, C_j)$ denotes the distance between the centers of the two classes C_i and C_j. The best partition therefore is the one that minimizes the average value calculated for each class (Davies and Bouldin, 1979).

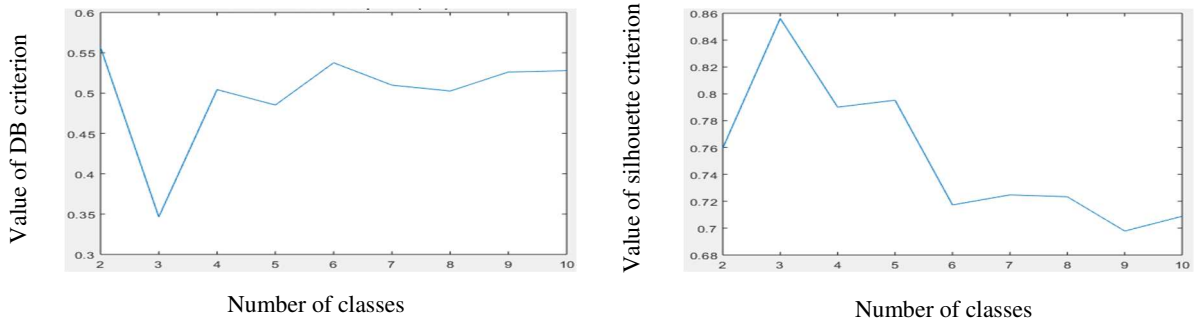


Figure 4: Optimal number of classes in the automatic classification during the morning for the secondary school in 2015 using both DB and silhouette

The DB criterion (Fig. 4) is minimized here for a distribution into 3 classes. The same optimal number of classes is obtained when maximizing the silhouette criterion. These two criteria have validated the finding that 3 is the optimal number for classifying the mornings during the

year 2015 by water consumption level, with each class containing individual days segmented according to their level of water consumption.

The graph in Figure 5 exposes the distribution of mornings in the 3 classes broken down by water consumption level.

3.1.1. Description of the classes found by K-Means

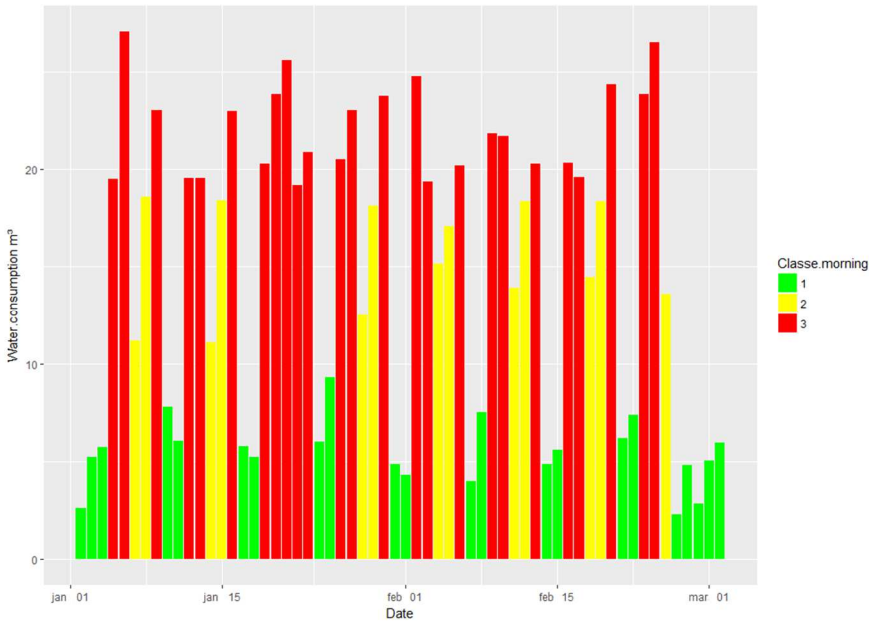


Figure 5: Depiction of the 3 classes identified using the K-Means algorithm in relying on a bar graph for the first two months of 2015 (morning)

In Figure 5, the total water consumption in m³ is presented for each morning during the first two months of year 2015. The values corresponding to Families 1, 2 and 3 are colored respectively in green, yellow and red. This bar graph shows that the lower consumption mornings have been combined into Family 1, with the average consumption mornings in Family 2 and, lastly, the higher consumption mornings lying in Family 3.

To describe in greater detail the applicable thresholds and distributions in each family class of the classification derived by K-Means, a series of boxplots have been generated.

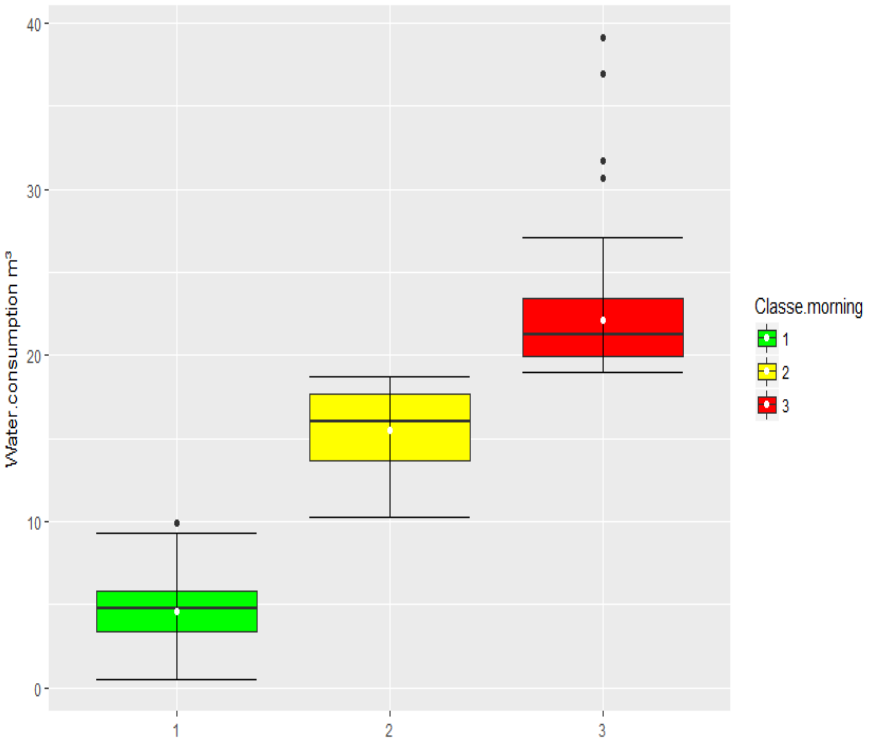


Figure 6: Boxplots of the 3 classes found by K-Means broken down by water consumption level (morning)

According to Figure 6, a water consumption interval can be defined for each class. For the mornings classified in Family 1, water consumption lies between 0 and 10 m³, whereas for Family 2 this consumption figure varies between 10 and 19 m³ and then any total morning water usage level above 19 m³ included in Family 3.

For the following graphs, a color was associated with each class, namely: green for the family of "low" consumption days, yellow for "average" consumption days, and red for "high" days (please note that the blue color still corresponds to those days when data are unavailable). This nomenclature results in an equivalent representation for each of the 9 data series.

The calendar display (Fig. 7) of the 3 classes yielded by this classification protocol offers a visualization of the school's weekly operating scenario as well as the seasonality factor. The *y*-axis of the calendar graph represents the weekdays from Monday to Sunday (from top to bottom) while the *x*-axis aligns all 52 weeks of the year 2015.

3.1.2. Classes for each day of 2015

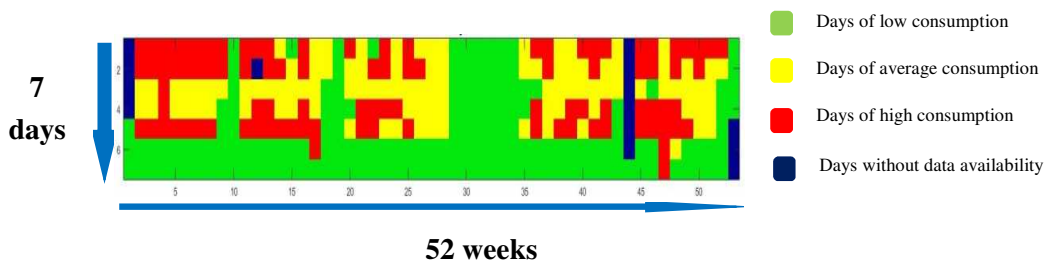


Figure 7: Automatic classification with K-Means of water consumption during the morning for secondary school in 2015

Figure 7 shows the results of a classification by K-Means, with k=3, of water consumption for the morning period (5 am to 12 pm) plotted as a calendar view, which proves to be more legible. A number of trends can be easily observed, such as weekend and holiday days (green class), revealing a grouping primarily in the low-consumption class.

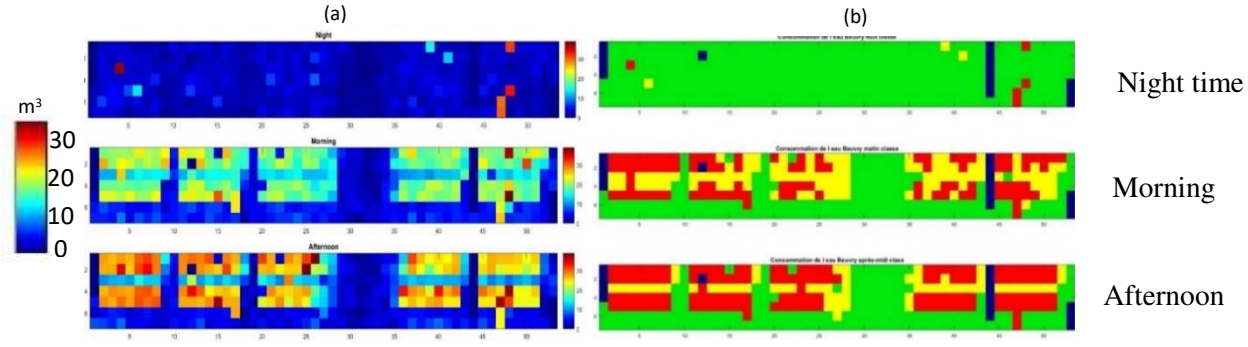


Figure 8: (a) Calendars of raw water consumption values, (b) their crosschecking with the automatic K-Means classification of secondary school in 2015 (the colors are the same as those in Fig. 7)

In Figure 8a, each calendar represents the total water consumption coded onto a color scale, extending from blue for the lowest consumption to red for the highest. The right panel (b) displays the result of the classification performed by K-Means, remembering that the dark blue color of figure 8(b) represents the days without data available for the study. The days of the nighttime calendar (a), which are similar by their blue color, are grouped into the same family by K-Means and are depicted by a green color (b). The class of "low" consumption days is then represented in green, while the "average" class is in yellow and the "high" class in red. From a visual perspective, we can initiate an analysis that allows distinguishing the days

of activity, weekend days, holiday periods and every Wednesday. Logically speaking, it can be remarked that daily water consumption is strongly correlated with the presence of students in the building. During holiday periods, this consumption level decrease to reach its minimum, i.e. zero.

The main advantage of this first step lies in its ability to propose an overview for the school building facility manager. However, a more fine-tuned analysis is required in order to manually identify each data element that could be categorized as either normal or abnormal.

In the next part, a decision tree will be created in order to simultaneously analyze the classifications obtained for the 3 per-period measurements.

3.2. Analysis of the classification combinations

As a reminder, the objective of this study is to develop a method that allows identifying those periods (morning, afternoon or night) with a rare appearance frequency, i.e. periods featuring a rare combination of water, gas and electricity consumption. In this aim, the approach consists of combining the classifications obtained separately for the 3 magnitudes. For an initial preview, Figure 9 exhibits the classification results from all the mornings of year 2015 compiled for the 3 consumption measurements. Through comparative observation of the 3 calendars, an attempt can be made at detecting the low-occurrence combinations.

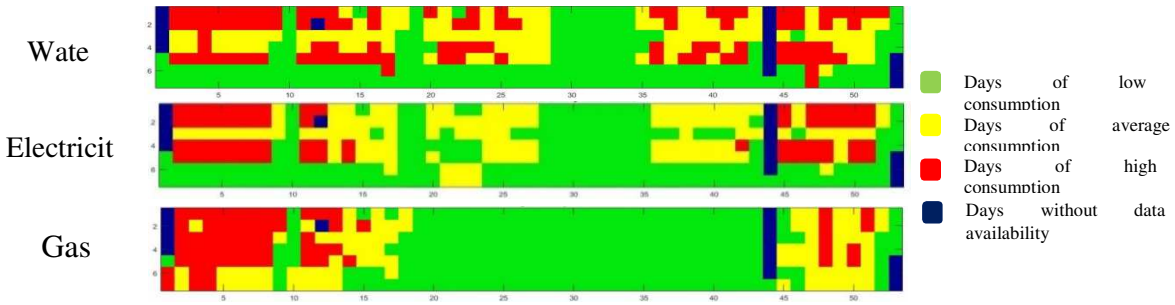


Figure 9: Automatic classification using K-Means of water, electricity and gas consumption during the mornings of 2015

In Figure 9, it can be observed in the lower left, for example, the weekend of the first week of year 2015. Over this period, high gas consumption is combined with low water and electricity consumption. This situation is only encountered on rare occasion during the other weekends. These particular days, highlighted in the crosschecking study, would require a more elaborate analysis (the first weekend may correspond to restart of the heating system at the end of the holiday period). This visual approach is instructive, yet a generalization to all instrumented buildings remains difficult.

In order to generalize this analysis, the clustering step is complemented by building a decision tree that combines the results obtained on the classifications across the 3 consumption sectors. The tree proposed herein, owing to its effective graphical modeling, is easy to interpret and explains the classes identified in the clustering part. Moreover, this technique derives the proportions for every day-specific operating situation and distinguishes the states of low occurrence.

3.2.1. Spotlight on the low-occurrence combinations

A decision tree is a non-parametric classifier that does not require any *a priori* statistical assumptions regarding data distribution. The process of building such a tree has been presented in (Quinlan, 1986) and (Brodley and Utgoff, 1992). The basic decision tree structure comprises a root node, a specific number of internal nodes and, lastly, a set of terminal nodes. The data are distributed recursively in the decision tree according to the established classification framework. Each node requires a decision rule, which may be implemented through introduction of a division test, often of the form:

$$F(\mathbf{x}) = \begin{cases} \mathbf{x_j} > \boldsymbol{c} & \text{for the univariate decision trees} \\ \sum_{j=1}^{n} a_j \mathbf{x_j} \leq \boldsymbol{c} & \text{for the multivariate decision trees} \end{cases} \quad \textbf{(4)}$$

where x_j denotes the measurement vectors on the n selected entities, a_j a vector of linear discrimination coefficients and c the decision threshold (Otukei and Blaschke, 2010).

Figure 10 shows the result of processing electricity consumption for the morning period obtained automatically at the secondary school by means of deducing decision-making rules on the other descriptive variables, i.e. water and gas. The ultimate decision tree developed allows determining the low-occurrence classification combinations.

- **L: Low consumption**
- **M: Medium consumption**
- **H: High consumption**

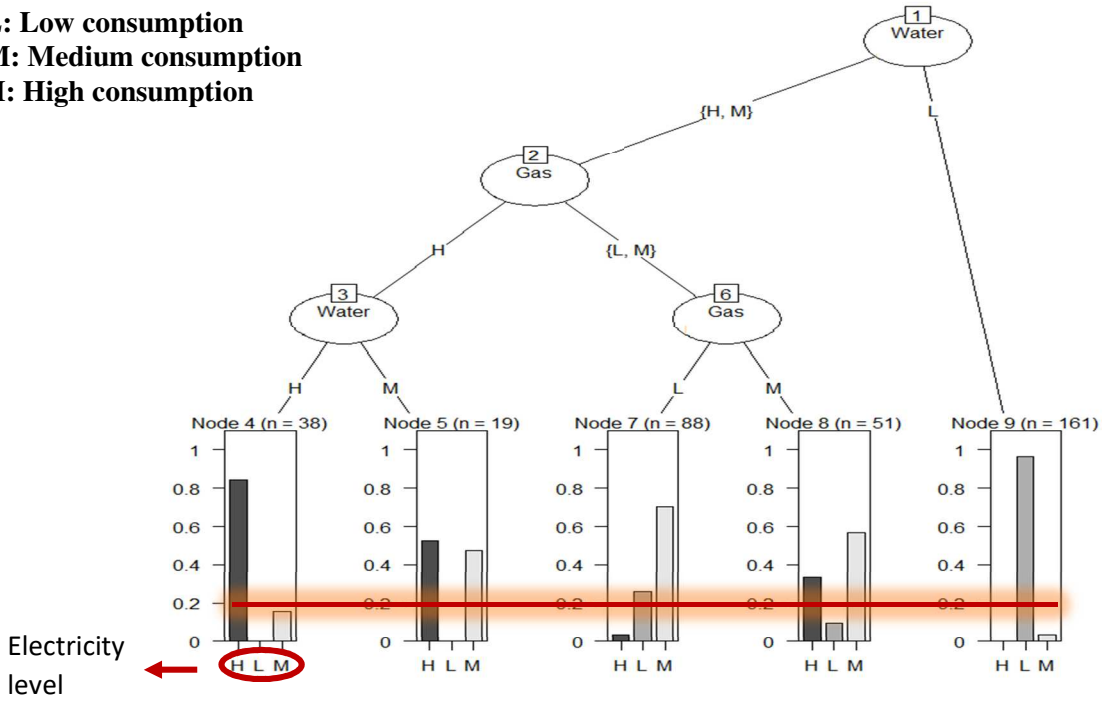


Figure 10: Decision tree for the morning hours of year 2015

According to Figure 10, certain consumption combinations appear in smaller proportions. For example, by taking the first branch on the right of the tree, 161 mornings during 2015 are observed to display low water consumption. Among these individual statistical points, nearly 98% are associated with a low level of electricity consumption as well, and this finding holds regardless of the gas consumption level. Only 2% of these days indicate an "average" level of electricity consumption. These "low-percentage" situations are repeated in several cases, as shown in Figure 10.

3.2.2. Definition of the "rare" and "frequent" statuses

In referring to Figure 10, in an arbitrary manner, it has been considered herein that all cases with a proportion of less than 20% shall be considered as "rare" operating days. This percentage was derived following consultation with the Department agencies and has made it possible to highlight operational anomalies. Moreover, such a parameter can be adjusted by the user. The other cases therefore are assumed to be "frequent" operating days. This percentage remains constant across all tree branches and should be defined in coordination with the facility manager.

When a day is classified in the "rare" category, an alarm can be activated and the individual in charge of monitoring has the option of conducting a more refined analysis of the corresponding consumption trends and, if necessary, take the appropriate actions.

Figure 11 offers a close-up of the calendar day representation corresponding to a "rare" combination situation during the year 2015.

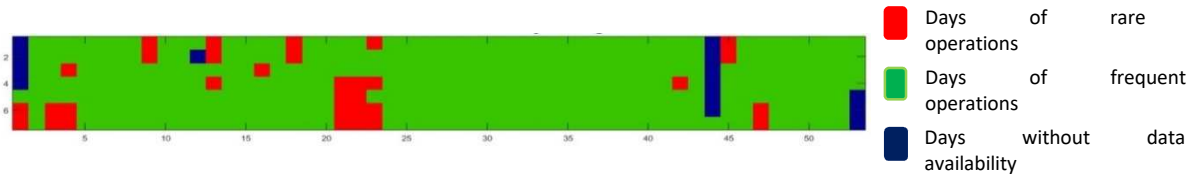


Figure 11: Rare and frequent operating days at the secondary school during the mornings of 2015

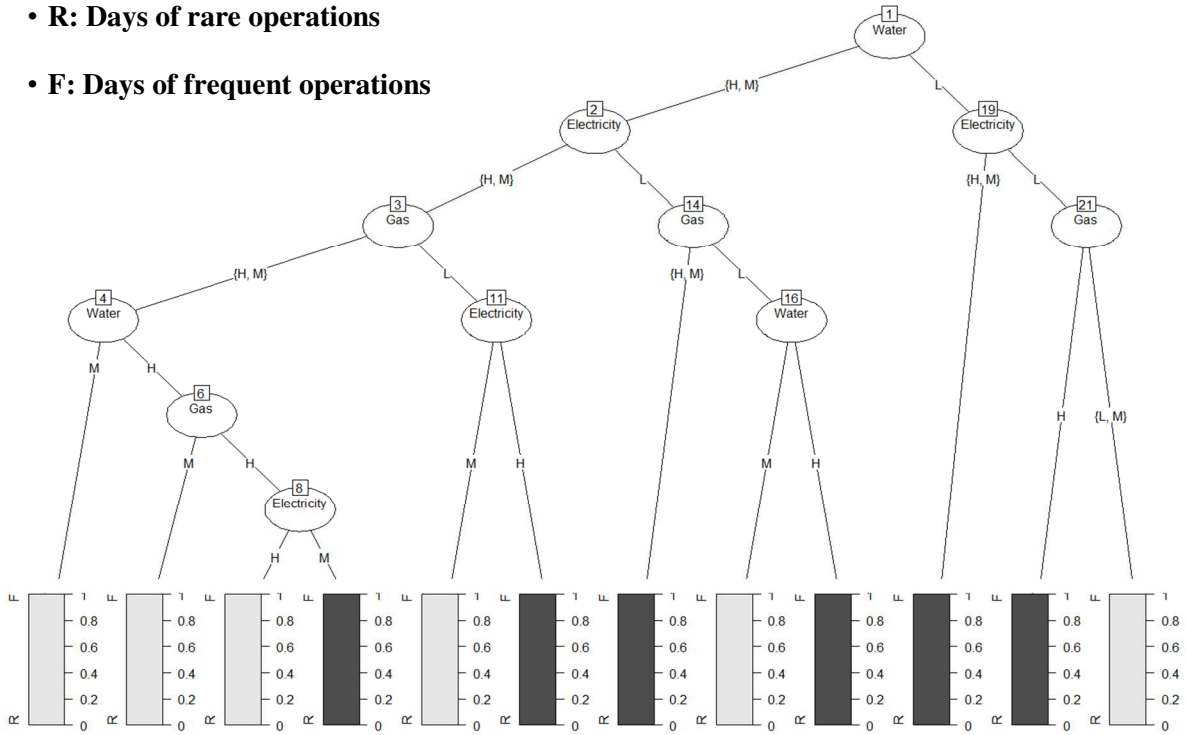
Another tree was built for year 2016 in the aim of confirming the distribution of combinations obtained in 2015. Both trees produced similar proportions for the various branches. The very good level of agreement noted allows validating the statistical model derived. Ultimately, among the 27 (3^3) possible combinations of 3 consumption sectors, 15 could be considered as rare cases and 12 as frequent cases.

3.3. Daily status assignment

After defining the rare and frequent cases, these results will be used to automatically assign a status to each of the days in 2017. For this step, a new decision tree will be created to

352 determine the status of each day of the new period based on the 27 (3^3) possible combinations
353 of the three consumption sectors.
354 The resulting tree (Fig. 12) summarizes the output of this analysis and yields a decision-
355 making aid for defining a daily status throughout the new period.
356

- **R: Days of rare operations**
- **F: Days of frequent operations**



357
358 *Figure 12: The model created by accounting for 27 combinations*

359 To validate its performance, this model was applied to the data from 2015 and 2016 in the aim
360 of determining whether it would group the days into rare and frequent cases in the same way
361 output by the decision trees.
362 Let's recall that for year 2015, a total of 357 days were available for the study, according to
363 the results of the 2015 tree (Fig. 10), with 325 days being defined as the "frequent" case and
364 32 in the "rare" case. Now, applying our model to the year 2015 data reveals whether the
365 same distribution can be obtained.

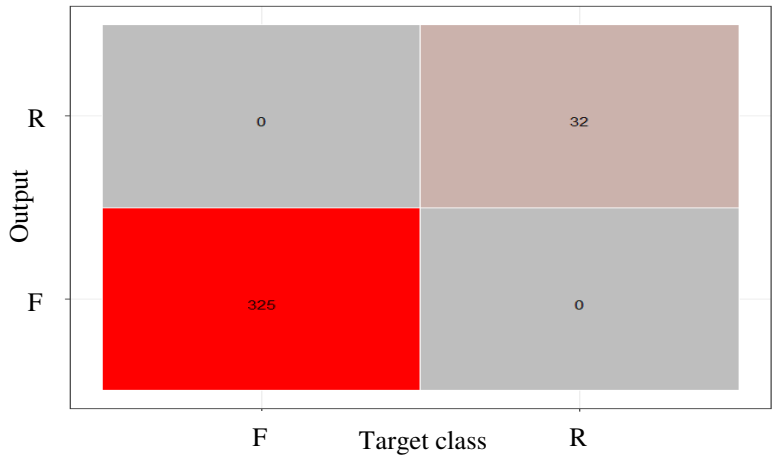


Figure 13: Confusion matrix for the combined data from year 2015

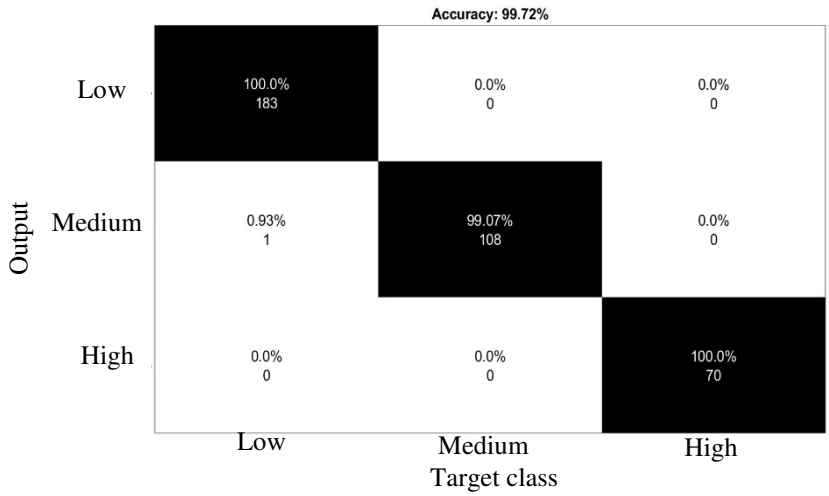
According to Figure 13, it is shown that the distribution of days according to our model is indeed identical to that of the 2015 tree. It can be stated therefore that our model functions with a 0% error.

4. Model application with per-period 2017 consumption data

4.1. Automatic classification of the days in year 2017

After grouping the individual days of 2015 and 2016 into different consumption families, we proceeded to create a model based on these results so as to classify the individual days of year 2017 into existing groups like for the previous years. This step entailed seeking to establish consumption level thresholds between the classes obtained by K-Means and then classifying the days of the year 2017 relative to these thresholds. This strategy necessitates choosing the class boundaries arbitrarily and moreover can lead to classification uncertainties for days lying close to the designated class boundaries. For this purpose, a supervised approach was implemented to classify the data using the classes identified from the previous section as learning data in order to classify the classification outcomes for the days of year 2017. The SVM algorithm (Cortes and Vapnik, 1995), Innocent Gapes (NB) (Behnoud far, Hosseini and Azizi, 2017) or Linear Discriminant Analysis (LDA) (Juuti, Corona and

385 Karhunen, 2018) were all applied. These supervised approaches aims to group classes during
386 the daytime of year 2017 through reliance on the classes derived for 2015 and 2016.
387 SVM and two other methods were applied to classify each day of year 2017 according to the 3
388 consumption sectors for every period of the day. The 18 data series spanning 2015 and 2016
389 (i.e. 3 types of measurements for 3 periods of the day over 2 years) were then combined into 9
390 series, such that each of these series was composed of both series for the same sector and the
391 same period in both 2015 and 2016. Every series therefore consisted of 722 values,
392 representing the number of days in 2015 and 2016 included in the study. Each of these 9
393 series was considered to be entered into the learning phase for the 3 methods. For every
394 series, we assigned a vector label to represent the class identified by K-Means for each day.
395 To validate this learning phase, we randomly collected 50% of the data for learning (training
396 dataset) and the remaining 50% for testing (test dataset). To verify the quality of our results,
397 we generated the confusion matrix by comparing data classified using the reference dataset
398 that were needed to run the test (Ruuska *et al.*, 2018).



399
400 *Figure 14: Confusion matrix for electricity consumption data during the morning for the SVM method*
401 According to Figure 14, it can be noted that the average accuracy of SVM equals 99.7%,
402 which surpasses that of the other methods (LDA: 98.91%; NB: 99.41%). As such, the one-vs.-

one SVM (OVO SVM) method will be used herein in order to classify the daily data of year 2017.

SVM refer to non-parametric classifiers, in an effort to obtain the optimal separating hyperplane between binary classes by respecting the maximized margin criterion. Our work however entails a multiclass classification case, which necessitates a one-to-one strategy applicable to any binary classifier introduced to solve a multiclass classification problem (Liu, Bi and Fan, 2017).

By examining the confusion matrix with SVM, 100% of the days of "low" and "high" electricity consumption, and 99.07% of the "average" consumption days have been correctly classified.

We are now in a position to classify the individual days of year 2017 according to their consumption levels, as output by SVM from running our model. Let's recall herein that the data are not available from December 4th 2017 forward.

4.2. Automatic assignment of each day's status by the decision tree

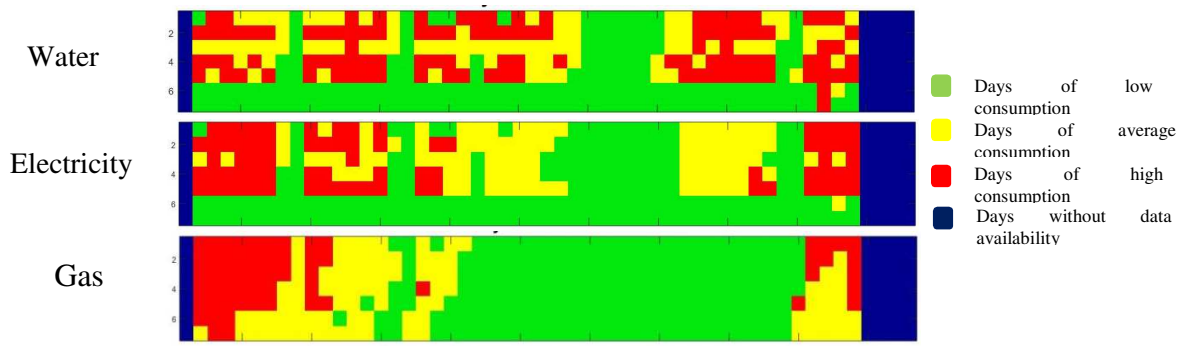


Figure 15: Supervised classification with SVM of water, electricity and gas consumption during the morning hours for year 2017

Before evaluating the grouping of days for year 2017 based on our model's assignment of rare vs. frequent case, let's attempt to extract the days of rare operations visually from Figure 15. In the bottom left, for example, the weekend of the second week of 2017 indicates high gas consumption combined with low water and electricity consumption. Moreover, during the

weekend prior to the final two weeks, high water consumption is combined with low electricity consumption and a medium gas consumption. These situations are rarely found during other weekends and thus considered rare in our model.

Let's now determine whether the model has accurately classified the days of year 2017 as rare vs. frequent.

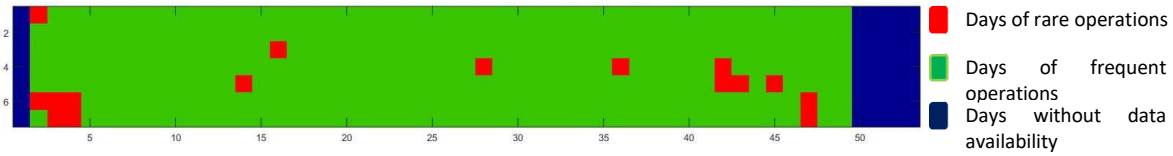


Figure 16: Rare vs. frequent days of operations at the secondary school for 2017 during the morning hours (the colors are the same as those in Fig. 11)

From Figure 16, we conclude that the days classified as rare are indeed rarely repeated in the previous figure; moreover, these days are considered rare in our model. In focusing on Sunday of Week 47 (i.e. November 19th 2017), Figure 15 indicates that during the morning hours, gas consumption was moderate while only a low amount of electricity was being consumed. This scenario is observed several times during the cold season, yet the high water consumption triggered a classification of this day as being rare. Such a combination only occurs once for all Sundays in 2017 (see Fig. 15). The status of rare day leads to analyzing the consumption profiles of this morning, as the expert flow manager might be expected to do. As an example, it could be useful to compare these profiles with those of both the previous Sunday (i.e. November 12th) and subsequent one (November 26th).

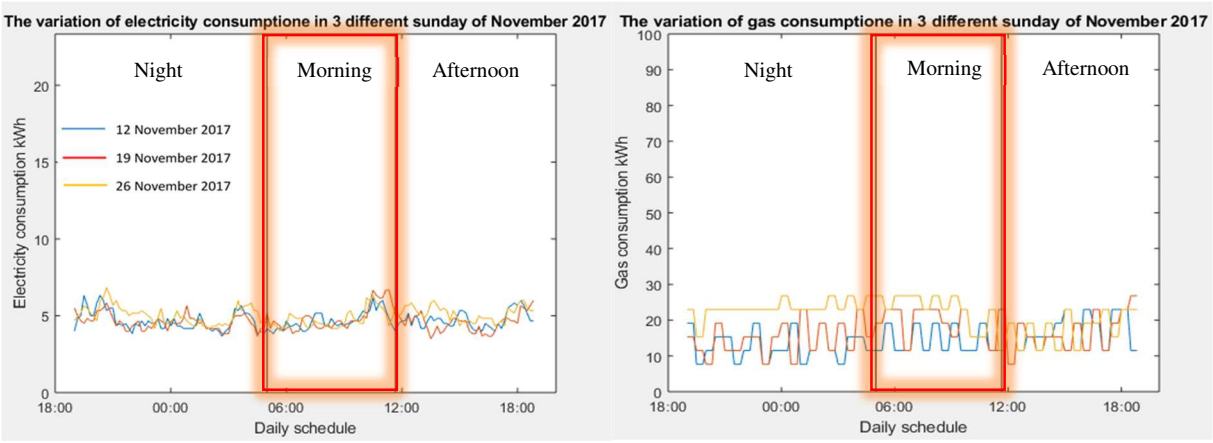


Figure 17: Evolution in electricity and gas consumption during 3 consecutive Sundays in November 2017

Figure 17 shows the electricity and gas consumption profiles for the morning classified as rare, as well as for the mornings of the other two Sundays. It can be observed that these 3 days exhibit practically the same behavior as regards the two morning consumption measurements. The following figure presents the corresponding water consumption profiles.

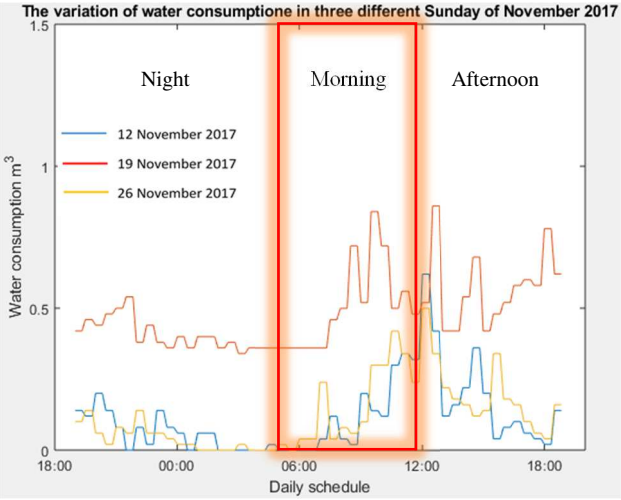


Figure 18: Evolution in water consumption during 3 consecutive Sundays in November 2017

A higher water consumption level is found during the morning of November 19th than that of the other two Sundays. The facility manager is then prompted to look for the cause of this atypical behavior (e.g. unforeseen event, water leak). It can also be stated that the main purpose herein, namely automatically detecting potential operating anomalies, has been accomplished. It is now possible to perform a classification (e.g. with SVM) in near real-time (i.e. at the end of every period of the day) according to 3 consumption sectors. Having established the consumption level classes, the next step consists of applying our protocol to the timeline and defining whether the period features a rare or frequent form of operations.

If the period is deemed "rare", this would not necessarily suggest abnormality or building system dysfunction. The cause might instead be a special event organized on the school grounds (e.g. an open house day during the weekend), yet our system is still capable of alerting the facility manager, who examines the curves daily and communicates with all

building managers, who in turn would be able to make a decision based on the data presented in order to prevent this problem from recurring.

5. Discussion

In such a context of data analysis, on of main challenges resides in the ability to define an exhaustive methodology that could help any facility manager with his daily work : identify and understand the behavior of buildings. To achieve this goal, the scope of statistical modeling is a source of many algorithms and methods for decision making support. We combined some of them : k-means clustering, decision tree, SVM classification in order to implement a tool that could assist the manager. The criteria to identify such algorithms resides not only in their performance of automatic data analysis, but also in their simplicity of tuning. Since the final application has to be useful for non statistician experts, the process does imply the fewest number of options as possible. An algorithm is then chose regarding their generalization capacity and their ability to face with large set of data. For example SVM is able to deal with high volume of data, as well as decision tree.

On top of that, the interpretability of the results, that is to say the ability to visual the result in a simple manner is a major point of interest. For example, a decision tree is well adapted for decision helping, since the decision is made up of a set of simple rules in cascade. As a perspective, such tool could also be envisaged in an interactive way, for example to combine the decision with a dynamic view of the original measurements.

6. Conclusion

The objective of this work has been to assist building facility managers identify operating anomalies by creating a data-oriented methodology. All work performed herein has been based on data mining and machine learning methods, which issue alerts regarding daily

operations of the instrumented secondary school. Our methodology was divided into several stages. First of all, we grouped the days of learning years 2015 and 2016 into families according to their similarities in terms of consumption across 3 sectors by use of K-Means. The days of evaluation year 2017 were then classified according to their daily consumption patterns, based on the classes identified during the previous two years, by SVM. Next, a decision tree combined results obtained on the 3 consumption sector classifications from 2015 and 2016, and the proportions for every daily operating category could be derived. "Rare" and "frequent" days were defined by means of a specific criterion. Lastly, once the consumption statuses had been defined, we built a model with 27 combinations in the aim of determining the status of each individual day for year 2017. The expert flow manager is then in a position to analyze the situation in greater detail and adopt the necessary corrective actions should an anomaly be detected.

The advantage of our method lies in the fact that it can be applied with no *a priori* knowledge on the building studied (floor area, heating schedules, users' habits, etc.). In follow-up work, we will attempt to achieve the same objective of defining alerts for facility managers, yet this time by use of statistical modeling methods (linear regression). Such an approach will serve to identify other days of rare operations through modeling gas consumption along with the other two consumption sectors studied herein plus other measurements (e.g. outdoor temperature, indoor temperature), in the aim of seeking a correlation between gas consumption and these other measurements.

7. Acknowledgments

This work was conducted within the framework of collaboration with the Pas-de-Calais Department (France) council. The authors particularly wish to thank the department’s energy service for its contribution.

8. References

Behnoud far, Pouria, Pantea Hosseini, et Ali Azizi. 2017. « Permeability determination of cores based on their apparent attributes in the Persian Gulf region using Navie Bayesian and Random forest algorithms ». *Journal of Natural Gas Science and Engineering* 37 (January): 52-68. https://doi.org/10.1016/j.jngse.2016.11.036.

Brodley, Carla E., et Paul E. Utgoff. 1992. « Multivariate Versus Univariate Decision Trees ». Amherst, MA, USA: University of Massachusetts.

Capozzoli, Alfonso, Fiorella Lauro, and Imran Khan. 2015. « Fault detection analysis using data mining techniques for a cluster of smart office buildings ». *Expert Systems with Applications* 42 (9): 4324-38. https://doi.org/10.1016/j.eswa.2015.01.010.

Chicco, G., R. Napoli, F. Piglione, P. Postolache, M. Scutariu, et C. Toader. 2004. « Load pattern-based classification of electricity customers ». *IEEE Transactions on Power Systems* 19 (2): 1232-39. https://doi.org/10.1109/TPWRS.2004.826810.

Cortes, Corinna, et Vladimir Vapnik. 1995. « Support-Vector Networks ». *Machine Learning* 20 (3): 273-97. https://doi.org/10.1007/BF00994018.

Davies, D. L., et D. W. Bouldin. 1979. « A Cluster Separation Measure ». *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1 (2): 224-27. https://doi.org/10.1109/TPAMI.1979.4766909.

Farrelly, Colleen M., Seth J. Schwartz, Anna Lisa Amodeo, Daniel J. Feaster, Douglas L. Steinley, Alan Meca, et Simona Picariello. 2017. « The analysis of bridging constructs with hierarchical clustering methods: An application to identity ». *Journal of Research in Personality* 70 (October): 93-106. https://doi.org/10.1016/j.jrp.2017.06.005.

Joshi, Aastha, et Rajneet Kaur. 2013. « Comparative Study of Various Clustering Techniques in Data Mining ». *International Journal of Advanced Research in Computer Science and Software Engineering*, 3.

Juuti, Mika, Francesco Corona, and Juha Karhunen. 2018. « Stochastic discriminant analysis for linear supervised dimension reduction ». *Neurocomputing* 291 (May): 136-50. https://doi.org/10.1016/j.neucom.2018.02.064.

Li, Guannan, Yunpeng Hu, Huanxin Chen, Haorong Li, Min Hu, Yabin Guo, Jiangyan Liu, Shaobo Sun, and Miao Sun. 2017. « Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions ». *Applied Energy* 185 (January): 846-61. https://doi.org/10.1016/j.apenergy.2016.10.091.

Liu, Yang, Jian-Wu Bi, et Zhi-Ping Fan. 2017. « A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm ». *Information Sciences* 394-395 (July): 38-52. https://doi.org/10.1016/j.ins.2017.02.016.

Morbitzer, Christoph, Paul Strachan, et Catherine Simpson. 2004. « Data mining analysis of building simulation performance data ». *Building Services Engineering Research and Technology* 25 (August): 253-67. https://doi.org/10.1191/0143624404bt098oa.

Otukei, J. R., et T. Blaschke. 2010. « Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms ». *International Journal of Applied Earth Observation and Geoinformation*, Supplement Issue on « Remote Sensing for Africa – A Special Collection from the African Association for Remote Sensing of the Environment (AARSE) », 12 (February): S27-31. https://doi.org/10.1016/j.jag.2009.11.002.

Petrovic, Slobodan. 2018. « A comparison between the silhouette index and the Davies-Bouldin index in labelling ids clusters », September.

Quinlan, J. R. 1986. « Induction of Decision Trees ». *Machine Learning* 1 (1): 81-106. https://doi.org/10.1007/BF00116251.

Ruuska, Salla, Wilhelmiina Hämäläinen, Sari Kajava, Mikaela Mughal, Pekka Matilainen, et Jaakko Mononen. 2018. « Evaluation of the confusion matrix method in the validation of an

automated system for measuring feeding behaviour of cattle ». *Behavioural Processes* 148 (mars): 56-62. https://doi.org/10.1016/j.beproc.2018.01.004.
Subbalakshmi, Chatti, G. Rama Krishna, S. Krishna Mohan Rao, et P. Venketeswa Rao. 2015. « A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set ». *Procedia Computer Science*, Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India, 46 (January): 346-53. https://doi.org/10.1016/j.procs.2015.02.030.
Usman, Ghousia, Usman Ahmad, et Mudassar Ahmad. 2013. « Improved K-Means Clustering Algorithm by Getting Initial Centroids », 9.
Xiao, Fu, et Cheng Fan. 2014. « Data mining in building automation system for improving building operational performance ». *Energy and Buildings* 75 (June): 109-18. https://doi.org/10.1016/j.enbuild.2014.02.005.
Yu, Zhun, Fariborz Haghighat, Benjamin C.M. Fung, et Hiroshi Yoshino. 2010. « A Decision Tree Method for Building Energy Demand Modeling ». *Energy and Buildings* 42 (10): 1637-46. https://doi.org/10.1016/j.enbuild.2010.04.006.