



HAL
open science

Identification of indoor air quality events using a K-means clustering analysis of gas sensors data

Alexandre Caron, Nathalie Redon, Patrice Coddeville, Benjamin Hanoune

► **To cite this version:**

Alexandre Caron, Nathalie Redon, Patrice Coddeville, Benjamin Hanoune. Identification of indoor air quality events using a K-means clustering analysis of gas sensors data. *Sensors and Actuators B: Chemical*, 2019, *Sensors and Actuators B: Chemical*, 297, pp.126709. 10.1016/j.snb.2019.126709 . hal-02276311

HAL Id: hal-02276311

<https://hal.univ-lille.fr/hal-02276311v1>

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 **Identification of indoor air quality events using a K-means clustering analysis of gas sensors** 2 **data**

3 Alexandre Caron^{1,2}, Nathalie Redon², Patrice Coddeville², Benjamin Hanoune¹

4 1. Univ. Lille, CNRS, UMR 8522 – PC2A – Physicochimie des Processus de Combustion et de
5 l’Atmosphère, F-59000 Lille, France.

6 2. IMT Lille Douai, Univ. Lille, SAGE - Département Sciences de l’Atmosphère et Génie de
7 l’Environnement, F-59000 Lille, France.

8 * Corresponding author: nathalie.redon@imt-lille-douai.fr

9 Tel: (+33)3 27 71 24 77

10 Fax: (+33)3 27 71 29 14

11 IMT Lille Douai, SAGE, 941 rue Charles Bourseul, CS10838, F-59508 Douai, France

12

13

14 **Abstract:**

15 Commercial miniature gas sensors, because they are smaller and cheaper than conventional
16 instruments, can be deployed in large numbers to investigate indoor air quality, for research and
17 operational purposes. To compensate for their limited metrological performances, it is necessary to
18 develop relevant data treatment procedures. We applied an unsupervised classification approach
19 based on the bisecting K-means algorithm to data acquired by online gas analyzers and by miniature
20 sensors during a measurement campaign in a low energy school building. This procedure, applied to
21 the analyzers measurements, was able to distinguish the ventilation status and the specific air
22 quality events taking place in the classroom. The same procedure applied to the data from the
23 sensors, even though they were not calibrated beforehand, was also able to identify the same events.
24 The good agreement between the two sets of results validates the methodology and opens up new
25 perspectives for a massive deployment of sensors inside buildings.

26

27 **Key words:** (indoor) air pollution, electronic gas sensors, unsupervised classification, k-means
28 clustering

29 **1. Introduction**

30 Indoor environments, where people in developed countries spend up to 90% of their time, present
31 high specific pollutant concentrations [1,2], inducing a risk for human health [3,4]. Indoor
32 pollutants, especially volatile organic compounds (VOCs), are emitted from building materials,
33 furniture, consumer products, from the occupants themselves and their activities. The air transferred
34 from outdoors also has a significant impact on the pollutants breathed indoors [5]. There is a need
35 for large scale and continuous measurements of the indoor air quality (IAQ) in various domains: (i)
36 for research purposes, in order to increase the understanding of the determinants of indoor air
37 pollution, such as the identification of pollution sources and of the pollutants trends and temporal
38 and spatial evolution, (ii) to allow mandatory or voluntary IAQ assessments of buildings, (iii) to
39 communicate helpful information to the occupants on the relationships between their daily activities
40 and the induced pollution levels, and also to alert them and implement corrective actions when
41 critical thresholds are exceeded, (iv) to control the operation of ventilation or air treatment systems
42 in order to reach the best compromise between health and energy consumption considerations. The
43 conventional gas analyzers used for laboratory research and regulatory outdoor air monitoring can
44 be deployed during research oriented measurement campaigns but not for real-time monitoring of
45 occupied indoor environments. These bulky analyzers generate many nuisances, such as noise or
46 vibrations, induce a considerable electrical consumption, and are too expensive to be
47 simultaneously deployed in many places.

48 In recent years, gas micro-sensors emerged as alternative relevant tools for air quality monitoring
49 [6–9]. New sensitive materials are constantly being developed in order to achieve better sensitivity,
50 selectivity and stability [10,11]. More and more micro-sensors are commercially available [12,13],
51 prompting many recent studies where their performances are investigated [14–17]. Among these
52 sensors, a distinction must be made between intrinsically non-selective sensors and selective or
53 nearly-selective sensors. Non-selective sensors are commonly based on metal oxide semiconductive
54 materials which respond to multiple compounds in the air, and are generally used in combinations

55 or arrays of sensors, also known as e-noses. Selective sensors include the electrochemical sensors
56 targeting compounds such as carbon monoxide, ozone, nitrogen monoxide or dioxide, sulfur
57 dioxide, hydrogen sulfide, the NDIR sensors used for carbon dioxide, as well as the PID sensors for
58 total volatile organic compounds measurements. In spite of the technological progresses, the
59 currently available sensors still require a complete metrological characterization [18], with in
60 particular the assessment of their reliability over time. Even if they are very sensitive, the response
61 of most sensors shows interferences with other compounds than the targeted one, depends upon the
62 temperature and humidity, and drifts with time [19–21]. Such a preliminary step of characterization
63 and calibration is not necessary for arrays of MOS sensors when used in association with pattern
64 recognition algorithms [22–27], which are methods used in data mining for the extraction of useful
65 information and the exploration of data correlation. Supervised classification approaches [28] are
66 based on a classifier built from a training set with a collection of labeled data, and then used to
67 assign new unlabeled data instances. Unsupervised classification [29] refers to algorithms that
68 require no training set (blind partitioning), no a priori knowledge of the structure of the dataset, and
69 automatically define the different classes. However, the physical meaning of these classes needs to
70 be a posteriori interpreted or verified by the expert.

71 In the present study, we investigate the potential of unsupervised classification, or clustering, to
72 analyze the output of selective gas sensors. The dataset used for the analysis has been acquired
73 during a field campaign aiming to investigate the drivers and dynamics of IAQ in a low energy
74 building [5].

75 Many clustering algorithms have been developed for data mining, such as reviewed in [29].
76 Different clustering algorithms, or even different ways to use them on the same dataset, can lead to
77 different partition results. None of them have proved to be the best technique in a large amount of
78 configurations. Some of these algorithms have been applied to electronic nose data clustering [30],
79 each with its respective possibilities and limitations [25,31,32], depending on the application. For

80 electronic nose data, hierarchical clustering is commonly used [33–37]. It results in a hierarchical
81 structure of the dataset that is more informative about the link between each group of data than the
82 unstructured clusters provided by other techniques. In addition, its representation (dendrograms)
83 allows the selection of the number of clusters. Centroid-based algorithms such as K-means [38] are
84 less considered in the micro gas sensor field [39,40] and especially for air quality investigation,
85 though K-means is a simple partitional algorithm and one of the most widely used techniques in
86 data mining thanks to its performances [41]. K-means algorithms combine simplicity, ease of
87 implementation and of use, speed of convergence, even with a large number of variables and
88 clusters, and ability to process datasets with missing values. For these reasons, we have chosen this
89 method to analyze the data from gas sensors in the investigation of indoor air quality.

90

91 **2. Materials and methods**

92 **2.1. Instruments and measurements settings**

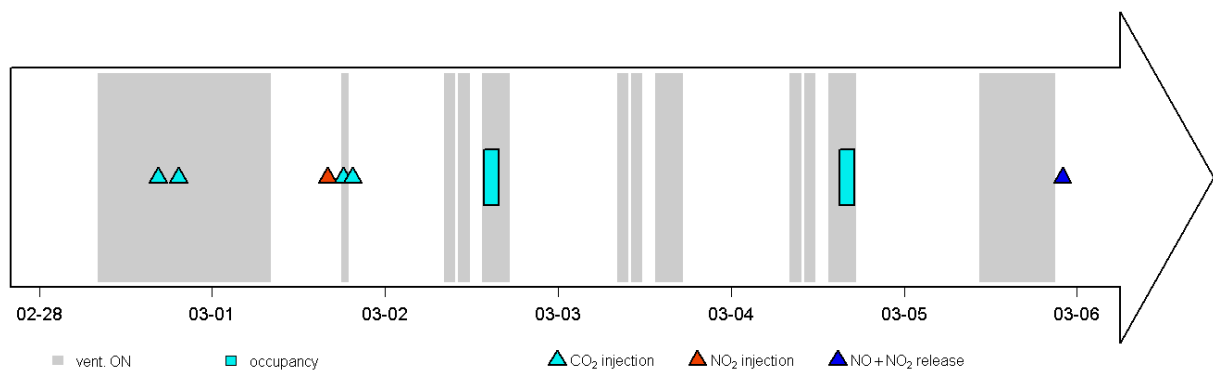
93 Measurements used for this work were performed during the Mermaid study [5]. This field
94 campaign has been carried out in February and March 2015 in a 51 m² (139 m³) classroom of an
95 energy efficient junior high school building in Northern France. Details on the analytical
96 instruments used for this study can be found in [5]. Only the data from the pollutants from outdoors
97 (NO_x and O₃), measured with online analyzers (Thermo 42i and Thermo 49i), and of CO₂, measured
98 with Testo 480 probe located at the center of the room are considered here. In addition to these
99 instruments, miniature sensors were installed in the center of the classroom, 90 cm above the floor.
100 They monitor nitrogen oxide (electrochemical, Alphasense NO-B4), nitrogen dioxide
101 (electrochemical, Alphasense NO₂-B4), ozone (electrochemical, Alphasense O₃-B4) and carbon
102 dioxide (NDIR, Alphasense CO₂-IRC A1). A Raspberry Pi B+ and an Arduino board are used to
103 collect, store and transmit the data. No correction or calibration of the sensors has been performed

104 prior to their installation in the room, and the raw output (voltage) is analyzed with the bisecting K-
105 means procedure.

106

107 2.2. Experimental datasets

108 The data used for this work corresponds to a 6-day continuous measurement (28 February 2015 to 6
109 March 2015). Fig. 1 presents the ventilation status and specific events taking place in the room
110 during this period. These events include three CO₂ injections with ventilation ON, one CO₂
111 injection with ventilation OFF, for the determination of the air exchange rate of the room, two
112 periods with people in the room, one NO₂ injection, and an accidental release of both NO and NO₂.



113

114 **Fig. 1.** Specific conditions and events occurring in the classroom

115

116 Dataset 1 (reference instruments) is a 4 x 8160 matrix consisting of 4 variables which are the
117 concentrations of ozone, nitrogen oxide and nitrogen dioxide (in ppb) measured by the online
118 analyzers and of carbon dioxide concentration (in ppm) measured by the Testo 480 probe, with a
119 one minute resolution. The analyzers and CO₂ probe were calibrated at the beginning of the
120 campaign.

121 Dataset 2 (electrochemical and NDIR sensors) is also a 4 x 8160 matrix, consisting of the voltage
122 (mV) outputs of the 3 electrochemical sensors and of the carbon dioxide concentration (in ppm)

123 provided by the NDIR sensor, with a 1 min resolution. As previously mentioned, no correction or
124 calibration of the sensors signals was performed.

125

126 **2.3 K-means implementation**

127 Numerous extensions of the basic K-means algorithm have been developed in order to improve its
128 partitioning abilities for dedicated applications. Classical K-means performs a direct classification
129 of the full dataset into a number of K clusters, whereas other methods, such as the bisecting K-
130 means method [38], use a hierarchical method and split the data by iteration. In 2000, Steinbach et
131 al. have shown that bisecting K-means, computed as an hybrid approach between the run-time
132 efficiency of conventionnal K-means and the quality efficiency of an agglomerative hierarchical
133 clustering, has higher performances than the conventional algorithm [44]. It as also been
134 demonstrated that bisecting K-means is relatively insensitive to the initialisation of the clusters
135 centers and has a higher computational efficiency than the conventional K-means algorithm [45]. In
136 spite of its simplicity, and of its wide use, the reasons behind the efficiency of the K-means
137 algorithm still need to be fully understood. The only required input from the user is the number of
138 clusters into which the dataset must be split. In the present work, we used the K-means procedure
139 with a program written in Julia 0.4.0.

140 The raw time series data of each dataset is used for input, except for the carbon dioxide
141 concentration, measured either by the Testo probe or the NDIR sensor, for which the logarithm
142 (base 10) is used, because of the much wider range over which this concentration can vary.
143 Preliminary calculations, not reported here, with the direct CO₂ concentration, were performed but
144 the clustering was less efficient than with the log values.

145 The only other parameter for input is the number K of clusters. Values of K ranging from 2 to 10
146 have been investigated for each dataset. Some mathematical criteria to determine the optimal

147 number of clusters have been proposed [44–46], but these may have no physical meanings, and we
148 have chosen to leave this determination to the judgment of the expert a posteriori.

149 The program first normalizes all the observations in the dataset, in order to standardize their
150 respective weight on the cluster partition. The clustering process is then initialized by splitting the
151 normalized dataset into two subsets. Then, each datapoint is assigned to its nearest cluster center
152 (centroid) according to the euclidian distance. The process is repeated until the association of all the
153 observations of the dataset to its respective cluster does not change anymore and the sum of the
154 squared errors of the distance is minimized. The process is iterated by splitting into two new
155 clusters the cluster with the highest sum of squared residuals, following the same procedure, until
156 the required number of clusters is reached. Tests have shown that changing the initial partition does
157 not influence the final results. The output file of the bisecting K-means clustering consists of the
158 raw data tagged with their respective cluster, and of the coordinates of the centroids of each of the K
159 clusters.

160 To compare the outputs of the K-means procedure applied to the two dataset (reference dataset and
161 sensors dataset), we will consider the overlap ratio between a cluster from dataset 2 and its
162 equivalent cluster from the reference dataset 1. It is expressed as the number of correct matches
163 between each datapoint in a defined cluster of each dataset normalized by the number of data points
164 in the reference cluster. Thus, a value of 1 indicates a perfect overlap, while a value of 0 means no
165 overlap between the two clusters.

166

167

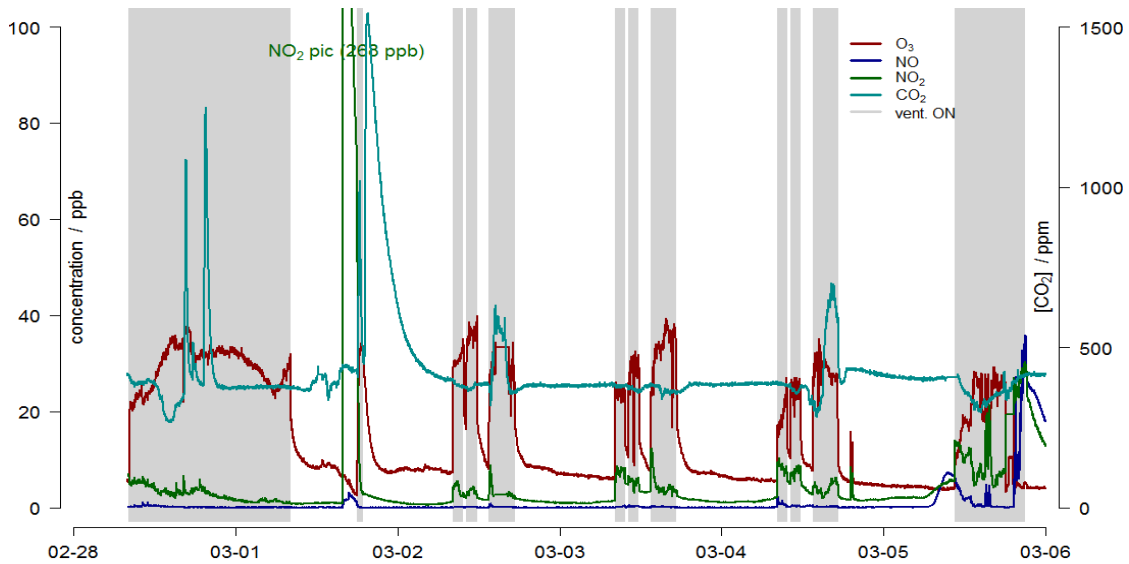
168

169 **3 Results and discussion**

170 **3.1 Clustering of the reference online analyzers measurements (dataset 1)**

171 Dataset 1 (concentrations of NO, NO₂, O₃ and CO₂ measured by the reference instruments) are
172 presented on Fig. 2, together with the ventilation status (ON/OFF) which has been found to be the
173 main driver of the chemistry in the room [5]. The specific events described on Fig.1 are clearly
174 distinguished on the concentration time chart of Fig. 2..

175

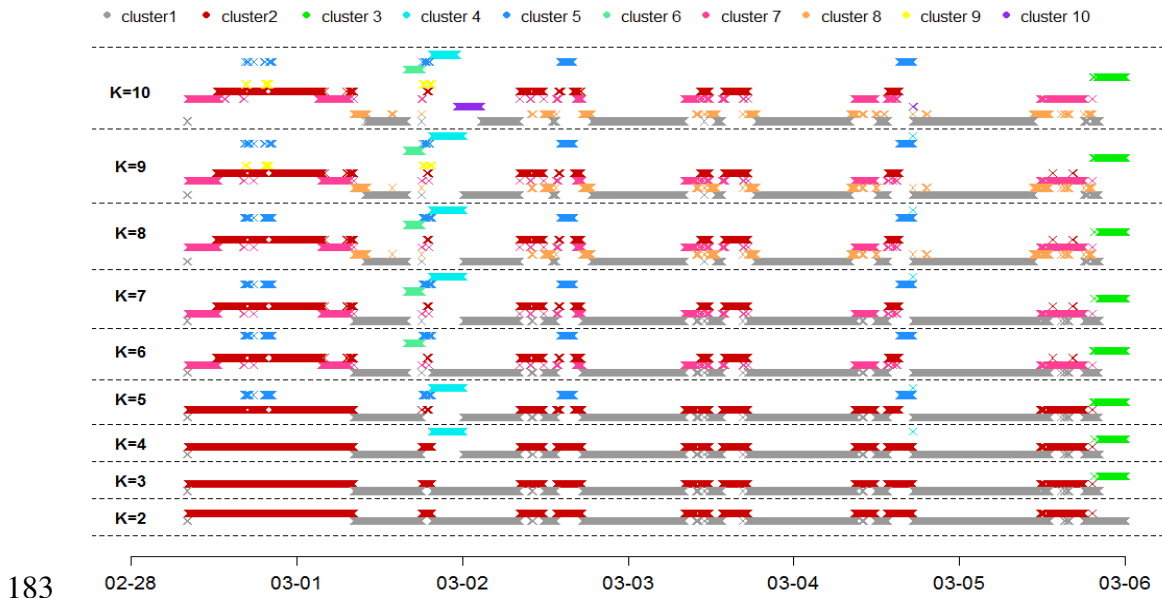


176

177 **Fig. 2.** Concentration time series measured by the reference analyzers. The ventilation status is
178 indicated by the background color, white for ventilation OFF, grey for ventilation ON.

179

180 The results of the clustering process, applied considering a number of clusters ranging from 2 to 10,
181 are displayed on Fig. 3. This chart clearly illustrates the hierarchical structure of the bisecting K-
182 means process.



183
184 **Fig. 3.** Dataset 1 clustering process output, for K values from 2 to 10

185

186 Data are successively grouped into new clusters, with the 2 clusters divided into smaller sub-
 187 clusters when K increases. For the initial partition (K=2), the bisecting K-means separates the
 188 dataset according to the ozone concentration, with cluster 1 corresponding to the periods with no
 189 ozone, i.e. when the ventilation is OFF, while cluster 2 refers to the periods where the ozone
 190 concentration is significant, i.e. when the ventilation is ON. This result agrees with the result of the
 191 MERMAID campaign [5], where the main driver of the IAQ within the room was found to be the
 192 ventilation status. When K increases from 3 to 5, the newly defined clusters can be related to the
 193 specific known events occurring in the room presented in Fig. 1. Cluster 3 corresponds to the
 194 accidental release of NO and NO₂ in the classroom, cluster 4 corresponds to the injection of carbon
 195 dioxide when the ventilation system is OFF, and cluster 5 corresponds to the moments when the
 196 ventilation system is ON and with CO₂ concentration higher than the background, due to either
 197 controlled injection of CO₂ in the room or to human presence.

198 Interestingly, the results for K=6 differ drastically from the previous cases. The new cluster (#6)
 199 corresponds to the injection of NO₂, the previously found cluster 4 (injection of CO₂ during
 200 ventilation OFF conditions) disappears, and cluster 2 (ventilation ON) is divided into two sub-

201 clusters. However, cluster 4 reappears when considering the $K=7$ partitioning, with no change on
202 the previously determined clusters. This indicates that a criterion of stability of the clusters when
203 increasing their number must be considered to correctly interpret the results of unsupervised
204 classification. Higher values of K up to 10 induce a refinement in the previously determined
205 clusters, according to different levels of ozone. Cluster 8 corresponds to the transition period when
206 the pollutants from outdoor air slowly decrease due to their reactivity, with no compensation from
207 the ventilation, cluster 9 represents the transient periods with elevated ozone and CO_2
208 concentration, and cluster 10 corresponds to a moderate CO_2 concentration with no ozone. These
209 transitions periods, and in particular the mixing between indoor and outdoor pollutants, are
210 generally overlooked during standard analysis.

211 While clusters 8 to 10 correspond to actual, well-defined conditions in the room, their interpretation
212 is less forward than the previous 7 clusters, and we will consider henceforth that $K=7$ provides the
213 best description of the air quality events in the room. A lower K value will miss some events, and a
214 higher K value will only split classes with physical meaning according to the levels of
215 concentration. $K=7$ leads to the best compromise between the lowest numbers of cluster and the
216 correct and separate description of every known event.

217

218 **Table 1** Summary of clusters size and coordinates of centroids for K=7 clustering (dataset 1)

Cluster	Events	n. obs	O ₃ / ppb	NO / ppb	NO ₂ /ppb	CO ₂ /ppm
C1	Vent. OFF	4480	7.7	0.5	2.8	396
C2	Vent. ON	1180	24.4	0.4	5.6	369
C3	Vent. ON	1467	32.4	0.2	3.1	367
C4	NO+NO ₂	228	4.3	23.8	20.5	416
C5	NO ₂	128	4.7	1.9	122.3	435
C6	CO ₂ (vent. OFF)	277	9.4	0.1	1.8	965
C7	CO ₂ (vent. ON)	400	31.1	0.2	4.0	666

219

220 The output results from the partition of dataset 1 into 7 clusters are presented in Table 1, which
 221 summarizes the size (number of data points in the cluster) and centroid coordinates of each cluster,
 222 together with their assignment. Cluster 1 groups the majority of the datapoints, with 4480
 223 observations, and is characterized by low levels of every pollutant, as it corresponds to the periods
 224 with no ventilation, i.e. no intake of outdoor pollutants. A high ozone level is the dominant
 225 parameter that influences cluster 2 (1180 obs.) and cluster 3 (1467 obs.). These two clusters
 226 correspond to periods when the ozone concentration increases in the classroom due to the activation
 227 of the ventilation system. Cluster 4 (228 obs.) data are characterized by background CO₂ (416 ppm)
 228 and O₃ (4.3 ppb) level, together with higher NO and NO₂ levels (> 20 ppb). This partition
 229 corresponds to the short event that takes place at the end of the measurement period (Fig. 1, Fig. 2)
 230 which is an accidental release of NO and NO₂. Cluster 5 (128 obs.) data are characterized by low
 231 level of pollutants in the classroom, except for NO₂ (112.3 ppb), and corresponds to a voluntary
 232 injection of nitrogen dioxide in the classroom. Cluster 6 represents the CO₂ injection when the
 233 ventilation is off, with low O₃, NO and NO₂ concentrations. Cluster 7 represents the moments with

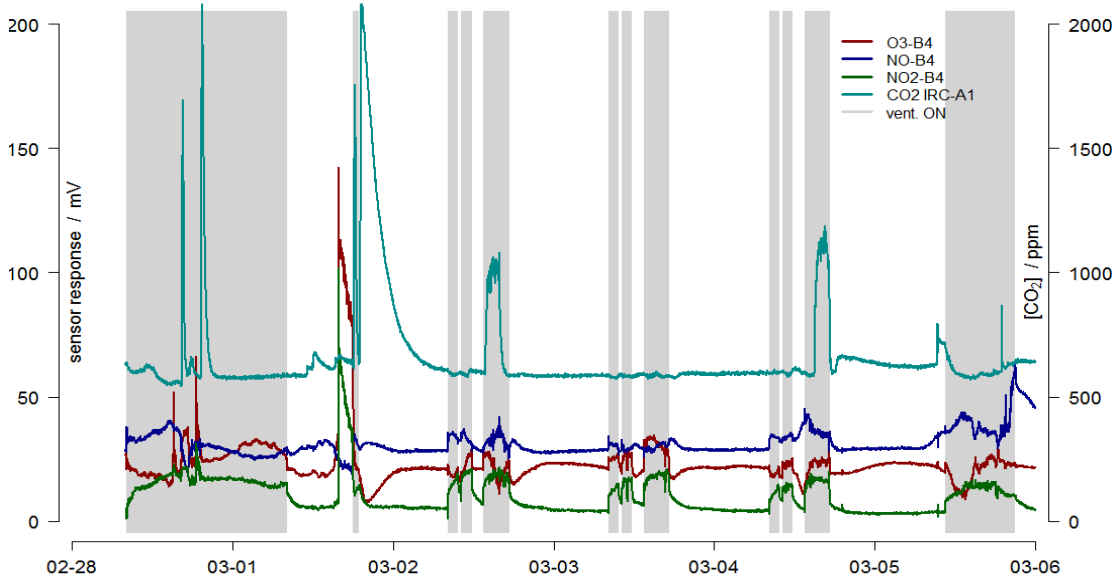
234 high O₃ concentration, that is during ventilation ON, together with high CO₂ concentration, due to
235 either a CO₂ injection or the presence of people in the room.

236

237 3.2 Clustering of the electrochemical and NDIR sensors measurements (dataset 2)

238 The time series evolution of dataset 2 is presented on Fig. 4. The direct interpretation of the signals
239 from the sensors must be done with caution, because of possible chemical interferences, such as the
240 cross response of NO₂ and O₃ on electrochemical sensors [47] and because the sensors were not
241 calibrated before taking the measurements. This is where unsupervised classification algorithms can
242 really help analyzing the data.

243



244

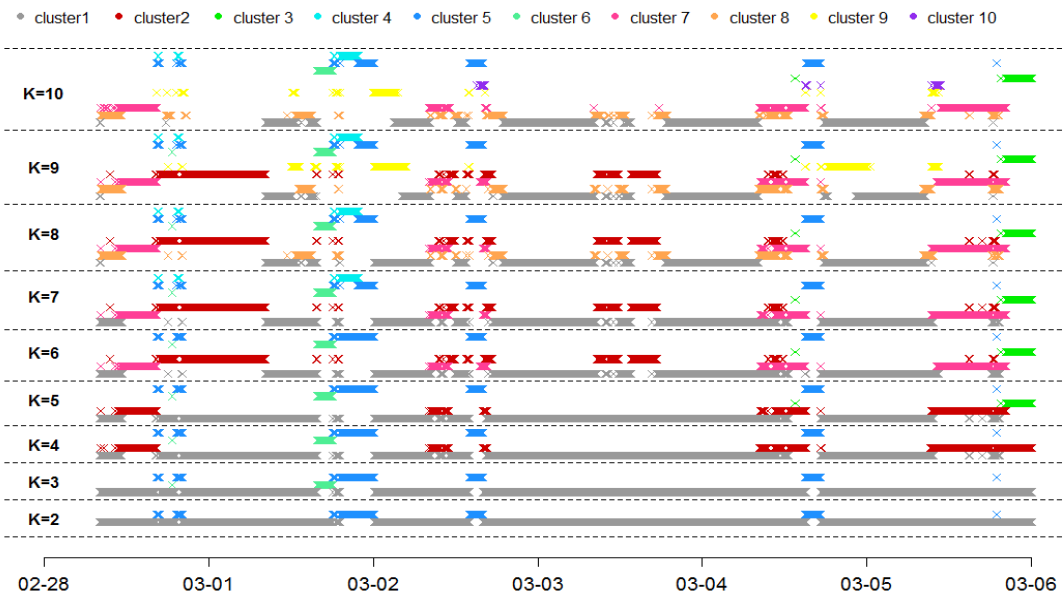
245 **Fig. 4.** Time series of the sensors signals. The ventilation status is indicated by the background
246 color, white for ventilation OFF, grey for ventilation ON.

247

248 The results of the process for each K value between 2 and 10 are displayed on Fig. 5, illustrating, as
249 for dataset 1, the hierarchical structure induced by the bisecting K-means process. However, the
250 different clusters do not appear in the same order as for dataset 1. K=2 separates the datapoints with
251 high CO₂ concentration. For K=3, the new cluster can be related to the injection of NO₂ in the
252 classroom. The accidental release of NO and NO₂ is identified in dataset 2 at K=5. The efficiency of

253 the ventilation, impacting the O₃ concentrations (oxydants coming from outdoor), is identified by
 254 clusters built from K=4 and K=6. At K=6, every known event can be related to a specific cluster,
 255 except the distinction between the CO₂ injections during the ventilation period or without
 256 ventilation, which appears only for the classification into 7 clusters. The clusters obtained for
 257 dataset 2 when increasing the value of K up to 10 are difficult to put in regard to the actual events
 258 taking place in the room. For instance, cluster 8 cannot be related to any event in the room. Also,
 259 cluster 9 is not stable, and partly disappears for K=10. Still, as the clusters do not appear in the same
 260 order when considering the reference dataset or the sensor dataset, it is necessary to explore the
 261 analysis with high K values, in order not to miss a specific event.

262
 263
 264



265
 266 **Fig. 5.** Dataset 2 (sensors) clustering process output for K values up to 10. To help the comparison
 267 with the results of dataset 1, the cluster numbering and color coding does not follow the increasing
 268 K order, but is taken at the K=7 level.

269
 270 This analysis demonstrates that the bisecting K-means procedure can also be used for sensors, when
 271 the activities in the room are known, even when the sensors are not calibrated before the

272 experiment. This comes from the fact that the information does not lie in the absolute value of the
273 measurements, but in the evolution of the intensity of the signals. This holds true also in spite of the
274 poor selectivity, as in the case of the NO₂ and O₃ sensors. However, the K-means processing of
275 dataset 2 also points to the difficulty of determining the optimal number of clusters, which in the
276 present case could be 7 or 8, depending if we consider the physical meaning of the classes, that is if
277 we use contextual information to supplement the measured data, or 8 if we consider the stability of
278 the clusters, without contextual information.

279 As discussed when treating previously dataset 1, and considering our goal of comparing the results
280 obtained from the two datasets, we will restrict hereafter our discussion to K=7. The results from
281 the partition of dataset 2 into 7 clusters are presented in Table 2, summarizing the size and centroid
282 coordinates of each cluster. Cluster 1 (4394 obs.) is characterized by a background level of the O₃-
283 B4 and NO-B4 response, of respectively 21.2 and 29.54 mV, a background (uncorrected) CO₂
284 concentration value of 614 ppm, and a low NO₂-B4 value of 5.8 mV. This partition matches well
285 with the period when the ventilation is turned off. Cluster 2 (1246 obs.) is characterized by a higher
286 response from the NO-B4 sensor (37.20 mV) and from the NO₂-B4 sensor (13.90 mV). Cluster 3
287 (1515 obs.) is characterized by a higher response from the three electrochemical sensors. Both
288 cluster 2 and cluster 3 describe the period where the outdoor pollutant contribution is significant,
289 and can be related to the ventilation ON status. The two clusters differ by the higher values of O₃ in
290 cluster 3. Cluster 4 (225 obs.) mainly appears at the end of the measurement period and is
291 characterized by a significant increase of NO-B4 sensor response (51.49 mV). It corresponds to the
292 simultaneous increase of NO and NO₂ concentration. Cluster 5 (129 obs.) corresponds to the
293 maximal values from the O₃-B4 and NO₂-B4 sensors, respectively of 95.6 mV and 49.3 mV. This
294 data partition can be linked to the injection of nitrogen dioxide into the room. Clusters 6 (208 obs.)
295 and 7 (443 obs.) data match with the high CO₂ concentration periods, with (uncorrected)
296 concentrations of 1658 and 1035 ppm respectively.

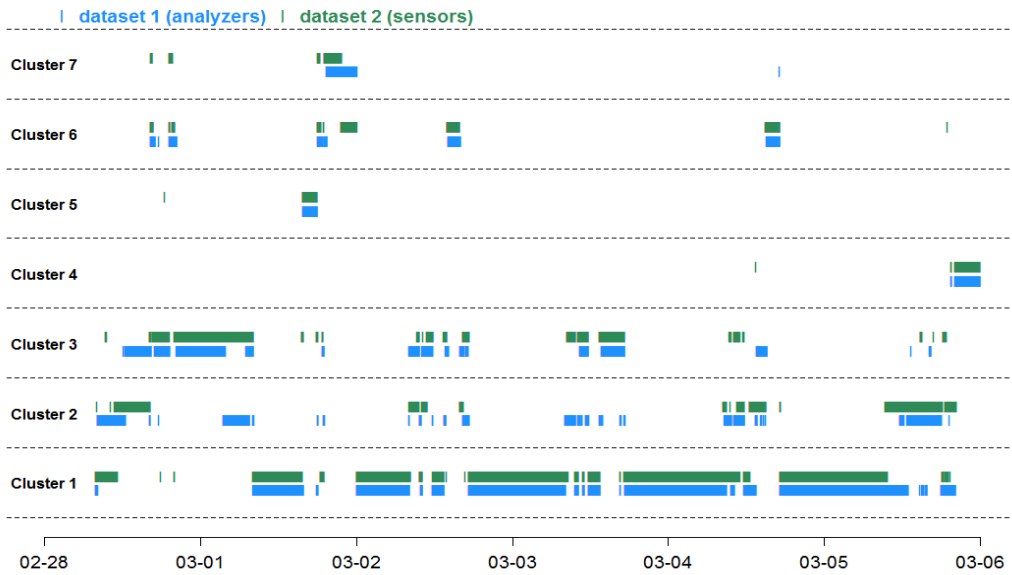
298 **Table 2** Summary of clusters size and coordinates of centroids for K=7 clustering (dataset 2)

Cluster	Events	n. obs	O ₃ -B4/ mV	NO-B4 / mV	NO ₂ -B4/mV	CO ₂ IRC-A1/ppm
C1	Vent. OFF	4394	21.20	29.54	5.80	614
C2	Vent. ON	1246	19.19	37.20	13.90	601
C3	Vent. ON	1515	28.32	28.82	16.53	592
C4	NO+NO ₂	225	22.13	51.49	7.34	642
C5	NO ₂	129	95.63	22.23	49.28	650
C6	CO ₂ (vent. OFF)	208	14.22	31.09	9.39	1658
C7	CO ₂ (vent. ON)	443	21.26	32.48	13.65	1035

299

300 **3.3 Comparison of both clustering results**

301 Fig. 6 depicts graphically the clusters obtained from the two datasets. The overlap between the two
302 sets of clusters is summarized in Table 3 (percentage of datapoints from each cluster from the
303 sensors dataset that are assigned to the cluster from the analyzers dataset). There is a general
304 agreement between the clusters from the analyzers and the ones from the sensors. Each cluster pair
305 contains roughly the same number of points (Table 1 and Table 2). This 1-to-1 correspondence
306 allows a direct comparison of the two sets of results.



307

308 **Fig. 6.** Graphical comparison of clustering results ($K=7$) between dataset 1 (analyzers) and dataset 2
 309 (sensors)

310

311 For instance, cluster 1 (vent. OFF), cluster 4 (NO+NO₂) and cluster 5 (NO₂) match very well each
 312 other, with only a few mismatched points. These clusters are associated to a well defined event,
 313 respectively ventilation OFF, NO+NO₂ release, and NO₂ injection, that is perfectly singled out by
 314 the K-means procedure, leading to an overlap ratio between dataset 2 and dataset 1 higher than 91
 315 % for cluster 1, and of 98% for clusters 4 and 5. Only about 8% of the datapoints of sensors cluster
 316 1 are associated to the reference cluster 2 (ventilation ON), though no explanation can be advanced.
 317 The agreement between the two datasets is slightly poorer for cluster 2 and 3 (vent. ON). The data
 318 are split between their respective reference clusters. This might be explained by the relatively low
 319 NO₂ and O₃ concentration amplitudes and the cross sensitivity of the NO-B4, NO₂-B4 and O₃-B4
 320 sensors [47]. 393 observations are mismatched and associated to the period without ventilation
 321 (cluster 1). Cluster 6 (CO₂ vent. OFF) and cluster 7 (CO₂ vent. ON) from the sensors dataset are
 322 also mainly divided through the two clusters of dataset 1 that describe the high CO₂ concentration
 323 periods (analyzers clusters 6 and 7), leading to overlaps of 66% at most. We assigned this weaker

324 agreement to the low NO₂ and O₃ concentration, which do not allow for a good differentiation
 325 between the ON and OFF ventilation conditions.

326

327 **Table 3** Overlap between the two sets of clusters

Reference Sensors	Cluster 1 (vent. OFF)	Cluster 2 (vent. ON)	Cluster 3 (vent. ON)	Cluster 4 (NO+NO ₂)	Cluster 5 (NO ₂)	Cluster 6 (CO ₂ OFF)	Cluster 7 (CO ₂ ON)
Cluster 1 (vent. OFF)	91.1%	23.0%	1.5%	0.0%	0.0%	0.4%	5.0%
Cluster 2 (vent. ON)	7.9%	37.1%	30.4%	2.2%	0.0%	0.0%	1.3%
Cluster 3 (vent. ON)	0.9%	39.4%	66.1%	0.0%	1.6%	0.0%	9.3%
Cluster 4 (NO+NO ₂)	0.0%	0.0%	0.0%	97.8%	0.0%	0.0%	0.0%
Cluster 5 (NO ₂)	0.0%	0.0%	0.1%	0.0%	98.4%	0.0%	0.0%
Cluster 6 (CO ₂ OFF)	0.0%	0.1%	0.0%	0.0%	0.0%	48.4%	18.3%
Cluster 7 (CO ₂ ON)	0.1%	0.4%	1.9%	0.0%	0.0%	51.3%	66.3%

328

329 Because of this similarities and links between clusters 2 and 3 and clusters 6 and 7 respectively, it is
 330 natural to simplify the data distribution into 5 different classes by grouping some clusters together.
 331 Thus, class 1 describes the indoor condition when the ventilation is off, with only the data from
 332 cluster 1. Class 2 is composed of cluster 2 and cluster 3, corresponding to the ventilated periods.
 333 Class 3 describes the combined increase of NO and NO₂ concentration of cluster 4. Class 4
 334 describes the injection of NO₂ inside the room (cluster 5). Finally, class 5 (cluster 6 + cluster 7)

335 describes the period with high CO₂ concentration, either from controlled CO₂ injection or from the
 336 presence of people in the room. Table 4 summarizes the overlap ratio calculated between the 5
 337 classes defined by the clustering results on reference analyzers (dataset 1) and sensor measurements
 338 (dataset 2). The overlap ratios are at least 87.6% for class 2, and up to 98.4% for class 4. Only class
 339 1 and 2 still are slightly mingled, with up to 11% of mismatched data. Class 5 also presents about
 340 6% of mismatched data, principally falling on class 2. This illustrates that the K-means procedure,
 341 while allowing to identify the events, is not able to pick up correctly the transition between the two
 342 baseline cases (ventilation ON and ventilation OFF), probably because this transition actually could
 343 or should be represented by an additional cluster, as already discussed in the analysis of dataset 1,
 344 nor the difference between the end of the CO₂ injection events, with only residual concentration that
 345 are also not clearly distinguished from the baseline case.

346

347 **Table 4** Overlap ratio between 5 classes of dataset 2 and dataset 1

Reference Sensors	Class 1 (vent. OFF)	Class 2 (vent ON)	Class 3 (NO+NO ₂)	Class 4 (NO ₂)	Class 5 (CO ₂)
Class 1 (vent. OFF)	91.1%	11.1%	0.0%	0.0%	3.1%
Class 2 (vent ON)	8.8%	87.6%	2.2%	1.6%	6.2%
Class 3 (NO+NO ₂)	0.0%	0.0%	97.8%	0.0%	0.0%
Class 4 (NO ₂)	0.0%	0.0%	0.0%	98.4%	0.0%
Class 5 (CO ₂)	0.1%	1.3%	0.0%	0.0%	90.7%

348

349

350 **4 Conclusions**

351 The present work demonstrates that low-cost sensors are able to detect specific air quality events
352 occurring inside a room, and that these events can be classified without supervision using a K-
353 means clustering procedure. The identification of these events requires some outside knowledge, as
354 provided by the log of the experiments, or by the expertise of the user. When applying the K-means
355 classification procedure to the data from the reference online gas analyzers, we were able to
356 discriminate the normal ventilation pattern ON/OFF inside the room, and the artificial events such
357 as CO₂ or NO₂ voluntary injections, as well as a NO and NO₂ unintentional spill. Applying the K-
358 means procedure to the signals from the sensors as input leads to the same results, with a really
359 good agreement with the analyzers, as shown by the overlap analysis. Based on the measured
360 components (CO₂, NO, NO₂ and O₃), two sets of two clusters were not so well determined, possibly
361 because of low NO₂ and O₃ concentration. Merging these two groups of clusters into two classes
362 provides a much better agreement between the reference data (analyzers) and the sensors, with an
363 overlap ratio higher than 88%.

364 The unsupervised classification does not require that the sensors be calibrated before the
365 experiment, or that the chemical interferences be studied beforehand. This is a definite advantage
366 towards a generalized deployment of sensors in buildings for the operative management of the
367 ventilation and filtration systems, when quantitative measurements are not critical. Should absolute
368 quantitative measurements be needed, the present methodology would still be applicable, but would
369 require that the metrological performances of the sensors be established prior to the deployment.

370 The current efforts from manufacturers and research groups to improve the performances of the
371 sensors, both on the short and on the long term, will be determinant to reach this objective.

372 The mathematical procedure we have developed could certainly be improved, in particular with
373 respect to the automatic determination of the optimal number of classes, which so far needs to be
374 defined beforehand or to be adjusted by the expert, using either criteria about the stability of the

375 clusters when their number is increased, or contextual information to supplement the signals from
376 the sensors. In addition, using this unsupervised analysis of the signals from the sensors in different
377 real or realistic conditions, it could be possible to construct a set of classes representative of various
378 IAQ events. The resulting database would be used as a guide for the interpretation of the monitored
379 events, and as input for a supervised classification model, which would render easy and efficient the
380 management of IAQ with sensors installed in buildings. Measurements in real conditions in various
381 buildings are currently underway, and will be used to further validate the classification methodology
382 proposed here, and to establish such a database of “chemical signature” associated with specific
383 IAQ events.

384

385

386 **Acknowledgments**

387 The authors would like to thank the French Environment and Energy Management Agency ADEME
388 (Agence de l'Environnement et de la Maîtrise de l'Energie) for their financial support of the
389 MERMAID project (PRIMEQUAL Program). This work is a contribution to the CaPPA project
390 (Chemical and Physical Properties of the Atmosphere), funded by the French National Research
391 Agency (ANR) through the PIA (Programme d'Investissement d'Avenir) under contract ANR-11-
392 LABX-005-01, and a contribution to the CPER research project CLIMIBIO, with financial support
393 from the French Ministère de l'Enseignement Supérieur et de la Recherche, the Hauts de France
394 Region and the European Funds for Regional Economical Development.

395

396

397

398

399 **References**

- 400 [1] G.A. Ayoko, H. Wang, *Volatile Organic Compounds in Indoor Environments*, Indoor Air
401 Pollution, Springer, Berlin, Heidelberg, 2014: pp. 69–107. doi:10.1007/698_2014_259.
- 402 [2] G. Buonanno, L. Morawska, L. Stabile, Particle emission factors during cooking activities,
403 Atmospheric Environment 43 (2009) 3235–3242. doi:10.1016/j.atmosenv.2009.03.044.
- 404 [3] K.W. Tham, Indoor air quality and its effects on humans—A review of challenges and
405 developments in the last 30 years, Energy and Buildings 130 (2016) 637–650.
406 doi:10.1016/j.enbuild.2016.08.071.
- 407 [4] G. Buonanno, F.C. Fuoco, A. Russi, L. Stabile, Individual exposure of women to fine and
408 coarse PM, Environmental Engineering and Management Journal 14 (2015) 827–836.
409 doi:10.30638/eemj.2015.092.
- 410 [5] M. Verrielle, C. Schoemaeker, B. Hanoune, N. Leclerc, S. Germain, V. Gaudion, N. Locoge,
411 The Mermaid Study: Indoor and Outdoor Average Pollutant Concentrations in 10 Low-Energy
412 School Buildings in France, Indoor Air. 26 (2016) 702–713. doi:10.1111/ina.12258.
- 413 [6] N. Castell, F.R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A.
414 Bartonova, Can commercial low-cost sensor platforms contribute to air quality monitoring and
415 exposure estimates?, Environment International. 99 (2017) 293–302.
416 doi:10.1016/j.envint.2016.12.007.
- 417 [7] A.C. Lewis, J.D. Lee, P.M. Edwards, M.D. Shaw, M.J. Evans, S.J. Moller, K.R. Smith, J.W.
418 Buckley, M. Ellis, S.R. Gillot, A. White, Evaluating the performance of low cost chemical
419 sensors for air pollution research, Faraday Discuss. 189 (2016) 85–103.
420 doi:10.1039/C5FD00201J.
- 421 [8] P. Kumar, A.N. Skouloudis, M. Bell, M. Viana, M.C. Carotta, G. Biskos, L. Morawska, Real-
422 time sensors for indoor air monitoring and challenges ahead in deploying them to urban
423 buildings, Science of The Total Environment. 560–561 (2016) 150–159.
424 doi:10.1016/j.scitotenv.2016.04.032.

- 425 [9] A. Caron, B. Hanoune, N. Redon, P. Coddeville, Gas sensor networks: relevant tools for real-
426 time indoor air quality indicators in low energy buildings, in: Proceedings of the Healthy
427 Buildings Europe 2015 Conference, 2015.
- 428 [10] F. Röck, N. Barsan, U. Weimar, Electronic nose: current status and future trends, *Chemical*
429 *Reviews*. 108 (2008) 705–725.
- 430 [11] G. Neri, First Fifty Years of Chemosistive Gas Sensors, *Chemosensors*. 3 (2015) 1–20.
431 doi:10.3390/chemosensors3010001.
- 432 [12] M. Aleixandre, M. Gerboles, Review of small commercial sensors for indicative monitoring of
433 ambient gas, *Chemical Engineering Transactions*. 30 (2012) 169–174.
- 434 [13] B. Szulczyński, J. Gębicki, Currently Commercially Available Chemical Sensors Employed
435 for Detection of Volatile Organic Compounds in Outdoor and Indoor Air, *Environments*. 4
436 (2017) 21. doi:10.3390/environments4010021.
- 437 [14] A. Caron, N. Redon, F. Thevenet, B. Hanoune, P. Coddeville, Performances and limitations of
438 electronic gas sensors to investigate an indoor air quality event, *Building and Environment*.
439 107 (2016) 19–28. doi:10.1016/j.buildenv.2016.07.006.
- 440 [15] X. Pang, M.D. Shaw, A.C. Lewis, L.J. Carpenter, T. Batchellier, Electrochemical ozone
441 sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-
442 quality monitoring, *Sensors and Actuators B: Chemical*. 240 (2017) 829–837.
443 doi:10.1016/j.snb.2016.09.020.
- 444 [16] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a
445 cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO
446 and CO₂, *Sensors and Actuators B: Chemical*. 238 (2017) 706–715.
447 doi:10.1016/j.snb.2016.07.036.
- 448 [17] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J.
449 Rickard, M. Davis, L. Weinstock, S. Zimmer-Dauphinee, K. Buckley, Community Air Sensor

- 450 Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban
451 environment in the southeastern United States, *Atmospheric Measurement Techniques*. 9
452 (2016) 5281–5292. doi:10.5194/amt-9-5281-2016.
- 453 [18] L. Spinelle, M. Aleixandre, M. Gerboles, Protocol of Evaluation and Calibration of Low-Cos
454 Gas Sensors for the Monitoring of Air Pollution, (2013).
455 [http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83791/eur%20report%20protocol](http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83791/eur%20report%20protocol%20evaluation.pdf)
456 [%20evaluation.pdf](http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83791/eur%20report%20protocol%20evaluation.pdf).
- 457 [19] P. Van Geloven, M. Honore, J. Roggen, S. Leppavuori, T. Rantala, The influence of relative
458 humidity on the response of tin oxide gas sensors to carbon monoxide, *Sensors and Actuators*
459 *B: Chemical*. 4 (1991) 185–188. doi:10.1016/0925-4005(91)80196-Q.
- 460 [20] J.H. Sohn, M. Atzeni, L. Zeller, G. Pioggia, Characterisation of humidity dependence of a
461 metal oxide semiconductor sensor array using partial least squares, *Sensors and Actuators B:*
462 *Chemical*. 131 (2008) 230–235. doi:10.1016/j.snb.2007.11.009.
- 463 [21] J.-E. Haugen, O. Tomic, K. Kvaal, A calibration method for handling the temporal drift of
464 solid state gas-sensors, *Analytica Chimica Acta*. 407 (2000) 23–39. doi:10.1016/S0003-
465 2670(99)00784-9.
- 466 [22] M.J. Fernández, J.L. Fontecha, I. Sayago, M. Aleixandre, J. Lozano, J. Gutiérrez, I. Gràcia, C.
467 Cané, M. del C. Horrillo, Discrimination of volatile compounds through an electronic nose
468 based on ZnO SAW sensors, *Sensors and Actuators B: Chemical*. 127 (2007) 277–283.
469 doi:10.1016/j.snb.2007.07.054.
- 470 [23] A. Szczurek, M. Maciejewska, Recognition of benzene, toluene and xylene using TGS array
471 integrated with linear and non-linear classifier, *Talanta*. 64 (2004) 609–617.
472 doi:10.1016/j.talanta.2004.03.036.
- 473 [24] C. Bur, M. Bastuck, A. Lloyd Spetz, M. Andersson, A. Schütze, Selectivity enhancement of
474 SIC-FET gas sensors by combining temperature and gate bias cycled operation using

- 475 multivariate statistics, *Sensors and Actuators B:Chemical* 193 (2014) 931-940.
476 doi:10.1016/j.snb.2013.12.030.
- 477 [25] Y. González Martín, M.C. Cerrato Oliveros, J.L. Pérez Pavón, C. García Pinto, B. Moreno
478 Cordero, Electronic nose based on metal oxide semiconductor sensors and pattern recognition
479 techniques: characterisation of vegetable oils, *Analytica Chimica Acta.* 449 (2001) 69–80.
480 doi:10.1016/S0003-2670(01)01355-1.
- 481 [26] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, G. Di Francia,
482 Calibrating chemical mutlisensory devices for real world applications: An in-depth comparison
483 of quantitative machine learning approaches, *Sensors and Actuators B: Chemical.* 255 (2018)
484 1191–1210. doi:10.1016/j.snb.2017.07.155.
- 485 [27] H.-K. Hong, H.W. Shin, H.S. Park, D.H. Yun, C.H. Kwon, K. Lee, S.-T. Kim, T. Moriizumi,
486 Gas identification using micro gas sensor array and neural-network pattern recognition,
487 *Sensors and Actuators B: Chemical.* 33 (1996) 68–71. doi:10.1016/0925-4005(96)01892-8.
- 488 [28] S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas, Machine learning: a review of classification and
489 combining techniques, *Artif Intell Rev.* 26 (2006) 159–190. doi:10.1007/s10462-007-9052-3.
- 490 [29] A. Olaode, G. Naghdy, C. Todd, Unsupervised classification of images: A review, *International*
491 *Journal of Image Processing (IJIP).* 8 (2014) 325–342.
- 492 [30] S. Marco, A. Gutierrez-Galvez, Signal and data processing for machine olfaction and chemical
493 sensing:A review, *IEEE Sensors Journal* 12 (2012) 3189-3214 doi:10.1109/jsen.2012.2192920.
- 494 [31] A. Hierlemann, R. Gutierrez-Osuna, Higher-order chemical sensing, *Chemical Reviews* 108.
495 (2008) 563–613. doi:10.1021.cr068116m.
- 496 [32] M. Bicego, G. Tessari, G. Tecchiolli, M. Bettinelli, A comparative analysis of basic pattern
497 recognition techniques for the development of small size electronic nose, *Sensors and*
498 *Actuators B: Chemical.* 85 (2002) 137–144. doi:10.1016/S0925-4005(02)00065-5.

- 499 [33] T. Alizadeh, Chemiresistor sensors array optimization by using the method of coupled
500 statistical techniques and its application as an electronic nose for some organic vapors
501 recognition, *Sensors and Actuators B: Chemical*. 143 (2010) 740–749.
502 doi:10.1016/j.snb.2009.10.018.
- 503 [34] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive
504 feature elimination, *Sensors and Actuators B: Chemical*. 212 (2015) 353–363.
505 doi:10.1016/j.snb.2015.02.025.
- 506 [35] J. Lei, C. Hou, D. Huo, Y. Li, X. Luo, M. Yang, H. Fa, M. Bao, J. Li, B. Deng, Detection of
507 ammonia based on a novel fluorescent artificial nose and pattern recognition, *Atmospheric
508 Pollution Research*. 7 (2016) 431–437. doi:10.1016/j.apr.2015.10.019.
- 509 [36] Z. Haddi, S. Mabrouk, M. Bougrini, K. Tahri, K. Sghaier, H. Barhoumi, N. El Bari, A. Maaref,
510 N. Jaffrezic-Renault, B. Bouchikhi, E-Nose and e-Tongue combination for improved
511 recognition of fruit juice samples, *Food Chemistry*. 150 (2014) 246–253.
512 doi:10.1016/j.foodchem.2013.10.105.
- 513 [37] C. Hou, J. Li, D. Huo, X. Luo, J. Dong, M. Yang, X. Shi, A portable embedded toxic gas
514 detection device based on a cross-responsive sensor array, *Sensors and Actuators B: Chemical*.
515 161 (2012) 244–250. doi:10.1016/j.snb.2011.10.026.
- 516 [38] A.K. Jain, Data Clustering: 50 Years Beyond K-Means, *Pattern Recognition Letters*. 31 (2010)
517 651–666. doi:10.1016/j.patrec.2009.09.011.
- 518 [39] M. Falasconi, M. Pardo, M. Vezzoli, G. Sberveglieri, Cluster validation for electronic nose
519 data, *Sensors and Actuators B: Chemical*. 125 (2007) 596–606. doi:10.1016/j.snb.2007.03.004.
- 520 [40] H. Wu, T. Yue, Z. Xu, C. Zhang, Sensor array optimization and discrimination of apple juices
521 according to variety by an electronic nose, *Anal. Methods*. 9 (2017) 921–928.
522 doi:10.1039/C6AY02610A.

- 523 [41] M. Verma, M. Srivastava, N. Chack, A.K. Diswar, N. Gupta, A comparative study of various
524 clustering algorithms in data mining, *International Journal of Engineering Research and*
525 *Applications (IJERA)*. 2 (2012) 1379–1384.
- 526 [42] M. Steinbach, G. Karypis, V. Kumar, others, A Comparison of Document Clustering
527 Techniques, in: *KDD Workshop on Text Mining*, Boston, 2000: pp. 525–526.
- 528 [43] S. Marshall, M.E. Celebi, Comparison of Conventional and Bisecting K-Means Algorithms on
529 Color Quantization, in: *14th IASTED International Conference on Signal and Image*
530 *Processing*, 2012. doi:10.2316/P.2012.786-041.
- 531 [44] J. Shen, S.I. Chang, E.S. Lee, Y. Deng, S.J. Brown, Determination of cluster number in
532 clustering microarray data, *Applied Mathematics and Computation*. 169 (2005) 1172–1185.
533 doi:10.1016/j.amc.2004.10.076.
- 534 [45] H. Yu, Z. Liu, G. Wang, An automatic method to determine the number of clusters using
535 decision-theoretic rough set, *International Journal of Approximate Reasoning*. 55 (2014) 101–
536 115. doi:10.1016/j.ijar.2013.03.018.
- 537 [46] M.M.-T. Chiang, B. Mirkin, Intelligent Choice of the Number of Clusters in K-Means
538 Clustering: An Experimental Study with Different Cluster Spreads, *J Classif*. 27 (2010) 3–40.
539 doi:10.1007/s00357-010-9049-5.
- 540 [47] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a
541 cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen
542 dioxide, *Sensors and Actuators B: Chemical*. 215 (2015) 249–257.
543 doi:10.1016/j.snb.2015.03.031.
- 544
- 545
- 546