DOI: 10.2429/proc.2017.11(2)040 2017;11(2)

Dariusz MAJEREK[1], Monika GARBACZ[2], Sylwia DUDA[2]
and Małgorzata NABRDALIK[3]

# MACHINE LEARNING ANALYSIS OF E-NOSE SIGNAL IN EARLY DETECTION OF MOLD CONTAMINATION IN BUILDINGS

## ZASTOSOWANIE UCZENIA MASZYNOWEGO DO ANALIZY SYGNAŁU E-NOSA WE WCZESNYM WYKRYWANIU PORAŻENIA BUDYNKÓW

**Abstract:** Mould that develops on moistened building barriers is a major cause of the Sick Building Syndrome (SBS). Fungi emit Volatile Organic Compounds (VOC) that can be detected in the indoor air using several techniques of detection e.g. chromatography but also using gas sensors arrays. All array sensors generate particular electric signals that ought to be analysed using properly selected statistical methods of interpretation. This work is focused on the attempt to apply unsupervised and supervised statistical classifying models in the evaluation of signals from gas sensors matrix to analyse the air sampled from the headspace of various types of the building materials at the different level of contamination but also clean reference materials.

**Keywords:** electronic nose, mould contamination, classification, confusion matrix, multidimensional scaling

## Introduction

World Health Organization (WHO) reports that the quality of indoor air has a greater impact on our health than outdoor air. The major cause of poor quality of indoor air is mould threat that develops on building barriers. It is caused by poor quality of ventilation systems and improper horizontal and vertical isolation against water [1]. Mold that develops on building materials not only affects their mechanical properties but also have a negative impact on health of people in mouldy rooms. Exposure to these negative conditions may be related to Sick Building Syndrome (SBS). Many studies have shown that indoor mould growth is common in the dwellings, especially where dampness in the buildings occurs [2-6]. The increase in the SBS symptoms is related to the research of building dampness. It approximately doubles the risk of health effects [3, 7, 8]. Non-specific symptoms included in SBS mainly concern: headache, throat, muscles, memory and sleep disorders, general weakness, irritability, skin irritation, irritation of the mucous membranes of the eyes and nose [3, 9]. However, the most serious hazards caused by moulds include allergies, mycoses, mycotoxicosis, lung hemosiderosis and immunological reactions. They are caused by the mould that comes from the buildings and identified as moulds belonging to the following genera *Aspergillus*, *Penicillium*, *Cladosporium*, *Alternaria*, *Acremonium*, *Ulocladium*, *Stachybotrys*, *Rhizopus*, *Mucor* [2, 3, 6, 9]. It has been recorded that people staying in the mouldy buildings for a longer period of time might suffer from allergies to moulds present therein. Most common among them are

[1] Department of Applied Mathematics, Fundamentals of Technology Faculty, Lublin University of Technology, ul. Nadbystrzycka 38, 20-618 Lublin, Poland, phone +48 81 538 45 62, email: d.majerek@pollub.pl
[2] Environmental Engineering Faculty, Lublin University of Technology, ul. Nadbystrzycka 40B, 20-618 Lublin, Poland
[3] Chair of Biotechnology and Molecular Biology, University of Opole, ul. kard. B. Kominka 6, 45-032 Opole, Poland
Contribution was presented during ECOpole'17 Conference, Polanica Zdroj, 4-7.10.2017

the allergies to antigens of *Penicillium chrysogenum, P.expansum, Alternaria alternata, Cladosporium cladosporioides, Aspergillus niger* and *A. flavus*. Mould allergies have been confirmed in skin tests and in tests for the presence of anti-mould antibodies asIgE against the moulds [3]. According to the other research papers, the biologically active agents produced by fungi are (1 → 3) -β-glucan (a component of the fungal cell wall) and volatile compounds - Microbial Volatile Organic Compounds (MVOCs) which are the products of the microbes' primary and secondary metabolism [3, 10]. The research showed that indoor levels of some MVOCs were positively related with SBS. The levels of airborne microorganisms and some MVOCs were higher in dwellings with dampness and moulds [7]. MVOCs include a variety of chemical compounds, e.g. alcohols, aldehydes, ketones, amines, terpenes, sulphur compounds, chlorinated hydrocarbons. MVOCs emitted by fungi can be markers indicating the mould development indoors. In contaminated materials, numerous MVOC compounds produced specifically in fungal metabolism were detected and have been identified as: 2-ethylhexanol, 1-octen-3-ol, 3-heptanol, 3-methyl-1-butanol, 2-methyl-1-butanol, 1,3-octadiene, 2-(5H)-furanone, 2-heptene, limonene, α-pinene, 2-methylisoborneol, 4-heptanone, 2-methylfuran, 3-methylfuran, dimethyldisulfide, methoxybenzene, camphor, terpenoid and sesquiterpenes [11-13]. Some MVOCs may also be precursors of mycotoxins. A specific set of sesquiterpene hydrocarbons, including aristolochenean, were investigated in biosynthesis of PR toxin by *Penicillium roqueforti*. Therefore also the sesquiterpene hydrocarbon pattern and especially aristolochene can be used as volatile markers for detecting the process of undergoing biosynthesis of PR toxin by *P. roqueforti* [14]. Detecting this specific MVOCs emitted by moulds demonstrated that this approach is both reliable and quick as fungal growth can be detected before any visible signs of contamination occur [12]. The type of produced MVOCs depends on the growth medium of the fungus. There were significant differences in the spectrum of MVOCs produced during the mould growth on paper, building materials and microbiological media. It was found that building materials are media for MVOCs production by moulds [3].

Fungal contamination is normally evaluated using standard mycological or molecular methods, such as Polymerase Chain Reaction, Gas Chromatography-Mass Spectrometry, High-Performance Liquid Chromatography-Mass Spectrometry, but they are time-consuming and require a lot of manual labour. Moreover, there are numerous mycelial hyphae fragments in the indoor air, smaller than 1 μm, which cannot be measured with commonly applied methods for air microflora analysis [3]. Early detection techniques allow to quickly estimate the mould contamination. They usually involve the application of gas sensor arrays, i.e. electronic noses.

Signals conducted from the electronic noses are the vectors of resistances measured in time on particular number of sensors. In order to classify signals, there were used the supervised and unsupervised techniques of machine learning. Supervised learning means that in the process of building model we are able to correct predicted value based on the knowledge from the “supervisor” or the “teacher”. The whole sample is divided into two parts - training sample and test sample. In both of them, there is an information about proper class membership which was used to verify model prediction. Unsupervised methods of machine learning don’t use this information and they simply group observations into the homogeneous classes.

The aim of this paper is to show, that both types of machine learning techniques are appropriate to assess mould contamination in buildings.

## Materials and methods

The measurements were conducted using eight MOS type, resistance semiconductor sensors. The applied sensors were produced by TGS Figaro, series 2600: TGS2600-B00, TGS2602-B00, TGS2610-C00, TGS2610-D00, TGS2611-C00, TGS2611-E00, TGS2612-D00, TGS2620-C00. They are used in many implementations of gas sensors arrays, because they are cheap, reliable, small and with low electric power consumption [15, 16].

The experiment was conducted in four selected rooms with a different level of mould threat (bedroom, wardrobe, basement, house). To produce the reference material there were collected samples of clean and synthetic air and also non-stricken building materials with the mass of 100 g: gypsum board, aerated concrete, and brick. In Suchorab et al. [17] there is more information about collecting the data.

A dataset of readouts from the e-nose consists of resistance levels measured in time. For our analysis was chosen last 30 seconds of each signal since they are stable in sense of the electric signal level.

In this paper, there were used two methods of unsupervised learning such as hierarchical cluster analysis and self-organizing map, also known as Kohonen Neural Network [18]. Those techniques are usually used in the wide range of classification problems without knowledge about the number of classes that have to be analyzed in the particular set of observations and without information on the real memberships of them [19]. Assessing a number of homogenous groups was conducted based on Ward method of agglomeration and Euclidean metric.

The supervised methods of machine learning used in this research were Partial Least Squares Discriminant Analysis (PLS-DA) [20] and Generalized Linear Models with Regularized Path (GLMNET) [21]. The use of these models was dictated by the fact that there were some redundancies between the sensors. Both methods deal with this problem in a different way, the first by reducing the dimensionality of space with the PLS method, and the other by introducing parameters that remove the singularity of the matrix of the discriminant model.

In order to find the best fit in both cases, 10-fold cross-validation was performed on different sets of tuning parameters. All measurements were scaled for the analysis purposes.

## Results and discussion

All the calculations and visualizations were done in R environment, which is the very popular language of programming adapted for statistical analysis [22].

The dispersion of signals within environments for all sensors expressed by the coefficient of variation (*CV*) are very small (Table 1), so even tiny differences in electric signal level between environments could be significant. This enables homogeneity of the individual observations due to the level of resistance readings on the individual sensors, and thus they can be classified on this basis.

Table 1

Coefficients of variation of particular measurements

| Environment | bedroom | wardrobe | house | basement | gypsum board | aerated concrete | brick | decayed timber | clean air1 | clean air2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2600_b00 | 0.87 | 0.47 | 0.02 | 0.02 | 0.64 | 0.48 | 0.36 | 0.08 | 0.39 | 0.04 |
| 2602_b00 | 0.19 | 0.13 | 0.07 | 0.05 | 0.17 | 0.19 | 0.05 | 0.15 | 1.04 | 0.11 |
| 2610_c00 | 1.11 | 0.48 | 0.02 | 0.04 | 0.26 | 0.14 | 0.11 | 0.03 | 0.13 | 0.05 |
| 2610_d00 | 0.51 | 0.27 | 0.03 | 0.06 | 0.39 | 0.27 | 0.19 | 0.71 | 0.50 | 0.44 |
| 2611_c00 | 1.09 | 0.52 | 0.02 | 0.06 | 0.19 | 0.08 | 0.07 | 0.22 | 0.08 | 0.04 |
| 2611_e00 | 0.31 | 0.22 | 0.09 | 0.17 | 0.24 | 0.20 | 0.15 | 0.20 | 0.04 | 0.02 |
| 2612_d00 | 0.49 | 0.27 | 0.03 | 0.07 | 0.33 | 0.24 | 0.17 | 0.55 | 0.44 | 0.49 |
| 2620_c00 | 1.24 | 0.57 | 0.03 | 0.02 | 0.58 | 0.43 | 0.33 | 0.03 | 0.33 | 0.04 |

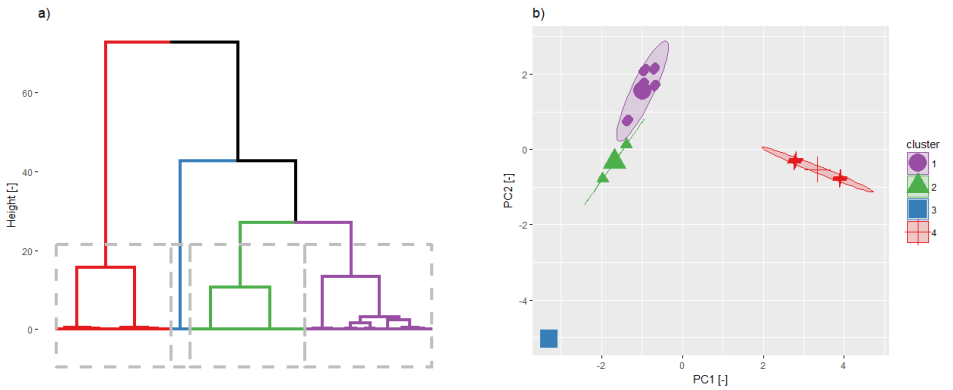


Fig. 1. Graphical signal interpretation in the form of: a) dendogram and b) clusters on PCA space. Source: Own elaboration

The method of hierarchical cluster analysis groups the observations into four classes. The number of clusters was obtained by the analysis of dendrogram (Fig. 1a), which is tree type method of visualization of the differences between observations. The choice of four clusters is quite obvious and gave rather homogenous groups of readings. In the first cluster, there are all observations from bedroom, wardrobe, gypsum board, aerated concrete and the brick, so this cluster consists non-stricken samples or with a low level of mould contamination. The second cluster contains all medium or high contaminated environments like house and basement. Totally stricken samples from decayed timber are grouped in the third cluster and the last group contains reference samples with clean air. The above clustering was visualized in two-dimensional space of principal components (Fig. 1b).

The self-organizing map was performed on 20×20 mesh of neurons. All observations from eight-dimensional space were mapped on this grid via neural network and then grouped into six homogenous class shown below (Fig. 2). The membership of particular clusters is the following: cluster 1 - aerated concrete, gypsum board, brick and bedroom, cluster 2 - wardrobe, cluster 3 - house and basement, cluster 4 - clean air 1, cluster 5 - clean air 2 and cluster 6 - decayed timber. This classification is even more homogenous than the obtained from hierarchical cluster analysis. The distinction between clean air 1 and clean air 2 is noted because one them is a fresh air and the other is a synthetic one. There is

a group of non-stricken samples (cluster 1), low-level contaminated samples from the wardrobe, medium and highly stricken - house and the basement. The last group contains the totally stricken samples - decayed timber.

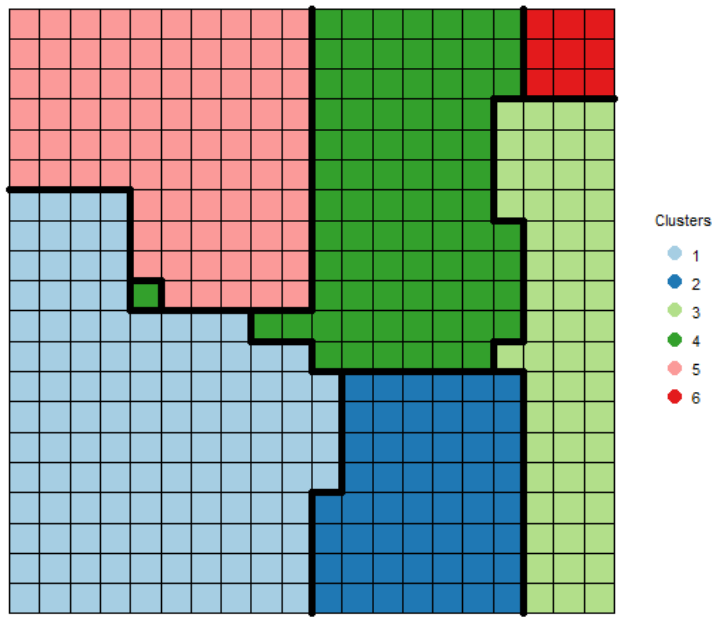


Fig. 2. Self-organizing map with six clusters. Source: Own elaboration

For the purpose of supervised machine learning methods, a certain organoleptic evaluation of the state of mould contamination of mentioned samples was introduced. It's quite similar to the distinction made by SOM, with two remarks. Samples of clean air are together and cluster 3 of the medium and high level of mould contamination is divided into two class. This new variable was used in training and tuning models. The sample was divided into two parts in 2/3 proportion for train sample.

A common method for describing the performance of a classification model is the confusion matrix [23]. This is a simple cross-tabulation of the observed and predicted classes for the data. Diagonal cells denote cases where the classes are correctly predicted while the off-diagonals illustrate the number of errors for each possible case. The simplest metric is the overall accuracy rate, which reflects the agreement between the observed and predicted classes.

Based on 10-fold cross-validation with 5 repeats was determined that the number of partial coordinates in PLS-DA needed for best fit is 3. Performance of PLS-DA classification measured by confusion matrix is perfect. This means that there is no any observation which was misclassified, both in the training and test sample. 95% confidence interval for accuracy in this model equals (0.979, 1) which shows very good classification potential.

The second model (GLMNET) was also build and tuned on the training sample via cross-validation. The best fit was obtained with $\alpha$ = 0 (mixing percentage) and $\lambda$ = 0.107 (regularization parameter).

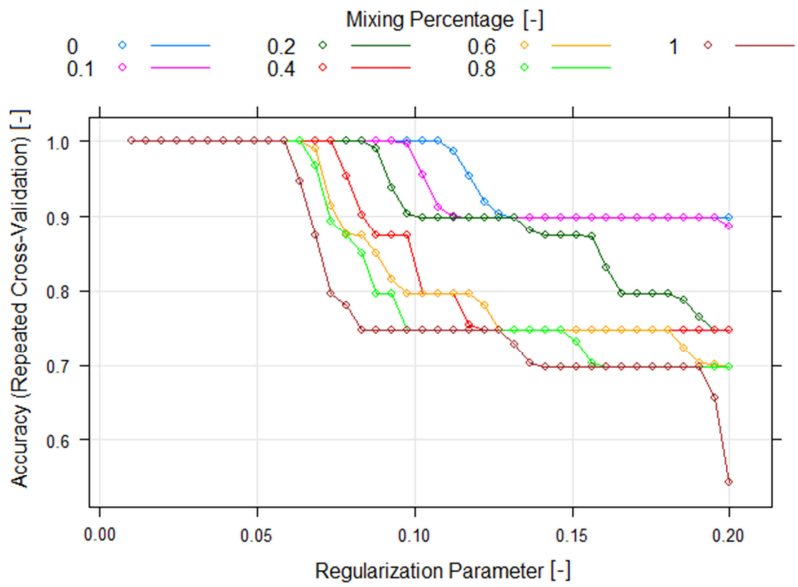


Fig. 3. Relations between accuracy and model parameters. Source: Own elaboration

The accuracy of this model was perfect with 95% confidence interval (0.98, 1). Obviously, true positive rate and true negative rate for all classes known as sensitivity and specificity, are one, which means that there is a perfect prediction in all classes.

## Summary and conclusions

The electronic nose can be used in early detection techniques of mould contamination in the buildings. Patterns with varying degrees of mould contamination are properly recognized by classification models. Both types of model supervised and unsupervised show very good classification quality. The results of this study look very optimistic since signals were characterized by very small dispersion. To verify the performance of the tested models and check their applicability it is planned to make similar investigations using completely new and different data congaing more noise in signals.

Tested methods are applicable and can be used for early detection of mould threat. They are perfect in area of data presented in this paper.

## References

[1] Łagód G, Suchorab Z, Guz Ł, Sobczuk H. Classification of buildings mold threat using electronic nose. AIP Conf Proc. 2017;1866:030002. DOI: 10.1063/1.4994478.

[2] Haleem Khan AA, Mohan Karuppayil S, Manoharachary C, Kunwar IK, Waghray S. Aerobiologia. 2009;25:119-123. DOI: 10.1007/s10453-009-9114-x.
[3] Gutarowska B. Grzyby strzępkowe zasiedlające materiały budowlane: wzrost oraz produkcja mikotoksyn i alergenów. Łódź: Wyd Politechniki Łódzkiej; 2010.
[4] Ryan TJ, Beaucham C. Chemosphere. 2013;90(3):977-985. DOI: 10.1016/j.chemosphere.2012.06.066.
[5] Benammar L, Menasria T, Chergui A, Benfiala s, Ayachi A. Int Biodeter Biodegr. 2017;117:115-122. DOI: 10.1016/j.ibiod.2016.12.004.
[6] Andersen B, Dosen I, Lewinska AM, Nielsen KF. Indoor Air. 2017;27:6-12. DOI: 10.1111/ina.12298.
[7] Sahlberg B, Gunnbjörnsdottir M, Soon A, Jogi R, Gislason T, Wieslander G, et al. Sci Total Environ. 2013;444:433-440. DOI: 10.1016/j.scitotenv.2012.10.114.
[8] Engvall K, Norrby C, Norbäck D. Int Arch Occup Environ Health. 2001;74(4):270-278. https://www.ncbi.nlm.nih.gov/pubmed/11401019.
[9] Cesuroglu O, Colakoglu GT. J Yeast Fungal Res. 2017;8(1):1-10. DOI: 10.5897/JYFR2017.0176.
[10] Kukec A, Dovjak M. Int J Sanit Eng Res. 2014:8(1):16-40. https://journal.institut-isi.si/prevention-and-control-of-sick-building-syndrome-sbs-part-1-identification-of-risk-factors/.
[11] Moularat S, Robine E, Ramalho O, Oturan MA. Sci Total Environ. 2008;407(1):139-146. DOI: 10.1016/j.scitotenv.2008.08.023.
[12] Moularat S, Robine E, Ramalho O, Oturan MA. Chemosphere. 2008;72(2):224-232. DOI: 10.1016/j.chemosphere.2008.01.057.
[13] Schleibinger H, Keller R, Ruden H. The Handbook of Environmental Chemistry. 2004;4:149-177. DOI: 10.1007/b94834.
[14] Jeleń HH. J Agric Food Chem. 2002;50(22):6569-6574. DOI: 10.1021/jf020311o.
[15] Guz Ł, Łagód G, Jaromin-Gleń K, Suchorab Z, Sobczuk H, Bieganowski A. Application of gas sensor arrays in assessment of wastewater purification effects. Sensors. 2015;15:1-21. DOI: 10.3390/s150100001.
[16] Bieganowski A, Jaromin-Glen K, Guz Ł, Łagód G, Jozefaciuk G, Franus W, et al. Evaluating soil moisture status using an e-nose. Sensors. 2016;16(6):886; DOI: 10.3390/s16060886.
[17] Suchorab ZH, Sobczuk Ł, Guz LM, Łagód G. Gas sensors array as a device to classify mold threat of the buildings. In: Pawłowska M, Pawłowski L, editors. Environmental Engineering. London: Taylor & Francis Group; 2017.
[18] Kohonen T. Neurocomputing. 1998;21(1-3):1-6. DOI: 10.1016/S0925-2312(98)00030-7
[19] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley Sons; 2009. ISBN: 9780471878766.
[20] Brereton RG, Lloyd GR. J Chemometr. 2014;28(4):213-225. DOI: 10.1002/cem.2609.
[21] Friedman JH, Hastie T, Tibshirani R. J Stat Softw. 2010;33(1):1-22. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/.
[22] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2017. https://www.R-project.org/.
[23] Khun M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.

## ZASTOSOWANIE UCZENIA MASZYNOWEGO DO ANALIZY SYGNAŁU E-NOSA WE WCZESNYM WYKRYWANIU PORAŻENIA BUDYNKÓW

[1] Katedra Matematyki Stosowanej, Wydział Podstaw Techniki, Politechnika Lubelska, Lublin
[2] Wydział Ochrony Środowiska, Politechnika Lubelska, Lublin
[3] Samodzielna Katedra Biotechnologii i Biologii Molekularnej, Uniwersytet Opolski, Opole

**Abstrakt:** Grzyb rozwijający się na ścianach budynków jest głównym powodem zjawiska, które nazwano Syndromem Chorego Budynku. Wolne związki organiczne emitowane przez grzyby mogą być wykryte różnymi metodami, m.in. na podstawie chromatografii, ale także za pomocą matryc czujników gazowych. Wszystkie tego typu narzędzia generują sygnały elektryczne, które można analizować za pomocą odpowiednich technik statystycznych. Praca skupia się na zastosowaniu nadzorowanych i nienadzorowanych technik uczenia maszynowego w ocenie sygnału pochodzącego z elektronicznego nosa.

**Słowa kluczowe:** elektroniczny nos, porażenie grzybem, klasyfikacja, macierz błędnych klasyfikacji, skalowanie wielowymiarowe