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

Alexandre Caron, Nathalie Redon, Patrice Coddeville, Benjamin Hanoune

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

Alexandre Caron[1,2], Nathalie Redon[2], Patrice Coddeville[2], Benjamin Hanoune[1]

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

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

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

Tel: (+33)3 27 71 24 77

Fax: (+33)3 27 71 29 14

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

Abstract:

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

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

1. Introduction

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

In recent years, gas micro-sensors emerged as alternative relevant tools for air quality monitoring [6–9]. New sensitive materials are constantly being developed in order to achieve better sensitivity, selectivity and stability [10,11]. More and more micro-sensors are commercially available [12,13], prompting many recent studies where their performances are investigated [14–17]. Among these sensors, a distinction must be made between intrinsically non-selective sensors and selective or nearly-selective sensors. Non-selective sensors are commonly based on metal oxide semiconductive materials which respond to multiple compounds in the air, and are generally used in combinations or arrays of sensors, also known as e-noses. Selective sensors include the electrochemical sensors targeting compounds such as carbon monoxide, ozone, nitrogen monoxide or dioxide, sulfur dioxide, hydrogen sulfide, the NDIR sensors used for carbon dioxide, as well as the PID sensors for total volatile organic compounds measurements. In spite of the technological progresses, the currently available sensors still require a complete metrological characterization [18], with in particular the assessment of their reliability over time. Even if they are very sensitive, the response of most sensors shows interferences with other compounds than the targeted one, depends upon the temperature and humidity, and drifts with time [19–21]. Such a preliminary step of characterization and calibration is not necessary for arrays of MOS sensors when used in association with pattern recognition algorithms [22–27], which are methods used in data mining for the extraction of useful information and the exploration of data correlation. Supervised classification approaches [28] are based on a classifier built from a training set with a collection of labeled data, and then used to assign new unlabeled data instances. Unsupervised classification [29] refers to algorithms that require no training set (blind partitioning), no a priori knowledge of the structure of the dataset, and automatically define the different classes. However, the physical meaning of these classes needs to be a posteriori interpreted or verified by the expert.

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

Many clustering algorithms have been developed for data mining, such as reviewed in [29]. Different clustering algorithms, or even different ways to use them on the same dataset, can lead to different partition results. None of them have proved to be the best technique in a large amount of configurations. Some of these algorithms have been applied to electronic nose data clustering [30], each with its respective possibilities and limitations [25,31,32], depending on the application. For electronic nose data, hierarchical clustering is commonly used [33–37]. It results in a hierarchical structure of the dataset that is more informative about the link between each group of data than the unstructured clusters provided by other techniques. In addition, its representation (dendrograms) allows the selection of the number of clusters. Centroid-based algorithms such as K-means [38] are less considered in the micro gas sensor field [39,40] and especially for air quality investigation, though K-means is a simple partitional algorithm and one of the most widely used techniques in data mining thanks to its performances [41]. K-means algorithms combine simplicity, ease of implementation and of use, speed of convergence, even with a large number of variables and clusters, and ability to process datasets with missing values. For these reasons, we have chosen this method to analyze the data from gas sensors in the investigation of indoor air quality.

2. Materials and methods

2.1. Instruments and measurements settings

Measurements used for this work were performed during the Mermaid study [5]. This field campaign has been carried out in February and March 2015 in a 51 m² (139 m^3) classroom of an energy efficient junior high school building in Northern France. Details on the analytical instruments used for this study can be found in [5]. Only the data from the pollutants from outdoors (NO_x and O_3), measured with online analyzers (Thermo 42i and Thermo 49i), and of CO_2, measured with Testo 480 probe located at the center of the room are considered here. In addition to these instruments, miniature sensors were installed in the center of the classroom, 90 cm above the floor. They monitor nitrogen oxide (electrochemical, Alphasense NO-B4), nitrogen dioxide (electrochemical, Alphasense NO_2-B4), ozone (electrochemical, Alphasense O_3-B4) and carbon dioxide (NDIR, Alphasense CO_2-IRC A1). A Raspberry Pi B+ and an Arduino board are used to collect, store and transmit the data. No correction or calibration of the sensors has been performed prior to their installation in the room, and the raw output (voltage) is analyzed with the bisecting K-means procedure.

2.2. Experimental datasets

The data used for this work corresponds to a 6-day continuous measurement (28 February 2015 to 6 March 2015). Fig. 1 presents the ventilation status and specific events taking place in the room during this period. These events include three CO_2 injections with ventilation ON, one CO_2 injection with ventilation OFF, for the determination of the air exchange rate of the room, two periods with people in the room, one NO_2 injection, and an accidental release of both NO and NO_2.

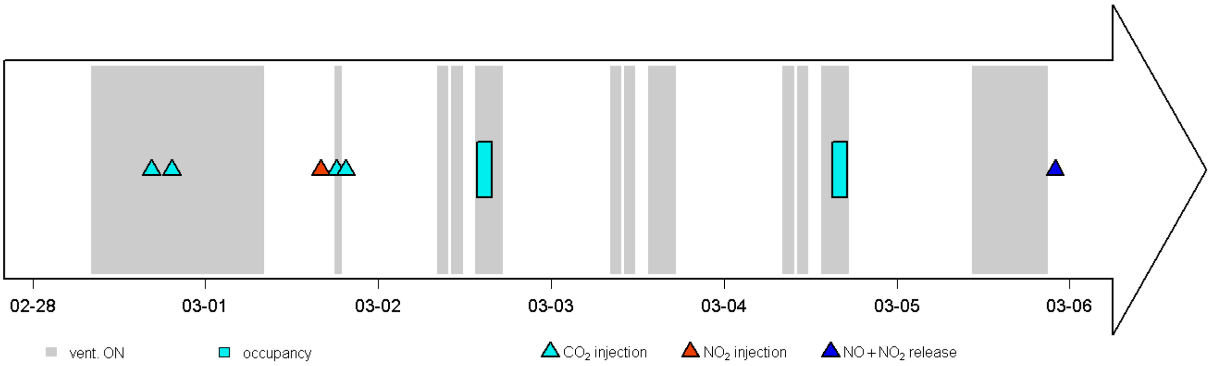

Fig. 1. Specific conditions and events occurring in the classroom

Dataset 1 (reference instruments) is a 4 x 8160 matrix consisting of 4 variables which are the concentrations of ozone, nitrogen oxide and nitrogen dioxide (in ppb) measured by the online analyzers and of carbon dioxide concentration (in ppm) measured by the Testo 480 probe, with a one minute resolution. The analyzers and CO_2 probe were calibrated at the beginning of the campaign.

Dataset 2 (electrochemical and NDIR sensors) is also a 4 x 8160 matrix, consisting of the voltage (mV) outputs of the 3 electrochemical sensors and of the carbon dioxide concentration (in ppm) provided by the NDIR sensor, with a 1 min resolution. As previously mentioned, no correction or calibration of the sensors signals was performed.

2.3 K-means implementation

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

The raw time series data of each dataset is used for input, except for the carbon dioxide concentration, measured either by the Testo probe or the NDIR sensor, for which the logarithm (base 10) is used, because of the much wider range over which this concentration can vary. Preliminary calculations, not reported here, with the direct CO_2 concentration, were performed but the clustering was less efficient than with the log values.

The only other parameter for input is the number K of clusters. Values of K ranging from 2 to 10 have been investigated for each dataset. Some mathematical criteria to determine the optimal number of clusters have been proposed [44–46], but these may have no physical meanings, and we have chosen to leave this determination to the judgment of the expert a posteriori.

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

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

3 Results and discussion

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

Dataset 1 (concentrations of NO, NO_2, O_3 and CO_2 measured by the reference instruments) are presented on Fig. 2, together with the ventilation status (ON/OFF) which has been found to be the main driver of the chemistry in the room [5]. The specific events described on Fig.1 are clearly distinguished on the concentration time chart of Fig. 2..

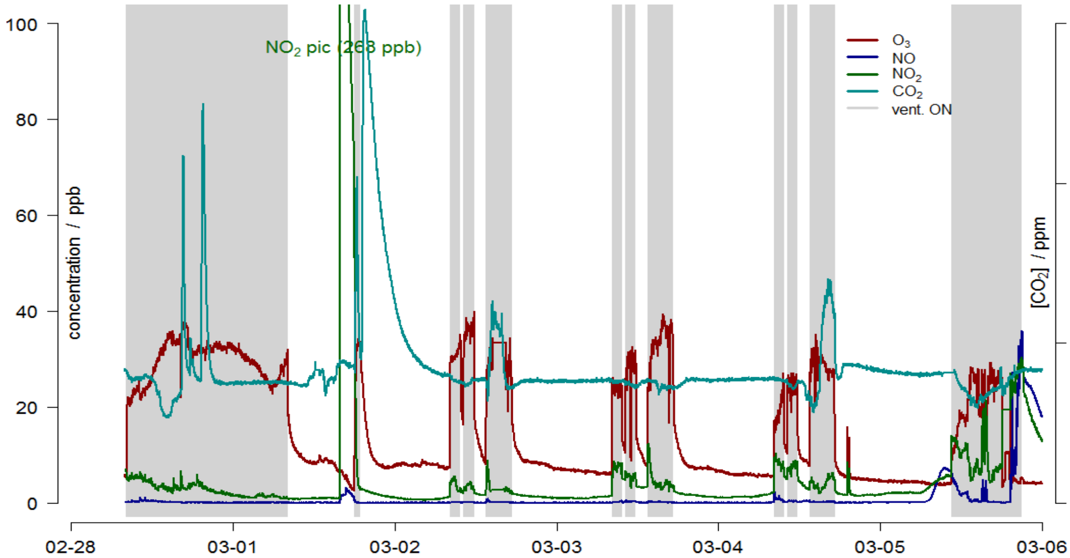

Fig. 2. Concentration time series measured by the reference analyzers. The ventilation status is indicated by the background color, white for ventilation OFF, grey for ventilation ON.

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

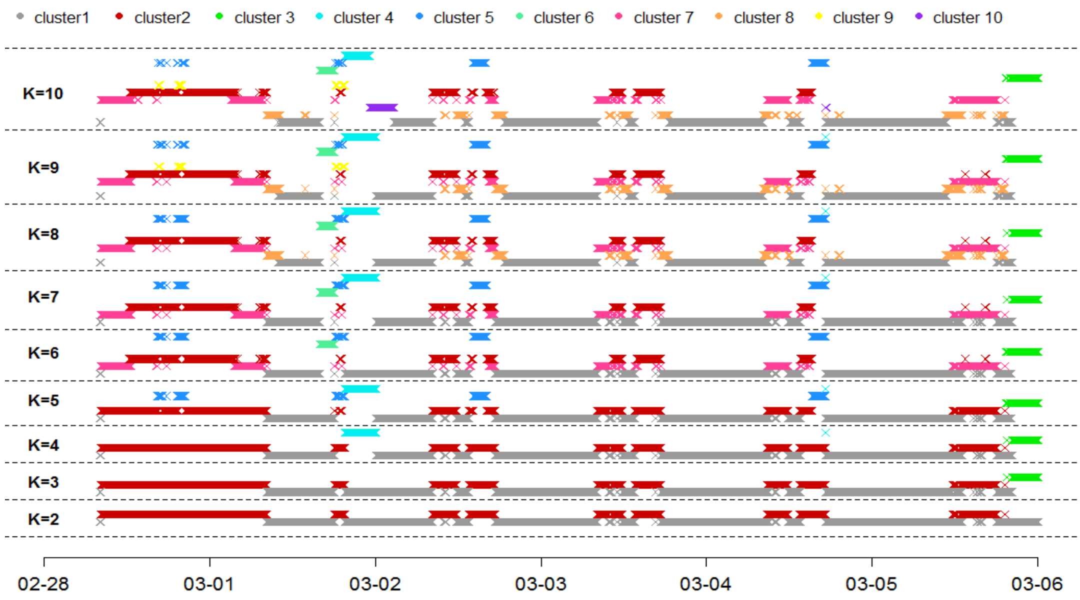

Fig. 3. Dataset 1 clustering process output, for K values from 2 to 10

Data are successively grouped into new clusters, with the 2 clusters divided into smaller sub-clusters when K increases. For the initial partition (K=2), the bisecting K-means separates the dataset according to the ozone concentration, with cluster 1 corresponding to the periods with no ozone, i.e. when the ventilation is OFF, while cluster 2 refers to the periods where the ozone concentration is significant, i.e. when the ventilation is ON. This result agrees with the result of the MERMAID campaign [5], where the main driver of the IAQ within the room was found to be the ventilation status. When K increases from 3 to 5, the newly defined clusters can be related to the specific known events occurring in the room presented in Fig. 1. Cluster 3 corresponds to the accidental release of NO and NO_2 in the classroom, cluster 4 corresponds to the injection of carbon dioxide when the ventilation system is OFF, and cluster 5 corresponds to the moments when the ventilation system is ON and with CO_2 concentration higher than the background, due to either controlled injection of CO_2 in the room or to human presence.

Interestingly, the results for K=6 differ drastically from the previous cases. The new cluster (#6) corresponds to the injection of NO_2, the previously found cluster 4 (injection of CO_2 during ventilation OFF conditions) disappears, and cluster 2 (ventilation ON) is divided into two sub-clusters. However, cluster 4 reappears when considering the K=7 partitioning, with no change on the previously determined clusters. This indicates that a criterion of stability of the clusters when increasing their number must be considered to correctly interpret the results of unsupervised classification. Higher values of K up to 10 induce a refinement in the previously determined clusters, according to different levels of ozone. Cluster 8 corresponds to the transition period when the pollutants from outdoor air slowly decrease due to their reactivity, with no compensation from the ventilation, cluster 9 represents the transient periods with elevated ozone and CO_2 concentration, and cluster 10 corresponds to a moderate CO_2 concentration with no ozone. These transitions periods, and in particular the mixing between indoor and outdoor pollutants, are generally overlooked during standard analysis.

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

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

The output results from the partition of dataset 1 into 7 clusters are presented in Table 1, which summarizes the size (number of data points in the cluster) and centroid coordinates of each cluster, together with their assignment. Cluster 1 groups the majority of the datapoints, with 4480 observations, and is characterized by low levels of every pollutant, as it corresponds to the periods with no ventilation, i.e. no intake of outdoor pollutants. A high ozone level is the dominant parameter that influences cluster 2 (1180 obs.) and cluster 3 (1467 obs.). These two clusters correspond to periods when the ozone concentration increases in the classroom due to the activation of the ventilation system. Cluster 4 (228 obs.) data are characterized by background CO_2 (416 ppm) and O_3 (4.3 ppb) level, together with higher NO and NO_2 levels (> 20 ppb). This partition corresponds to the short event that takes place at the end of the measurement period (Fig. 1, Fig. 2) which is an accidental release of NO and NO_2. Cluster 5 (128 obs.) data are characterized by low level of pollutants in the classroom, except for NO_2 (112.3 ppb), and corresponds to a voluntary injection of nitrogen dioxide in the classroom. Cluster 6 represents the CO_2 injection when the ventilation is off, with low O_3, NO and NO_2 concentrations. Cluster 7 represents the moments with high O_3 concentration, that is during ventilation ON, together with high CO_2 concentration, due to either a CO_2 injection or the presence of people in the room.

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

The time series evolution of dataset 2 is presented on Fig. 4. The direct interpretation of the signals from the sensors must be done with caution, because of possible chemical interferences, such as the cross response of NO_2 and O_3 on electrochemical sensors [47] and because the sensors were not calibrated before taking the measurements. This is where unsupervised classification algorithms can really help analyzing the data.

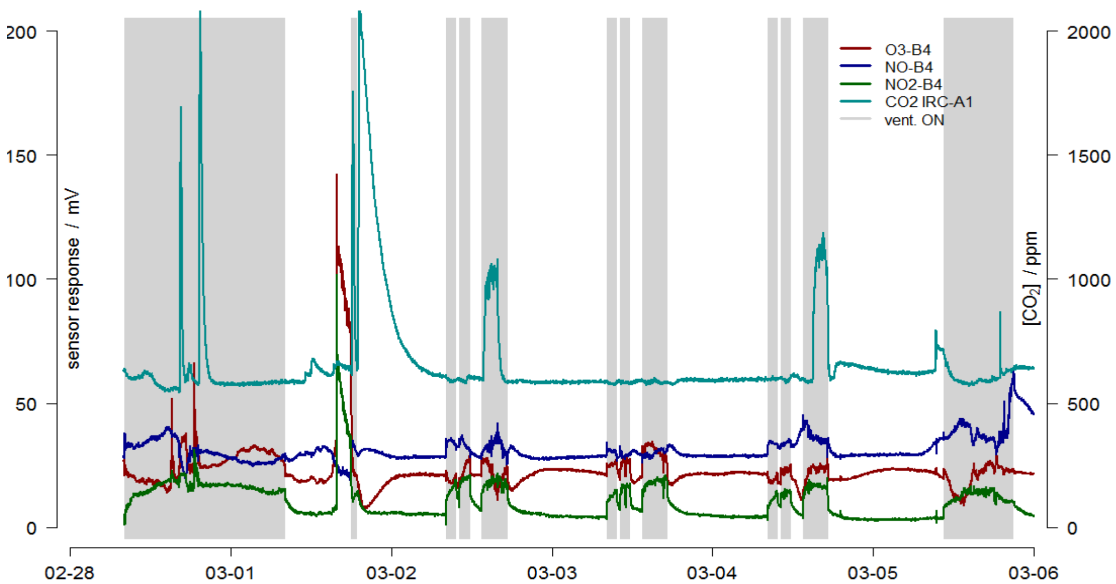

Fig. 4. Time series of the sensors signals. The ventilation status is indicated by the background color, white for ventilation OFF, grey for ventilation ON.

The results of the process for each K value between 2 and 10 are displayed on Fig. 5, illustrating, as for dataset 1, the hierarchical structure induced by the bisecting K-means process. However, the different clusters do not appear in the same order as for dataset 1. K=2 separates the datapoints with high CO_2 concentration. For K=3, the new cluster can be related to the injection of NO_2 in the classroom. The accidental release of NO and NO_2 is identified in dataset 2 at K=5. The efficiency of the ventilation, impacting the O_3 concentrations (oxydants coming from outdoor), is identified by clusters built from K=4 and K=6. At K=6, every known event can be related to a specific cluster, except the distinction between the CO_2 injections during the ventilation period or without ventilation, which appears only for the classification into 7 clusters. The clusters obtained for dataset 2 when increasing the value of K up to 10 are difficult to put in regard to the actual events taking place in the room. For instance, cluster 8 cannot be related to any event in the room. Also, cluster 9 is not stable, and partly disappears for K=10. Still, as the clusters do not appear in the same order when considering the reference dataset or the sensor dataset, it is necessary to explore the analysis with high K values, in order not to miss a specific event.

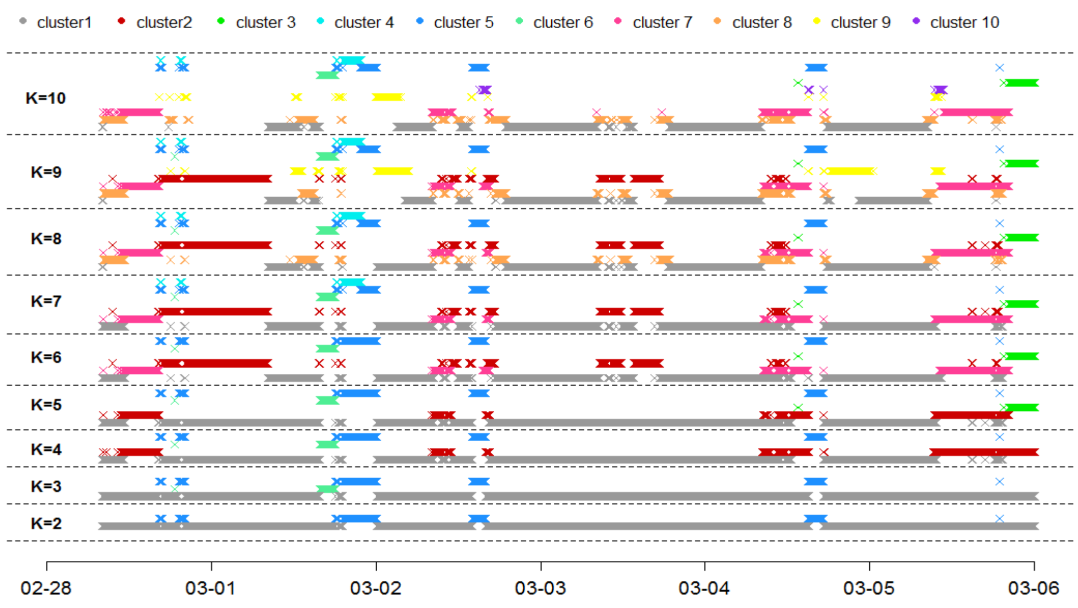

Fig. 5. Dataset 2 (sensors) clustering process output for K values up to 10. To help the comparison with the results of dataset 1, the cluster numbering and color coding does not follow the increasing K order, but is taken at the K=7 level.

This analysis demonstrates that the bisecting K-means procedure can also be used for sensors, when the activities in the room are known, even when the sensors are not calibrated before the experiment. This comes from the fact that the information does not lie in the absolute value of the measurements, but in the evolution of the intensity of the signals. This holds true also in spite of the poor selectivity, as in the case of the NO_2 and O_3 sensors. However, the K-means processing of dataset 2 also points to the difficulty of determining the optimal number of clusters, which in the present case could be 7 or 8, depending if we consider the physical meaning of the classes, that is if we use contextual information to supplement the measured data, or 8 if we consider the stability of the clusters, without contextual information.

As discussed when treating previously dataset 1, and considering our goal of comparing the results obtained from the two datasets, we will restrict hereafter our discussion to K=7. The results from the partition of dataset 2 into 7 clusters are presented in Table 2, summarizing the size and centroid coordinates of each cluster. Cluster 1 (4394 obs.) is characterized by a background level of the O3-B4 and NO-B4 response, of respectively 21.2 and 29.54 mV, a background (uncorrected) CO_2 concentration value of 614 ppm, and a low NO_2-B4 value of 5.8 mV. This partition matches well with the period when the ventilation is turned off. Cluster 2 (1246 obs.) is characterized by a higher response from the NO-B4 sensor (37.20 mV) and from the NO2-B4 sensor (13.90 mV). Cluster 3 (1515 obs.) is characterized by a higher response from the three electrochemical sensors. Both cluster 2 and cluster 3 describe the period where the outdoor pollutant contribution is significant, and can be related to the ventilation ON status. The two clusters differ by the higher values of O_3 in cluster 3. Cluster 4 (225 obs.) mainly appears at the end of the measurement period and is characterized by a significant increase of NO-B4 sensor response (51.49 mV). It corresponds to the simultaneous increase of NO and NO_2 concentration. Cluster 5 (129 obs.) corresponds to the maximal values from the O_3-B4 and NO_2-B4 sensors, respectively of 95.6 mV and 49.3 mV. This data partition can be linked to the injection of nitrogen dioxide into the room. Clusters 6 (208 obs.) and 7 (443 obs.) data match with the high CO_2 concentration periods, with (uncorrected) concentrations of 1658 and 1035 ppm respectively.

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

3.3 Comparison of both clustering results

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

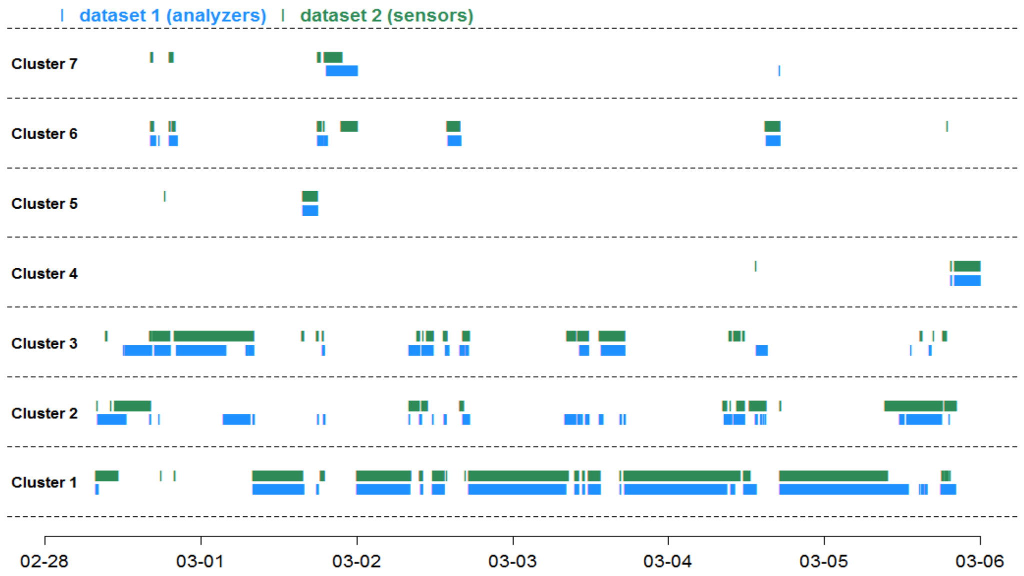

Fig. 6. Graphical comparison of clustering results (K=7) between dataset 1 (analyzers) and dataset 2 (sensors)

For instance, cluster 1 (vent. OFF), cluster 4 ($NO+NO_2$) and cluster 5 (NO_2) match very well each other, with only a few mismatched points. These clusters are associated to a well defined event, respectively ventilation OFF, $NO+NO_2$ release, and NO_2 injection, that is perfectly singled out by the K-means procedure, leading to an overlap ratio between dataset 2 and dataset 1 higher than 91 % for cluster 1, and of 98% for clusters 4 and 5. Only about 8% of the datapoints of sensors cluster 1 are associated to the reference cluster 2 (ventilation ON), though no explanation can be advanced. The agreement between the two datasets is slightly poorer for cluster 2 and 3 (vent. ON). The data are split between their respective reference clusters. This might be explained by the relatively low NO_2 and O_3 concentration amplitudes and the cross sensitivity of the NO-B4, NO2-B4 and O3-B4 sensors [47]. 393 observations are mismatched and associated to the period without ventilation (cluster 1). Cluster 6 (CO_2 vent. OFF) and cluster 7 (CO_2 vent. ON) from the sensors dataset are also mainly divided through the two clusters of dataset 1 that describe the high CO_2 concentration periods (analyzers clusters 6 and 7), leading to overlaps of 66% at most. We assigned this weaker agreement to the low NO_2 and O_3 concentration, which do not allow for a good differentiation between the ON and OFF ventilation conditions.

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%**

Because of this similarities and links between clusters 2 and 3 and clusters 6 and 7 respectively, it is natural to simplify the data distribution into 5 different classes by grouping some clusters together. Thus, class 1 describes the indoor condition when the ventilation is off, with only the data from cluster 1. Class 2 is composed of cluster 2 and cluster 3, corresponding to the ventilated periods. Class 3 describes the combined increase of NO and NO_2 concentration of cluster 4. Class 4 describes the injection of NO_2 inside the room (cluster 5). Finally, class 5 (cluster 6 + cluster 7) describes the period with high CO_2 concentration, either from controlled CO_2 injection or from the presence of people in the room. Table 4 summarizes the overlap ratio calculated between the 5 classes defined by the clustering results on reference analyzers (dataset 1) and sensor measurements (dataset 2). The overlap ratios are at least 87.6% for class 2, and up to 98.4% for class 4. Only class 1 and 2 still are slightly mingled, with up to 11% of mismatched data. Class 5 also presents about 6% of mismatched data, principally falling on class 2. This illustrates that the K-means procedure, while allowing to identify the events, is not able to pick up correctly the transition between the two baseline cases (ventilation ON and ventilation OFF), probably because this transition actually could or should be represented by an additional cluster, as already discussed in the analysis of dataset 1, nor the difference between the end of the CO_2 injection events, with only residual concentration that are also not clearly distinguished from the baseline case.

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%**

4 Conclusions

The present work demonstrates that low-cost sensors are able to detect specific air quality events occurring inside a room, and that these events can be classified without supervision using a K-means clustering procedure. The identification of these events requires some outside knowledge, as provided by the log of the experiments, or by the expertise of the user. When applying the K-means classification procedure to the data from the reference online gas analyzers, we were able to discriminate the normal ventilation pattern ON/OFF inside the room, and the artificial events such as CO_2 or NO_2 voluntary injections, as well as a NO and NO_2 unintentional spill. Applying the K-means procedure to the signals from the sensors as input leads to the same results, with a really good agreement with the analyzers, as shown by the overlap analysis. Based on the measured components (CO_2, NO, NO_2 and O_3), two sets of two clusters were not so well determined, possibly because of low NO_2 and O_3 concentration. Merging these two groups of clusters into two classes provides a much better agreement between the reference data (analyzers) and the sensors, with an overlap ratio higher than 88%.

The unsupervised classification does not require that the sensors be calibrated before the experiment, or that the chemical interferences be studied beforehand. This is a definite advantage towards a generalized deployment of sensors in buildings for the operative management of the ventilation and filtration systems, when quantitative measurements are not critical. Should absolute quantitative measurements be needed, the present methodology would still be applicable, but would require that the metrological performances of the sensors be established prior to the deployment. The current efforts from manufacturers and research groups to improve the performances of the sensors, both on the short and on the long term, will be determinant to reach this objective.

The mathematical procedure we have developed could certainly be improved, in particular with respect to the automatic determination of the optimal number of classes, which so far needs to be defined beforehand or to be adjusted by the expert, using either criteria about the stability of the clusters when their number is increased, or contextual information to supplement the signals from the sensors. In addition, using this unsupervised analysis of the signals from the sensors in different real or realistic conditions, it could be possible to construct a set of classes representative of various IAQ events. The resulting database would be used as a guide for the interpretation of the monitored events, and as input for a supervised classification model, which would render easy and efficient the management of IAQ with sensors installed in buildings. Measurements in real conditions in various buildings are currently underway, and will be used to further validate the classification methodology proposed here, and to establish such a database of "chemical signature" associated with specific IAQ events.

Acknowledgments

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

References

[1] G.A. Ayoko, H. Wang, Volatile Organic Compounds in Indoor Environments, Indoor Air Pollution, Springer, Berlin, Heidelberg, 2014: pp. 69–107. doi:10.1007/698_2014_259.

[2] G. Buonanno, L. Morawska, L. Stabile, Particle emission factors during cooking activies, Atmospheric Environment 43 (2009) 3235–3242. doi:10.1016.j.atmosenv.2009.03.044.

[3] K.W. Tham, Indoor air quality and its effects on humans—A review of challenges and developments in the last 30 years, Energy and Buildings 130 (2016) 637–650. doi:10.1016/j.enbuild.2016.08.071.

[4] G. Buonanno, F.C. Fuoco, A. Russi, L. Stabile, Individual exposure of women to fine and coarse PM, Environmental Engineering and Management Journal 14 (2015) 827–836. doi:10.30638/eemj.2015.092.

[5] M. Verriele, C. Schoemaecker, B. Hanoune, N. Leclerc, S. Germain, V. Gaudion, N. Locoge, The Mermaid Study: Indoor and Outdoor Average Pollutant Concentrations in 10 Low-Energy School Buildings in France, Indoor Air. 26 (2016) 702–713. doi:10.1111/ina.12258.

[6] N. Castell, F.R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A. Bartonova, Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, Environment International. 99 (2017) 293–302. doi:10.1016/j.envint.2016.12.007.

[7] A.C. Lewis, J.D. Lee, P.M. Edwards, M.D. Shaw, M.J. Evans, S.J. Moller, K.R. Smith, J.W. Buckley, M. Ellis, S.R. Gillot, A. White, Evaluating the performance of low cost chemical sensors for air pollution research, Faraday Discuss. 189 (2016) 85–103. doi:10.1039/C5FD00201J.

[8] P. Kumar, A.N. Skouloudis, M. Bell, M. Viana, M.C. Carotta, G. Biskos, L. Morawska, Real-time sensors for indoor air monitoring and challenges ahead in deploying them to urban buildings, Science of The Total Environment. 560–561 (2016) 150–159. doi:10.1016/j.scitotenv.2016.04.032.

[9] A. Caron, B. Hanoune, N. Redon, P. Coddeville, Gas sensor networks: relevant tools for real-time indoor air quality indicators in low energy buildings, in: Proceedings of the Healthy Buildings Europe 2015 Conference, 2015.

[10] F. Röck, N. Barsan, U. Weimar, Electronic nose: current status and future trends, Chemical Reviews. 108 (2008) 705–725.

[11] G. Neri, First Fifty Years of Chemoresistive Gas Sensors, Chemosensors. 3 (2015) 1–20. doi:10.3390/chemosensors3010001.

[12] M. Aleixandre, M. Gerboles, Review of small commercial sensors for indicative monitoring of ambient gas, Chemical Engineering Transactions. 30 (2012) 169–174.

[13] B. Szulczyński, J. Gębicki, Currently Commercially Available Chemical Sensors Employed for Detection of Volatile Organic Compounds in Outdoor and Indoor Air, Environments. 4 (2017) 21. doi:10.3390/environments4010021.

[14] A. Caron, N. Redon, F. Thevenet, B. Hanoune, P. Coddeville, Performances and limitations of electronic gas sensors to investigate an indoor air quality event, Building and Environment. 107 (2016) 19–28. doi:10.1016/j.buildenv.2016.07.006.

[15] X. Pang, M.D. Shaw, A.C. Lewis, L.J. Carpenter, T. Batchellier, Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring, Sensors and Actuators B: Chemical. 240 (2017) 829–837. doi:10.1016/j.snb.2016.09.020.

[16] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2, Sensors and Actuators B: Chemical. 238 (2017) 706–715. doi:10.1016/j.snb.2016.07.036.

[17] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis, L. Weinstock, S. Zimmer-Dauphinee, K. Buckley, Community Air Sensor

Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, Atmospheric Measurement Techniques. 9 (2016) 5281–5292. doi:10.5194/amt-9-5281-2016.

[18] L. Spinelle, M. Aleixandre, M. Gerboles, Protocol of Evaluation and Calibration of Low-Cos Gas Sensors for the Monitoring of Air Pollution, (2013). http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83791/eur%20report%20protocol%20evaluation.pdf.

[19] P. Van Geloven, M. Honore, J. Roggen, S. Leppavuori, T. Rantala, The influence of relative humidity on the response of tin oxide gas sensors to carbon monoxide, Sensors and Actuators B: Chemical. 4 (1991) 185–188. doi:10.1016/0925-4005(91)80196-Q.

[20] J.H. Sohn, M. Atzeni, L. Zeller, G. Pioggia, Characterisation of humidity dependence of a metal oxide semiconductor sensor array using partial least squares, Sensors and Actuators B: Chemical. 131 (2008) 230–235. doi:10.1016/j.snb.2007.11.009.

[21] J.-E. Haugen, O. Tomic, K. Kvaal, A calibration method for handling the temporal drift of solid state gas-sensors, Analytica Chimica Acta. 407 (2000) 23–39. doi:10.1016/S0003-2670(99)00784-9.

[22] M.J. Fernández, J.L. Fontecha, I. Sayago, M. Aleixandre, J. Lozano, J. Gutiérrez, I. Gràcia, C. Cané, M. del C. Horrillo, Discrimination of volatile compounds through an electronic nose based on ZnO SAW sensors, Sensors and Actuators B: Chemical. 127 (2007) 277–283. doi:10.1016/j.snb.2007.07.054.

[23] A. Szczurek, M. Maciejewska, Recognition of benzene, toluene and xylene using TGS array integrated with linear and non-linear classifier, Talanta. 64 (2004) 609–617. doi:10.1016/j.talanta.2004.03.036.

[24] C. Bur, M. Bastuck, A. Lloyd Spetz, M. Andersson, A. Schütze, Selectivity enhancement of SIC-FET gas sensors by combining temperature and gate bias cycled operation using

multivariate statistics, Sensors and Actuators B:Chemical 193 (2014) 931-940. doi:10.1016/j.snb.2013.12.030.

[25] Y. González Martín, M.C. Cerrato Oliveros, J.L. Pérez Pavón, C. García Pinto, B. Moreno Cordero, Electronic nose based on metal oxide semiconductor sensors and pattern recognition techniques: characterisation of vegetable oils, Analytica Chimica Acta. 449 (2001) 69–80. doi:10.1016/S0003-2670(01)01355-1.

[26] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, G. Di Francia, Calibrating chemical mutlisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sensors and Actuators B: Chemical. 255 (2018) 1191–1210. doi:10.1016/j.snb.2017.07.155.

[27] H.-K. Hong, H.W. Shin, H.S. Park, D.H. Yun, C.H. Kwon, K. Lee, S.-T. Kim, T. Moriizumi, Gas identification using micro gas sensor array and neural-network pattern recognition, Sensors and Actuators B: Chemical. 33 (1996) 68–71. doi:10.1016/0925-4005(96)01892-8.

[28] S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas, Machine learning: a review of classification and combining techniques, Artif Intell Rev. 26 (2006) 159–190. doi:10.1007/s10462-007-9052-3.

[29] A. Olaode, G. Naghdy, C. Todd, Unsupervised classification of images: A review, International Journal of Image Processing (IJIP). 8 (2014) 325–342.

[30] S. Marco, A. Gutierrez-Galvez, Signal and data processing for machine olfaction and chemical sensing:A review, IEEE Sensors Journal 12 (2012) 3189-3214 doi:10.1109.jsen.2012.2192920.

[31] A. Hierlemann, R. Gutierrez-Osuna, Higher-order chemical sensing, Chemical Reviews 108. (2008) 563–613. doi:10.1021.cr068116m.

[32] M. Bicego, G. Tessari, G. Tecchiolli, M. Bettinelli, A comparative analysis of basic pattern recognition techniques for the development of small size electronic nose, Sensors and Actuators B: Chemical. 85 (2002) 137–144. doi:10.1016/S0925-4005(02)00065-5.

[33] T. Alizadeh, Chemiresistor sensors array optimization by using the method of coupled statistical techniques and its application as an electronic nose for some organic vapors recognition, Sensors and Actuators B: Chemical. 143 (2010) 740–749. doi:10.1016/j.snb.2009.10.018.

[34] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination, Sensors and Actuators B: Chemical. 212 (2015) 353–363. doi:10.1016/j.snb.2015.02.025.

[35] J. Lei, C. Hou, D. Huo, Y. Li, X. Luo, M. Yang, H. Fa, M. Bao, J. Li, B. Deng, Detection of ammonia based on a novel fluorescent artificial nose and pattern recognition, Atmospheric Pollution Research. 7 (2016) 431–437. doi:10.1016/j.apr.2015.10.019.

[36] Z. Haddi, S. Mabrouk, M. Bougrini, K. Tahri, K. Sghaier, H. Barhoumi, N. El Bari, A. Maaref, N. Jaffrezic-Renault, B. Bouchikhi, E-Nose and e-Tongue combination for improved recognition of fruit juice samples, Food Chemistry. 150 (2014) 246–253. doi:10.1016/j.foodchem.2013.10.105.

[37] C. Hou, J. Li, D. Huo, X. Luo, J. Dong, M. Yang, X. Shi, A portable embedded toxic gas detection device based on a cross-responsive sensor array, Sensors and Actuators B: Chemical. 161 (2012) 244–250. doi:10.1016/j.snb.2011.10.026.

[38] A.K. Jain, Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters. 31 (2010) 651–666. doi:10.1016/j.patrec.2009.09.011.

[39] M. Falasconi, M. Pardo, M. Vezzoli, G. Sberveglieri, Cluster validation for electronic nose data, Sensors and Actuators B: Chemical. 125 (2007) 596–606. doi:10.1016/j.snb.2007.03.004.

[40] H. Wu, T. Yue, Z. Xu, C. Zhang, Sensor array optimization and discrimination of apple juices according to variety by an electronic nose, Anal. Methods. 9 (2017) 921–928. doi:10.1039/C6AY02610A.

[41] M. Verma, M. Srivastava, N. Chack, A.K. Diswar, N. Gupta, A comparative study of various clustering algorithms in data mining, International Journal of Engineering Research and Applications (IJERA). 2 (2012) 1379–1384.

[42] M. Steinbach, G. Karypis, V. Kumar, others, A Comparison of Document Clustering Techniques, in: KDD Workshop on Text Mining, Boston, 2000: pp. 525–526.

[43] S. Marshall, M.E. Celebi, Comparison of Conventional and Bisecting K-Means Algorithms on Color Quantization, in: 14th IASTED International Conference on Signal and Image Processing, 2012. doi:10.2316/P.2012.786-041.

[44] J. Shen, S.I. Chang, E.S. Lee, Y. Deng, S.J. Brown, Determination of cluster number in clustering microarray data, Applied Mathematics and Computation. 169 (2005) 1172–1185. doi:10.1016/j.amc.2004.10.076.

[45] H. Yu, Z. Liu, G. Wang, An automatic method to determine the number of clusters using decision-theoretic rough set, International Journal of Approximate Reasoning. 55 (2014) 101–115. doi:10.1016/j.ijar.2013.03.018.

[46] M.M.-T. Chiang, B. Mirkin, Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads, J Classif. 27 (2010) 3–40. doi:10.1007/s00357-010-9049-5.

[47] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide, Sensors and Actuators B: Chemical. 215 (2015) 249–257. doi:10.1016/j.snb.2015.03.031.