CHEMICAL ENGINEERING TRANSACTIONS
VOL. 68, 2018
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Selena Sironi, Laura Capelli

ISBN 978-88-95608-65-5; **ISSN** 2283-9216
DOI: 10.3303/CET1868053

Self-Organizing Map Algorithm as a Tool for Analysis, Visualization and Interpretation of Electronic Nose High Dimensional Raw Data

Sabina Licen [a,*], Sergio Cozzutto[b], Monica Angelucci[c], Pierluigi Barbieri[a]

aDept. of Chemical and Pharmaceutical Sciences, University of Trieste, Via L. Giorgieri 1,34127 Trieste , Italy
b ARCO SolutionS s.r.l., spin-off company of the Dept. of Chemical and Pharmaceutical Sciences, University of Trieste, Via L. Giorgieri 1, 34127 Trieste , Italy
c Agenzia Regionale per la Protezione Ambientale dell'Umbria (Arpa Umbria), Via Pievaiola, 207/B-3 Loc.S.Sisto 06132 Perugia, Italy
slicen@units.it

Electronic noses used for outdoor ambient air characterization to assess odor impacts on population can produce large datasets since usually the sampling is conducted with high frequency (e.g. data per minute) for periods that can reach several months, with a number of sensors that ranges usually from four-six as a minimum, up to above thirty. The environmental analyst has thus to deal with large datasets (millions of data) that have to be properly elaborated for obtaining meaningful interpretation of the instrumental signals. A recent review questioned the capability of some classic statistical elaboration tools for application to e-noses, highlighting how very few in field application are present in scientific literature. In the present work we describe: (i) the use of Self-Organizing Map (SOM) algorithm as a tool for analysis and visualization of e-nose raw data collected at a receptor site near a bio-waste composting facility; (ii) a second level clusterization using k-means clustering algorithm to identify "air types" that can be detected at the receptor and (iii) the use of e-nose data related to the plant odour sources as well as odour measurements of ambient air collected at the receptor site, to classify the air types. Eventually we evaluate the frequency and duration of the air type/s identified as malodorous.

1. Introduction

Electronic noses used for outdoor ambient air characterization to assess odour impacts on population can produce large datasets since usually the sampling is conducted with high frequency (e.g. data per minute) for periods that can reach several months, with a number of sensors that ranges usually from four-six as a minimum, up to above thirty. The environmental analyst has thus to deal with large datasets (millions of data) that have to be properly elaborated for obtaining meaningful interpretation of the instrumental signals. A recent review (Boeker, 2014) questioned the capability of some classic statistical elaboration tools for application to e-noses, highlighting how very few in field application are present in scientific literature.
Moreover in 2015 the European Committee for Standardization a working group (CEN/TC264/WG41) was established to draft a European Standard document about the "Instrumental odour measurement". Among three main issues the group focuses on criteria for developing and validating mathematical models linking instrument metrics to odour (Guillot, 2016).
Recently we proposed an approach (SISISCON - Smart Integrated System for Instrumental and Sensorial Characterization of Olfactory Nuisances, Licen et al. 2016) which integrates different tools (i.e., e-nose real time monitoring, air sampling and olfactometric analysis, citizens' complaints registration, meteorological data monitoring, source sampling and olfactometric and e-nose analysis) to deal with industrial olfactory nuisances, producing a model to support authorities in the operative control and appropriate decisions about malodorous emissions. A first application of the method has been recently published (Licen et al., 2018) in which however e-nose data concerning the odour sources were not available.

Please cite this article as: Licen S., Cozzutto S., Angelucci M., Barbieri P., 2018, Self organizing map algorithm as a tool for analysis, visualization and interpretation of electronic nose high dimensional raw data, Chemical Engineering Transactions, 68, 313-318
DOI: 10.3303/CET1868053

In the present work we describe: (i) the use of Self-Organizing Map (SOM) algorithm (Himberg et al., 2001) as a tool for analysis and visualization of e-nose raw data collected at a receptor site near a bio-waste composting facility; (ii) a second level clusterization using k-means clustering algorithm to identify "air types" that can be detected at the receptor and (iii) the use of e-nose data related to the plant odour sources as well as odour measurements of ambient air collected at the receptor site, to classify the air types.

2. Materials and methods

2.1 Site

The investigated site is located in central Italy. The plant is a bio-waste composting facility. Two clusters of dwellings are positioned near to the plant, approximatively 100 m and 300 m far from the plant odor sources respectively (see details in Figure 1).

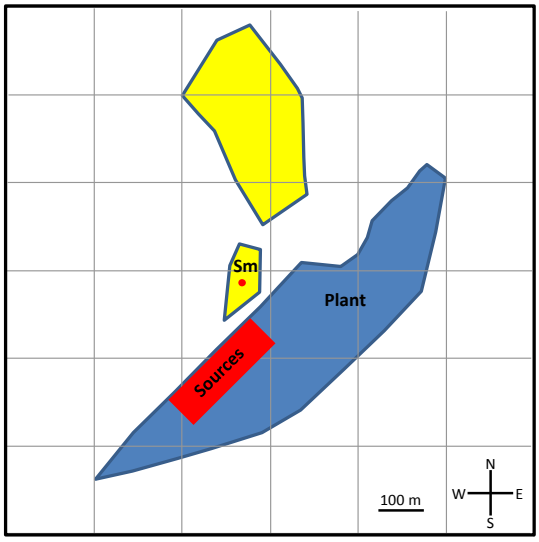

Figure 1: Map of the investigated area (Sm=sampling site; blue area=plant; red area= plant odor sources; yellow areas=civil dwellings).

2.2 Ambient air sampling

The air samples were collected in 8L bags prepared in-house using Nalophan^{TM} for the bag and Teflon^{TM} for the pipes. A remotely activated automatic sampling system (OdorPrep, by Lab Service Analytica S.r.L., Italy) has been used to collect ambient air at the receptor. The sampler has been activated by SMS after a citizen complaint reception. The sampling lasted two minutes. The sampler was placed in Sm (Figure. 1).

2.3 Source sampling

The samples of three different sources were collected in bags prepared as described in the previous paragraph using a manually activated “lung principle” sampler.

2.4 Odor concentration measurements

The air samples were analyzed in compliance with EN 13725:2003. The analysis was carried out by use of a dynamic olfactometer (WOLF by ArcoSolutions s.r.l., Italy) (Brattoli et al. 2014).

2.5 Electronic nose continuous monitoring and source samples characterization

A hybrid electronic nose implementing an array of 33 sensors (MOS, polymer/black carbon NCA and electrochemical sensors) purchased by Sensigent (Baldwin Park CA – U.S.A.) was placed in Sm (Figure 1). The monitoring period started Tuesday, 8th August 2017 and ended Monday, 18th December 2017.
The e-nose registered data-per-minute for every sensor obtaining a vector of 33 values for every minute. The sensors will be named in the text as S3 to S18, S21 to S23, S58 to S60, S62 to S68, "TVOC", "H2S", "NH3", "SH" as reported in the output files of the instrument. The bags containing the source samples were submitted to the e-nose without dilution recording the e-nose signal for 10 minutes.

2.6 Self-Organizing Maps

The experimental data collected by an e-nose can have complex and statistically non-linear relationships, thus we decided to use Self-Organizing Map algorithm to process the data, as, for example Principal Component Analysis, which is another unsupervised approach, is able to detect linear relationships only. In brief the SOM algorithm allows to reduce the experimental data set obtaining a new set of vectors (known as neurons or units) which still represent the variability of the processed data. The neuron vectors show simple geometric relationships (distances) and have the same number of variables as the experimental data. The neurons can be represented as hexagons stuck together in a bi-dimensional map allowing visual exploration of the data. (Vesanto, 1999). The SOM algorithm works as follows: (i) an experimental vector (sample) is presented to the initialized SOM algorithm (it is fundamental to establish the SOM map dimensions, some heuristic rules are proposed by Himberg et al. (2001) (ii) the algorithm identifies (in terms of distance) the Best Matching Unit (i.e. neuron) for the sample; (iii) the Best Matching Unit (BMU) adjusts itself (decreasing the distance) according to the vector presented, thus it "learns" from the experimental data; (iv) after all the samples are presented to the SOM one epoch is finished. The process can be iterated for a wisely selected number of epochs avoiding overfitting.
A second abstraction level can be obtained grouping neurons in clusters according to similarity, using hierarchical clustering algorithms (for example k-means clustering). The "slices" which form the SOM map are the so-called heatmaps which show the distribution of the values of every single experimental variable among the neurons on the map as modelled by the algorithm, showing how each one of the experimental variables relates to the others. Figure 2 depicts the way in which the SOM algorithm works as well as the hierarchical clustering method.

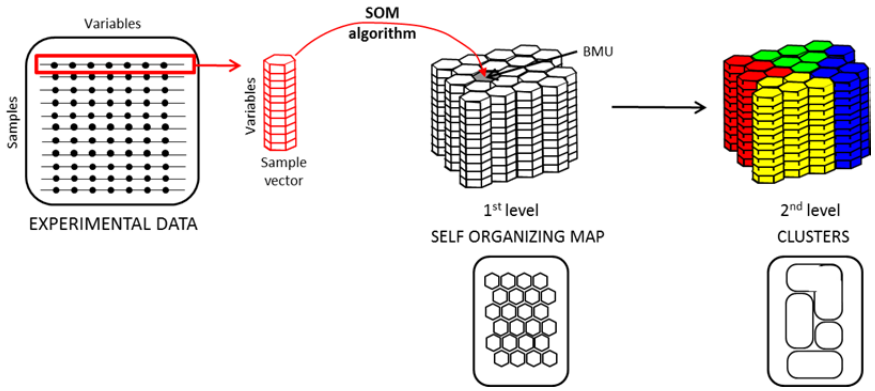

Figure 2: Representation of the experimental data elaboration using SOM algorithm and a hierarchical clustering method.

2.7 Calculations

SOM calculations and clustering classification were performed in the Matlab 6.5 (MathWorks, Inc.) computing environment, implementing the SOM toolbox (Vesanto, 2000). A desktop computer implementing an Intel i7 processor and 8 Gigabyte RAM was used, allowing training phase of few seconds for avoiding overfitting (Lampinen,1999), while handling data of dimensionality as the one considered in this paper and described in the next chapter. SOM outputs exploration and SOM visualization were performed using in-house scripts in R software environment (R. Core Team, 2016) implemented by the “openair” package (Carslaw and Ropkins, 2012)

3. Results and discussion

3.1 E-nose experimental data processing by SOM algorithm

A Self-Organizing Map is built for identifying recurrent sensor data patterns at the receptor site, that can be further clustered in few air types being characteristic for the e-nose position. Rule extraction for the clusters can be performed, leading to explicit definition of sensors, ranges of values and external conditions that allow to discriminate among air types. For each cluster a correlation among data from different sensors can be detected. Rule extraction from identified clusters (Malone, 2005) permits in principle to inductively identify

explicit relationships among sensors variables but it will not be discussed in the present paper.
The algorithm initialization and other parameters were set as reported in Licen et al. (2018), the SOM map dimensions resulted in a 34x14 hexagonal lattice. Starting from 141632 experimental vectors (which means a total number of sensor values above 4 million), we obtained 476 neurons each one consisting in a vector with a value for every variable (sensor signal). Every neuron represented 298 experimental vectors on average. The three neurons representing the highest number of experimental vectors are shown in Figure 3 (on the left) using black hexagons labelled by a white letter (a=4521, b=3152, c=3077). In order to group neurons we used the k-mean clustering algorithm, obtaining five clusters represented with different colors in Figure 3, in which the cluster numbers label the cluster centroids. In Figure 3 the heatmaps (see par.1.6) are represented as well. The sequence of heatmaps in the figure derives from the hierarchical clustering of the variables (see in Licen et al., 2018 for details on the method): S23,S4,S6-S17 (group 1); S18 (singleton); S60,S65,S58,TVOC (group 2); S68,S3,S5,SH,S62-S64,S66,S59,S67 (group 3); S21,H2S (group 4).

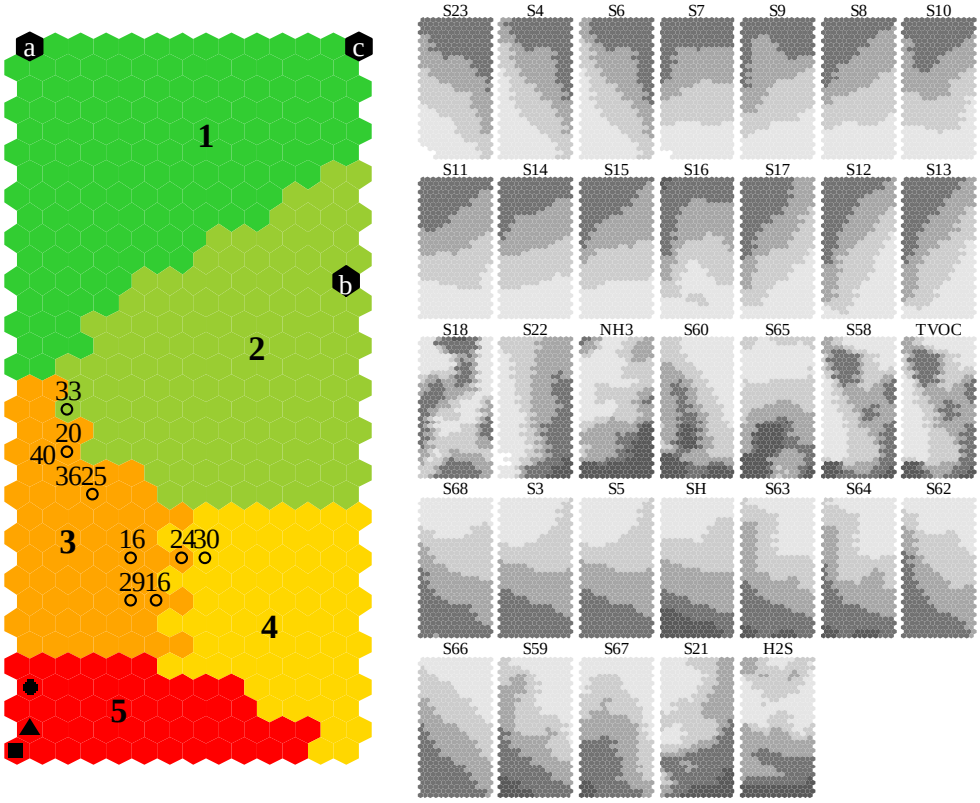

Figure 3: On the left: SOM map. Legend: bold numbers: cluster numbers plotted on the centroid of the cluster; Circles and numbers: olfactometric measurements at the receptor; filled point, square and triangle: odor sources; filled hexagons: units with the highest number of experimental vectors (a=4521, b=3152, c=3077). On the right: SOM heatmaps ordered according to the hierarchical clustering of the experimental variables, the filling of the hexagons represents the basic statistics (grayscale, light gray = low values, dark gray=high values) of each sensor.

3.2 Air types classification

The profiles of the variables (sensor signals) as modeled by the SOM algorithm and grouped by cluster can be depicted using boxplots as shown in Figure 4. The variables have been normalized to allow the comparison between cluster profiles. The sequence of presentation derives from the hierarchical clustering (see par. 2.1).
A cluster profiles exploration (Figure 4) showed that cluster 1 and 2 were very different from cluster 5. Moreover cluster 1 and 2 comprised the three neurons representing the higher number of experimental vectors (see previous paragraph). Considering that malodor events are usually transient this information was a first evidence for labelling these clusters as "not odorous". To classify the other clusters we used the data collected by source sampling and ambient air sampling.

The e-nose data corresponding to the source submission were not used to build the model. The sensor vectors were projected onto the map as external data obtaining an independent odor assignment for clusters. In particular the three sources were projected onto three different neurons laying in cluster 5.
The olfactometric measurements were used as follows, considering that the only common variable between e-nose data and odor concentration was the sampling date/time: (i) the air sampling date/time was identified (usually 2-3 minutes); (ii) the e-nose data vectors corresponding to the same date/time were identified; (iii) the map neurons representing the above mentioned data vectors were identified;(iv) the odor concentration values obtained by the olfactometric analysis were directly depicted onto the SOM map over the above mentioned neuron.
The olfactometric measurements, which showed not extremely high values, with a maximum of 40 OUe m^{-3} and a mean of 27 OUe m^{-3} (for a comparison see Licen et al. 2018), laid gathered in cluster 3, with two exceptions at the cluster edge (Figure 3). Surprisingly we obtained no olfactometric measurements laying in cluster 5. Exploring the data we observed that, according to date/time, usually cluster 5 followed cluster 3 and, as the activation of the air sampler was triggered by a citizen complaint (see par. 1.2), we supposed that the citizens which live nearby the plant were sensitized and they called right after they started to perceive a "modest" nuisance because they "know" that after that usually the malodour began to worsen.
Considering all the above mentioned outcomes we classified cluster 5 as "malodorous", cluster 3 as "modestly malodorous" and cluster 4 as "not determined".

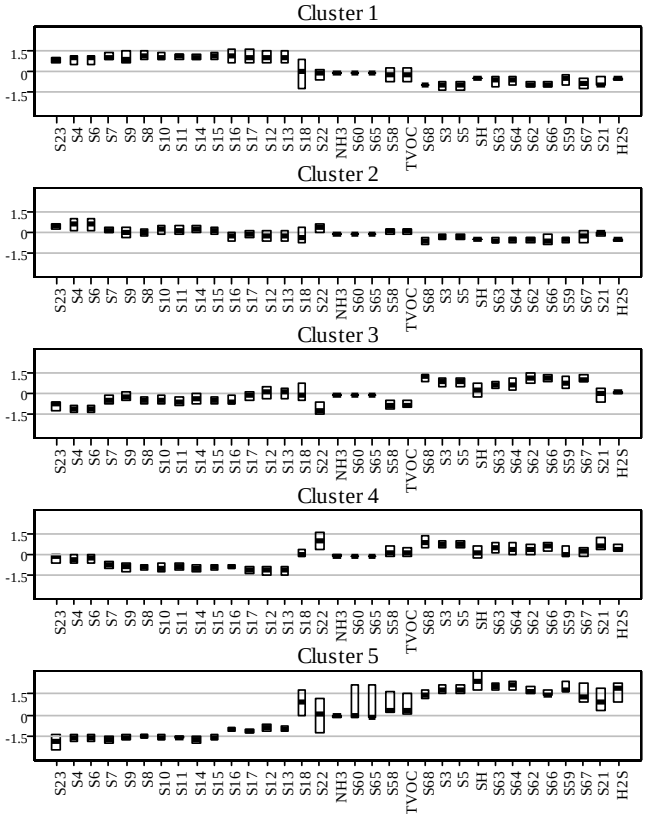

Figure 4: Cluster profiles as modelled by the algorithm, normalized by variable and represented by boxplots ordered according to hierarchical clustering.

3.3 Evaluation of odor frequency and duration

After air types classification we evaluated the frequency and duration over the considered period for every cluster. We obtain frequency percentages as follows: cluster-1 39 %, cluster-2 25%, cluser-3 12%, cluster-4 12 % and cluster-5 12%. The duration results are reported in Table 1. The cluster durations were separately evaluated thus the sum of every row in table 1 accounts for 100 %.

Table 1: Cluster duration percentage results

Cluster	0-1 h	1-2 h	2-4 h	4-8 h	8-12 h	12-24 h	24-48 h	48-60 h
1	71	8	4	4	6	5	1	1
2	82	9	4	3	1	1	0	0
3	86	5	4	4	1	0	0	0
4	93	3	1	2	1	0	0	0
5	84	4	4	4	4	0	0	0

4. Conclusions

The use of SOM algorithm allowed to deal with millions of single sensor values thus obtaining a classification of the data representing five "air types" perceived at the receptor site. The integration of the model with information collected by independent methods (odor sources, olfactometric measurements) allowed to classify the air types as "malodorous" or "not odorous" and to evaluate two important parameters for odor impact assessment: frequency and duration. The present study also clearly shows that the way to involve citizens in the sampling campaigns has to be mused on to obtain results as completed and clear as possible.
Finally we think that, starting from the outcomes of this work, the challenge is to explore the possibilities of the SOM algorithm to obtain a fine odor intensity evaluation as well as odor prediction linked to the air types experienced at the receptor site.

References

Boeker, P. 2014 “On ‘Electronic Nose’ methodology” Sensors and Actuators B 204, 2–17.

Brattoli M., Barbieri G., Barbieri P., Cozzutto S., De Gennaro G., Fabbris A., Spaccavento R., 2014, Development and technology assessment of the analytical performance of an eight position dynamic olfactometer, Chemical Engineering Transactions, 40, 115-120 DOI: 10.3303/CET1440020

Carslaw, D.C., Ropkins, K. 2012 Openair – an R package for air quality data analysis, Environ. Model. Softw. 27–28, 52–61.

Guillot, J.-M. 2016, E-noses: Actual limitations and perspectives for environmental odour analysis, Chemical Engineering Transactions, 54, 223-228. DOI: 10.3303/CET1654038

Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., Simula, O., 2001. The Self-Organizing Map as a Tool in Knowledge Engineering, in: Pattern Recognition in Soft Computing Paradigm, Fuzzy Logic Systems Institute (FLSI) Soft Computing Series. WORLD SCIENTIFIC, pp. 38–65. doi:10.1142/9789812811691_0002.

Lampinen, J., Kostiainen T. 1999. Overtraining and model selection with the self-organizing map. Neural Networks. 3. 1911 - 1915 vol.3. 10.1109/IJCNN.1999.832673.

Licen, S., Barbieri, G., Stel, F., Strappini, M., Barbieri, P. 2016 Molestie olfattive, nasi elettronici e sistemi integrati intelligenti per la caratterizzazione strumentale e sensoriale degli impatti odorigeni nell’aria ambiente (Olfactory nuisances, electronic noses and smart integrated systems for sensory and instrumental characterization of odor impacts on ambient air) BEA-Il bollettino degli esperti ambientali 2, 7-18. English abstract available at www.unideaweb.it/html/pubblicazioni/pdf_BEA/BEA_2016/BEA_2016_02.pdf

Licen, S., Barbieri, G., Fabbris, A., Briguglio, S.C., Pillon, A., Stel, F., Barbieri, P., 2018. Odor control map: Self organizing map built from electronic nose signals and integrated by different instrumental and sensorial data to obtain an assessment tool for real environmental scenarios Sensors and Actuators B: Chemical, 263, 476-485, https://doi.org/10.1016/j.snb.2018.02.144.

Malone, J, McGarry, K., Wermter S., Bowerman C., 2005. Data mining using rule extraction from Kohonen self-organising maps Neural Comput & Applic,15: 9–17

R. Core Team, 2016 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, R Core Team, Vienna, Austria.

Vesanto, J., 1999. SOM-based data visualization methods. Intelligent Data Analysis 3, 111–126. doi:10.1016/S1088-467X(99)00013-X.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankagas, J. 2000, SOM Toolbox for Matlab 5, Report A57, Available at: www.cis.hut.fi/projects/somtoolbox/ package/papers/techrep.pdf.