Multi-unit calibration rejects inherent device variability of chemical sensor arrays

Ana Solórzano [ɸ 1,2], Raquel Rodríguez-Pérez [ɸ 2], Marta Padilla[2], Thorsten Graunke[3], Luis Fernandez[1,2], Santiago Marco[1,2], Jordi Fonollosa[1,2,4,5]

[1]Department of Electronics and Biomedical Engineering, Universitat de Barcelona, Marti i Franquès 1, 08028-Barcelona, Spain

[2]Signal and Information Processing for Sensing Systems, Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona Spain

[3]ams Sensors Germany GmbH, Gerhard-Kindler-Str. 8, 72770 Reutlingen, Germany.

[4]Department of ESAII, Center for Biomedical Engineering Research, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain

[5]Institut de Recerca Sant Joan de Déu, Santa Rosa 39-57, 08950 Esplugues de Llobregat, Spain.

ɸ: Contributed equally.

Abstract

Inherent sensor variability limits mass-production applications for metal oxide (MOX) gas sensor arrays because calibration for replicas of a sensor array needs to be performed individually. Recently, calibration transfer strategies have been proposed to alleviate calibration costs of new replicas, but they still require the acquisition of transfer samples. In this work, we present calibration models that can be extended to uncalibrated replicas of sensor arrays without acquiring new samples, i.e., *general* or *global* calibration models. The developed methodology consists in including multiple replicas of a sensor array in the calibration process such that sensor variability is rejected by the general model. Our approach was tested using replicas of a MOX sensor array in the classification task of six gases and synthetic air, presented at different background humidity and concentration levels. Results showed that direct transfer of individual calibration models provides poor classification accuracy. However, we also found that general calibration

models kept predictive performance when were applied directly to new copies of the sensor array. Moreover, we explored, through feature selection, whether particular combinations of sensors and operating temperatures can provide predictive performances equivalent to the calibration model with the complete array, favoring thereby the existence of more robust calibration models.

1. Introduction

In the modern era of ubiquitous sensing, cost-efficient gas sensors are expected to encounter many different applications, such as environmental monitoring, public safety, building automation and medical systems [1]. Among the different gas sensing technologies, metal oxide (MOX) sensors are a popular choice due to the ease of use, fast response, high sensitivity, miniaturization options and low-cost [2,3]. However, the calibration of MOX sensors faces two significant limitations due to the high correlation among features, sensor drift, scattering at different concentrations, and lack of reproducibility, among others [4,5]. First, calibration models only keep their predictive performance in the same conditions in which they were built. Thus, the utility of the model is compromised by the change of environmental or sampling conditions and the aging of the sensor. In consequence, calibration models require many calibration conditions to incorporate cross-sensitivities to the model. Second, inherent sensor variability requires specific calibration models for each system, even for replicas of gas sensor arrays with the same design and sensor types. Hence, the sensor variability and the constrained conditions of the calibration models hinder the generalization capability of the models. As a result, calibration is a costly and time-consuming process for MOX sensor arrays.

Substantial improvements in MOX sensor technology appeared in the last 20 years leading to significant performance enhancement [6–11]. Among the most relevant contributions to the amelioration of MOX sensing behavior we should highlight the development of new materials [12], the use of temperature modulation [13,14], and the integration of sensing layers in micromachined substrates, leading to an important power reduction [15–17]. While these technology improvements diminished the inter-device tolerances, calibration models for MOX gas sensor arrays cannot be transferred among replicas of the arrays and they are not robust against sensor replacements.

The widespread use of MOX sensor arrays in practical applications will rely on the improvement of the robustness of the calibration models and the reduction of calibration costs and time. Robust calibration models should capture the chemical variance of interest and, at the same time, reject non-relevant sources of variance. Some of the strategies that may improve calibration models and reduce costs are: i) data pre-processing, ii) robust feature selection and iii) calibration transfer techniques. Moreover, strict validation procedures with external validation test data are necessary to ascertain the performance of the predictive model and assess its applicability domain [18].

Calibration costs might be reduced with methods that limit the number of calibration conditions [19]. Data pre-processing strategies are widely used to reject some effects caused by changes in concentration, sampling, or environmental conditions. After pre-processing step, the number of calibration conditions that may be presented to the calibration models to obtain accurate predictions can be reduced [20]. Besides, data pre-processing avoids the inclusion of non-relevant variance, resulting in simpler and more general models (parsimonious models). Gutierrez-Osuna et al. explored different data pre-processing methodologies for odor discrimination. The figure of merit used was the classification rate of a k-NN in a Fisher linear discriminant subspace. Their results confirmed the benefit of data preprocessing in pattern recognition methodologies for gas sensors [21]. Other strategies to make calibration models more robust are based on orthogonal projection filters. Artursson et al. proposed Component Correction [22], Padilla et al. explored Orthogonal Signal Correction [4], and Zityadinov et al. proposed Common Principal Component Analysis [23]. For the above-mentioned approaches, the resulting calibration models were more robust.

Another strategy also employed to improve the robustness of the model is feature selection, which has been successfully applied in many occasions to identify the most informative features that are invariant with respect to changes in the conditions of the instrument or the environment [5]. Techniques such as Sequential Feature Selection (SFS), Sequential Backward Selection (SBS), and Genetic Algorithms (GA) have been previously applied for this task [24,25]. Feature selection is closely related to sensor-array optimization, a popular approach to reduce the number of sensors [26,27]. Therefore, this strategy may also be understood as the selection of the best combination of sensors that

successfully discriminate the target volatiles. In the presence of unwanted sources of variance (tolerance between units, sensor drift, etc.), the selected sensors will constitute a robust combination against those sources of noise. For instance, Raman et al. presented a MOX microhotplate sensor-array optimization problem, and they encountered that the best configuration included two materials and four replicas of each [28,29]. GA for the use of sensor selection was already proposed by Gardner et al., who also fixed the number of features to be selected to avoid the exploration of all the combinatorial space [30]. GA is currently one of the most popular approaches due to its ability to avoid locally optimum solutions compared to other techniques [5]. Hence, feature selection allows the development of calibration models using fewer parameters, resulting in simpler calibration models that may be better suited to predict new samples. Besides, a system with a reduced set of sensors may need fewer calibration conditions and calibration costs may be further reduced.

Over the years, different methodologies have also been proposed to reduce the number of calibration conditions. For example, Shmilovici et al. used Support Vector Regression to choose the best calibration points (support vectors) for the quantitative prediction of mixtures of three gases with a 12-MOX sensor array [31]. The performance of active control sampling has also been explored, aiming at the optimization of the order to present calibration conditions (gas class and concentration) to ensure, on the fly, the most relevant condition during the whole calibration sample acquisition [19].

Moreover, similarly to standardization techniques applied in spectroscopy [32–34], calibration transfer strategies have been explored to extend calibration models to replicas of a sensing instrument in order to reduce calibration costs. Briefly, calibration transfer strategies are based on building a calibration model with a complete set of calibration samples using a master (reference) instrument. Then, a smaller set of samples (transfer samples) is acquired with a slave instrument to determine a function that maps the responses of the slave and the master systems. This mapping of the response spaces is then used to transfer the calibration model between replicas.

In the field of chemical gas sensor arrays, the concept of data transferability between units in view of calibration transfer was pioneered by Balaban et al. They compared univariate regression between individual sensors, multivariate linear regression (MLR) and

Artificial Neural Networks (ANN) [35]. After the correction of the slave's signals, they applied Statistical Neural Networks to the master data and the transformed data. Multivariate Regression techniques (MLR) provided the most accurate results in terms of classification rate. In another interesting work, Zhang et al. compared the response of different MOX sensors and found out that sensitivity can differ up to 17% for sensors of the same type. This mismatch may lead to unreliable predictive models if the calibration model is directly extended to other copies of the same sensor. Hence, they proposed univariate linear functions to map the response of the sensors of the same type [36,37], in a similar way as Balaban did years before. However, they claim that results improve if simple linear regression is replaced by a robust weighted least squares regression. Fernández et al. compared four different techniques for calibration transfer between temperature-modulated sensor arrays: direct standardization (DS), piecewise direct standardization (PDS), orthogonal signal correction (OSC), and Generalized Least Squares Weighting (GLSW). They found that among all the tested techniques, PDS provided best results [38]. Very recently, DS was proposed to transfer a Support Vector Machine model from a master device to slave instruments and, interestingly, calibration transfer strategies were also applied to counteract drift [39]. In the literature, drift is usually connected to physical changes in the sensing layer attributed to, for example, sensor aging and poisoning (first-order drift); and to uncontrolled changes in the environmental conditions (second-order drift) [40,41]. Other recent works also used transfer learning strategies to compensate first-order drift in sensor arrays [42–44]. Actually, any system affected by first-order drift can be considered as another virtual replica of the original system with deviated sensitivity. Hence, calibration transfer algorithms can help to correct first-order drift. Additionally, calibration transfer techniques were proposed to counteract the variations in the behavior of a sensor array after sensor replacement. Tomic et al. developed a methodology based on Component Correction (CC) and Multiplicative Drift Correction (MDC) followed by a k-NN as classifier [45]. In a first stage, they acquired data with an array composed of 12 MOX sensors and 5 MOSFET sensors. One MOX sensor was replaced, and they acquired a second data set. The second dataset was transformed to use the calibration model built with the first dataset. Best classification rates were achieved after signals were transformed.

The mentioned calibration transfer strategies successfully reduce calibration (and recalibration) costs of system replicas. The ratio between the number of calibration samples and the number of transfer samples determines the savings in the calibration process. This analysis is sometimes overlooked in the literature, and different strategies make use of very different ratios, ranging from 5% to 60%. Nevertheless, calibration costs of MOX sensor arrays are still identified as a major burden towards large-scale deployment of such systems [18,46].

Finally, in the recent years, there has been an increasing amount of literature on calibration strategies for Distributed Environmental Sensory Networks (DESN) as they became economically feasible to monitor air quality in urban areas [47–50]. Advanced strategies for DESN calibration combine measurements acquired by different sensory nodes. These calibration strategies usually rely on data acquired with collocated nodes in time and space, or assumptions such that the nodes are exposed to similar concentration levels at night [51]. Fishbain and Moreno-Centeno presented a consensus methodology for uncalibrated wireless DESN that is also applicable when the aggregation of measurements is incomplete [52]. However, the mentioned strategies are designed for networks of sensing systems as they need data acquired by other units to provide calibration models. In this work, we will focus on systems that operate standalone systems and do not have data available from other sensing units. Actually, consensus algorithm strategies were also applied to correct sensor failures in sensor arrays and extend system lifetime before re-calibration or sensor replacement [53,54].

Specifically, we propose a calibration methodology for arrays of MOX gas sensors that takes advantage of the inherent sensor variability and, therefore, the obtained calibration model can be extended directly to other replicas of the array without the acquisition of new (transfer) samples. This has been achieved by building a general calibration model with different copies of a sensor array and searching a set of sensors that favors the simplicity of this general calibration model. We tested our methodology in the task of discrimination between six gases presented at three concentration levels over a background of synthetic air at three humidity levels. Different calibration strategies were compared to ascertain whether general models trained with several arrays provided similar predictive performances than models trained specifically for individual arrays.

2. Materials and methods

2.1 Dataset

Five replicas (units) of a 24-sensor array were assembled to generate a thorough dataset to test different calibration strategies. Each sensor array included three different types of MOX gas sensors, and each array included eight units of the three types of sensors. The eight elements of the same type within an array operated at different temperatures for a total of 24 sensors per array.

The sensors were provided by ams AG [55] with labels that identify the different sensing properties of the sensors: MLC sensors were specifically designed to be sensitive to carbon monoxide, MLK sensors targeted methane, and MLX targeted volatile organic compounds. The sensors were packaged in the standard TO-39 housing (4-pin header). A heater and interdigital electrode structures were placed on a 1 μm-thin membrane, on top of which a tin dioxide sensitive layer was deposited. The built-in heater enables the control of the sensor operating temperature, which in turn changes the performance characteristics of the sensor [56,57]. At typical operation temperature (270ºC for MLC and MLX, and 320ºC for MLK), MLK shows the highest sensitivity to methane in the concentration range of 0.01-4%, and MLC shows the highest sensitivity to CO in the range of 0.5-500 ppm. The typical recovery time of the sensors is in the range of seconds. For more details on sensor specifications, the reader is referred to the manufacturer.

The actual temperature of the sensing layer according to the power applied to the heater was determined after a two-step process using the heater as a resistive temperature sensor. First, the temperature coefficient of the heater resistance (TCR) was determined. The temperature was increased stepwise in an oven, and the resistance of the sensor heater was measured using a four-wire configuration. Then, the relationship between the applied heating power P and the measured heating resistance R was measured by gradually increasing the power using a 100-Ω series resistor. Hence, the relationship between heating power P and operating temperature T was obtained by comparing both measurements. Typical heater resistance at room temperature is 95 Ω, with 1700 ppm/K temperature coefficient. In this analysis we neglected temperature gradients that may occur in the sensitive material. Proper thermal design of the structure may confine the power dissipation on the hotplate and minimize the power dissipation on the suspension

elements. Additionally, it is desired that the temperature distribution is homogenous in the sensitive material and the temperature gradient is located just along the suspension membrane. Detailed discussion on the thermal design and behavior of hotplates can be found in Walden et al. [58]. On the other hand, we did not implement any active temperature control to reject changes in the room temperature. A proposal for active temperature control can be found in Kneer et al. [59] However, the temperature in the sensor chamber was continuously monitored, and laboratory temperature was also regulated. Specifically, to verify the experimental conditions (set-points), temperature and humidity were continuously measured using an SHT11 sensor (accuracy of ±0.4ºC and ±3%r.H.) provided by Sensirion (Switzerland). The recorded temperature was 25.5±0.5ºC, so we expect a minimal impact on the temperature of the hotplate.

The sensors of the same type, within the same array, operated at eight different temperatures. The sensor heater was adjusted via a constant power board (Analog Devices; Part No.: Eval-AD5380SDZ) to ensure a constant temperature. Specifically, the nominal values of the estimated operating temperatures were 245, 275, 303, 313, 330, 340, 356 and 381°C. We will refer to each combination of sensor type and operating temperature as a feature. Therefore, every sensor array provides patterns of 24 features (i.e., input space of 24 dimensions).

The sensor arrays were placed in a measurement chamber (volume = 286.9 cm^3) and exposed to different gas conditions, while sensor responses were acquired at a sampling frequency of 1Hz. A set of mass flow controllers (MFC) was used to control the composition of the sample gas, the humidity level, and the total flow through the chamber. Specifically, we used three MFCs (EL-FLOW Series provided by Bronkhorst) to control the mixture of synthetic dry air (range 0-200 sccm), humidified air (0-200 sccm), and the gas under test (0-100 sccm). The flows of three MFCs were adjusted to generate the desired concentration levels under the defined humidity level, while keeping the total flow constant.

The experimental protocol consisted in exposing the sensor arrays to synthetic air at a defined humidity level for 90 minutes. This first (cleaning) stage was used to purge the measurement chamber from the previous measurement and recover sensor baselines. Next, a gas sample was introduced to the measurement chamber for 30 minutes

(measurement cycle). The air flow was kept constant at 196 ml/min for all the duration of the experiment (cleaning and measurement stages). The humidity level only was changed after the experiment was completed.

For each measurement and sensor, we extracted the sensor conductance 28 minutes after the gas was presented (gas response), and 2 minutes before the sample was introduced (baseline levels). **Figure 1 (top)** shows the signals of MLK sensors operating at four different temperatures when exposed to CO. **Figure 1 (bottom)** presents the corresponding set-points for CO and humidity levels. Black and green points overlapping the set points represent the time at which the sensor signals were acquired for the construction of our dataset.

(see below) (new version of the figure)

Figure 1: Sensor responses to CO presentation under different humidity levels. Acquired signals for four MLK sensors (top panel), and CO and humidity set-points (bottom panel). Sensor conductivity changes according to the presented conditions. Blue (baseline) and green (measurement cycle) dots represent sampling points for the construction of the dataset.

Six different gases were presented to the sensor arrays: acetaldehyde, methane, ethanol, propane, nitrogen dioxide, and carbon monoxide. Each gas was presented at three concentrations and three humidity levels. **Table 1** details the set-points for the gas concentration and humidity levels for each of the presented gases. The temperature and humidity sensor (SHT11) provided actual measurements of the experimental conditions and its variability. The mean temperature during measurements was 25.5ºC, and the deviations were always within the accuracy of the temperature sensor. Similarly, the measured relative humidity levels at the measuring points were within the accuracy provided by the humidity sensor. Specifically, the measured values ranged in the intervals of (20,22)% r.H., (40,42)% r.H., and (60,61)% r.H. for the three set-points, respectively always at 25.5ºC. Only the measurements that correspond to the highest concentration level of the gases, at the highest humidity level, were discarded due to technical setback. Hence, in total, we considered 48 different gas conditions (six gases at eight humidity and gas concentration combinations). In order to consider a balanced dataset among all the gas conditions and the blank samples (background air), we randomly selected 8 of the

background measurements (baseline measurements). From now on, for simplicity, we call *air* to the class *background air*, although it contains three humidity levels. All in all, our dataset, includes 56 (gas and air) conditions, which were evaluated once for each of the five copies of the sensor array.

(see below)

Table 1: Volatiles, humidity and concentration levels to which every sensor array was exposed for the dataset acquisition.

2.2. Evaluation of calibration methods

In order to evaluate the performance of different calibration methods, and based on the dataset described previously, we built predictive models for the discrimination of the six gases regardless of concentration and humidity levels.

For benchmarking purposes, the performance of an individually calibrated sensor array was taken as reference. The performance of this individual calibration model was evaluated when applied to different replicas of the sensor array, in order to estimate the impact of sensor tolerances on the accuracy of the predictive model. As an alternative, we proposed to build a general calibration model that incorporates a number of replicas of the sensor array. We validated the model with test samples from different copies of the sensor array. Due to the characteristics of the used dataset, general calibration models were calibrated with up to four replicas and validated with the rest, but this can be changed in the general case.

Hence, we explored the prediction ability of calibration models in four different scenarios.
- Scenario 1. An individual calibration model is obtained with data from one sensor array, and it is tested with data from the same sensor array. This constitutes the reference condition for our analysis.
- Scenario 2. An individual calibration model is obtained with data from one sensor array, and it is tested with data from different sensor arrays. This case aims at evaluating whether a direct transfer of the calibration models is feasible or not.
- Scenario 3. A general calibration model is obtained with multiple copies of the sensor array, and it is applied to a different sensor array. This model makes use of

all the data captured with the units included in calibration. This scenario explores whether the model is able to extract the common behavior among a group of sensor arrays and, thereby it rejects sensor tolerances. Moreover, two other methodologies were explored to compare the performance of the general model with other training data. We also built a calibration model with the averaged response of different calibration units. Finally, we built calibration models with data from different sensing units, but keeping constant the total number of calibration measurements.

- Scenario 4. It is the same procedure as in Scenario 3, but including a feature selection step. In this scenario, we search for sensor-array optimization. Therefore, in this case, feature selection is directly related to sensor selection, in particular, sensor type and its operating temperature.

For the four scenarios, calibration models were built and evaluated for all the possible combinations of training/test data partitioning.

Classification models were based on a multiclass (all classes vs. all classes) Partial Least Squares - Discriminant Analysis (PLS-DA) [60], followed by k-Nearest Neighbours (k-NN) in the latent variable (LV) subspace. The optimization of the number of latent variables (LV) and neighbors (k) was performed considering the ranges LV = [1, 24] and k = [1, 19]. The combination of hyperparameters that provided maximum classification rate in internal cross-validation was used to build the final model with all the calibration samples. Then, the performance of the models was estimated with the corresponding test dataset. **Figure 2** shows the overall validation methodology.

(see below) (slightly modified figure)

Figure 2: Workflow of the methodology employed to build and test the calibration models. The dataset was divided into calibration and test sets. For each scenario, the calibration model for all the possible combinations of calibration/test data partitioning was built and evaluated.

The size of training and test sets for each of the considered scenarios is as follows (see **Table 2).**

- Scenario 1. One measurement of each volatile and one air measurement were randomly selected as the test set, whereas the rest of the measurements constituted the training set. The model parameters were selected after internal cross-

validation using Leave One Out (LOO). The process was repeated eight times until all the measurements were used for model testing (external validation). This methodology allowed to evaluate the performance of the calibration models when the models were built and tested with data from the same replica. The same process was applied to the five sensor array replicas. Therefore, classification accuracy was evaluated with a total of 280 measurements (40 measurements per class).

- Scenario 2. All measurements from one sensor array (56) were included in the training set, and the measurements from the rest of sensor arrays constituted the test set. This procedure was repeated such that all sensor arrays were used once for model training or calibration.

- Scenarios 3 and 4. All the measurements from four sensor array replicas (224) were used to train the models, and the measurements acquired with the remaining sensor array (56) composed the test set. The process was repeated such that all the sensor arrays were used once for testing the calibration models.

Even though scenario 4 is equivalent to scenario 3 in terms of training and test set size, scenario 4 includes feature (or sensor) selection as a part of the calibration process (i.e., only performed with training samples).

The employed feature selection strategy was based on GA [61]. First, different subsets of features (population) were randomly created, and their predictive performance was estimated through PLS-DA models. PLS-DA models were built using the internal training set, and the Fisher's ratio (computed in the LV subspace) of the internal test samples was the figure of merit for selecting the best subsets of features. The best performing subsets of features were then used to generate a new population through crossover among them or mutations [62,63]. This process was repeated until the objective function (i.e., Fisher's ratio) of the best individual did not increase more than 1% after 10 iterations. The process is summarized in **Figure 3**. GA assumes that the best-fitted individual in the last iteration is the final solution. However, different trials of GA do not converge to the same solution. Therefore, to prevent a locally optimum solution, GA were repeated and a frequency of selection for every feature was obtained as a result. Finally, the final calibration model was built considering the features that were systematically selected and thus their

frequency of selection was larger than expected form a random selection. The analysis was done with R, using, *class, pls,* and *GA* [64–66] libraries.

Figure 3: General scheme of Genetic Algorithms. An initial population is generated, and then there are several iterations during which the individuals are evaluated, selected, recombined (crossover) and mutated. The individual with the best fitness of the final iteration gives the selected features.

All the four scenarios were evaluated in terms of classification rate using a generalization of the area under the ROC curve (AUC) for multiclass classification problems proposed by Hand and Till [67]. To test whether the results were obtained by chance, all the models were trained with permuted labels 1000 times (permutation test). The null hypothesis is that the relationship between the data and the labels cannot be learned by the classifier during training. To reject the null hypothesis we accepted a risk of 0.05.

Table 2: Data partitioning used for each scenario. Number of sensor arrays and samples used in the different considered scenarios.

3. Results and discussion

This work aims at developing calibration models that reject sensor variability and can be extended to new replicas with no need of transfer samples. To meet this goal, general calibration models were built with samples from five sensor array replicas and were scrutinized in their ability to predict samples acquired with a sensor array not present in the training set. The performance of these general calibration models was compared with that of individual models (calibrated with only one sensor array). In addition, a small subset of features was selected to construct general models that are parsimonious and more robust.

3.1. Dataset visualization.

A Principal Component Analysis (PCA) provided an exploratory analysis of the captured dataset (see **Figure 4**). Projected data appears in regions for different gas classes, with some overlap among them. Different sensor responses, due to the different presented gas conditions, dominate the inherent sensor variability in this PCA projection. Although **Figure 4** provides a simple and yet informative visualization of the dataset in a 2-dimension space, in general, calibration models will make use of a higher number of dimensions.

(see below)

Figure 4: PCA scores of the different sensor array samples. Numbers correspond to the sensor array and color codes the gas (class).

3.2. Individual calibration models

Models consisting of PLS-DA followed by k-NN were built using data acquired with one sensor array. Then, the discrimination ability of the models was evaluated with data acquired with the same sensor array using data not presented in calibration (scenario 1) and using data captured with other copies of the sensor array (scenario 2).

Figures 5a and **5b** show the confusion matrices for scenario 1 and scenario 2, respectively. The built calibration models were able to perfectly predict the class of new samples if they were acquired with the same sensor array (100% classification rate). However, individual models substantially decreased their predictive ability for new samples measured with other replicas of the sensor array. In scenario 2, even though all the obtained models predicted propane, nitrogen dioxide and methane without error, only 62.5% of the CO samples were correctly classified. The rest of CO samples were misclassified as background air (40 samples), ethanol (16 samples) or acetaldehyde (4 samples). One-fourth of the air samples acquired during the baseline cycle was wrongly predicted as CO. In addition, one of the models misclassified one acetaldehyde sample as CO, giving a total of 41 false positives and 60 false negatives for the task of CO detection. The overall classification rate (CR) was 91%. We run a Fisher's exact test (two-tailed) [68] to assess the significance of the observed differences between the CR obtained for the model tested with the same sensor array, and the CR obtained for the model tested

with different sensor arrays. The null hypothesis that establishes that both distributions are the same was rejected (p-value<0.0001).

Thus, our results confirmed that individual models were local to the sensor array employed for calibration. To a certain extent, individual models were able to predict gas samples from other replicas, but generally failed in one of the most relevant discrimination tasks: CO detection [69]. For this gas, individual models were not able to extrapolate to data acquired with other arrays. The local behavior of models trained with one single array is well-known due to the already mentioned variability among sensors.

(see below)

Figure 5: Confusion matrix obtained when the model is tested with data acquired with the same sensor array present in calibration (top, scenario 1) and when it is tested with data acquired with different copies of the array (bottom, scenario 2). The classification accuracy decreases if the calibration model did not include data from the sensor array used for training.

3.3. General calibration models

General calibration models were built using data captured with four replicas of the sensor arrays, and they were then evaluated with the remaining replica. **Figure 6** shows the confusion matrix after all the iterations and sensor combinations were performed. Interestingly, results show that the models can be extended with high accuracy to new replicas, even though no measurements from the new sensor array were included in calibration. General models achieved a classification ratio of 99% of the test samples despite the perturbation due to varying humidity levels. From the 280 test samples, the model only confused two air samples with CO. We also performed Fisher's exact tests to evaluate the significance of the different CRs. The null hypothesis that the distribution of the general calibration model and the distribution of the individual calibration models (tested with data from the same array) are the same cannot be rejected (p-value=0.5). The hypothesis that the general model is the same than the individual model tested with data from new arrays was rejected (p-value<0.0001). Hence, although the slightly lower CR, statistically, the distribution of the CR obtained with the general model (scenario 3) cannot be considered different to the distribution of the predictions when a calibration model is built and tested with the same unit (scenario 1).

Figure 7 shows the PLS-DA score plot for a general calibration model. This representation highlights the presence of different clusters that correspond to the distinct gases. In general, test samples of every gas fall close to those of the calibration set in the PLS-DA subspace, thereby suggesting a good generalization of the model.

The inclusion of the sensor array variability to the general calibration model improved the ability of the model to predict the class of samples acquired with new sensor arrays. The model is able to extract the redundant information of the sensor arrays, reducing thereby the individual sources of noise. In order to ascertain to what extent sensor variability needs to be included in the calibration process, we evaluated the performance of a new model that uses averaged sensor signals.

Specifically, we compared the performance of our calibration method that incorporates, directly, all the variability of the sensor signals, with a methodology that averages the response of four calibration units. The calibration models obtained with the averaged responses were tested with data acquired with the sensor array not included in calibration (a unit was set aside for model evaluation, and we repeated the process until all the units were used for test). Classification rate (97.5%) was higher than the one observed for individual models applied to other sensor arrays replicas (91%), but did not reach high classification rate obtained with the general models that carry all inherent sensor variability (99%). Similarly, than for the global model, air samples were misclassified as CO. However, for the model built upon averaged signals, air samples were also misclassified as methane, and CO samples were wrongly classified as air and propane.

Actually, the average of the signals still carries information that is beneficial to reduce individual sources of noise [70]. Therefore, one expects the averaged model to perform better than the direct transfer of an individual calibration model. Hence, that for efficient rejection of the sensor variability, the calibration model needs to rely on a training dataset that includes measurements that are representative of the sensor variability, which is not the case in the case of averaged sensor responses across units. Additionally, averaging the sensor signals still requires the use of four sensor units, so we do not have any savings in terms of test effort. Results confirm that the averaged-signal model does not reach the classification rates obtained by the global calibration model, confirming thereby the

benefits of including all sensor diversity to the calibration dataset. In this sense, the variance corresponding to the intrinsic sensor variability was better rejected with the calibration model trained with the original signals.

Figure 6: Average of confusion matrices of the general calibration models. Models were trained with samples of four different arrays and evaluated with samples of a different array. The performance of the general model is similar to the performance of individual calibration models when were trained and tested with samples of the same array.

Figure 7: PLS-DA score plots for the general calibration model. Color indicates the gas type and marker type indicates calibration/test point. (left) LV 1 vs LV 2, (right) LV 1 vs LV 3. Calibration and test points appear in the same regions, confirming thereby a generalization of the models.

The superior performance of the general calibration models with respect to the individual models, when applied to new sensing units, may come from i) the higher data variability in the calibration dataset due to the inclusion of different replicates, and ii) the larger size of the calibration dataset. In order to investigate further these two effects, we built calibration models with data from different sensing units but keeping constant the total number of calibration measurements. **Figure 8** shows the model performance, in terms of classification rate, as calibration data from increasing number of units is used. Two strategies are shown in the mentioned figure: i) considering all available samples of the used arrays (each new unit adds 56 calibration measurements to the model), and ii) keeping constant the total number (56) of calibration samples. Results show that diversity from different sensor arrays is favorable for building general calibration models, although when all the available data is used, the performance of the calibration model increases more rapidly. Nevertheless, in practical calibration scenarios, one may expose the sensor arrays to the calibration conditions simultaneously. In this case, it is recommendable to use all data at hand. From **Figure 8** one can observe that adding a second array to the calibration model increases the prediction ability of the model significantly, and the performance reaches a plateau when three sensor arrays are used. Hence, one can conclude that, for the considered classification task, three copies of the sensor array would suffice to build general calibration models that can be extended successfully to new

arrays. However, more complex tasks or systems with different sensing elements may require a different number of units in calibration to provide reliable general models.

(see below)

(new figure in the manuscript, figure numbers are shifted from now on)

Figure 8: Performance of the models when evaluated with test samples from units not used in calibration, for different number of sensor arrays used in calibration. Keeping the number (56) of calibration samples constant (red), and using all the data available (56 per sensor array) for the units used in calibration (blue).

3.4 Sensor selection

We also explored whether a subset of the 24 considered features (24 combinations of sensor type and operating temperature) can be selected to model the chemical variation of interest and reject the calibration conditions, since this might help in model generalization. In particular, we selected a subset of features using GA with the purpose of building more robust general models. Since feature selection is directly related to the selection of sensing elements, it also results in more cost-efficient devices.

The methodology based on GA selected 4 features of each array after 96 executions: MLC sensor at 245ºC and 356ºC, MLK at 381ºC, and MLX at 386ºC. The average of the confusion matrices for scenario 4 is presented in **Figure 9**. General calibration models that only use data from 4 sensors provided good predictions of the gas classes. Specifically, all the acetaldehyde, methane, CO, ethanol, nitrogen dioxide, and propane samples were correctly classified. The confusion of CO with background air is the only one that appears in the general calibration models, and it happens at a higher frequency with models that incorporate feature selection. In particular, CO showed nine and two misclassifications in the general models with and without feature selection, respectively.

With a 97% of classification rate, general models with feature selection showed a good ability to predict samples of new arrays. Their classification accuracy was higher than individual models (91%) and very close to the general models considering the complete sensor array (99%). The Fisher's exact tests evaluated the significance of the CRs for the general model with, and without feature selection. The null hypothesis that establishes

that both distributions are the same cannot be rejected (p-value=0.062). These results seem to indicate that in order to build general models, sensor array diversity is an asset but, in the search for simpler arrays, the provided benefits get slightly degraded.

Figure 10 shows the PLS-DA projection of one general calibration model with feature selection. Similarly to the calibration model with all the features, the proximity among training and test datasets suggests the good generalization of the models with feature selection.

An optimum selection of sensors could basically keep constant the performance of general calibration models with the advantages that a simpler sensor array brings, such as miniaturization, cost reduction and a parsimonious model of easier interpretation (due to the reduction of sensing elements). In particular, feature selection allowed the identification of the common features that have high predictive performance. We avoided, thereby, including many variables in the model, which might be modeling other conditions that include unwanted sources of variance. It could be that the most accurate model, which is the general without feature selection, is also the one that degrades faster in time due to the inclusion of non-informative features. It is the application or the demands of the user that determine whether the drop in the performance (limited to one gas class only) compensates or not the reduced size of the sensor array after feature reduction, its cost-efficient design, and the simplicity of further re-calibration or sensor replacement. For a practical application, it would also be interesting to assess whether the model that incorporates feature selection shows reduced drift in time or rejects other sources of variability better than the general model with the complete sensor array.

(see below)

Figure 9: Average of confusion matrices of the general calibration models with selected features after GA. Models were trained with samples of four different arrays and evaluated with samples from a different array.

(see below)

Figure 10: PLS-DA score plots for the general calibration model with feature selection. Color indicates gas type and shape indicates train/test data. (A) LV 1 vs LV 2, (B) LV 1 vs LV 3. The model is capable of separating the seven different gases partially. Training and Test samples follow the same distribution.

3.5 Area Under the ROC Curve (AUC)

In order to compare further the performance of each model, we computed a generalization of the AUC for multiclass classification to estimate the robustness of the different models. **Figure 11** shows the distribution of AUC values for scenarios 2, 3, and 4. The averages of the AUC obtained for general calibration models with and without feature selection are 99%. On the other hand, the performance of individual calibration models evaluated with new samples drops to 97%, compared to 100% for the scenario in which the model was tested with samples from the same sensor array. In addition, results of permutation tests point out the statistical significance of the results. With a 95% of confidence, all the models rejected the null hypothesis that the obtained AUC values could come from a model trained with random labels (i.e., random distribution).

The obtained AUC values confirm that general calibration models show higher prediction ability to classify samples acquired with new sensor arrays than models built with one sensor array. Moreover, general calibration models with only four features are also able to predict samples from new arrays successfully. This indicates that the selected features are informative and resulted in a parsimonious model to improve the robustness of the calibration.

(see below) (adapted figure according to reviewer's comments)

Figure 11: Relative frequency of the area under the curve (AUC) computed for each model. General models are more accurate than individual models for the prediction of new samples from arrays not presented in training. The histograms present the relative frequency for all the possible combinations of arrays.

3.6 Model complexity

The complexity of the models was studied in terms of the number of latent variables, LV, and the number of neighbors, k, that were selected during model training. **Figure 12** shows the frequency of selection of optimum LV and k values for individual and general models. For both types of models, optimum k ranges from 1 to 3, whereas the optimum number of LV tends to be larger in individual models (LV = [5,7]) than in general models (LV = [3,5]). Higher number of LV means that individual models are more complex. This possibly indicates that the models are capturing better the particular behavior of the calibrated sensor array, but this information is specific for this sensor array and it becomes

a misrepresentation of the general sensor behavior. **Figure 12** also shows that general models with feature selection perform better when trained with fewer LV. This is a direct consequence of the previous dimensionality reduction. Hence, GA helps in reducing the complexity of the model.

(see below) (adapted figure according to reviewer's comments)

Figure 12: Relative frequency of the number of latent variables and nearest neighbors used to build general and individual models. General models use fewer latent variables than individual models, providing an easier interpretation and a better generalization of the model.

4. Conclusions

In this work, we presented a methodology to improve the robustness of calibration models and reduce calibration costs of standalone MOX gas sensor arrays. Our approach is operative, scalable to a larger number of calibration units, and can be readily applied to multiple application fields that MOX gas sensors have found and may find in the future.

Our methodology is compatible with any performance test that could be carried out before actual system calibration to remove faulty elements. In this work, however, we focused on the calibration process to provide a prediction for a new sample, given a task. MOX sensors can be incorporated in systems targeting very different applications, where each application will have its own target volatiles and interfering gases. Hence, given the broad response that MOX gas sensors produce, their non-specificity, and the different array configurations (number and types of sensors, operating temperatures, etc.), it is necessary to define a set of calibration conditions specific for each application.

Although our study shows that the best results are obtained when sensor arrays are calibrated individually, this approach is not feasible in mass-production applications due to high costs. Previous approaches to obtain robust models and reduce costs are mainly based on calibration transfer. However, these strategies still require the acquisition of calibration samples for every new system and, ultimately, they rely on calibration models built upon individual sensor systems. Herein, we showed that general -or global- calibration models can be obtained if several replicas of the sensor array are included in the calibration process. Moreover, feature selection was also applied to avoid including

sensors or features that overfit the model to certain sampling, environmental or instrumental conditions. All in all, the obtained calibration models were able to reject the intrinsic system variability: for the proposed classification task, the prediction ability of the models only decreased marginally with respect to specific calibration models. Hence, the variability present in the training set allowed building robust models that could be extended to new replicas without any further calibration, resulting in calibration models that are replica-independent.

We also showed that general calibration models tend to be simpler (fewer number of LV) than individual calibration models, especially after selecting a subset of informative features or sensors. Actually, individual models are trained for the particular array of sensors, and higher dimensions of the model account for the specific response of the set of sensors. This information captured at higher dimensions results in better system performance, but the same information misleads the calibration model when it is used for other arrays. Higher dimensions can be understood as a more detailed calibration model for one array, but at the same time, it could be considered as an overfitted model when used for other arrays. Similarly, general models without feature selection performed only slightly better than the ones with a subset of selected features. However, general models with feature selection reinforce the common features that systematically exhibit better predictive ability using other replicas, and feature selection may help to reject better other sources of unwanted variability. Therefore, even though, for the selected test task, the model performance after feature selection does not improve the model with the complete sensor array, the calibration of a new sensor array, or a future re-calibration, might work better after feature selection.

We tested our methodology on the classification task of six different volatiles at different humidity levels. We selected the volatiles in view of their relevance in potential applications for chemical sensing, such as fire detection (carbon monoxide), combustible gas leak detection (propane and methane), air quality monitoring (nitrogen dioxide), or common interfering volatile (ethanol). However, our calibration approach should be tested for scenarios that consider more complex tasks, closer to final applications, using the specifically-designed detection systems. Also, it may be necessary to investigate our approach to a regression task. While quantification is usually a requirement in many applications, testing our methodology in a regression task would provide more sensitive

measures on the performance of the models when extended to new replicas. Actually, calibration transfer strategies did show good results when tested in regression tasks [38,39].

Specifically, we showed that, for the proposed test scenario, using four copies of the developed sensor array, one can build models that generalize to other replicas. New copies of the sensor array would not need any specific calibration nor the acquisition of transfer samples if they are set to operate under the same task, saving, therefore, future calibration costs. Nevertheless, the number of copies required to build a general calibration model may change for different sensing systems and proposed tasks. The goodness of the model extension to new units depends on the complexity of the task and the inherent sensor variability. Or, in other words, the ability of the general model to successfully predict samples from new sensor arrays is conditioned by the responses in the sensor space of the presented stimuli with respect to the distribution of responses caused by the inherent sensor variability, given a stimulus. Hence, one may need to explore the feasibility of building general calibration models for more complex tasks or other technologies with different sensor tolerance. Nevertheless, if the calibration of a particular unit needs to meet higher standards or accuracy, one can always perform individual calibration. Calibration costs and system design can also be reduced by applying feature reduction techniques. We also showed that the number of sensing elements can be reduced while system accuracy remains. If the system is built with a fewer number of sensors, a larger number of sensor arrays can be calibrated simultaneously, scaling down calibration costs as well.

It is worth noting that the initial calibration process may be more expensive than individual calibrations. For instance, more units need to be built and exposed to calibration conditions, more sophisticated equipment may be required to acquire several units simultaneously, or more time may be necessary to complete the whole calibration dataset. However, in the long term, general calibration provides significant cost reduction for mass-production applications where a large number of units needs to be calibrated: only a few will be calibrated, and the others will benefit from the general calibration model. The savings in calibration costs will be more and more significant as the total number of sensor arrays produced becomes larger.

Acknowledgments:

References

[1]   R.A. Potyrailo, Multivariable Sensors for Ubiquitous Monitoring of Gases in the Era of Internet of Things and Industrial Internet, Chem. Rev. 116 (2016). doi:10.1021/acs.chemrev.6b00187.

[2]   G.F. Fine, L.M. Cavanagh, A. Afonja, R. Binions, Metal oxide semi-conductor gas sensors in environmental monitoring, Sensors. 10 (2010) 5469–5502.

[3]   H. Gao, H. Jia, B. Bierer, J. Wöllenstein, Y. Lu, S. Palzer, Scalable gas sensors fabrication to integrate metal oxide nanoparticles with well-defined shape and size, Sensors Actuators B Chem. 249 (2017) 639–646. doi:https://doi.org/10.1016/j.snb.2017.04.031.

[4]   M. Padilla, a. Perera, I. Montoliu, a. Chaudry, K. Persaud, S. Marco, Drift compensation of gas sensor array data by Orthogonal Signal Correction, Chemom. Intell. Lab. Syst. 100 (2010) 28–35. doi:10.1016/j.chemolab.2009.10.002.

[5]   S. Marco, A. Gutierrez-Galvez, Signal and data processing for machine olfaction and chemical sensing: A review, IEEE Sens. J. 12 (2012) 3189–3214. doi:10.1109/JSEN.2012.2192920.

[6]   O. Monereo, J.D. Prades, A. Cirera, Self-heating effects in large arrangements of randomly oriented carbon nanofibers: Application to gas sensors, Sensors Actuators, B Chem. 211 (2015) 489–497. doi:10.1016/j.snb.2015.01.095.

[7]   F. Herrero-Carrón, D.J. Yáñez, F.D.B. Rodríguez, P. Varona, An active, inverse temperature modulation strategy for single sensor odorant classification, Sensors

Actuators, B Chem. 206 (2015) 555–563. doi:10.1016/j.snb.2014.09.085.

[8]   M. Drobek, J.H. Kim, M. Bechelany, C. Vallicari, A. Julbe, S.S. Kim, MOF-Based Membrane Encapsulated ZnO Nanowires for Enhanced Gas Sensor Selectivity, ACS Appl. Mater. Interfaces. 8 (2016) 8323–8328. doi:10.1021/acsami.5b12062.

[9]   G. Korotcenkov, B.K. Cho, Metal oxide composites in conductometric gas sensors: Achievements and challenges, Sensors Actuators, B Chem. 244 (2017) 182–210. doi:10.1016/j.snb.2016.12.117.

[10]   V. Krivetskiy, I. Malkov, A. Garshev, N. Mordvinova, O.I. Lebedev, S. Dolenko, A. Efitorov, T. Grigoriev, M. Rumyantseva, A. Gaskov, Chemically modified nanocrystalline SnO2-based materials for nitrogen-containing gases detection using gas sensor array, J. Alloys Compd. 691 (2017) 514–523. doi:10.1016/j.jallcom.2016.08.275.

[11]   V. Krivetskiy, A. Efitorov, A. Arkhipenko, S. Vladimirova, M. Rumyantseva, S. Dolenko, A. Gaskov, Selective detection of individual gases and CO/H2mixture at low concentrations in air by single semiconductor metal oxide sensors working in dynamic temperature mode, Sensors Actuators, B Chem. 254 (2018) 502–513. doi:10.1016/j.snb.2017.07.100.

[12]   G. Eranna, B.C. Joshi, D.P. Runthala, R.P. Gupta, Oxide Materials for Development of Integrated Gas Sensors—A Comprehensive Review, Crit. Rev. Solid State Mater. Sci. 29 (2004) 111–188. doi:10.1080/10408430490888977.

[13]   A.P. Lee, B.J. Reedy, Temperature modulation in semiconductor gas sensing, Sensors Actuators, B Chem. 60 (1999) 35–42. doi:10.1016/S0925-4005(99)00241-5.

[14]   A. Ortega, S. Marco, A. Perera, T. Šundic, A. Pardo, J. Samitier, An intelligent detector based on temperature modulation of a gas sensor with a digital signal processor, Sensors Actuators B Chem. 78 (2001) 32–39. doi:10.1016/S0925-4005(01)00788-2.

[15]   S. Zampolli, I. Elmi, E. Cozzani, G.C. Cardinali, A. Scorzoni, M. Cicioni, S. Marco, F. Palacio, J.M. Gómez-Cama, I. Sayhan, T. Becker, Ultra-low-power components for an RFID Tag with physical and chemical sensors, Microsyst. Technol. 14 (2008) 581–588. doi:10.1007/s00542-007-0444-8.

[16]   I. Sayhan, A. Helwig, T. Becker, G. Muller, I. Elmi, S. Zampolli, M. Padilla, S. Marco, Discontinuously Operated Metal Oxide Gas Sensors for Flexible Tag Microlab Applications, IEEE Sens. J. 8 (2008) 176–181. doi:10.1109/JSEN.2007.912791.

[17]   I. Simon, N. Bârsan, M. Bauer, U. Weimar, Micromachined metal oxide gas sensors: Opportunities to improve sensor performance, Sensors Actuators, B Chem. 73 (2001) 1–26. doi:10.1016/S0925-4005(00)00639-0.

[18]   S. Marco, The need for external validation in machine olfaction: Emphasis on health-related applications Chemosensors and Chemoreception, Anal. Bioanal. Chem. 406 (2014) 3941–3956. doi:10.1007/s00216-014-7807-7.

[19] I. Rodriguez-Lujan, J. Fonollosa, A. Vergara, M. Homer, R. Huerta, On the calibration of sensor arrays for pattern recognition using the minimal number of experiments, Chemom. Intell. Lab. Syst. 130 (2014). doi:10.1016/j.chemolab.2013.10.012.

[20] J.W. Gardner, P.N. Bartlett, Pattern recognition in odour sensing, in: Sensors Sens. Syst. an Electron. Nose, Kluwer Academic Publishers, Dordretch, 1992: pp. 161–179.

[21] R. Gutierrez-Osuna, H.T. Nagle, A method for evaluating data-preprocessing techniques for odour classification with an array of gas sensors., IEEE Trans. Syst. Man. Cybern. B. Cybern. 29 (1999) 626–632.

[22] T. Artursson, T. Eklöv, I. Lundström, P. Mårtensson, M. Sjöström, M. Holmberg, T. Eklo, I. Lundstro, M. Sjo, Drift correction for gas sensors using multivariate methods, J. Chemom. 14 (2000) 711–723. doi:10.1002/1099-128X(200009/12)14:5/6<711::AID-CEM607>3.0.CO;2-4.

[23] A. Ziyatdinov, S. Marco, A. Chaudry, K. Persaud, P. Caminal, A. Perera, Drift compensation of gas sensor array data by common principal component analysis, Sensors Actuators B Chem. 146 (2010) 460–465. doi:10.1016/j.snb.2009.11.034.

[24] A. Perera, T. Sundic, A. Pardo, R. Gutierrez-Osuna, S. Marco, A portable electronic nose based on embedded PC technology and GNU/Linux: hardware, software and applications, IEEE Sens. J. 2 (2002) 235–246. doi:10.1109/JSEN.2002.800683.

[25] M. Pardo, L.G. Kwong, G. Sberveglieri, K. Brubaker, J.F. Schneider, W.R. Penrose, J.R. Stetter, Data analysis for a hybrid sensor array, 106 (2005) 136–143. doi:10.1016/j.snb.2004.05.045.

[26] A. Pardo, S. Marco, C. Calaza, A. Ortega, A. Perera, T. Sundic, J. Samitier, Methods for sensors selection in pattern recognition, in: J.W. Gardner, K.C. Persaud (Eds.), Proc. Seventh Int. Symp. Olfaction Electron. Noses., CRC Press, Bristol, 2001: pp. 83–88.

[27] A. Vergara, E. Llobet, Sensor Selection and Chemo-Sensory Optimization: Toward an Adaptable Chemo-Sensory System, Front. Neuroeng. 4 (2012) 19. doi:10.3389/fneng.2011.00019.

[28] B. Raman, D.C. Meier, J.K. Evju, S. Semancik, Designing and optimizing microsensor arrays for recognizing chemical hazards in complex environments☆, Sensors Actuators B Chem. 137 (2009) 617–629. doi:10.1016/j.snb.2008.11.053.

[29] B. Raman, M. Douglas C., S. Semancik, A Statistical Approach to Materials Evaluation and Selection for Chemical Sensor Arrays, Springer, New York, NY, 2009.

[30] J.W. Gardner, P. Boilot, E.L. Hines, Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach, 106 (2005) 114–121. doi:10.1016/j.snb.2004.05.043.

[31] A. Shmilovici, G. Bakir, S. Marco, A. Perera, Finding the Best Calibration Points for a Gas Sensor Array with Support Vector Regression, in: 2nd Int. IEEE Conf., 2004: pp.

174–177.

[32] B. Walczak, E. Bouveresse, D.L. Massart, Standardization of near-infrared spectra in the wavelet domain, Chemom. Intell. Lab. Syst. 36 (1997) 41–51. doi:10.1016/S0169-7439(96)00075-5.

[33] E. Bouveresse, D.L. Massart, Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration, Chemom. Intell. Lab. Syst. 32 (1996) 201–213. doi:10.1016/0169-7439(95)00074-7.

[34] K.D.T.M. Milanez, T.C. Araújo Nóbrega, D. Silva Nascimento, R.K.H. Galvão, M.J.C. Pontes, Selection of robust variables for transfer of classification models employing the successive projections algorithm, Anal. Chim. Acta. 984 (2017) 76–85. doi:10.1016/j.aca.2017.07.037.

[35] M.O. Balaban, F. Korel, A.Z. Odabasi, G. Folkes, Transportability of data between electronic noses: Mathematical methods, Sensors Actuators, B Chem. 71 (2000) 203–211. doi:10.1016/S0925-4005(00)00617-1.

[36] L. Zhang, F.C. Tian, X.W. Peng, X. Yin, A rapid discreteness correction scheme for reproducibility enhancement among a batch of MOS gas sensors, Sensors Actuators, A Phys. 205 (2014) 170–176. doi:10.1016/j.sna.2013.11.015.

[37] L. Zhang, F. Tian, C. Kadri, B. Xiao, H. Li, L. Pan, H. Zhou, Sensors and Actuators B : Chemical On-line sensor calibration transfer among electronic nose instruments for monitoring volatile organic chemicals in indoor air quality, Sensors Actuators B. Chem. 160 (2011) 899–909. doi:10.1016/j.snb.2011.08.079.

[38] L. Fernandez, S. Guney, A. Gutierrez-Galvez, S. Marco, Calibration transfer in temperature modulated gas sensor arrays, Sensors Actuators, B Chem. 231 (2016) 276–284. doi:10.1016/j.snb.2016.02.131.

[39] J. Fonollosa, L. Fernández, A. Gutiérrez-Gálvez, R. Huerta, S. Marco, Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization, Sensors Actuators, B Chem. (2016). doi:10.1016/j.snb.2016.05.089.

[40] M. Zuppa, C. Distante, P. Siciliano, K.C. Persaud, Drift counteraction with multiple self-organising maps for an electronic nose, Sensors Actuators, B Chem. 98 (2004) 305–317. doi:10.1016/j.snb.2003.10.029.

[41] A. Vergara, S. Vembu, T. Ayhan, M. a. Ryan, M.L. Homer, R. Huerta, Chemical gas sensor drift compensation using classifier ensembles, Sensors Actuators B Chem. 166–167 (2012) 320–329. doi:10.1016/j.snb.2012.01.074.

[42] K. Yan, D. Zhang, Y. Xu, Correcting Instrumental Variation and Time-Varying Drift Using Parallel and Serial Multitask Learning, IEEE Trans. Instrum. Meas. 66 (2017) 2306–2316. doi:10.1109/TIM.2017.2707898.

[43] L. Zhang, Y. Liu, Z. He, J. Liu, P. Deng, X. Zhou, Anti-drift in E-nose: A subspace

projection approach with drift reduction, Sensors Actuators, B Chem. 253 (2017) 407–417. doi:10.1016/j.snb.2017.06.156.

[44]   L. Zhang, D. Zhang, Domain Adaptation Extreme Learning Machines for Drift Compensation in E-Nose Systems, IEEE Trans. Instrum. Meas. 64 (2015) 1790–1801. doi:10.1109/TIM.2014.2367775.

[45]   O. Tomic, T. Eklöv, K. Kvaal, J.-E. Haugen, Recalibration of a gas-sensor array system related to sensor replacement, Anal. Chim. Acta. 512 (2004) 199–206. doi:10.1016/j.aca.2004.03.001.

[46]   P. Boeker, On "Electronic Nose" methodology, Sensors Actuators, B Chem. 204 (2014) 2–17. doi:10.1016/j.snb.2014.07.087.

[47]   W. Tsujita, A. Yoshino, H. Ishida, T. Moriizumi, Gas sensor network for air-pollution monitoring, Sensors Actuators, B Chem. 110 (2005) 304–311. doi:10.1016/j.snb.2005.02.008.

[48]   S. De Vito, P. Di Palma, C. Ambrosino, E. Massera, G. Burrasca, M.L. Miglietta, G. Di Francia, Wireless sensor networks for distributed chemical sensing: Addressing power consumption limits with on-board intelligence, IEEE Sens. J. 11 (2011) 947–955. doi:10.1109/JSEN.2010.2077277.

[49]   M.I. Mead, O. a M. Popoola, G.B. Stewart, P. Landshoff, M. Calleja, M. Hayes, J.J. Baldovi, M.W. McLeod, T.F. Hodgson, J. Dicks, a. Lewis, J. Cohen, R. Baron, J.R. Saffell, R.L. Jones, The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks, Atmos. Environ. 70 (2013) 186–203. doi:10.1016/j.atmosenv.2012.11.060.

[50]   R. Piedrahita, Y. Xiang, N. Masson, J. Ortega, A. Collier, Y. Jiang, K. Li, R.P. Dick, Q. Lv, M. Hannigan, L. Shang, The next generation of low-cost personal air quality sensors for quantitative exposure monitoring, Atmos. Meas. Tech. 7 (2014) 3325–3336. doi:10.5194/amt-7-3325-2014.

[51]   S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D.M. Broday, B. Fishbain, On the feasibility of measuring urban air pollution by wireless distributed sensor networks, Sci. Total Environ. 502 (2015) 537–547. doi:10.1016/j.scitotenv.2014.09.059.

[52]   B. Fishbain, E. Moreno-Centeno, Self Calibrated Wireless Distributed Environmental Sensory Networks, Sci. Rep. 6 (2016) 1–10. doi:10.1038/srep24382.

[53]   J. Fonollosa, A. Vergara, R. Huerta, Algorithmic mitigation of sensor failure: Is sensor replacement really necessary?, Sensors Actuators, B Chem. 183 (2013). doi:10.1016/j.snb.2013.03.034.

[54]   E. Martinelli, G. Magna, A. Vergara, C. Di Natale, Cooperative classifiers for reconfigurable sensor arrays, Sensors Actuators, B Chem. 199 (2014) 83–92. doi:10.1016/j.snb.2014.03.070.

[55] ams AG, (n.d.).

[56] J. Fonollosa, L. Fernandez, R. Huerta, A. Gutierrez-Galvez, S. Marco, Temperature optimization of metal oxide sensor arrays using Mutual Information, Sensors Actuators, B Chem. 187 (2013) 331–339.

[57] J. Kneer, J. Wöllenstein, S. Palzer, Manipulating the gas-surface interaction between copper(II) oxide and mono-nitrogen oxides using temperature, Sensors Actuators, B Chem. 229 (2016) 57–62. doi:10.1016/j.snb.2016.01.104.

[58] P. Walden, J. Kneer, S. Knobelspies, W. Kronast, U. Mescheder, S. Palzer, Micromachined Hotplate Platform for the Investigation of Ink-Jet Printed, Functionalized Metal Oxide Nanoparticles, J. Microelectromechanical Syst. 24 (2015) 1384–1390. doi:10.1109/JMEMS.2015.2399696.

[59] J. Kneer, A. Eberhardt, P. Walden, A. Ortiz Pérez, J. Wöllenstein, S. Palzer, Apparatus to characterize gas sensor response under real-world conditions in the lab, Rev. Sci. Instrum. 85 (2014) 55006. doi:10.1063/1.4878717.

[60] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790–3798. doi:10.1039/C3AY40582F.

[61] D.E. (David E. Goldberg, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Longman Publishing Co., Inc., 1989.

[62] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: How and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207. doi:10.1016/S0169-7439(98)00051-3.

[63] R. Leardi, Genetic algorithms in chemistry, J. Chromatogr. A. 1158 (2007) 226–233. doi:10.1016/j.chroma.2007.04.025.

[64] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Springer-Verlag New York, 2002. doi:10.1007/978-0-387-21706-2.

[65] B.-H. Mevik, R. Wehrens, The pls Package: Principal Component and Partial Least, J. Stat. Softw. 18 (2007).

[66] L. Scrucca, GA: A Package for Genetic Algorithms in R, J. Stat. Softw. 53 (2013) 1–37. doi:10.1359/JBMR.0301229.

[67] D.J. Hand, R.J. Till, A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, Mach. Learn. 45 (2001) 171–186. doi:10.1023/A:1010920819831.

[68] A. Agresti, A Survey of Exact Inference for Contingency Tables, Stat. Sci. 7 (1992) 131–153. doi:10.1214/ss/1177011454.

[69] J.J. Rose, L. Wang, Q. Xu, C.F. Mctiernan, S. Shiva, J. Tejero, M.T. Gladwin, Carbon Monoxide Poisoning : Pathogenesis , Management , and Future Directions of Therapy, 195 (2017) 596–606. doi:10.1164/rccm.201606-1275CI.

[70] T.C. Pearce, M.A. Sánchez-Montañés, Chemical Sensor Array Optimization: Geometric and Information Theoretic Approaches, in: Handb. Mach. Olfaction Electron. Nose Technol., 2004: pp. 347–375. doi:10.1002/3527601597.ch14.