DOI 10.5162/sensor2015/D5.1

Quantification of Volatile Organic Compounds in the ppb-range using Partial Least Squares Regression

M. Bastuck[1], C. Bur[1,2], T. Sauerwald[1], A. Lloyd Spetz[2], Mike Andersson[2], A. Schütze[1]
[1] Lab for Measurement Technology, Saarland University, D-66123 Saarbrücken, Germany
m.bastuck@lmt.uni-saarland.de
[2] Division of Applied Sensor Science, IFM, Linköping University, SE-58183 Linköping, Sweden

Abstract:
Gas-sensitive, silicon carbide based field-effect transistors (SiC-FETs) with platinum gate are operated dynamically using temperature cycled operation (TCO) to quantify three hazardous volatile organic compounds (VOCs), i.e. benzene, naphthalene and formaldehyde, in the low ppb range. A suitable temperature cycle is developed based on static response measurements, and a linear model is built employing Partial Least Squares Regression (PLSR) with features extracted from the temperature cycle. Additionally, a strategy for cycle optimization using a t-test on the model coefficients is presented.

Key words: Volatile Organic Compounds (VOCs), quantification, SiC field-effect transistor (SiC-FET), temperature cycled operation (TCO), Partial Least Squares Regression (PLSR)

Introduction

Humans spend almost 90 % of their time indoors [1] which makes indoor air quality (IAQ) a crucial matter for human health. After a prolonged stay in buildings, many people experience symptoms like headache and discomfort, generally summarized as Sick Building Syndrome (SBS) [2]. These symptoms have been attributed to Volatile Organic Compounds (VOCs) present in the indoor air. Moreover, some VOCs are carcinogenic and, thus, have a severe negative impact on human health even in very small concentrations (ppb and sub-ppb). The three most relevant VOCs are investigated in this work: benzene, naphthalene and formaldehyde.

The World Health Organization (WHO) [3] and the INDEX project [4] have published potential hazards and guidelines for safe exposure limits of these three VOCs, amongst others. Benzene (C_6H_6) is genotoxic and carcinogenic at any concentration. Its occurrence in fuel raises outdoor levels, but more critical are indoor concentrations up to 2 ppb caused by solvents and cigarette smoke. Naphthalene (C_8H_{10}) is a suspected human carcinogen with a long-term exposure limit of 1.9 ppb. Apart from being an additive in gasoline which affects outdoor levels, it is contained in cigarette smoke and insect repellants. Measured values in German homes vary between 0.1 and 2.6 ppb [5]. Formaldehyde (CH_2O) is used in disinfectants, resins and polymers. Values between 4 and 200 ppb have been measured in residential homes [4]. Chronic exposure can possibly cause cancer. The limit for long-term exposure is 80 ppb.

Due to these health risks, monitoring the levels of hazardous VOCs in buildings would be desirable. Demand-controlled ventilation based on on-line measurements would decrease energy consumption while maintaining a healthy environment. Currently, there is no system on the market that can perform on-line identification and quantification of different VOCs at such low concentrations. While gas chromatography with subsequent mass spectrometry (GC/MS) can obtain the necessary resolution, it is expensive, requires a sampling procedure and it can usually not be used on-site. On the other hand, on-line capable methods like flame ionization detectors (FID) can only measure the amount of total VOCs (TVOC) [6]. This value does, however, not quantify the risk of exposure for specific carcinogenic compounds and might not even be a proper indicator for unspecific symptoms like SBS [2].

It has been shown that metal-oxide semiconductor sensors and gas-sensitive silicon carbide based field-effect transistors (SiC-FET) are sensitive enough to detect sub-ppb levels of VOCs [7],[8],[9]. Improved

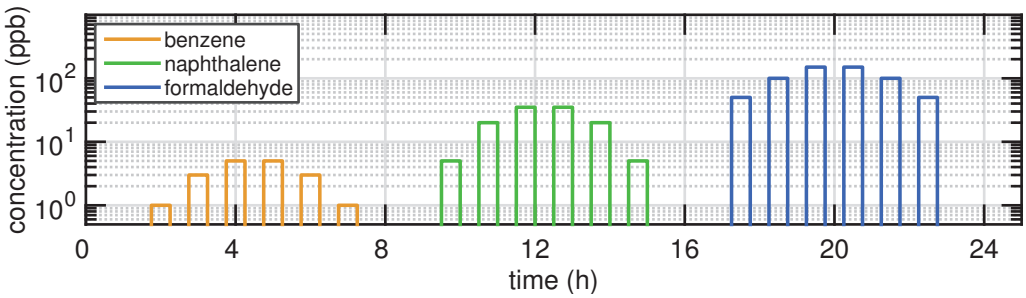

Fig. 1. Gas profile, 30 min gas pulses with 30 min background (humid air) in between.

selectivity is obtained by temperature cycled operation (TCO) [7],[9].

This work uses data from a platinum-gated SiC-FET with temperature cycled operation to quantify different concentrations of benzene, naphthalene and formaldehyde, three main pollutants in indoor air. Partial Least Squares Regression (PLSR) [10] is employed to build a linear model for quantification from features of the periodic sensor signal. Based on these models, a strategy for cycle optimization using a t-test on the model's coefficients is presented.

Experimental details

A custom-made gas mixing apparatus (GMA) [11] is used to provide very low and well-defined concentrations of benzene (1, 3, 5 ppb), naphthalene (5, 20, 35 ppb) and formaldehyde (50, 100, 150 ppb) in background, i.e. humid synthetic air (25 % r.h., Fig. 1). VOC spiked air is applied for 30 min with a pause of 30 min (in clean air) after each pulse. Each concentration is applied twice to test for possible sensor drift. At least 20 temperature cycles per gas exposure are taken into account for evaluation, i.e. at least 40 cycles for each concentration in total. The “zero concentration” group comprises approx. 220 cycles in background from the long pauses around 8 and 16 h (Fig. 1).

A SiC field-effect transistor (SiC-FET) with porous platinum gate [12] is mounted in a sealed measurement chamber and exposed to the gas flow while its temperature is changed periodically according to the cycle shown in Fig. 2. The sensor signal is the drain current at 4 V drain-source voltage and zero gate bias. The data acquisition rate is 10 Hz.

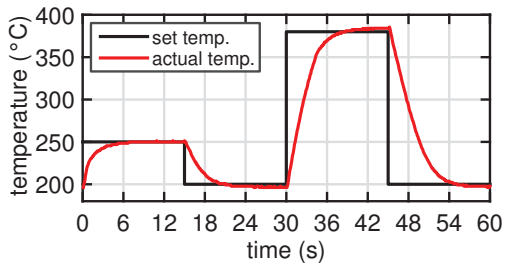

Fig. 2. Temperature cycle.

The shape of the temperature cycle is derived from measurements at static temperatures (Fig. 3) in order to provide sensitive, selective and stable signals. The sensor shows good sensitivity to formaldehyde and naphthalene at various temperatures.

175 °C cannot be considered since some devices will give a signal out of measurement range (1 mA) due to production tolerances. Then, the region from 200 to 250 °C is promising for the discrimination of naphthalene and formaldehyde since the sensitivities to both gases, and especially the ratio, varies strongly. Within this region also the highest sensitivity for benzene is obtained at 225 °C and is thus also included in the cycle as a transient temperature. The plateau at 380 °C is used to clean the sensor surface from adsorbed species to enhance the signal stability.

Data treatment

Each cycle represents one concentration measurement. However, one cycle is described by 600 data points, which can lead to problems, especially overfitting, in multivariate analysis. Thus, as a first step of dimensionality reduction, each temperature cycle is divided into ten equal parts (“feature ranges”) of 6 s length each. In each interval “features”, i.e. signal mean value (“mv”) and slope (“bfl”, best fit line), are computed from the sensor signal. These 20 shape-describing features are standardized, so that each feature has a mean of zero and a standard deviation of one, and then fed to the Partial Least Squares Regression (PLSR) algorithm [10]. This algorithm tries to establish a linear relationship between a linear combination of the features and the concentration of a target gas by projecting the data in a suitable way before applying linear regression. Thus, one coefficient for each feature is obtained, so that a linear combination gives the estimated concentration. The data is projected into a new, usually lower-dimensional space which itself is built from linear combinations of the features, so that the best compromise between explained variance and covariance to the concentration is

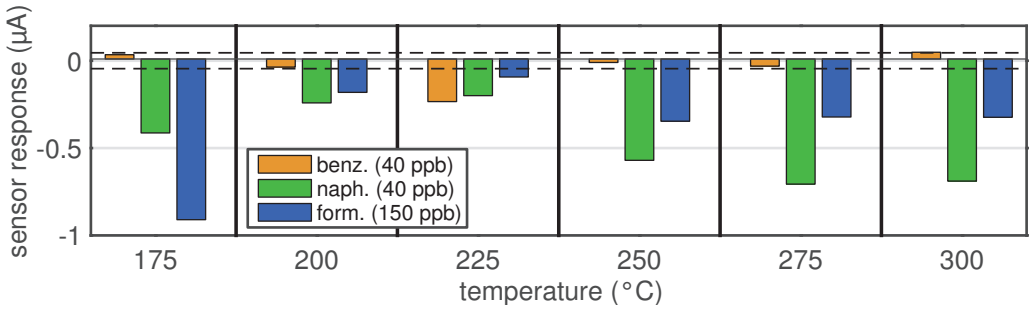

Fig. 3. Static gas response at different temperatures.

obtained. The number of dimensions of this new space is called “components” of the model.

Taking too few components into account will result in a poor model because information is missing. However, a model with too many components will have poor prediction ability although it fits the training data very well. This is due to overfitting, an effect where insignificant information from the signal, often noise, is used by the model as if it was a real feature. It is hence important to choose the optimal number of components, which is achieved by minimizing the Root Mean Squared Error of Prediction (RMSEP). Here, the RMSEP is calculated using 10-fold cross-validation [13], where a tenth of the data from each concentration group is randomly selected and not used for building the model, but is later projected as “unknown” data for which the Root Mean Squared Error (RMSE) is determined. This is done ten times so that all data is once used for validation, and the mean RMSEP is calculated.

With the data in this work, overfitting did not occur. Instead, the RMSEP becomes almost constant above a certain number of components, i.e. additional components do not add further information. Hence, very small fluctuations in RMSEP, depending on the random cross-validation sets, can lead to rather strong variations of the optimal number of components. Therefore we determine the mean of the absolute lowest RMSEP, add 10 % of its standard error, and subsequently chose the model with fewest components which produces a RMSEP just below that value. The meaning of the absolute RMSEP of a model is easier to interpret when compared to other models; hence, its change in percent is also given, based on the full model (cf. Tab. 1).

Pearson’s correlation coefficient is used to measure the covariance between model output and known response. A value of 100 % means perfect linear correlation.

Another important parameter of the model is its resolution, expressed as uncertainty. The uncertainty is here defined as $2\times2\sigma_{max}$, which corresponds to the end-to-end length of the largest 2σ error bar. For normally distributed outputs this means that 95 % of all model outputs for a given concentration are within this boundary. Hence, another concentration must be $4\sigma_{max}$ above or below for being recognized almost certainly as a different concentration.

The t-test [14] is used to check the features’ coefficients for statistical significance. If the coefficient has a high probability (here 95 %) of being zero, its contribution to the model is negligible. Thus, the feature can be omitted which can be helpful when trying to optimize and shorten the cycle.

Results and discussion

A PLSR model was built for each test gas. The respective cycles in this test gas were extracted from the whole measurement shown in Fig. 1.

In a first step, all models were built using all 20 of the available features: ten mean values and ten best fit lines. Parameters of the resulting models are summarized in Tab. 1, and a plot of the model is shown in Fig. 4a for benzene. The best models are obtained with 11 (benzene) to 14 (formaldehyde) components, and their correlation coefficients are between 98.3 % (naphthalene, Fig. 5) and 99.0 % (formaldehyde, Fig. 6). For benzene and formaldehyde, linear models are evidently an appropriate approach for quantification using SiC-FETs. Naphthalene exhibits a slightly non-linear behavior and will thus additionally be examined using non-linear methods in a future paper.

The resolution of the three different models varies strongly. For example, the uncertainty is 1.3 ppb for benzene (Fig. 4a), equal to the error bar of the 3 ppb concentration. The uncertainty of the models for naphthalene, 7.0 ppb, and formaldehyde, 40.0 ppb, were determined accordingly. Setting uncertainty and model span into relation, it is obvious that the uncertainty increases when the model spans a wider concentration range. The quotient of the uncertainty divided by highest concentration is around 26 % for benzene and formaldehyde,

Tab. 1. Model parameters.

gas	parameter	full model	t-test model	inv. t-test model
benz.	components	11	10	-
	features	20	14	-
	Pearson's R	98.4 %	98.4 %	-
	uncertainty (rel.)	1.3 ppb	1.3 ppb (0 %)	-
	RMSEP (rel.)	0.32 ppb	0.32 ppb (0 %)	
naph.	components	13	8	7
	features	20	10	10
	Pearson's R	98.3 %	98.2 %	97.4 %
	uncertainty (rel.)	7.0 ppb	6.8 ppb (–3 %)	11.2 ppb (+60 %)
	RMSEP (rel.)	2.51 ppb	2.53 ppb (+1 %)	2.98 ppb (+19 %)
form.	components	14	8	-
	features	20	12	-
	Pearson's R	99.0 %	98.8 %	-
	uncertainty (rel.)	40.0 ppb	50.3 ppb (+26 %)	-
	RMSEP (rel.)	8.29 ppb	8.86 ppb (+7 %)	

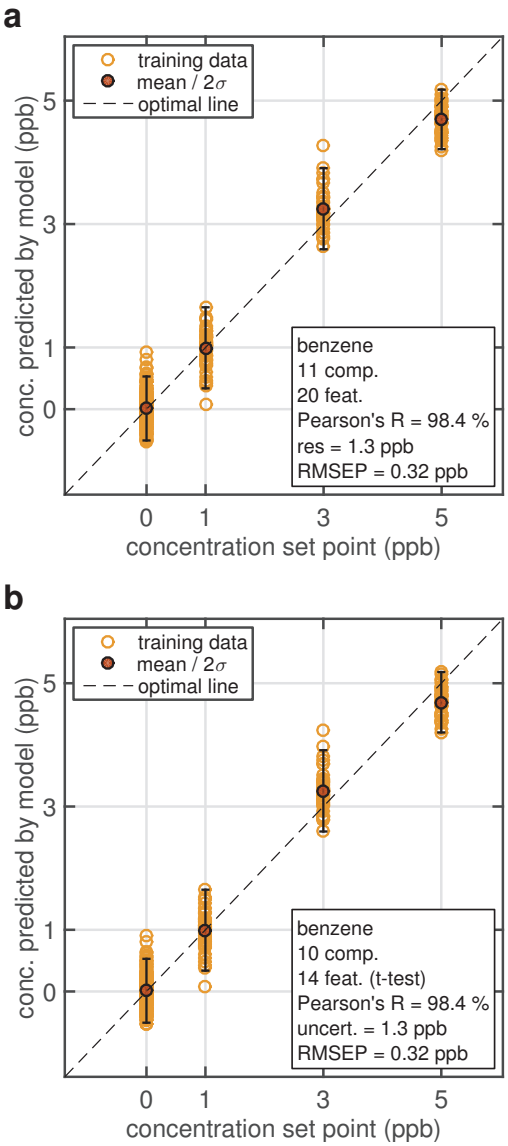

Fig. 4. PLSR model for benzene with (a) all features and (b) 14 significant features selected by t-test.

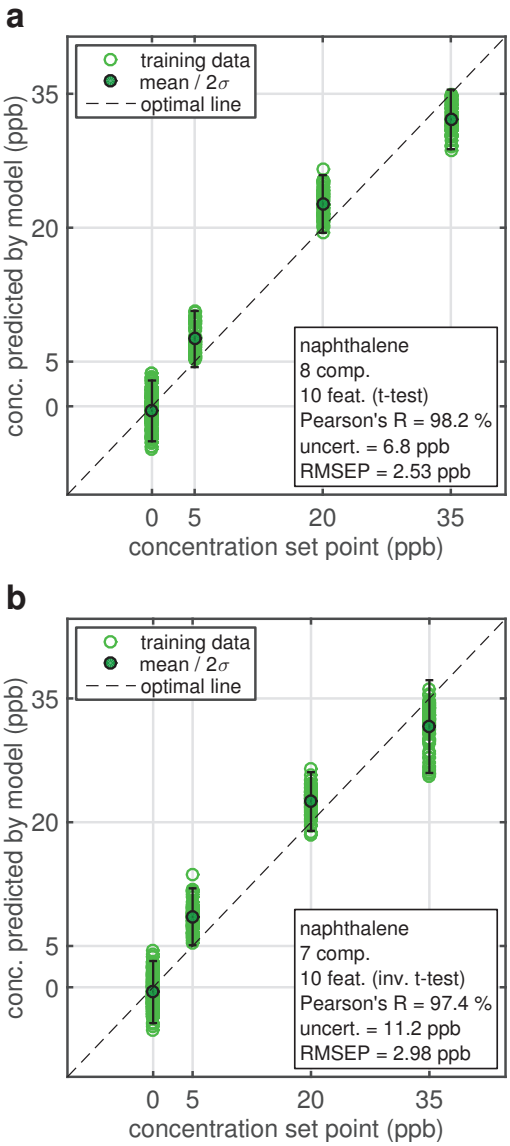

Fig. 5. PLSR model for naphthalene with (a) ten significant features selected by t-test and (b) with the other, non-significant features.

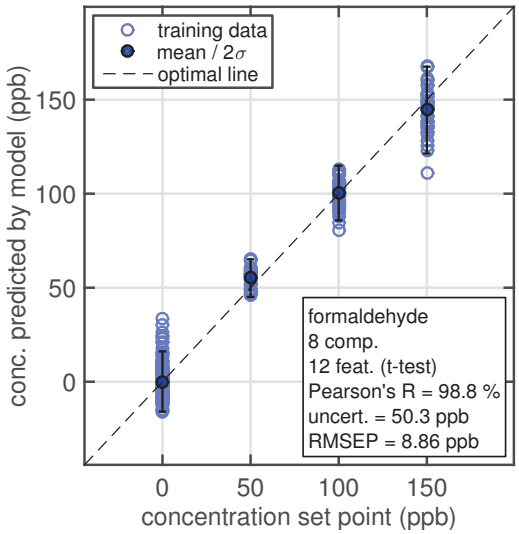

Fig. 6. PLSR model for formaldehyde with 14 features selected by t-test.

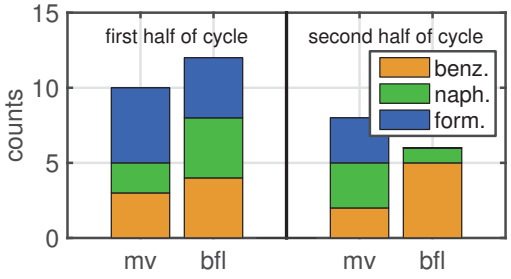

Fig. 7. Number of features that are considered significant by the t-test, for the first and second part of the cycle shown in Fig. 2.

and 20 % for naphthalene. Therefore we assume that the resolution of each model can be improved employing a hierarchical approach, similar to those used for classification with Linear Discriminant Analysis in [15] and [16]: first, the concentration is determined roughly with a wide-range model, and in a second step with a narrow-range model around this concentration. This approach is similar to local variants of the PLSR algorithm, like Locally Weighted PLSR (LW-PLSR) [17], which could therefore be another possibility to improve resolution (especially for naphthalene, as this also takes non-linearity into account).

PLSR assigns one coefficient to each feature. The value of this coefficient often represents the importance of the respective feature. This is interesting especially for optimizing the temperature cycle, e.g. shorten the cycle by omitting parts which do not contribute to the result Furthermore, we can verify our approach for the derivation of the cycle. While the first half of the temperature cycle was specifically designed to detect the target gases, the second half is mainly for cleaning purposes. Thus, the features from the first half should be more significant than those from the second half. This is verified by the t-test shown in Fig. 7. Especially the significance of the slope diminishes strongly in the second half because potential gas signals are overlaid by the strong signal change caused by the large temperature drop from 380 to 200 °C. It can thus be concluded that feature selection using t-test is able to identify significant parts of the cycle. Distinct trends regarding superiority of one feature or the other cannot be seen here.

A reduced feature set is determined by applying the t-test of the coefficients of the respective full model with all 20 features. Here, six (benzene, formaldehyde) to ten features (naphthalene) can be excluded. Another model is then built using this reduced feature set (Fig. 4b, Fig. 5a, and Fig. 6). The optimal number of components for these models is always lower, between 8 and 10, than for the full model because non-significant information must not be filtered by the algorithm. The correlation coefficient remains nearly unaffected and does not decrease more than 0.2 % compared to the full model. However, while the uncertainty remains constant for benzene, it decreases about 3 % for naphthalene and increases 26 % for formaldehyde. The numbers for their respective prediction ability, expressed in RMSEP compared to the full model, behave mostly similar. No difference in prediction ability is seen for benzene, while the RMSEP increases about 1 % and 7 % for naphthalene and formaldehyde, respectively.

In the reduced model for naphthalene, 10 out of 20 features are excluded. It is therefore easy to validate the t-test's choice by building a model based only on the excluded features. Indeed, the resulting "inverse" model performs comparatively worse than the model built using the t-test (Tab. 1, compare Fig. 5a and Fig. 5b). Compared to the full model its correlation coefficient decreases by 0.8 % which does, however, not affect the linearity much. More importantly, its uncertainty increases more than 60 % compared to the full model, while it even decreased slightly for the t-test model. Furthermore, the RMSEP of the inverse model increases by about 19 % compared to the full model, whereas the increase is only 1 % for the model built with the t-test. This leads to the conclusion that feature selection can strongly affect the prediction ability and verifies again that the t-test identifies important features. The presented method is thus suitable for efficient feature selection and, on this basis, temperature cycle optimization.

Conclusion and outlook

We have shown that SiC-FETs with temperature cycled operation can be used to quantify benzene, naphthalene and formaldehyde at health relevant concentrations in the ppb-range. The extracted features are “sufficiently linear” to use the linear Partial Least Squares Regression (PLSR) algorithm to build a quantification model. The model resolution seems to depend on the model range: for benzene, concentration differences of 1.3 ppb can be discriminated between 0 and 5 ppb, while for formaldehyde the uncertainty is 40 ppb in the range 0 to 150 ppb. Because of this dependence, approaches like hierarchical modeling are expected to improve the result.

Another optimization strategy is using non-linear algorithms, e.g. LW-PLSR. However, unlike the simple approach used in this work, this approach does not return a global set of coefficients. With these coefficients, a t-test can be used to identify futile features and parts of the temperature cycle and thus help in cycle optimization.

Acknowledgements

This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement No 604311.

References

[1] P.L. Jenkins, T.J. Phillips, E.J. Mulberg, Steve.P. Hui, Activity patterns of Californians: Use of and proximity to indoor pollutant sources, *Atmospheric Env.* 26A (12), 2148–2148 (1992); doi: 10.1016/0960-1686(92)90402-7

[2] J.T. Brinke, S. Selvin, A.T. Hodgson, W.J. Fisk, M.J. Mendell, C.P. Koshland, J.M. Daisey, Development of New Volatile Organic Compound (VOC) Exposure Metrics and their Relationship to "Sick Building Syndrome" Symptoms, *Indoor Air* 8 (3), 140-152 (1998); doi: 10.1111/j.1600-0668.1998.t01-1-00002.x

[3] WHO Regional Office for Europe, *WHO guidelines for indoor air quality* 9 (2010); ISBN: 978 92 890 0213 4

[4] K. Koistinen, D. Kotzias, S. Kephalopoulos, C. Schlitt, P. Carrer, M. Jantunen, S. Kirchner, J. McLaughlin, L. Mølhave, E. O. Fernandes, B. Seifert, The INDEX project: Executive summary of a European Union project on indoor air pollutants, *Allergy Eur. J. Allergy Clin. Immunol.* 63 (7), 810–819 (2008); doi: 10.1111/j.1398-9995.2008.01740.x

[5] R. Preuss, J. Angerer, H. Drexler, Naphthalene - An environmental and occupational toxicant, *Int. Arch. Occup. Environ. Health*, 76 (8), 556–576 (2003); doi: 10.1007/s00420-003-0458-1

[6] L. Mølhave, G. Clausen, B. Berglund, J. De Ceaurriz, A. Kettrup, T. Lindvall, M. Maroni, A. C. Pickering, U. Risse, H. Rothweiler, B. Seifer, M. Younes, Total Volatile Organic Compounds (TVOC) in Indoor Air Quality Investigations, *Indoor Air* 7 (4), 225-240 (1997); doi: 10.1111/j.1600-0668.1997.00002.x

[7] M. Leidinger, T. Sauerwald, W. Reimringer, and G. Ventura, Selective detection of hazardous VOCs for indoor air quality applications using a virtual gas sensor array,*J. Sensors Sens. Syst.* 3, 253–263 (2014); doi: 10.5194/jsss-3-253-2014

[8] D. Puglisi, J. Eriksson, C. Bur, A. Schütze, A. L. Spetz, and M. Andersson, Catalytic metal-gate field effect transistors based on SiC, *J. Sensors Sens. Syst.* 4, 1–8 (2015); doi: 10.5194/jsss-4-1-2015

[9] C. Bur, M. Andersson, A. L. Spetz, N. Helwig, A. Schütze, Detecting Volatile Organic Compounds in the ppb range with platinum-gate SiC-Field Effect Transistors, *IEEE Sens. J.* 14 (9), 3221–3228 (2014); doi: 10.1109/ICSENS.2013.6688279.

[10] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58.(2), 109–130 (2001); doi: 10.1016/S0169-7439(01)00155-1

[11] N. Helwig, M. Schüler, C. Bur, A. Schütze, T. Sauerwald, Gas mixing apparatus for automated gas sensor characterization *, Meas. Sci. Technol.* 25, 055903 (2014); doi: 10.1088/0957-0233/25/5/055903

[12] M. Andersson, R. Pearce, A. Lloyd Spetz, New generation SiC based field effect transistor gas sensors, *Sensors and Actuators B Chemical* 179, 95–106 (2013); doi: 10.1016/j.snb.2012.12.059.

[13] M. Browne, Cross-Validation Methods, *J. Math. Psychol.* 44 (1), 108–132 (2000); doi: 10.1006/jmps.1999.1279

[14] N. R. Draper and H. Smith, Applied regression analysis, 1st ed., no. 766. *Wiley Series in Probability and Mathematical Statistics*, 1966.

[15] A. Schütze, A. Gramm, T. Rühl, Identification of Organic Solvents by a Virtual Multisensor System With Hierarchical Classification. *IEEE Sensors Journal* 4 (6), 857–863. (2004); doi: 10.1109/JSEN.2004.833514

[16] Z. Darmastuti, C. Bur, N. Lindqvist, M. Andersson, A. Schütze, A. Lloyd Spetz, Hierarchical methods to improve the performance of the SiC-FET as SO2 sensors in flue gas desulphurization systems , *Sensors and Actuators B* 206, 609–616 (2015); doi: 10.1016/j.snb.2014.09.113

[17] D. Perez-Guaita, J. Kuligowski, G. Quintás, S. Garrigues, M. D. La Guardia, Modified locally weighted-Partial least squares regression improving clinical predictions from infrared spectra of human serum samples, *Talanta* 107, 368–375 (2013); doi: 10.1016/j.talanta.2013.01.035