



HAL
open science

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

Mostafa Akil, Pierre Tittlein, Didier Defer, Frédéric Suard

► To cite this version:

Mostafa Akil, Pierre Tittlein, Didier Defer, Frédéric Suard. Statistical indicator for the detection of anomalies in gas, electricity and water consumption: Application of smart monitoring for educational buildings. *Energy and Buildings*, 2019, 199, pp.512-522. 10.1016/j.enbuild.2019.07.025 . hal-03168147

HAL Id: hal-03168147

<https://hal.science/hal-03168147v1>

Submitted on 25 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

4 Mostafa Akil*¹, Pierre Tittlein¹, Didier Defer¹, Frédéric Suard²

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

7 ² WiseBIM, F-73375 Bourget-du-Lac

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

9 **HIGHLIGHTS:**

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

19 **ABSTRACT**

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

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

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

38

39 1. Introduction

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

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

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

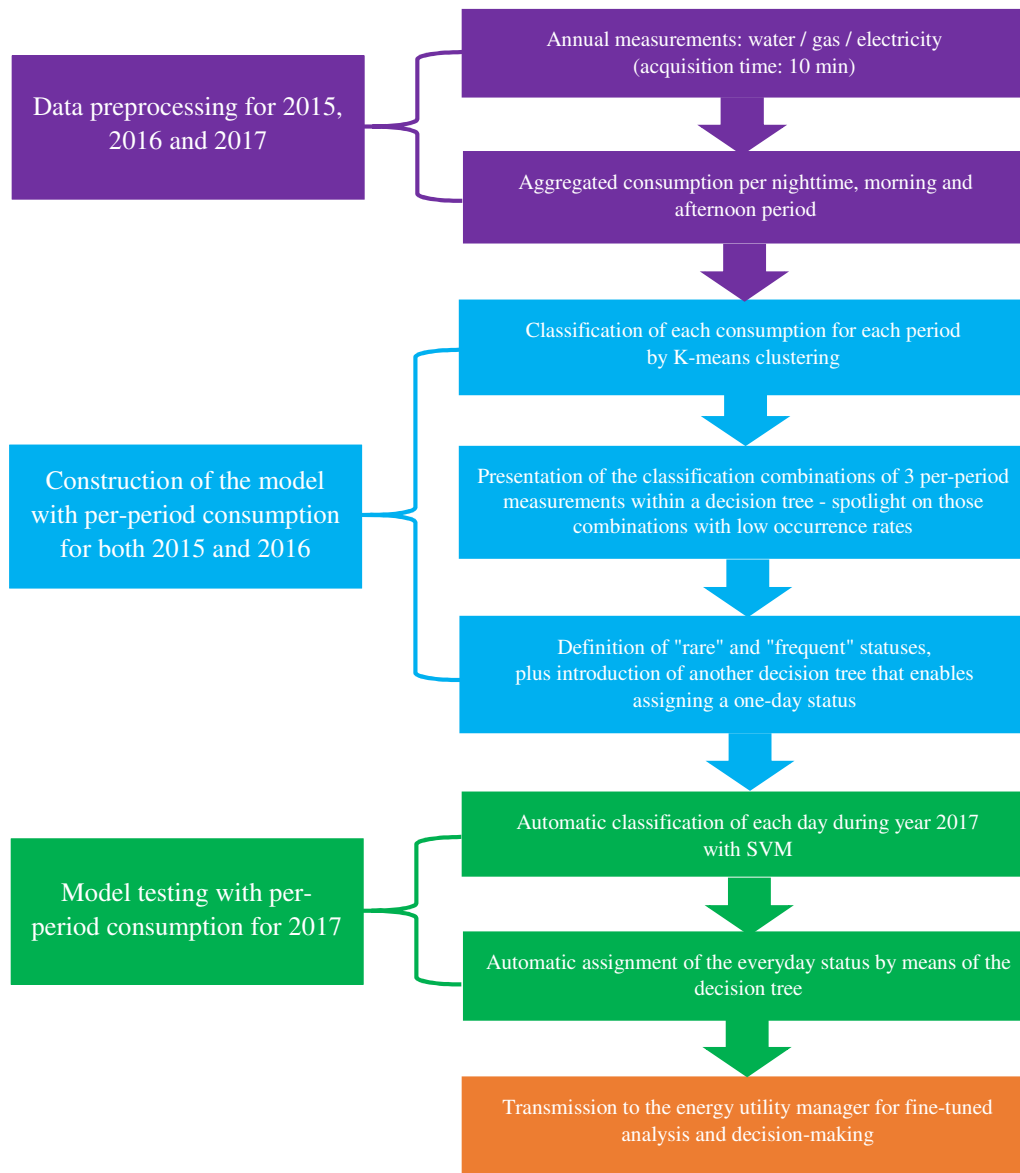
64 Let's note that the objectives behind these works consisted of either predicting or processing
65 the data and then classifying it by means of clustering methods. The underlying notion
66 therefore is to develop an analytical methodology based on a mix of data processing and
67 classification through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

73

74

75

76



77

78

Figure 1: Flowchart of our work program

79

Figure 1 summarizes the overall approach presented in this document: meter readings of daily

80

water, gas and electricity consumption at 10-minute intervals split into 3 periods (nighttime,

81

morning and afternoon) were aggregated into total consumption levels for each period,

82

yielding 9 data streams. First, for each measurement, the periods were combined into families

83

sorted by similar consumption level through applying the K-Means algorithm. By visualizing

84

the classification combinations obtained for the 3 measurements in a decision tree, some of

85

those with a low occurrence rate could be highlighted. In defining a specific criterion, the rare

86

and frequent operating days were identified by period. On this basis, a new decision tree could

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

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

101

102 2. Data Processing

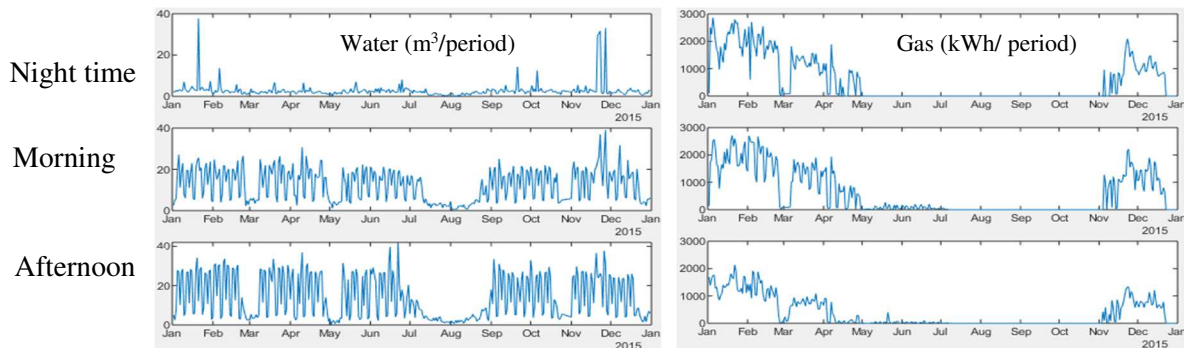
103 This work is illustrated for one of the secondary schools over the period 2015 to 2017.
104 Some periods are unusable due to measurement problems (on March 17th 2015, the interval
105 from Oct. 26-31, 2015, and December 2017). These periods however only concern less than
106 2% of all the data, so the data quality does not require a thorough corrective processing.

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

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

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

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

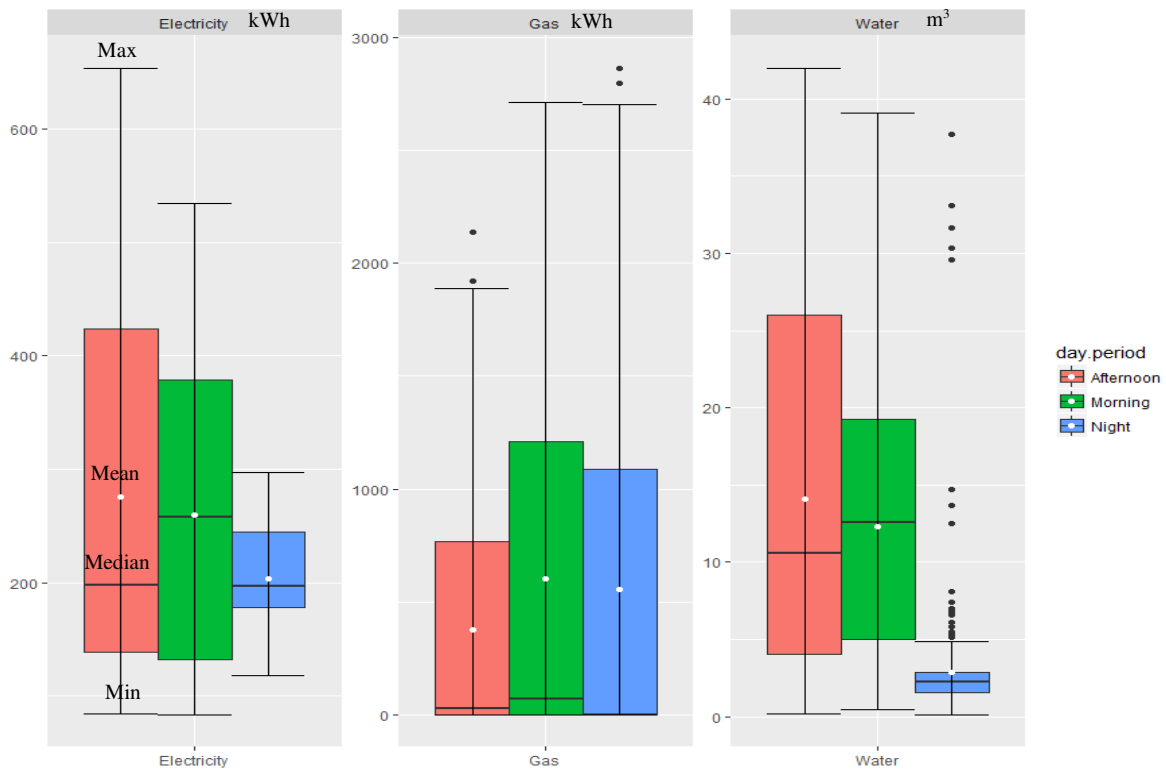


125
126 *Figure 2: Overall consumption of water and gas in each division of the day in 2015*

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

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

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



140

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

142 According to Figure 3, these boxplots serve to visualize several distribution parameters
143 associated with the consumption measurements used in our study, namely the median, the
144 interquartile interval and the maximum and minimum distribution values. A tremendous
145 variability in both water and electricity consumption initially stands out during the mornings
146 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

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

174 1. First step: Applied to the full set of elements to be classified, a sorting routine
175 serves to randomly choose k elements constituting the initial centers of the k
176 classes.

177 2. Class assignment: Each individual or element in the set is assigned to the class
178 whose center lies closest. Various measurements enable quantifying this proximity.
179 The Euclidian distance selected for this exercise expresses the distance between
180 element $x_i^{(j)}$ and the center of class c_j by: $\|x_i^{(j)} - c_j\|^2$.

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

$$183 \quad J = \sum_{i=1}^n \sum_{j=1}^k \|x_i^{(j)} - c_j\|^2$$

184 (1)

185 3. In each class that has been assembled in this manner, the new barycenter is
186 determined and designated as the new class center. Two cases become possible:

- 187 ○ The k new class centers remain unchanged, hence the classification step is
188 over.
- 189 ○ The new group of class centers has been modified, hence phase 2 of the
190 assignment step is repeated with the new group.

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

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

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

199 The DB criterion is calculated as follows:

$$200 \quad DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left\{ \frac{I(C_i) + I(C_j)}{I(C_i, C_j)} \right\}$$

201 (2)

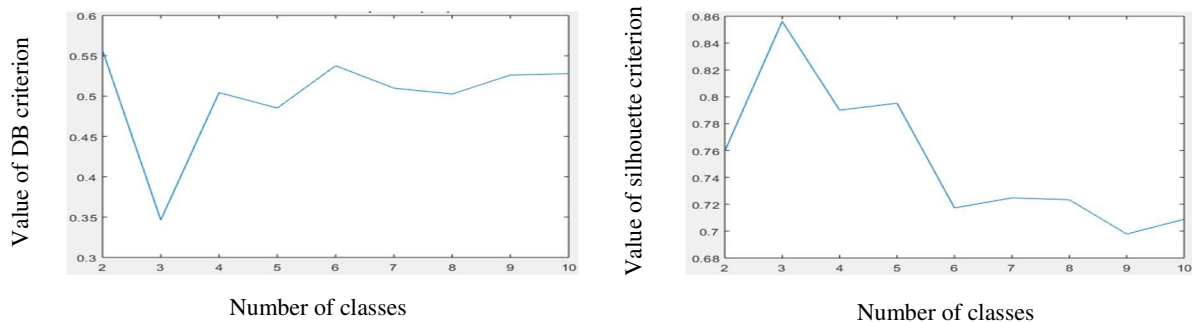
202 For each class i of the partition, the focus consists of identifying the class j that maximizes the
 203 "similarity index", as described below:

$$204 \quad R_{ij} = \frac{I(C_i) + I(C_j)}{I(C_i, C_j)}$$

205 (3)

206 $I(C_i)$ denotes the average distance between the individuals belonging to class C_i and the class
 207 center, whereas $I(C_i, C_j)$ denotes the distance between the centers of the two classes C_i and C_j .

208 The best partition therefore is the one that minimizes the average value calculated for each
 209 class (Davies and Bouldin, 1979).



210

211 *Figure 4: Optimal number of classes in the automatic classification during the morning*

212 *for the secondary school in 2015 using both DB and silhouette*

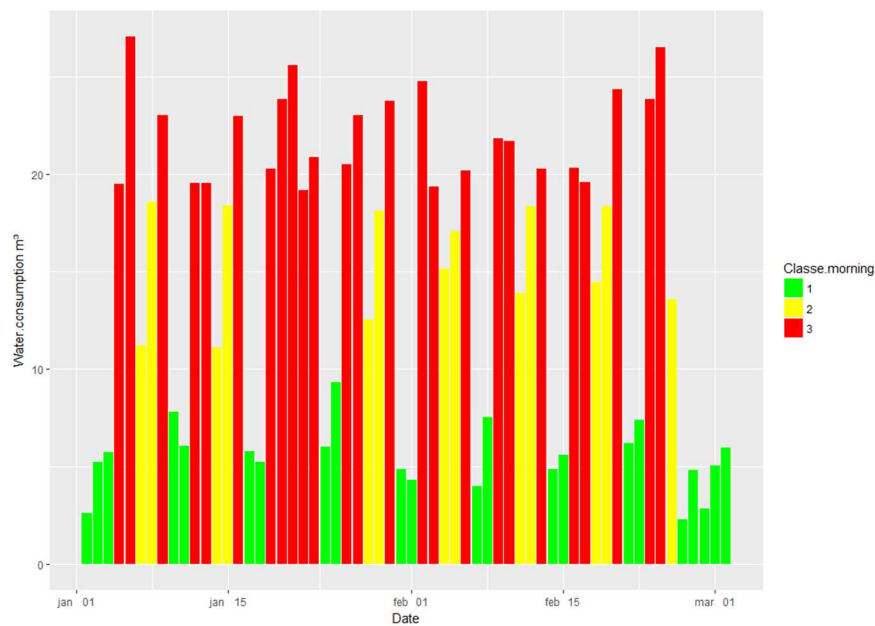
213 The DB criterion (Fig. 4) is minimized here for a distribution into 3 classes. The same optimal
 214 number of classes is obtained when maximizing the silhouette criterion. These two criteria
 215 have validated the finding that 3 is the optimal number for classifying the mornings during the

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

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

220

221 3.1.1. Description of the classes found by K-Means

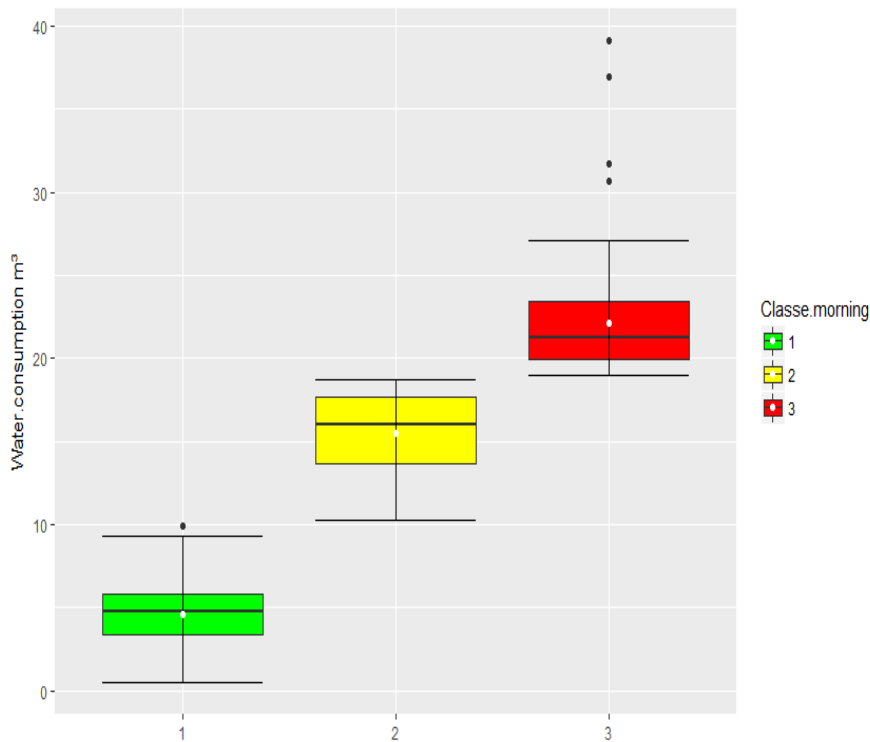


222

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

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

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



232

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

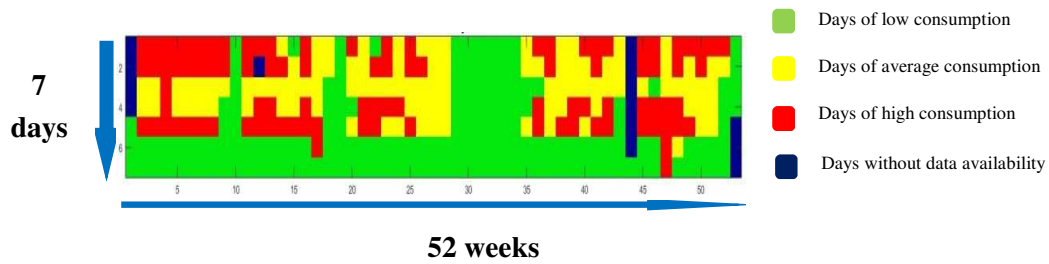
235 According to Figure 6, a water consumption interval can be defined for each class. For the
 236 mornings classified in Family 1, water consumption lies between 0 and 10 m³, whereas for
 237 Family 2 this consumption figure varies between 10 and 19 m³ and then any total morning
 238 water usage level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

239 For the following graphs, a color was associated with each class, namely: green for the family
 240 of "low" consumption days, yellow for "average" consumption days, and red for "high" days
 241 (please note that the blue color still corresponds to those days when data are unavailable).

242 This nomenclature results in an equivalent representation for each of the 9 data series.

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

247 3.1.2. Classes for each day of 2015

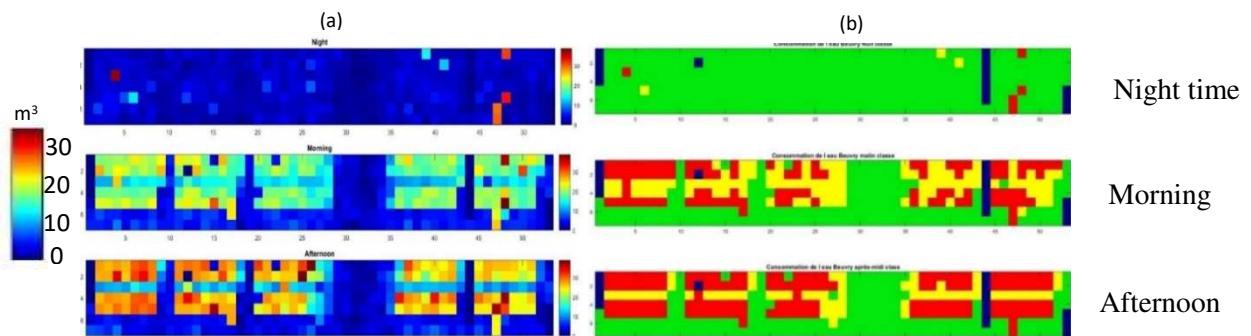


248

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

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

255



256 *Figure 8: (a) Calendars of raw water consumption values, (b) their crosschecking with the automatic*
 257 *K-Means classification of secondary school in 2015 (the colors are the same as those in Fig. 7)*
 258

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

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

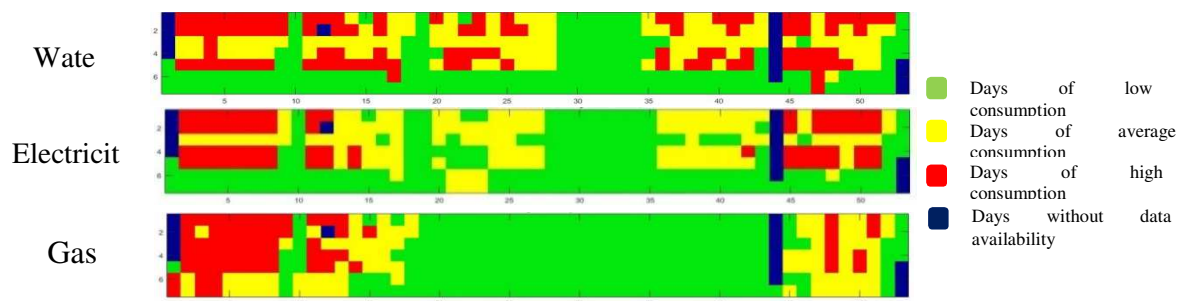
271 The main advantage of this first step lies in its ability to propose an overview for the school
272 building facility manager. However, a more fine-tuned analysis is required in order to
273 manually identify each data element that could be categorized as either normal or abnormal.
274 In the next part, a decision tree will be created in order to simultaneously analyze the
275 classifications obtained for the 3 per-period measurements.

276

277 3.2. Analysis of the classification combinations

278 As a reminder, the objective of this study is to develop a method that allows identifying
279 those periods (morning, afternoon or night) with a rare appearance frequency, i.e. periods
280 featuring a rare combination of water, gas and electricity consumption. In this aim, the
281 approach consists of combining the classifications obtained separately for the 3 magnitudes.
282 For an initial preview, Figure 9 exhibits the classification results from all the mornings of year
283 2015 compiled for the 3 consumption measurements. Through comparative observation of the
284 3 calendars, an attempt can be made at detecting the low-occurrence combinations.

285



286

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

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

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

302

303 3.2.1. Spotlight on the low-occurrence combinations

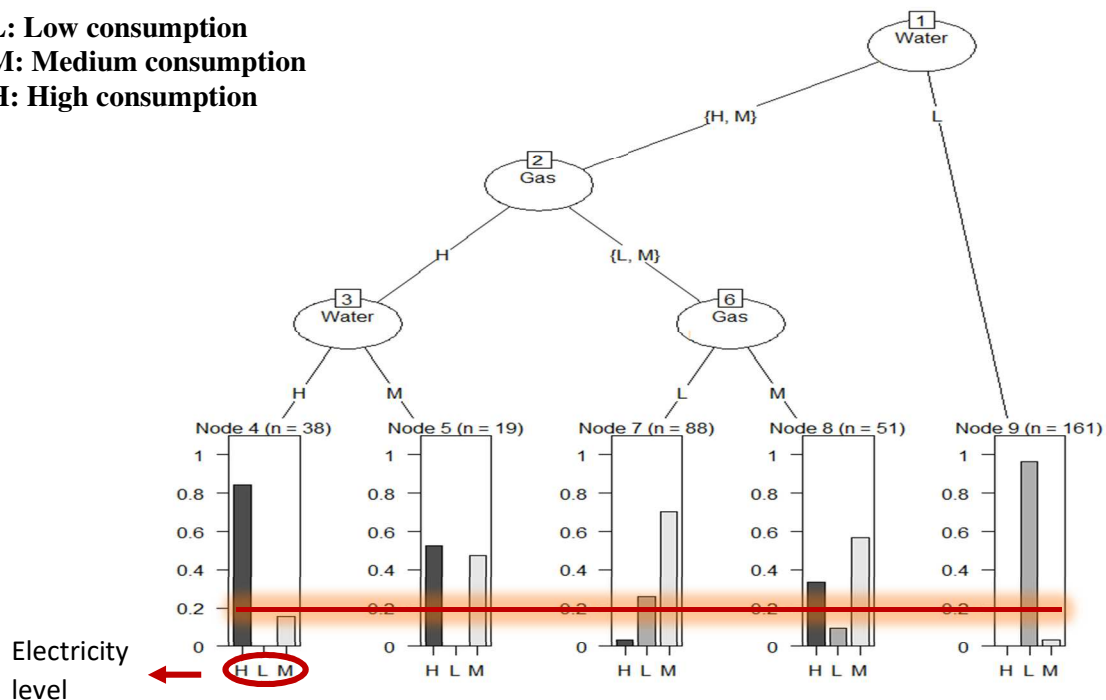
304 A decision tree is a non-parametric classifier that does not require any *a priori* statistical
305 assumptions regarding data distribution. The process of building such a tree has been
306 presented in (Quinlan, 1986) and (Brodley and Utgoff, 1992). The basic decision tree
307 structure comprises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

$$311 \quad F(\mathbf{x}) = \begin{cases} \mathbf{x}_j > \mathbf{c} & \text{for the univariate decision trees} \\ \sum_{j=1}^n a_j \mathbf{x}_j \leq \mathbf{c} & \text{for the multivariate decision trees} \end{cases}$$

312 (4)

313 where x_j denotes the measurement vectors on the n selected entities, a_j a vector of linear
 314 discrimination coefficients and c the decision threshold (Otukey and Blaschke, 2010).
 315 Figure 10 shows the result of processing electricity consumption for the morning period
 316 obtained automatically at the secondary school by means of deducing decision-making rules
 317 on the other descriptive variables, i.e. water and gas. The ultimate decision tree developed
 318 allows determining the low-occurrence classification combinations.

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



319
 320 *Figure 10: Decision tree for the morning hours of year 2015*

321 According to Figure 10, certain consumption combinations appear in smaller proportions. For
 322 example, by taking the first branch on the right of the tree, 161 mornings during 2015 are
 323 observed to display low water consumption. Among these individual statistical points, nearly
 324 98% are associated with a low level of electricity consumption as well, and this finding holds
 325 regardless of the gas consumption level. Only 2% of these days indicate an "average" level of
 326 electricity consumption. These "low-percentage" situations are repeated in several cases, as
 327 shown in Figure 10.

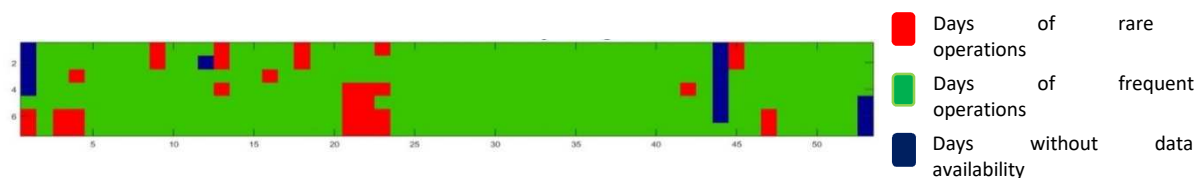
328

329 3.2.2. Definition of the "rare" and "frequent" statuses

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

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

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



342

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

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

349 3.3. Daily status assignment

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

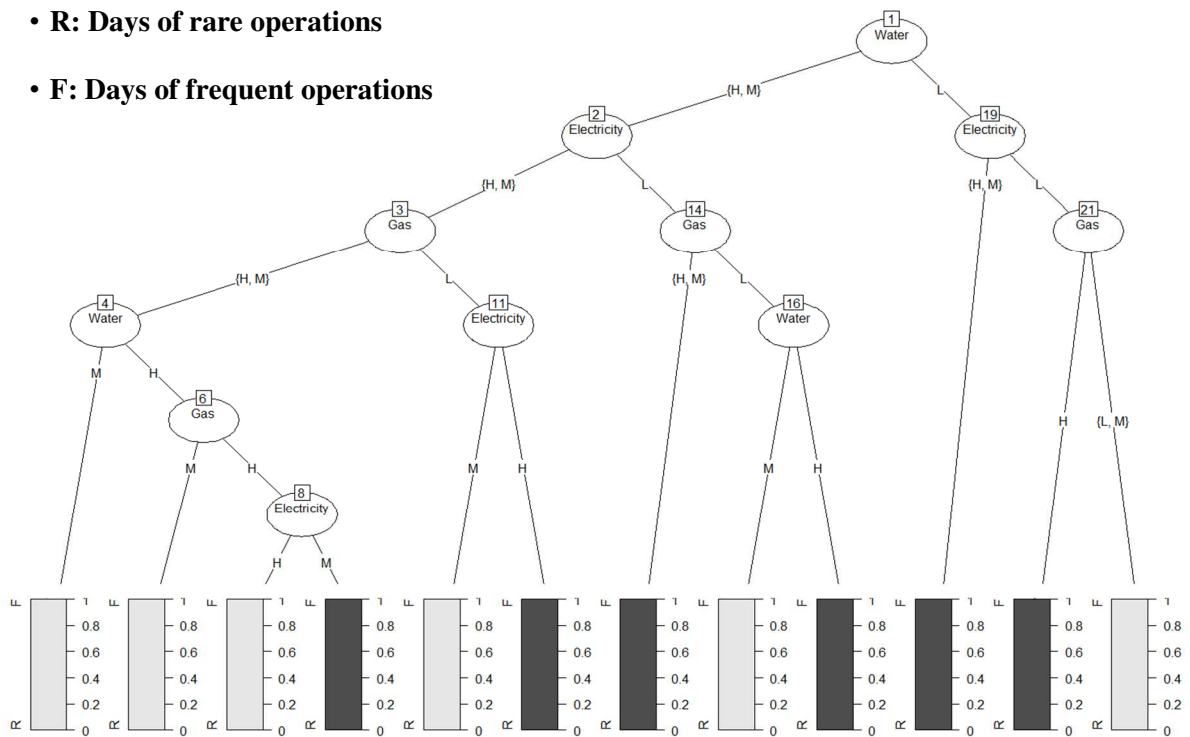
352 determine the status of each day of the new period based on the 27 (3^3) possible combinations
 353 of the three consumption sectors.

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.

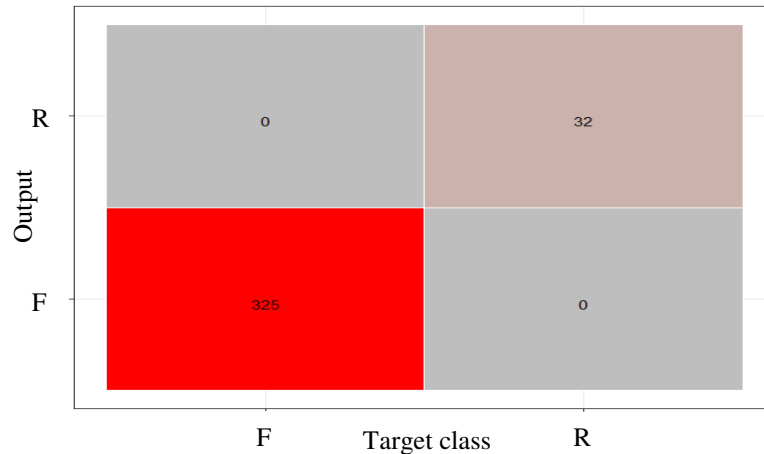


Figure 13: Confusion matrix for the combined data from year 2015

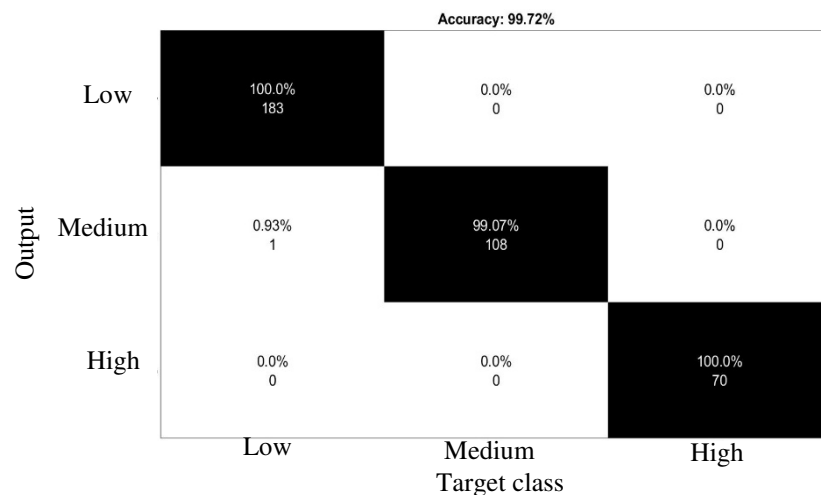
According to Figure 13, it is shown that the distribution of days according to our model is indeed identical to that of the 2015 tree. It can be stated therefore that our model functions with a 0% error.

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.-

403 one SVM (OVO SVM) method will be used herein in order to classify the daily data of year
404 2017.

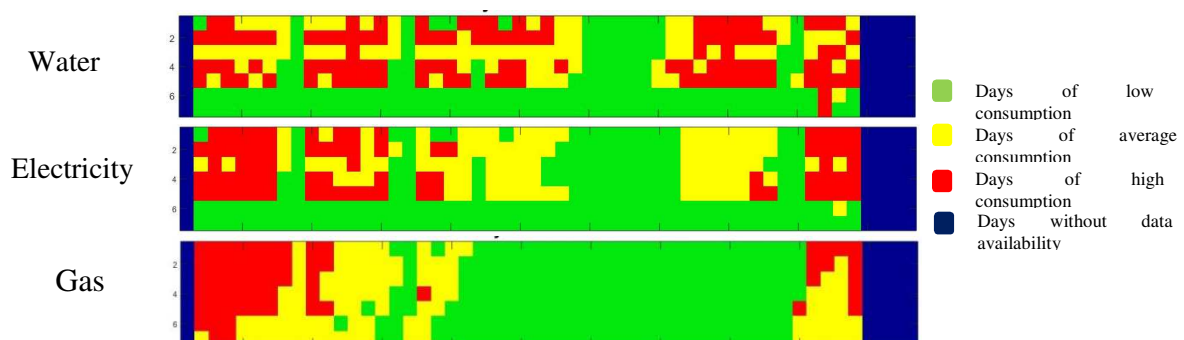
405 SVM refer to non-parametric classifiers, in an effort to obtain the optimal separating
406 hyperplane between binary classes by respecting the maximized margin criterion. Our work
407 however entails a multiclass classification case, which necessitates a one-to-one strategy
408 applicable to any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

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

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

416

417 4.2. Automatic assignment of each day's status by the decision tree



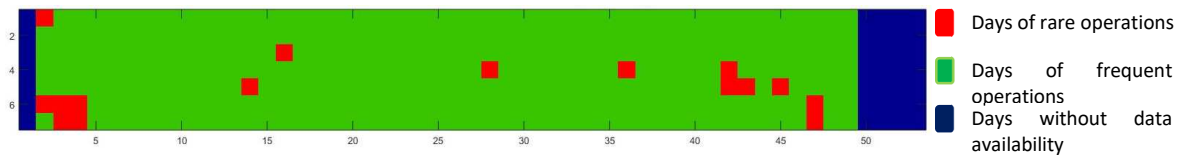
418

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

421 Before evaluating the grouping of days for year 2017 based on our model's assignment of rare
422 vs. frequent case, let's attempt to extract the days of rare operations visually from Figure 15.

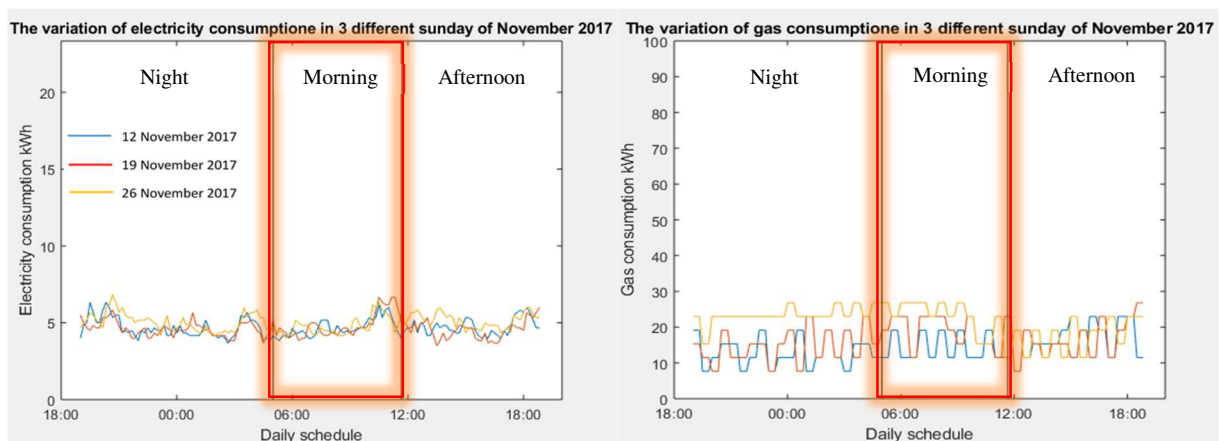
423 In the bottom left, for example, the weekend of the second week of 2017 indicates high gas
424 consumption combined with low water and electricity consumption. Moreover, during the

425 weekend prior to the final two weeks, high water consumption is combined with low
 426 electricity consumption and a medium gas consumption. These situations are rarely found
 427 during other weekends and thus considered rare in our model.
 428 Let's now determine whether the model has accurately classified the days of year 2017 as rare
 429 vs. frequent.



430
 431 *Figure 16: Rare vs. frequent days of operations at the secondary school for 2017*
 432 *during the morning hours (the colors are the same as those in Fig. 11)*

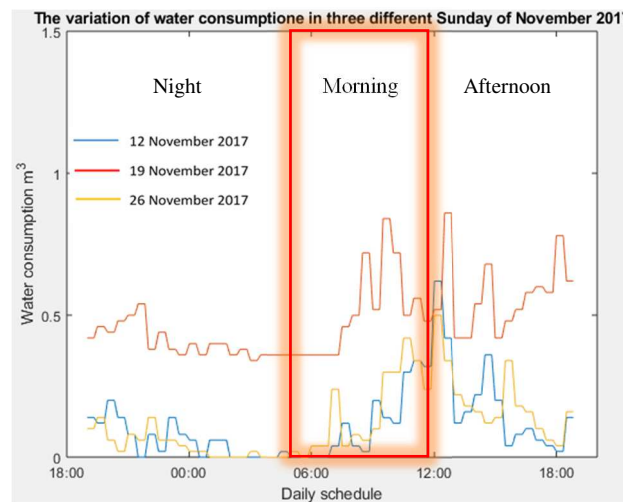
433 From Figure 16, we conclude that the days classified as rare are indeed rarely repeated in the
 434 previous figure; moreover, these days are considered rare in our model. In focusing on
 435 Sunday of Week 47 (i.e. November 19th 2017), Figure 15 indicates that during the morning
 436 hours, gas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



443

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

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



450
451 *Figure 18: Evolution in water consumption during 3 consecutive Sundays in November 2017*

452 A higher water consumption level is found during the morning of November 19th than that of
453 the other two Sundays. The facility manager is then prompted to look for the cause of this
454 atypical behavior (e.g. unforeseen event, water leak). It can also be stated that the main
455 purpose herein, namely automatically detecting potential operating anomalies, has been
456 accomplished. It is now possible to perform a classification (e.g. with SVM) in near real-time
457 (i.e. at the end of every period of the day) according to 3 consumption sectors. Having
458 established the consumption level classes, the next step consists of applying our protocol to
459 the timeline and defining whether the period features a rare or frequent form of operations.
460 If the period is deemed "rare", this would not necessarily suggest abnormality or building
461 system dysfunction. The cause might instead be a special event organized on the school
462 grounds (e.g. an open house day during the weekend), yet our system is still capable of
463 alerting the facility manager, who examines the curves daily and communicates with all

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

466

467 5. Discussion

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

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

484

485 6. Conclusion

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

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

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

509

510 7. Acknowledgments

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

514 8. References

- 515 Behnoud far, Pouria, Pantea Hosseini, et Ali Azizi. 2017. « Permeability determination of cores based
516 on their apparent attributes in the Persian Gulf region using Navie Bayesian and Random
517 forest algorithms ». *Journal of Natural Gas Science and Engineering* 37 (January): 52-68.
518 <https://doi.org/10.1016/j.jngse.2016.11.036>.
- 519 Brodley, Carla E., et Paul E. Utgoff. 1992. « Multivariate Versus Univariate Decision Trees ». Amherst,
520 MA, USA: University of Massachusetts.
- 521 Capozzoli, Alfonso, Fiorella Lauro, and Imran Khan. 2015. « Fault detection analysis using data mining
522 techniques for a cluster of smart office buildings ». *Expert Systems with Applications* 42 (9):
523 4324-38. <https://doi.org/10.1016/j.eswa.2015.01.010>.
- 524 Chicco, G., R. Napoli, F. Pigliione, P. Postolache, M. Scutariu, et C. Toader. 2004. « Load pattern-based
525 classification of electricity customers ». *IEEE Transactions on Power Systems* 19 (2): 1232-39.
526 <https://doi.org/10.1109/TPWRS.2004.826810>.
- 527 Cortes, Corinna, et Vladimir Vapnik. 1995. « Support-Vector Networks ». *Machine Learning* 20 (3):
528 273-97. <https://doi.org/10.1007/BF00994018>.
- 529 Davies, D. L., et D. W. Bouldin. 1979. « A Cluster Separation Measure ». *IEEE Transactions on Pattern
530 Analysis and Machine Intelligence PAMI-1* (2): 224-27.
531 <https://doi.org/10.1109/TPAMI.1979.4766909>.
- 532 Farrelly, Colleen M., Seth J. Schwartz, Anna Lisa Amodeo, Daniel J. Feaster, Douglas L. Steinley, Alan
533 Meca, et Simona Picariello. 2017. « The analysis of bridging constructs with hierarchical
534 clustering methods: An application to identity ». *Journal of Research in Personality* 70
535 (October): 93-106. <https://doi.org/10.1016/j.jrp.2017.06.005>.
- 536 Joshi, Aastha, et Rajneet Kaur. 2013. « Comparative Study of Various Clustering Techniques in Data
537 Mining ». *International Journal of Advanced Research in Computer Science and Software
538 Engineering*, 3.
- 539 Juuti, Mika, Francesco Corona, and Juha Karhunen. 2018. « Stochastic discriminant analysis for linear
540 supervised dimension reduction ». *Neurocomputing* 291 (May): 136-50.
541 <https://doi.org/10.1016/j.neucom.2018.02.064>.
- 542 Li, Guannan, Yunpeng Hu, Huanxin Chen, Haorong Li, Min Hu, Yabin Guo, Jiangyan Liu, Shaobo Sun,
543 and Miao Sun. 2017. « Data partitioning and association mining for identifying VRF energy
544 consumption patterns under various part loads and refrigerant charge conditions ». *Applied
545 Energy* 185 (January): 846-61. <https://doi.org/10.1016/j.apenergy.2016.10.091>.
- 546 Liu, Yang, Jian-Wu Bi, et Zhi-Ping Fan. 2017. « A method for multi-class sentiment classification based
547 on an improved one-vs-one (OVO) strategy and the support vector machine (SVM)
548 algorithm ». *Information Sciences* 394-395 (July): 38-52.
549 <https://doi.org/10.1016/j.ins.2017.02.016>.
- 550 Morbitzer, Christoph, Paul Strachan, et Catherine Simpson. 2004. « Data mining analysis of building
551 simulation performance data ». *Building Services Engineering Research and Technology* 25
552 (August): 253-67. <https://doi.org/10.1191/0143624404bt098oa>.
- 553 Otukei, J. R., et T. Blaschke. 2010. « Land cover change assessment using decision trees, support
554 vector machines and maximum likelihood classification algorithms ». *International Journal of
555 Applied Earth Observation and Geoinformation*, Supplement Issue on « Remote Sensing for
556 Africa – A Special Collection from the African Association for Remote Sensing of the
557 Environment (AARSE) », 12 (February): S27-31. <https://doi.org/10.1016/j.jag.2009.11.002>.
- 558 Petrovic, Slobodan. 2018. « A comparison between the silhouette index and the Davies-Bouldin index
559 in labelling ids clusters », September.
- 560 Quinlan, J. R. 1986. « Induction of Decision Trees ». *Machine Learning* 1 (1): 81-106.
561 <https://doi.org/10.1007/BF00116251>.
- 562 Ruuska, Salla, Wilhelmiina Hämäläinen, Sari Kajava, Mikaela Mughal, Pekka Matilainen, et Jaakko
563 Mononen. 2018. « Evaluation of the confusion matrix method in the validation of an

564 automated system for measuring feeding behaviour of cattle ». *Behavioural Processes* 148
565 (mars): 56-62. <https://doi.org/10.1016/j.beproc.2018.01.004>.
566 Subbalakshmi, Chatti, G. Rama Krishna, S. Krishna Mohan Rao, et P. Venketeswa Rao. 2015. « A
567 Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data
568 Set ». *Procedia Computer Science*, Proceedings of the International Conference on
569 Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty
570 Palace & Island Resort, Kochi, India, 46 (January): 346-53.
571 <https://doi.org/10.1016/j.procs.2015.02.030>.
572 Usman, Ghousia, Usman Ahmad, et Mudassar Ahmad. 2013. « Improved K-Means Clustering
573 Algorithm by Getting Initial Centroids », 9.
574 Xiao, Fu, et Cheng Fan. 2014. « Data mining in building automation system for improving building
575 operational performance ». *Energy and Buildings* 75 (June): 109-18.
576 <https://doi.org/10.1016/j.enbuild.2014.02.005>.
577 Yu, Zhun, Fariborz Haghighat, Benjamin C.M. Fung, et Hiroshi Yoshino. 2010. « A Decision Tree
578 Method for Building Energy Demand Modeling ». *Energy and Buildings* 42 (10): 1637-46.
579 <https://doi.org/10.1016/j.enbuild.2010.04.006>.
580